Lecture 07 - Propensity Score Matching Tutorial

Samuel DeWitt

Propensity Score Matching - Tutorial Overview

There are a few things we will cover in this lecture:

1) Basic setup of data
2) Estimating the propensity score
3) Matching treated and untreated cases
4) Evaluating common support assumptions
5) Estimating and interpreting treatment effects

Basic Setup of Data

To run a propensity score matching procedure you need a few things in your data:

1) A dichotomous (0/1) treatment indicator
2) Multiple pre-treatment variables to help generate the propensity score
3) An outcome measured after treatment

Luckily, we have all of the above in the NLSY97 data sets I have provided for you.

Defining Treatment

For the purposes of this tutorial, we will be defining treatment as engagement in any type of delinquency in the 1998 wave of data collection.

This indicator is represented in the data by the delinq98 variable that I create below:

```
NLSY97HW1 <- NLSY97HW1  %>%
  mutate(varscore98=cargun98+destprop98+stl_lt5098+stl_gt5098+
           othprop98+attack98+selldrugs98)

NLSY97HW1$delinq98<-with(NLSY97HW1, ifelse(varscore98>=1,1,0))
```

Defining Matching Variables

Next, we need a series of variables we **think** are related to our treatment indicator.

Here's a list of the indicators we will be using in this tutorial:

1) Antisocial peers
2) Gender (Male=1)
3) Age at 1st wave

A code chunk in the .rmd document (not shown here) creates some of these variables.

Defining Matching Variables

An important caveat - I am keeping this purposefully simple for this example.

In fact, this is almost embarassingly simple. Most applications of propensity score matching use substantially more matching variables (generally several dozen or more).

So, it is best to treat this as a simple example. I doubt you'll see applications of matching that look like this (if you do, you should *immediately* doubt the **conditional independence assumption**!!)

Defining the Outcome Variable

We want to define a reasonable outcome we think might be associated with engagement in at least one type of delinquency. Obviously levels of delinquency won't work because we know the control group has a value of zero.

In other words, our outcome cannot also be included in the *treatment* variable.

Here, we will use drug use, since both treated and untreated youth can engage in using drugs and this measure is not included in the calculation for treatment.

```r
NLSY97HW1 <- NLSY97HW1 %>%
  mutate(drug_use99=num_coc99+days_marij99+days5drinks99)
```

Defining Variables - Check, Check, and Check

We now have our three sets of variables: 1) pre-treatment matching variables, 2) the dichotomous treatment variable, and 3) the outcome variable.

Now, we need to implement the matching procedure which, as we covered in a prior lecture, begins with a logistic regression.

As a (brief) means of review, the logistic regression predicts the 0/1 treatment variable using the pre-treatment matching variables. It then assigns a probability for each individual to **have** been treated, even when they weren't.

We use **that** probability (i.e., **propensity**) to match treated and untreated youth.

Running the Matching

To estimate the propensity score match we will use the **MatchIt** package in R - **Hint** you will need to install the package before you try to run this code!!

The *matchit* function has the following format:

matchit(formula, data, method="nearest", distance="logit", caliper=0.05, discard="none")

Running the Matching

I want to go over each individual part of this function one by one:

- The *formula* section
  - This is where you write the formula for the logistic regression. It follows the same format as we have used before.
  - $DV \sim IV_1 + IV_2 + IV_3 + \dots IV_k$

Running the Matching

I want to go over each individual part of this function one by one:

- The *data* section
  - We have also seen this before; this is simply where we identify the data frame where the variables in the *formula* function come from.
  - $data = NLSY97HW1$ in this example.

Running the Matching

I want to go over each individual part of this function one by one:

- The *method* section
  - This automatically defaults to **nearest** which means the function will match a treated case to the **nearest neighbor** propensity score.
  - In practice, this means that the **untreated** case with the *closest* (nearest) propensity score will be chosen as a match for each **treated** case.

Running the Matching

I want to go over each individual part of this function one by one:

- The *distance* section
  - This section defines the model used to estimate the propensity score. It will default to **logit** which stands for **logistic regression**.

Running the Matching

I want to go over each individual part of this function one by one:

- The *caliper* section
  - This provides direction for the **nearest neighbor** matching procedure by defining an upper limit for how different a propensity score can be within a matched pair.
  - Taken literally, a **caliper** of .05 means that no matches will be permitted where the difference within a matched pair with respect to the **propensity** to be treated is larger than .05 (or, more than a 5 percentage point difference).

Running the Matching

I want to go over each individual part of this function one by one:

- The *discard* section
  - This option relates to whether units should be **discarded** if their propensity scores lie outside of some pre-determined range or are not within **common support**
  - This option defaults to **none** but we will replace it with **both** in our examples, since this will *trim* propensity scores for both groups that do not overlap.
  - That is, specifying **both** will generally drop **untreated** cases with non-overlapping low propensities to have been treated and **treated** cases with non-overlapping high propensities to have been treated.

Running the Matching

Now we will estimate the propensity score with our data.

We define treatment as the youth reporting engaging in deliquent activity during their 1998 interview. This is measured as 0 for no delinquency and 1 for at least one type of delinquent behavior.

Covariates predicting selection include 1) Antisocial peers, 2) Gender, 3) Age at 1st wave.

Our outcome of interest is the frequency of recent (last 30 days) drug use (including marijuana, cocaine or other hard drugs, and excessive alcohol use).

Running the Matching

First, the command that produces the match:

```
match1<-matchit(delinq98~bad_peers+as.factor(male)+age,
                data=NLSY97HW1, method="nearest",
                distance="logit", caliper=.05,
                discard="both")
```

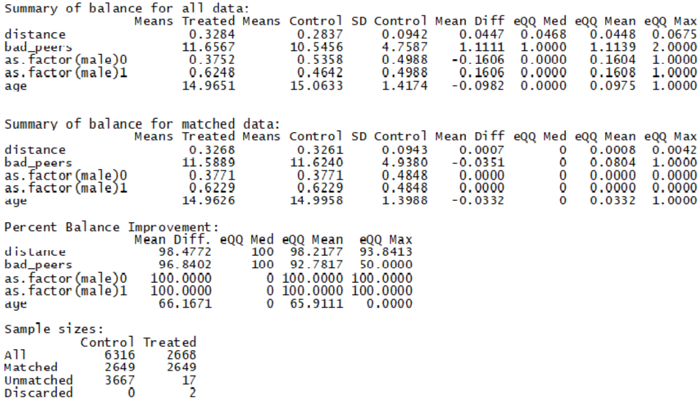
Running the Matching

Then, the output:

```
##
## Call:
## matchit(formula = delinq98 ~ bad_peers + as.factor(male) + age,
##     data = NLSY97HW1, method = "nearest", distance = "logit",
##     discard = "both", caliper = 0.05)
##
## Summary of balance for all data:
##                  Means Treated Means Control SD Control Mean Diff eQQ Med
## distance                0.3284        0.2837     0.0942    0.0447  0.0468
## bad_peers              11.6567       10.5456     4.7587    1.1111  1.0000
## as.factor(male)0        0.3752        0.5358     0.4988   -0.1606  0.0000
## as.factor(male)1        0.6248        0.4642     0.4988    0.1606  0.0000
## age                    14.9651       15.0633     1.4174   -0.0982  0.0000
##                  eQQ Mean eQQ Max
## distance           0.0448  0.0675
## bad_peers          1.1139  2.0000
## as.factor(male)0   0.1604  1.0000
## as.factor(male)1   0.1608  1.0000
## age                0.0975  1.0000
##
##
## Summary of balance for matched data:
##                  Means Treated Means Control SD Control Mean Diff eQQ Med
## distance                0.3267        0.3261     0.0942    0.0006       0
```

Running the Match

Summary is far too large to fit on a single page - here's a picture:



```
Summary of balance for all data:
                    Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance                   0.3284        0.2837     0.0942    0.0447  0.0468   0.0448  0.0675
bad_peers                 11.6567       10.5456     4.7587    1.1111  1.0000   1.1139  2.0000
as.factor(male)0           0.3752        0.5358     0.4988   -0.1606  0.0000   0.1604  1.0000
as.factor(male)1           0.6248        0.4642     0.4988    0.1606  0.0000   0.1608  1.0000
age                       14.9651       15.0633     1.4174   -0.0982  0.0000   0.0975  1.0000


Summary of balance for matched data:
                    Means Treated Means Control SD Control Mean Diff eQQ Med eQQ Mean eQQ Max
distance                   0.3268        0.3261     0.0943    0.0007       0   0.0008  0.0042
bad_peers                 11.5889       11.6240     4.9380   -0.0351       0   0.0804  1.0000
as.factor(male)0           0.3771        0.3771     0.4848    0.0000       0   0.0000  0.0000
as.factor(male)1           0.6229        0.6229     0.4848    0.0000       0   0.0000  0.0000
age                       14.9626       14.9958     1.3988   -0.0332       0   0.0332  1.0000

Percent Balance Improvement:
                    Mean Diff. eQQ Med eQQ Mean  eQQ Max
distance               98.4772     100  98.2177  93.8413
bad_peers              96.8402     100  92.7817  50.0000
as.factor(male)0      100.0000       0 100.0000 100.0000
as.factor(male)1      100.0000       0 100.0000 100.0000
age                    66.1671       0  65.9111   0.0000

Sample sizes:
          Control Treated
All          6316    2668
Matched      2649    2649
Unmatched    3667      17
Discarded       0       2
```

Running the Match

Just highlighting some obvious differences in text (I'll discuss in lecture)

1) The *distance* variable is the average propensity score - it's reassuring that it's higher for the **treated** group that actually reports being delinquent.

2) All matching variables exhibit small or large differences across treated and control samples. The most pronounced difference is the proportion of males in the treated group (0.6248) compared to the proportions of males in the untreated group (0.4642).

3) These differences are made substantially smaller in the matched sample indicating we have achieved **covariate balance**.

4) Our sample is reduced to 2,649 matched pairs, from an original sample of 6,316 control cases and 2,668 treatment cases.

Evaluating the Model

One of the nice things about the *MatchIt* package is that both covariate balance and common support can be quickly evaluated using the **plot()** command and the stored matching model (here, the *match1* object we created above).

First, I will show you a series of covariate balance plots, then a histogram plot comparing propensity score distributions across the samples.

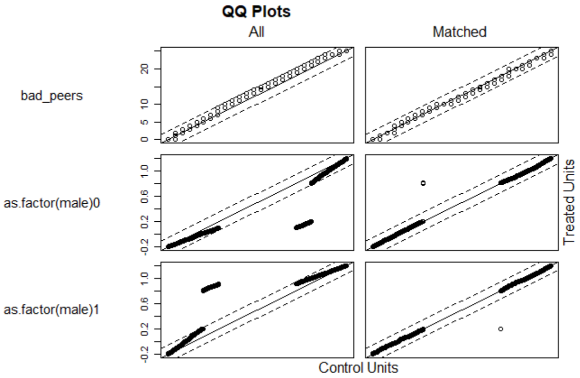
Evaluating Covariate Balance

```
plot(match1)
```

I do not print the plot here because it's too large to fit on a single slide without editing it by hand in another program.

The following slide includes the plot.

Evaluating Covariate Balance



Evaluating Covariate Balance

With the degree that the plot has been squished to fit it's difficult to make out the differences but I'll point out a few things.

1) We want the observations to align with the 45 degree line in each plot. Departues from perfect alignment indicate **covariate imbalance**.

2) The left column includes the full sample, while the right column includes just the matched pairs sample. Covariate **imbalance** in the left column should be corrected in the right column if the propensity score matching procedure was successful.

3) Consistent with the summary results above, this does appear to be true, particularly for the gender matching variable.

Evaluating Common Support

We can also plot histograms of the propensity scores across the groups to evaluate the **common support** assumption.

The next slide demonstrates how to accomplish this using the type=“hist” option within the **plot()** command.

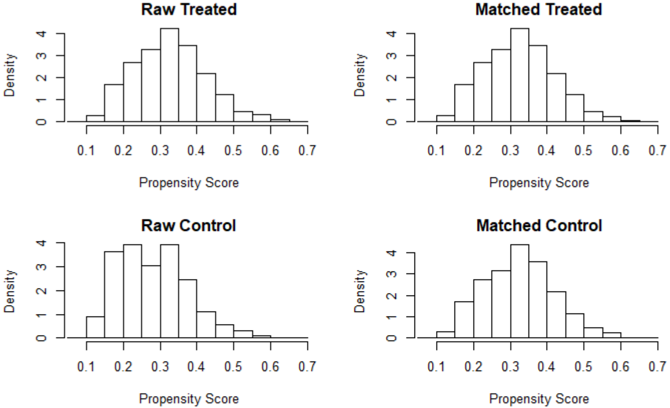
Evaluating Common Support

```
comm_supp<-plot(match1, type="hist")
```

I exclude the printout of the figure here since it's far too large to fit into the slide as is (fits much better into a pdf or html file, though).

Evaluating Common Support

And here's the plot:



Evaluating Common Support

Two things of note:

1) Notice the discrepancies in the left column between the treated and control cases in terms of their propensity score distributions. It's not apparent that the distributions are **very** different, but it is clear that treated youth have slightly higher probabilities of reporting delinquency in 1998 than does the control sample of youth.

2) The right column displays the propensity score distributions **after** we have excluded unmatched cases. Notice that the distributions are much more similar now than they were pre-matching. This is additional evidence that our propensity score matching procedure worked as intended.

Estimating Treatment Effects

The end goal of a propensity score matching procedure is to achieve **conditional independence** for treatment assignment and, thereby, to be able to evaluate outcomes as if assignment to treatment was actually random.

Typically, outcomes after a propensity score procedure are evaluated using t-tests, also known as mean comparison tests, which tell us whether group averages in outcomes are **significantly** different or not.

To accomplish this with the *MatchIt* package we need to use another function: **t.test()**

Estimating Treatment Effects

Step 1: Create a matched samples data frame:

```
matched<-match.data(match1)
```

For this step, we use the *match1* object we created above with the **match.data()** function.

This function creates a new data object that just includes the matched pairs sample, not the full data set. We will need the reduced data file for following steps.

Estimating Treatment Effects - Structure of t.test() Command

The structure of the **t.test()** command is as follows:

t.test(formula=outcome~treatment)

Where the outcome is the post-treatment numeric variable *drug_use99* and treatment is the dichotomous indicator for delinquency in the 1998 interview (*delinq98*).

Estimating Treatment Effects - Structure of t.test() Command

We will run two versions of this command - once for the full sample and another time for the matched sample.

The degree to which those t-test estimates differ indicates the amount of bias present in the initial treatment effect estimate for the full sample.

Estimating Treatment Effects

Step 2: Estimate the t-test for the full sample.

```r
with(NLSY97HW1, t.test(drug_use99~delinq98))
```

```
##
##  Welch Two Sample t-test
##
## data:  drug_use99 by delinq98
## t = -9.8107, df = 3611.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.917916 -4.613425
## sample estimates:
## mean in group 0 mean in group 1
##        2.046548        7.812219
```

Estimating Treatment Effects

```
with(matched, t.test(drug_use99~delinq98))
```

```
##
##  Welch Two Sample t-test
##
## data:  drug_use99 by delinq98
## t = -7.1967, df = 5064.6, p-value = 7.069e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.432184 -3.678088
## sample estimates:
## mean in group 0 mean in group 1
##        2.778323        7.833459
```

Estimating Treatment Effects

Our *full sample* treatment effect estimate is equivalent to **7.81-2.04 = 5.77**.

This implies that youth who report engaging in at least one type of delinquency in the 1998 interview use cocaine, marijuana, or drink excessively almost 6 days a month more than youth who did not report being delinquent in the 1998 interview.

Our *matched* treatment effect estimate is equivalent to **7.83-2.78 = 5.05**.

This implies that youth who report engaging in at least one type of delinquency in the 1998 interview use cocaine, marijuana, or drink excessively slightly more than 5 days a month more than youth who did not report being delinquent in the 1998 interview.

Estimating Treatment Effects

The difference in those estimates is **0.72 days**.

If this weren't a simple example, I might conclude that the original estimate was not *that* biased, but since the number of matching variables we used was very small, this likely just means we did a poor job of estimating the propensity score.

The End

That's it. . . .no more from me today.

Be well!