POLS3004 Causal Analysis: Tutorial Exercise 5

We'll use data from LaLonde's evaluation of economic training programs in the United States. The study used observational data from a treatment group of people who took part in the National Supported Work Demonstration (NSW), a job training program in the mid 1970s. The comparison group is a sample of people who did not go through training, taken from the Current Population Survey (CPS). Treatment assignment was not random. What can we learn from the data?

We get our data from the MatchIt package:

```
install.packages(''MatchIt'')
library(MatchIt)
ldata <- MatchIt::lalonde
```

Also, install and load the `Matching` package.

The dataset contains:

- N - 614 observations (185 treated, 429 control)
- *treat* - 1 = treated, 0 = control
- *age* - age in years
- *educ* - number of years of education
- *black* - 1 = African American, 0 otherwise
- *hispan* - 1 = Hispanic origin, 0 otherwise
- *married* - 1 = married, 0 otherwise
- *nodegree* - 1 = no high school degree, 0 otherwise
- *re*74 = income in 1974 (dollars)
- *re*75 = income in 1975, in U.S. dollars (dollars)
- *re*78 = income in 1978, in U.S. dollars (dollars) [the main outcome variable]

a) Regress income in 1978 on the treatment, controlling for *hispan*, *educ* and *married*. What is the estimated Average Treatment Effect of training? Is it statistically significant?

Code:

```
m1 <- lm(re78 ~ treat +hispan + educ + married, data = ldata)
summary(m1)
```

Answer:
The ATE is $300.04 and it is not significant at conventional levels (standard error=657.23).

b) Now, carry out exact matching to find the effect of treatment on income in 1978, matching on educ, hispan and married. Report the ATT and its standard error. Does the result differ from part (a)?

Code Hints:

- Use the `Match()` function, and look up the help file to discover how to do exact matching
- Supply `Match()` with the arguments Y (the outcome variable), TR (the treatment variable) and X (a matrix of the variables used for matching, created using `cbind()`)
- `summary.Match()` will summarise the results

Code:

```
matching.vars1 <- cbind(ldata$hispan,ldata$educ, ldata$married)
m1 <- Match(Y=ldata$re78, Tr=ldata$treat, X=matching.vars1, exact=TRUE)
summary.Match(m1)
```

Answer:
The ATT is $447.89 and its standard error is 725.92. The result is similar to part (a), but not exactly the same. This is partly because we are now focusing on a smaller number of observations.

c) Examine balance between treated and untreated units before and after matching, using t-tests for equality of means. How does balance differ before and after matching?

Code Hints:

- Use the `MatchBalance()` function. It requires two main arguments.
- The first is a formula in the format: `treatment` ~ `matchingvar1 + matchingvar2 + ...`
- The second is the name of your matching estimate in part (b)

Code:

```
mbal1  <- MatchBalance(treat ~ hispan + educ + married, match.out=m1, data=ldata)
```

Answer:
For two of the three variables, there are large and statistically significant imbalances in the

unmatched sample. 6% of the treated units are hispanic compared to 14% of the controls, and 19% of treated units are married compared to 51% of control units. Balance is better on education, though. In the matched sample, balance is completely perfect across all three variables because we are using exact matching. With only three variables, it is relatively easy to find exact matches in a fairly large sample like this one.

d) Now repeat parts (b) and (c) with the addition of re74 and re75 to the list of matching variables, using the Mahalanobis distance metric instead of exact matching. How good is balance after matching? Does the estimated ATT change?

Code Hint:

- Look at the help file to find out how to select the Mahalanobis distance metric

Code:

```
matching.vars2 <- cbind(ldata$hispan,ldata$educ, ldata$married, ldata$re74, ldata$re75)
m2 <- Match(Y=ldata$re78, Tr=ldata$treat, X=matching.vars2,Weight=2)
summary.Match(m2)
mbal2  <- MatchBalance(treat ~ hispan + educ + married + re74 + re75, match.out=m2,
data=ldata)
```

Answer:

The estimated ATT is now $697.12 with a standard error of 962.76. Excellent balance is achieved after matching, except on re75, where a significant difference of around $238 remains between treated and untreated units. Nonetheless, this is much smaller than the difference in the unmatched sample.

e) Estimate propensity scores for all units using a logistic regression of treatment on all variables except re78, and add them to the dataset as a new column

Code Hints:

- Remember that a propensity score is the estimated probability of treatment. This can be calculated using the `fitted.value()` function

Code:

```
propscores.reg <- glm(treat~age+educ+black+hispan+married+nodegree+re74+re75,
family=binomial, data=ldata)
propscores <- fitted.values(propscores.reg)
ldata <- cbind(ldata,propscores)
```

f) Repeat part (b) and (c) again, this time using only your estimated propensity scores to match. How good is balance after matching? Does the estimated ATT change?

Code:

```
m3 <- Match(Y=ldata$re78, Tr=ldata$treat, X=propscores,Weight=2)
summary.Match(m3)
mbal3  <- MatchBalance(treat ~ age+educ+black+hispan+married+nodegree+re74+re75,
match.out=m3, data=ldata)
```

The estimated ATT is now $1933.40 with a standard error of 1090.60, making the effect significant at the 10% significance level. Balance after matching is in general very good, based on the t-test results. There are statistically significant differences for *age* and *married*, but in practical terms the differences are small (and smaller than in the unmatched sample).

In general this question demonstrates the power of selection bias. In the unmatched sample (part a), the training program has a very small effect that is statistically indistinguishable from zero. This is probably due to people with low potential outcomes selecting into training. As we get closer to comparing "apples with apples" in part (f), we can see that training probably has a substantial positive effect. This is only visible when we compare the treated group to controls who were as similar as possible to the treatment group in every way except for taking the treatment.

g) An alternative to the matching estimator is to carry out matching, and then run a regression on the matched dataset. Do this using the following steps:

   i) Estimate the ATT from matching as in part (f), this time with the option `ties=FALSE`. Your estimate will be similar but not identical to (f)[1]

   ii) Create a dataset of only the observations used in matching in (i).
   Code Hint: You can extract the rows used from the dataset by adding `$index.treated` and `$indexcontrol` to the name if your matching estimate

   iii) Run a regression using only your dataset from (ii).

How similar is your regression in (iii) to the matching estimate in (i)?

Code:

```
m4 <- Match(Y=ldata$re78, Tr=ldata$treat, X=propscores,Weight=2,ties=F)
summary.Match(m4)

ldata.tr <- ldata[m4$index.treated,]
ldata.con <- ldata[m4$index.control,]
```

[1]This makes things less complicated for the sake of this example: without `ties=FALSE`, the matching estimator includes all ties and weights tied observations. To get a similar regression estimate, we would also have to weight our regression. With `ties=FALSE`, R instead breaks the ties at random, selecting only one of the tied observations, so our regression does not need to be weighted. This also means that everyone will get slightly different answers to this question, depending on how R randomly breaks the ties. In general it is more principled to use `ties=TRUE` since we do not arbitrarily throw away data in that scenario. Note that it is also possible to reduce the number of ties using the `distance.tolerance` option

```
ldata.reg <- rbind(ldata.tr,ldata.con)

summary(lm(re78 ~ treat ,data=ldata.reg))
summary(lm(re78 ~ treat + propscores ,data=ldata.reg))
```

Everyone will get a slightly different answer here due to the randomness of tie-breaking. But you will find that the two estimates are identical if you did not include controls. This makes sense because a regression with only a single binary variable is the same as a difference in means (see week 1). With controls, the two estimates will be very similar but not identical. Including controls can be helpful, just like in an experiment, if some imbalances remain after matching.