3. Matching methods (I)

LPO 8852: Regression II

Sean P. Corcoran

Treatment effects

Treatment effect for unit i using the potential outcomes framework:

$$\tau_i = Y_i(1) - Y_i(0)$$

where $Y_i(D_i)$ is the potential outcome for unit i. $D_i = 1$ if i is treated and $D_i = 0$ if not. Two *estimands* that may be of interest:

Population average treatment effect (ATE):

$$\tau_{ATE} = E(\tau) = E[Y(1) - Y(0)]$$

Average treatment effect on the treated (ATT):

$$\tau_{ATT} = E(\tau|D = 1) = E[Y(1)|D = 1] - \underbrace{E[Y(0)|D = 1]}_{\text{not observed}}$$

Treatment effects

When we use the mean of Y for the *untreated* in place of $E[Y(0)|D=1]$:

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

Selection bias reflects the difference in the average Y in the untreated state ($Y(0)$) between the treatment and control group.

Example 1: job training program

Person	Treat	Educ.	Age	Y(0)	Y(1)	Y
1	1	1	26	10	14	14
2	1	1	21	8	12	12
3	1	1	30	12	16	16
4	1	1	19	8	12	12
5	1	0	25	6	10	10
6	1	0	22	4	8	8
7	0	0	21	4	8	4
8	0	0	26	6	10	6
9	0	0	28	8	12	8
10	0	0	20	4	8	4
11	0	1	26	10	14	10
12	0	1	21	8	12	8
13	0	0	16	2	6	2
14	0	0	15	1	5	1

Source: Jennifer Hill (2011) lecture notes. Assume Y is earnings and D_i indicates participation in job training program.

Example 1: job training program

In the above example, $ATE = ATT = 4$. But:

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

$$12.0 - 5.4 = 4.0 + \underbrace{8.0 - 5.4}_{\text{selection bias}}$$

The treated group has a higher $Y(0)$ than the untreated group. Notice the treated group also has a higher average education and age, two things associated with higher earnings. Their Y would likely have been higher even in the absence of treatment.

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 private vs. public colleges example:

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

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

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).

Treatment effects

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, Group=A)=105,000
Avg(Y|D=0, Group=A)=110,000. Difference = 105,000 - 110,000 = -5,000

Avg(Y|D=1, Group=B)=60,000
Avg(Y|D=0, 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.

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

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.

Subclassification

The above example is a case of **subclassification**: grouping treated and untreated observations into strata, calculating differences within strata, and then weighting those differences to get a treatment effect estimate.

Subclassification is an example of a **selection on observables** design. Other examples include matching, weighting (and multiple regression!)

For these methods to yield valid estimates of an ATE/ATT, selection bias must be due entirely to the covariates you are conditioning on. Once you account for these, potential outcomes are no longer related to treatment—a big assumption, but reasonable in some applications.

Conditional independence assumption (CIA)

A setting in which potential outcomes are independent ($\perp\!\!\!\perp$) of treatment status is called *unconfoundedness*, *ignorability*, *selection on observables*, *exogeneity*, or the **conditional independence assumption** (CIA).

$$Y(0), Y(1) \perp\!\!\!\perp D|X$$

Note this is an assumption that is not easily validated. Another requirement of these methods is **common support** or *overlap*:

$$0 < P(D = 1|X) < 1$$

This is something that can be examined in the data.

Example 3: Subclassification

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)		
Label	Income Range	Sample Variance	Sample Mean		Public	Catholic (% of stratum total)	Public	Catholic	Diff.
			Public	Catholic					
Hi_Inc	$35,000 to $74,999	0.24	11.38	11.42	1,969	344 (*14.87%*)	53.60	55.72	2.12***,†
Med_Inc	$20,000 to $34,999	0.22	9.65	9.73	1,745	177 (*9.21%*)	50.34	53.86	3.52***,†
Lo_Inc	≤$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***

$^{\sim}p$ <0.10; $^{*}p$ <0.05; $^{**}p$ <0.01; $^{***}p$ <0.001
†One-sided test.

Example 3: Subclassification

The weighted average ATE uses total cell sizes as weights; the ATT uses counts of treated cases in each cell as weights. These TEs 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.

Example 3: Subclassification

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.
Hi_Inc	*Hi_Ach*	1,159	227	58.93	59.66	0.72
	MHi_Ach	432	73	49.18	50.71	1.53*,†
	MLo_Ach	321	38	42.75	44.23	1.48
	Lo_Ach	57	6	39.79	40.40	0.62
Med_Inc	*Hi_Ach*	790	93	57.42	59.42	2.00**,†
	MHi_Ach	469	49	47.95	50.14	2.19**,†
	MLo_Ach	390	33	41.92	44.56	2.64*,†
	Lo_Ach	96	2	37.94	39.77	1.83
Lo_Inc	*Hi_Ach*	405	36	56.12	56.59	0.47
	MHi_Ach	385	13	47.12	48.65	1.53
	MLo_Ach	433	21	40.99	41.70	0.71
	Lo_Ach	142	1	36.81	42.57	5.76
				Weighted Average ATE		***1.50***
				Weighted Average ATT		***1.31***

$^{\cdot}p$ <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 methods

A closely related approach to subclassification is **matching**. Rather than group observations into strata and averaging over strata, we "impute" counterfactuals by matching each treated (untreated) case to a similar untreated (treated) case based on one or more covariates.

- Exact matching
- Approximate matching (e.g., nearest neighbor, coarsened exact matching, propensity score)

Need not match to only one "counterfactual"—can match to multiple cases. It is also possible to match *with replacement* to promote better matches.

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:

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

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 relaxes the demand for an exact match and identifies "nearest neighbors" based on one or more covariates. How do we measure distance to find nearest neighbors?

- Easy with one covariate: absolute distance between x's
- With multiple covariates: 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 covariance between x's

Stata's `teffects nnmatch`

Stata's `teffects` implements a wide array of treatment effect estimators using matching, weighting, regression adjustment, etc. `teffects nnmatch` can use 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 simulated data.

Stata's teffects nnmatch

```
. teffects nnmatch (y age educ) (treat) , nneighbor(5) atet vce(iid) dmvariables

Treatment-effects estimation                   Number of obs      =        200
Estimator      : nearest-neighbor matching     Matches: requested =          5
Outcome model  : matching                                  min =          5
Distance metric: Mahalanobis                               max =         15
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
treat (1 vs 0)	**31.00002**	**.8976628**	**34.53**	**0.000**	**29.24063**	**32.75941**

```
Matching variables:
```

See the `teffects` documentation for standard error calculations, which are based on Abadie & Imbens (2006, 2011, 2012).

Abadie & Imbens (2008) do *not* recommend bootstrap estimation of standard errors when doing nearest neighbor matching.

Stata's tebalance

By design, nearest neighbor matching seeks to balance the confounding covariate(s) in the treated and untreated groups. You can see how you did in this regard using `tebalance summarize` following `teffects`:

```
. tebalance summarize
note: refitting the model using the generate() option

  Covariate balance summary
```

	Raw	Matched
Number of obs =	**200**	**168**
Treated obs =	**84**	**84**
Control obs =	**116**	**84**

	Standardized differences		Variance ratio	
	Raw	Matched	Raw	Matched
age	**.5124947**	**.0095797**	**.8829962**	**1.011965**
educ	**.1125516**	**.20222**	**1.038685**	**1.08452**

Note: the standardized difference is the difference in means between the treated and untreated groups, divided by the square root of a pooled variance. They can be interpreted in standard deviation units.

Stata's tebalance

Try `tebalance summarize, baseline` following `teffects` to see baseline differences in covariates in original units.

```
. tebalance summarize, baseline
note: refitting the model using the generate() option

  Covariate balance summary
                                              Raw              Matched
                      Number of obs =         750                  556
                      Treated obs   =         278                  278
                      Control obs   =         472                  278
```

	Means		Variances	
	Control	Treated	Control	Treated
age	27.49364	30.3705	41.56259	38.57342

Bias corrections

Recall that in this genre of estimators—in which the CIA is assumed to hold—the only source of bias comes from imbalance in the covariates (i.e., imperfect matches).

When there is imperfect matching, the treatment effect estimator is a combination of the "true" effect and differences in Y that are a byproduct of the imbalance in covariates.

Abadie & Imbens (2011) propose a consistent bias-corrected estimator. The idea here is that one can use OLS to estimate the relationship between Y and covariates X. The difference in (predicted) Y due to the differences in X (between the perfect and actual match) is used to adjust the treatment effect estimate. In `teffects`: use `biasadj(`*varnames*`)` option with *varnames* the list of continuous covariates.

Propensity scores

Rosenbaum & Rubin (1983) showed that if $Y(0), Y(1)$ are independent of D conditional on X, then they are also independent of D conditional on a **propensity score** constructed using X.

- Rather than stratifying or matching on all of the variables in X, it is sufficient to use the "one-number summary" of the relationship between treatment and X: $P(X) = Pr(D = 1|X)$
- $P(X)$ can be estimated using a logit, probit, or LPM regression from which one can obtain predicted probabilities $\widehat{P(X)}$. LPM is not advised if predicted probability falls outside of [0,1].

Stata also refers to the propensity score as the probability of treatment.

Propensity scores

The propensity score estimator for ATT can be written as:

$$E_{P(X)|D=1}\left(\underbrace{E[Y(1)|D=1,P(X)]}_{\text{treated}} - \underbrace{E[Y(0)|D=0,P(X)]}_{\text{untreated}}\right)$$

Effectively, *for each propensity score* we calculate the difference in mean outcomes for the treated and untreated with that $P(X)$. We then take a weighted average of these over the different propensity score values. The subscript $_{P(X)|D=1}$ means we are taking a weighted average over the area of common support.

Compare logic to Example 2 where we averaged the group differences in earnings across two groups with common support (A and C), weighting as appropriate.

Propensity scores

In practice $P(X)$ takes on a continuum of values and thus stratifying on $P(X)$ itself—in the manner we did with subclassification—is not feasible.

Thus, we can do other things with the propensity score, including matching and re-weighting. Even when propensity scores are not used to estimate treatment effects, they can be useful diagnostic tools since they force you to think about balance between the treated and untreated groups, and the model of selection into treatment.

Propensity scores

We will look at several uses for propensity scores:

- Matching
- Inverse probability weighting (IPW)

Note: King & Nelson (2019) critique of using propensity scores for matching. (See link to seminar video on Github). Preferred use these days is IPW.

Propensity scores in practice

Key considerations:

- Choice of model for estimating propensity score (logit, probit)
- Selection of covariates for estimating treatment model
- Algorithm used for matching, if matching: how matches are made, how many, how close
- Checking for overlap and common support
- Assessing match quality (balanced distribution of covariates)
- Treatment effect calculation
- Estimating standard errors for treatment effects

Source: Caliendo & Kopeinig (2008), a good guide for practice. Also see the text by Guo & Fraser (2015), *Propensity Score Analysis*.

Estimating the propensity score

Choice of model:

- For binary treatment, whether one uses a logit, probit, or LPM model is probably not that consequential.
- For multiple treatments, the choice may be more important (see Caliendo & Kopeinig, 2008)

Covariate selection:

- Goal: choose X's such that the unconfoundedness holds—should promote covariate balance.
- Should be correlated with treatment (D_i) and the outcome Y.
- Selection should be based on theory and contextual knowledge.
- X should be measured *before* treatment, and not affected by it (or by the anticipation of treatment).
- X's should not be "too good" at predicting treatment—we are relying on common support.

Matching algorithms

There are many approaches to identifying matches for treated cases:

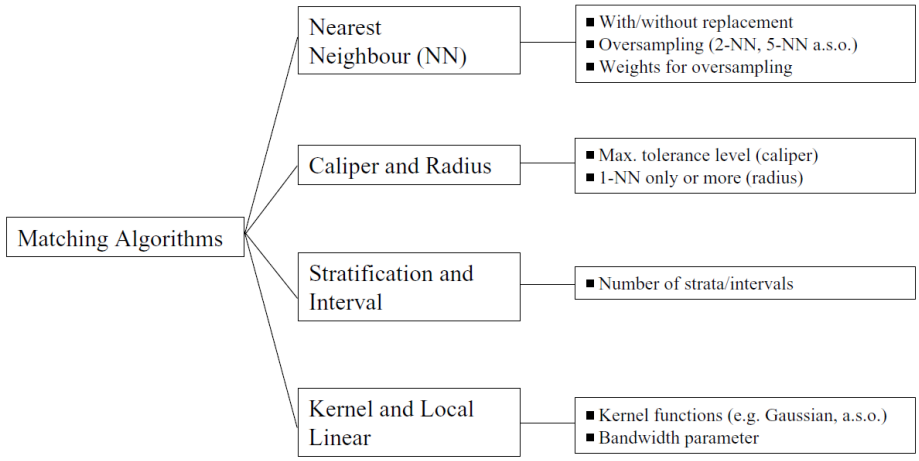


Figure 2. Different Matching Algorithms.

Matching algorithms

Nearest neighbor (NN):

- Treated cases are paired with one or more nearest (untreated) neighbors based on their $\widehat{P(X)}$.
- Matches can be *with* or *without* replacement
  - With replacement: better matches, possibly less bias, but higher standard errors (re: repeated use of same observations)
  - Without replacement: worse matches, possibly more bias, but lower standard errors (re: using more variation)
  - If matching without replacement, order matters, so sort randomly (and preserve sort order if you wish to replicate)
- "Oversampling": choosing > 1 match for each treated case

Matching algorithms

Caliper and radius matching:

- A "caliper" is a tolerance level for how different $\widehat{P(X)}$ can be for matched observations
- "Radius" matching defines a caliper and then uses all untreated neighbors in the caliper

Stratification and interval matching:

- This method partitions the common support into intervals (strata) and then calculates mean differences within these strata

Kernel (KM) and local linear matching (LLM):

- Weighting algorithm that uses weighted average of nearly all untreated observations. Weights may depend on how different the $\widehat{P(X)}$ are.

Checking for overlap and common support

Given $\widehat{P(X)}$, one can inspect the distributions for the treated and untreated observations to look for common support.

- Can compare the maxima and minima of $\widehat{P(X)}$ for the two groups
- Can formally compare the density distributions for each

The command `teffects overlap` (following `teffects psmatch`) produces densities of propensity scores.

Stata's `teffects psmatch`

`teffects psmatch` can estimate propensity scores and produce ATT and ATE via nearest neighbor matching using propensity scores.

`teffects psmatch` *(y) (t x, tmodel), options*

Again y is the outcome, x are the covariates, and t is the treatment indicator. *tmodel* is the type of propensity score model you would like to estimate (e.g., logit, probit). In the options can specify `ate` or `atet`, the number of nearest neighbors, the caliper, etc.

I also recommend the older user-written package `psmatch2`, which is useful for refining your propensity score model *before* requesting the ATT estimate. Alternatively, can "quietly" run `teffects` and then diagnose balance with `tebalance`. NOTE, however, that the treatment effect standard errors are incorrect in `psmatch2`. Use `teffects` for the final ATT calculation.

Stata's `teffects psmatch`

Can obtain predicted propensity scores after `teffects psmatch` using the `predict` command. Requires the `gen()` option in the `teffects psmatch` command, which creates variables containing the index of the nearest neighbor(s):

`predict` *(newvar)*`, ps` *options*

Can also predict *potential outcomes* (`po`), individual treatment effects given potential outcomes (`te`), and distance to nearest neighbor (`distance`).

In-class exercise

Using NSW data matched to CPS and PSID (Lalonde 1983 and others):

```
quietly teffects psmatch (re78) (treat age educ black hisp
re74 re75, probit), atet gen(mvar)
```

or

```
psmatch2 treat age educ black hisp re74 re75
```

- Estimates propensity scores (default for `psmatch2` is probit regression)
- Identifies nearest neighbor matches (default in `psmatch2` is matching with replacement).
- Use the option `ties` with `psmatch2` if you want to keep all matches with the same propensity score (the default in `teffects`).

Example

`psmatch2` will show you the probit estimates. Alternatively, could just use `probit` (or `logit`) command.

```
. psmatch2 treat age educ black hisp re74 re75

Probit regression                                 Number of obs     =     18,927
                                                  LR chi2(6)        =     861.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -609.54681                       Pseudo R2         =     0.4140

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0171489   .0041262    -4.16   0.000    -.0252362   -.0090616
   education |  -.0301524    .015144    -1.99   0.046    -.0598342   -.0004707
       black |   1.589191   .0958607    16.58   0.000     1.401308    1.777075
    hispanic |   .6522818   .1525644     4.28   0.000      .353261    .9513026
        re74 |   -.000023   .0000105    -2.19   0.029    -.0000436   -2.40e-06
        re75 |   -.000082   .0000133    -6.14   0.000    -.0001081   -.0000558
       _cons |  -1.677663   .2261594    -7.42   0.000    -2.120927   -1.234399
------------------------------------------------------------------------------
Note: 346 failures and 0 successes completely determined.
```

Example

`psmatch2` creates several variables in your dataset: `_pscore`, `_treated`, `_support`, `_weight`, `_id`, `_n1`, `_nn`, `_pdif`

- `_pscore`: estimated $P(X)$
- `_treated`: flags observations Stata recognized as treated
- `_support`: flags observations on common support
- `_weight`: weight for matched controls (untreated obs only)
- `_id`: id number assigned for identifying matches
- `_n1`: id of nearest neighbor (treated obs only)
- `_nn`: number of matched neighbors
- `_pdif`: absolute value of diff between $P(X)$ and $P(X)$ of NN

As noted earlier, `teffects psmatch` can be augmented with options (and used with the `predict` command to get similar information)

Example

Inspect the histograms of propensity scores for the treated and untreated observations: `psgraph` (uses *all* of the data, not just the matched sample)

Assessing balance on covariates

In this step we check whether the X have similar distributions in the matched sample (treated and untreated observations). Why? If the groups are "exchangeable" we would expect similar distributions of X.

- For this test there must be one control for every treatment observation. If there are multiple control observations for a given treatment observation, they should be weighted so that the sum of the weights is equal to 1.
- If the covariates are *not* balanced, this suggests the propensity score needs to be re-estimated, perhaps with interaction terms, quadratic, or higher-order terms, or by including additional covariates.
- Stata: can use `pstest` following `psmatch2`, or `tebalance summarize`

Example

```
pstest age educ black hisp re74 re75
```

```
. pstest age educ black hisp re74t re75t,both
```

Variable	Unmatched / Matched	Mean Treated	Mean Control	%bias	%reduct |bias|	t-test t	t-test p>|t|	V(T)/ V(C)
age	U	25.816	33.444	-82.3		-9.43	0.000	0.42*
	M	25.816	24.989	8.9	89.2	0.95	0.342	0.58*
education	U	10.346	12.04	-67.9		-7.92	0.000	0.48*
	M	10.346	10.811	-18.6	72.6	-1.95	0.053	0.62*
black	U	.84324	.09739	224.5		33.96	0.000	.
	M	.84324	.84865	-1.6	99.3	-0.14	0.886	.
hispanic	U	.05946	.06671	-3.0		-0.39	0.694	.
	M	.05946	.03784	8.9	-198.1	0.97	0.335	.
re74t	U	2.0956	14.746	-156.5		-16.63	0.000	0.22*
	M	2.0956	1.7488	4.3	97.3	0.79	0.433	1.96*
re75t	U	1.5321	14.38	-170.9		-17.24	0.000	0.10*
	M	1.5321	1.5778	-0.6	99.6	-0.14	0.891	1.03

* if variance ratio outside [0.75; 1.34] for U and [0.75; 1.34] for M

Sample	Ps R2	LR chi2	p>chi2	MeanBias	MedBias	B	R	%Var
Unmatched	0.463	961.39	0.000	117.5	119.4	266.1*	0.24*	100
Matched	0.013	6.66	0.354	7.2	6.6	27.0*	0.69	75

* if B>25%, R outside [0.5; 2]

pstest %bias

The column %bias provides the standardized percent bias: the difference in sample means between the treated and untreated observations as a percentage of the square root of the average of the sample variances in the treated and untreated groups.

$$\Delta_X = \frac{\bar{X}_1 - \bar{X}_0}{\sqrt{(s_0^2 + s_1^2)/2}}$$

Estimating treatment effect

Once the propensity scores are estimated/finalized, they can be used to estimate the treatment effect. Methods for estimating ATT:

- NN matching: calculate the average difference in Y between treated and (matched) untreated observations. Use weights if multiple matched observations.
- Inverse probability weighting: weighting observations appropriate by their inverse probability of treatment
- Interval matching: calculate the average difference in Y within each interval.
- Kernel matching: each treated observation has a "composite" match using the entire set of untreated observations. Each of the latter is weighted by a similarity measure.
- Other: using propensity score to construct a regression sample and/or weight observations in a regresssion.

Example

```
psmatch2 treat age educ black hisp re74 re75, outcome(re78)
```

```
. psmatch2 treat age educ black hisp re74 re75, outcome(re78)

Probit regression                                 Number of obs     =     18,927
                                                  LR chi2(6)        =     861.45
                                                  Prob > chi2       =     0.0000
Log likelihood = -609.54681                       Pseudo R2         =     0.4140

------------------------------------------------------------------------------
       treat |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
, bin(20) fcolor(none) lcolor(red)) (histogram _pscore if treat==0 [fweight=_weight], bin(20) fcolor(none) lcolor(red)
-------------+----------------------------------------------------------------
         age |  -.0171489   .0041262    -4.16   0.000    -.0252362   -.0090616
   education |  -.0301524    .015144    -1.99   0.046    -.0598342   -.0004707
       black |   1.589191   .0958607    16.58   0.000     1.401308    1.777075
    hispanic |   .6522818   .1525644     4.28   0.000      .353261    .9513026
        re74 |  -.000023   .0000105    -2.19   0.029    -.0000436   -2.40e-06
        re75 |  -.000082   .0000133    -6.14   0.000    -.0001081   -.0000558
       _cons |  -1.677663   .2261594    -7.42   0.000    -2.120927   -1.234399
------------------------------------------------------------------------------
Note: 346 failures and 0 successes completely determined.

------------------------------------------------------------------------------
    Variable     Sample |    Treated     Controls   Difference          S.E.   T-stat
-----------------------+------------------------------------------------------------
        re78  Unmatched |  6349.1435    15594.9895  -9245.84596    803.597327  -11.51
              ATT      |  6349.1435    5528.42643   820.717073    829.853344    0.99
------------------------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

                    psmatch2:
psmatch2:          Common
Treatment          support
assignment  | On suppor |     Total
------------+-----------+----------
  Untreated |    18,742 |    18,742
    Treated |       185 |       185
------------+-----------+----------
      Total |    18,927 |    18,927
```


Estimators thus far

Consider the matching estimators examined thus far:

- Exact matching: each treated (untreated) case is matched to one or more untreated (treated) cases with the same X.
- Nearest neighbor matching: each treated (untreated) case is matched to one or more untreated (treated) cases with the *closest* X.
- Propensity score matching: each treated (untreated) case is matched to one or more untreated (treated) cases with the *closest* $\widehat{P(X)}$.

The matched sample is used to calculate the ATE/ATT. Each relies on the CIA: conditional on X, treatment is as good as random.

Where do weights come in? Only to account for multiple matches (e.g., 1 treated observation may be matched to 10 neighbors, so each of the 10 neighbors get 1/10 weight). Data are "pruned" if they aren't matched.

Inverse probability weighting (IPW)

Inverse probability weighting uses all of the data, reweighting observations to create desired balance. Weights used are:

$$w_{ATT} = D_i + (1 - D_i)\frac{\widehat{P(X)}}{1 - \widehat{P(X)}}$$

$$w_{ATE} = \frac{D_i}{\widehat{P(X)}} + \frac{(1 - D_i)}{1 - \widehat{P(X)}}$$

Inverse probability weighting (IPW)

Intuition using a simple example:

	Treated ($D = 1$)	Untreated ($D = 0$)	P(D|X)
X=1	1	9	0.1
X=0	4	1	0.8

1 confounding covariate X, where the probability of treatment varies with X (0.1 for $X = 1$ and 0.8 for $X = 0$).

Inverse probability weighting (IPW)

Intuition using a simple example:

	Treated ($D = 1$)	Untreated ($D = 0$)	P(D|X)
X=1	1	9	0.1
X=0	4	1	0.8

Consider how exact and propensity score matching would work in this example. For ATT each treated case would be matched to one or more untreated cases. For the moment consider $nn = 1$:

- Among $X = 1$, the 1 treated case would be matched to 1 untreated case. The 1 untreated stands in for 9 cases.
- Among $X = 0$, each treated case would be matched to 1 untreated case. Each untreated stands in for 0.25 cases

Inverse probability weighting (IPW)

	Treated	Untreated	P(D|X)	IPW Treated	IPW Untreated
X=1	1	9	0.1	10.00	1.11
X=0	4	5	0.8	1.25	5.00

IPWs give more weight to treated cases with a low $P(X)$ and untreated cases with a high $P(X)$. After applying the weights, you have "effectively" the same sample size of treated and untreated cases with the same $P(X)$

Inverse probability weighting (IPW)

Note: IPW estimators become unstable when there is low overlap (cases with very low probability of treatment). Re: these observations get extremely high weight when using inverse probability.

Stata's `teffects ipw`

`teffects ipw` can estimate ATT and ATE using inverse probability weighting. Propensity scores (probability of treatment) are used in the weights. The syntax is very similar to `psmatch`:

`teffects ipw` *(y) (t x, tmodel), options*

y is the outcome, *x* are the covariates, and *t* is the treatment indicator. *tmodel* is the type of propensity score model you would like to estimate (e.g., logit, probit). In the options can specify `ate`, `atet`, or the potential outcome means `po`.

Matching vs. regression

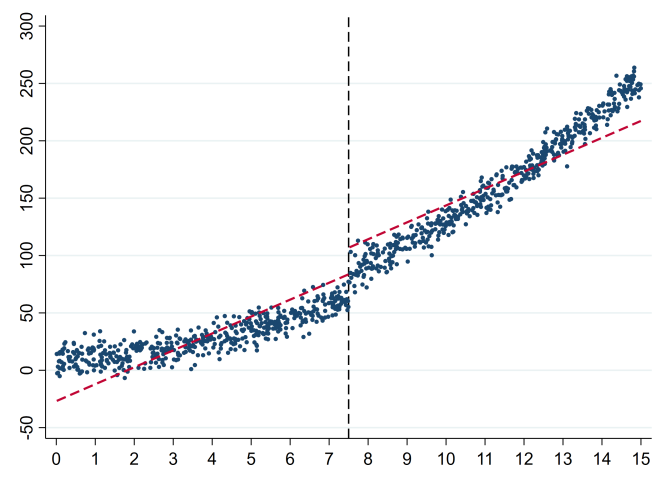
If potential outcomes are independent of treatment conditional on X, why not just estimate a regression controlling for X?

- Matching does not require assumptions about functional form for the outcome model (e.g., a linear relationship between Y and X).
- Regression runs the risk of extrapolating onto a space where there is little common support.
- Matching focuses our attention on balance and the degree of common support.

That said, propensity score matching "shifts the problem to the task of estimating the propensity score." If the model for the propensity score is mis-specified, the propensity score matching estimator will be biased.

Matching vs. regression

By making strong functional form assumptions, one can use regression to estimate treatment effects even when there is little overlap in X between the treated and untreated cases. But getting the functional form wrong can lead to poor inferences. We'll see this later with regression discontinuity:



Matching vs. regression

See also Murnane & Willett ch. 12 on the differences between matching strategies and regression (pp. 304-ff).

Assessing matching methods

How well does this method perform? Can we do better?

- Studies comparing impact estimates from matching methods to those from randomized experiments (e.g., Wilde & Hollister, 2007)
- Imbens (2015) recommendations for matching and subclassification

Matching vs. experimental impact estimates

Wilde & Hollister (2007) is one example of a study benchmarking propensity score matching estimators against estimates from randomized experiments.

- In this case, uses Tennessee STAR class size reduction program
- Other examples using PSM: Dehejia & Wahba (1999, 2002); Smith & Todd (2005); Agodini & Dynarski (2004); Diaz & Handa (2006); Michalopoulos, Bloom & Hill (2004)

Tennessee STAR

- Randomized controlled trial from 1985 in which students were randomized within schools to either small classes (13-17) or regular-sized classes (22-25)
- 79 participating schools statewide
- See Krueger (1999) and related papers

Matching vs. experimental impact estimates

Wilde & Hollister (2007) focus on 11 schools from 6 districts, and kindergarten only

- Each school is treated as a separate experiment
- Experimental treatment effect estimate calculated within school
- For treated students in each school, use propensity score to select matched comparisons from STAR control group in *other* schools
  - Trimmed untreated cases with pscores outside range of treated cases
  - Balanced the distribution of propensity scores by **stratifying** into bins and adjusting size of bins until no significant differences in each bin
  - Tested for balance in covariates, making adjustments as needed
  - Matched with replacement
  - In their case: the matched comparisons are actually more similar to the treated cases than are the actual control group students

Matching vs. experimental impact estimates

Table 4. Tennessee Project STAR mean percentile combined math and reading scores within school for students in treatment, experimental control, and nonexperimental PSM comparison group.

1	2	3	4
Project STAR School ID Number	Mean Percentile Combined Test Score Small Class	Mean Percentile Combined Test Score Control Group (Regular and Regular with Aide Classes)	Mean Percentile Combined Test Score Nonexperimental PSM Comparison Group
7	62.02	65.35	50.34
	(32.50)	(27.60)	(19.51)
9	70.04	60.20	48.07
	(21.79)	(24.61)	(23.71)
16	43.66	20.16	61.00
	(25.28)	(15.51)	(27.83)
22	68.94	44.73	55.07
	(29.30)	(24.33)	(29.85)
27	58.78	69.54	25.57
	(20.07)	(23.32)	(15.23)
28	44.53	41.71	51.37
	(29.41)	(24.46)	(29.43)
32	35.14	23.19	57.02
	(24.38)	(19.19)	(33.92)
33	40.46	29.08	60.26
	(16.30)	(20.35)	(29.47)
51	74.23	60.04	59.32
	(20.64)	(24.29)	(24.01)
63	80.25	62.08	50.67
	(17.77)	(25.66)	(23.96)
72	78.92	57.62	45.24
	(22.70)	(27.85)	(20.73)

Standard deviations in parentheses.

Matching vs. experimental impact estimates

Table 5. Project STAR regression adjusted estimates of program effect using experimental controls and nonexperimental comparison groups.

1	2	3	4	5
Project STAR School ID Number	Regression Adjusted Estimate of Program Effect with Experimental Controls	Regression Adjusted Estimate of Program Effect with Nonexperimental Comparisons	Are the Effects Opposite in Sign?	Is the Difference Between Experimental and Nonexperimental Estimates Significant?
7	−4.59	11.63		
	(7.07)	(6.86)	Yes	Yes
9	11.89	21.97		
	(3.87)*	(5.54)*	No	No
16	22.73	−17.35		
	(4.56)*	(11.23)	Yes	Yes
22	24.14	13.87		
	(10.06)*	(13.32)	No	No
27	−10.01	33.20		
	(8.07)	(5.68)*	Yes	Yes
28	0.79	−6.01		
	(8.87)	(12.38)	Yes	No
32	12.10	−21.89		
	(5.31)*	(15.26)	Yes	Yes
33	11.53	−19.80		
	(6.82)	(11.64)	Yes	Yes
51	14.00	15.20		
	(4.36)*	(2.53)*	No	No
63	15.17	29.59		
	(6.75)*	(5.68)*	No	No
72	18.50	33.69		
	(4.98)*	(5.87)*	No	Yes
Across schools weighted average	12.72	17.79		
	(1.70)*	(1.73)*	No	Yes

*Robust standard errors clustered at the classroom level in parentheses. Asterisk indicates significance at the 5 percent level. See Appendix B for discussion of the standard errors. (All appendices are available at the end of this article as it appears in JPAM online. Go to publisher's website and use search engine to locate article at: http://www3.interscience.wiley.com/cgi-bin/jhome/34787).

Matching vs. experimental impact estimates

Wilde & Hollister (2007) conclusions:

- They conclude that PSM performs poorly, although the weighted average across schools is not too far off the mark
- They argue that the question for study designers and decision-makers should be: "how close is *close enough*"? Would non-experimental estimates lead to a different decision than those from the randomized experiment?
- Strength of the study: a multi-site randomized experiment in which data were collected and measured the same way
- Weaknesses: kindergarten students did not have pre-treatment measurements on the outcome, inability to make finer geographic matches

Alternatives: can matches be improved upon?

There are many ways to identify matches and construct matching estimators, of which Wilde & Hollister (2007) chose one. Recall the decision points (Caliendo & Kopeinig, 2008):

- Number of nearest neighbors
- Radius and caliper
- Use of kernels
- Regression adjustment

Propensity scores can also be used for other purposes, e.g., subclassification.

Imbens (2015)

The review by Imbens (2015) offers some useful guidance on the use of matching methods, with examples.

1. Cases in which OLS estimators are likely to be especially problematic for estimating causal effects with non-experimental data
2. Two recommended methods: (1) subclassification and regression; or (2) matching
3. Trimming and other pre-processing steps to improve balance
4. Supplementary analyses for assessing plausibility of conditional independence assumption

A key takeaway: "there are no, and will not be, general results implying that in general some estimators are superior to all others"

Imbens (2015) on OLS

Imbens provides an example illustrating why and when OLS can be problematic. Key takeaway points:

1. The OLS regression provides the average of the potential control outcomes for the treated
2. Functional form assumptions can matter a lot: extrapolation and misspecification
3. This is especially true when the distribution of covariates differs between the treated and untreated cases
4. Extreme values can have a large influence on the OLS estimates: "regression models are not fundamentally robust to the substantial differences between treatment and control groups"

Imbens (2015) on analytic methods

Stages:

1. **Design stage**: trimming the full sample and balancing on covariates
2. **Supplementary analysis stage**: assessing unconfoundedness
3. **Analysis stage**: estimating treatment effect

Important: the outcome data are not used until the last stage.

Imbens (2015) on analytic methods

Tools:

1. **Normalized differences**: for assessing balance, use *normalized differences in mean covariates*. In Stata, this is the "% bias" in `pstest`. This is preferable to t-tests of significant differences.
2. **Propensity score**: Imbens uses logit, but notes the choice of probit or logit matters more when there are cases with pscores close to 0 or 1
   - "the propensity score plays a mechanical role in balancing the covariates ... In choosing a specification, there is therefore little role for theoretical substantive arguments. We are mainly looking for a specification that leads to an accurate approximation to the conditional expectation."
   - There is no harm in specification searches at this stage
   - Interactions and non-linearities are often important
   - There are data-driven algorithms for selecting covariates (e.g., stepwise approach of Imbens & Rubin, 2015; lasso methods)

Imbens (2015) on analytic methods

Tools:

3. **Propensity score**
   - "However, the point is again not to find a single method for estimating the propensity score that will outperform all others. Rather, the goal is to find a reasonable method for estimating the propensity score that will, in combination with the subsequent adjustment methods, lead to estimates for the treatment effects of interest that are similar to those based on other reasonable methods for estimating the propensity score"
   - Stepwise approach of Imbens & Rubin: first choose a set of predictors that will be in the model, regardless of other decisions (e.g., lagged measures of the outcome). Then determined a threshold for inclusion of other linear and quadratic terms. Finally, successively add predictors and compare to this threshold.

Imbens (2015) on analytic methods

Tools:

4. **Blocking method**: one recommended estimator uses the propensity score to block (or subclassify) observations and then use regression within blocks.
   - Partition the range of the propensity score into J intervals
   - Within each interval, estimate a linear regression with some covariates (all or a subset of those thought to be most important)
   - The J estimates of the treatment effects are then combined into one overall effect
   - 5 blocks is common, but Imbens & Rubin (2015) propose an algorithm for selecting this

Imbens (2015) on analytic methods

Tools:

4. **Matching method**: with replacement. Rather than using propensity scores, they use the Mahalanobis distance metric:

$$d = ||x, x'|| = (x - x')' \hat{\Omega}_X^{-1} (x - x')$$

where x and x' are two vectors of x-values, and $\hat{\Omega}_X^{-1}$ is the sample covariance matrix of the covariates (X).