Matching and Weighting in R Exercises

Statistical Horizons

Stephen Vaisey

Instructions

For these exercises, we are going to use one of the versions of the "Lalonde data," which is used in almost every paper on matching.[1] This is data on a job training program (the treatment) that was intended to raise future earnings (the outcome).

If you have made a project folder with the course materials in it, you can load the data by typing `load("exercise_data.Rdata")`. This will bring two objects into the global environment: `d_exper`, which is the experimental subset of the data and `d`, which comprises the treated cases and a sample of observational controls from the PSID. The treatment is `treat` and the outcome is `re78`, which is income in $1000s. We are going to use the experimental subset to set an experimental benchmark and then see how close we can get to this benchmark using various matching and weighting methods. The rest of the variables are as follows:

Variable	Description
`age`	Age in years
`educ`	Years of education
`black`	1 = Black; 0 otherwise
`hisp`	1 = Hispanic; 0 otherwise
`married`	1 = married; 0 otherwise
`nodegr`	1 = no degree; 0 otherwise
`re74`	1974 income in $1000s
`re75`	1975 income in $1000s
`u74`	1 = no '74 income; 0 otherwise
`u75`	1 = no '75 income; 0 otherwise

To make sure you have everything loaded that you need, put the following commands at the top of your Rmd file. If you're using the tidyverse commands, make sure to put `tidyverse` last, or other commands might mask `dplyr::select`.

```
library(survey)
library(broom)
library(cobalt)
library(MatchIt)
library(WeightIt)
library(tidyverse)
theme_set(theme_minimal()) # optional
```

Before starting the exercises, you may want to consider a few things that will make your life easier:

[1] Lalonde, R. (1986). "Evaluating the econometric evaluations of training programs with experimental data." *American Economic Review* 76: 604-620.

- add a factor version of the treatment to the data frame for easy plotting
- create formula objects that contain the propensity score (or matching) models with and without quadratic terms

You can get by without doing these steps but they avoid extra typing down the road.

You will begin by looking at the experimental data (`d_exper`). After that, you will conduct various forms of matching and weighting on the observational data (`d`). For each exercise *after the first four*, your basic workflow will be:

1. Match or weight, as directed
2. Check balance (overall, if applicable and by covariate) using graphical and numeric means
3. Estimate the ATT

Exercises

1. Use the experimental data to estimate the effect of the job training treatment. How much does it appear to affect 1978 income? Now look at the observational data (for all exercises from now on). How large is the raw difference in 1978 income between the treatment group and the PSID comparison group?

2. Try to estimate the effect of the treatment using regression. What does regression say the effect of the program is?

3. Begin by exact matching on all the dummy variables. How many treated cases cannot be matched? What is the (FS)ATT estimate?

4. Use the observational data to estimate each case's propensity to receive treatment using `glm()`. Use a logistic regression with quadratic terms for age, education, 1974 income, and 1975 income. Spend a few moments thinking about what this model says. If you are familiar with plotting in R, look at the density plots of the p-score for treated and untreated groups. (If not, you can move on. We'll do the same thing using `bal.plot()` in a bit.)

5. Conduct 1:1 nearest-neighbor matching on the log odds of the propensity score. Use `bal.plot()` to compare the overall propensity score distributions. Do once without replacement and once with replacement. Why do you think there's a difference? Try to figure it out. Estimate the ATT for each assumption (i.e., with or without replacement). If you achieve good overall balance on the propensity score, try checking individual covariate balance using `love.plot()`.

6. Estimate propensity scores and ATT weights using `weightit()`. Ignore the warning you get. We'll discuss that more in class. Estimate the ATT. Check for covariate balance.

7. Now do the same as above using covariate balancing propensity scores.

8. Try Mahalanobis distance matching with replacement and using a caliper of .1. How many unique control cases get matched?

9. Use entropy balancing to balance treatment and control. Confirm that you've achieved balance on the means and the variances of the covariates.

10. Now revisit questions 3 and 5-9. This time, instead of just using simple regressions to estimate the ATT, estimate full outcome regressions using the dataset you "preprocesssed" with matching or weighting. How does this affect the estimates?

11. **Bonus:** implement a bootstrap of your preferred estimate. What is the bootstrapped standard error?