002: Regression and matching

EC 607

Due *before* midnight on Tuesday, 19 May 2020

DUE Your solutions to this problem set are due *before* 11:59pm on Tuesday, 19 May 2019 on Canvas.

Your problem set **must be typed** with R code beneath your responses. *E.g.*, `knitr` and `R Markdown`.

OBJECTIVE This problem set has three purposes: (1) reinforce the econometrics topics we reviewed in class; (2) build your R toolset; (3) start building your intuition about causality within econometrics.

README This problem set uses data from LaLonde (1986), who compared the estimated effects of a randomized employment program—National Supported Work Demonstration (NSW)—to the estimated effects produced using non-experimental methods (*i.e.*, pretending treatment had not been randomized). You should read (at least the first few pages of) the paper. More here from Rajeev Dehejia.

01. Download and load two datasets: (1) data from the randomized employment program (we'll call this the **NSW data**) and (2) data on 2,490 potential 'control' individuals from the PSID (Panel Study of Income Dynamics) (we'll call this the **PSID data**).

The last page of the problem set describes the variables in these data.

Note Questions 02–07 use the *NSW data*.

02. Regress real earnings in 1975 (the year before treatment) on treatment (and an intercept, which we will always assume should be included unless otherwise stated). Why/how is this regression (and its outcome) informative? What does it tell us?

03. The program rolled out in 1976 and ended (at least for our purposes) in 1978, so we'll use earnings in 1978 to estimate whether the program had any sustained effect on earnings.

Regress 1978 earnings on treatment. What do you find?

04. What is required for us to interpret the estimated in **03.** as causal? Does our setting meet this requirement?

05. Add controls for age, education, race (black and Hispanic). How does your estimated treatment effect and its standard change. Why do you think this happense?

06. What is a "bad control"? Are any of the controls we added in **05.** "bad"? Briefly explain.

07. Since we have an experiment, can we interpret the coefficient on `nodegree` (not having a high-school diploma) as causal? What about its interaction with treatment? Briefly explain.

08. Compare a simple difference in means to your results in the regression results in **03.**.

Hint The `dplyr` functions `group_by()` and `summarize()` could be helpful.

09. Create a new dataset that combines **treated individuals from the *NSW data*** and **control individuals from the *PSID data***. We'll refer to this dataset as our **mixed dataset**.

Hint Remember our old friends `filter()` and `bind_rows()` from `dplyr`.

Note Questions from 10–13 use this *mixed dataset*, focusing on earnings in 1978.

10. Compare the difference in means from the **mixed dataset** to the difference from the **NSW dataset**.

11. Use our potential-outcomes (Rubin causal model) notation to explain how the difference with the mixed dataset may be biased. Does the sign of the difference across the two differences-in-means match what you would expect from our model of selection bias? Briefly explain.

12. Time for nearest-neighbor matching. Use all six covariates. **You must write write the code for this matching yourself.** You *can* use basic pre-written functions (*e.g.*, `mean`), but do not use matching-estimation packages that do all of the matching for you.

> **12A.** Estimate the average treatment effect on the treated by matching treated individuals to their nearest neighbor using a **Euclidean** metric.
>
> **12B.** Estimate the average treatment effect on the treated by matching treated individuals to their nearest neighbor using a **Mahalanobis** metric. *Hint:* The `StatMatch` package has a function named `mahalanobis.dist()` that finds the *Mahalanobis distance* between observations of two datasets.
>
> **12C.** How do your estimates in **12A.** and **12B.** compare to your previous estimates?
>
> **Extra credit** Use kernel matching (any kernel) to estimate the treatment effect.

13. Now for propensity-score methods. **Again: Do not use pre-written propensity-score packages. Write the code for the steps.** You *can* use the `glm()` function to estimate a logit regression (more below in **13A**).

> **13A.** Estimate the propensity score for each treated individual using the covariates using a logit model that is linear in the covariates. Which variables are predictive of treatment?
>
> *Hint:* The function `glm()` with `family = binomial` estimates a logit model.
>
> **13B.** Add the estimated propensity scores ($\hat{p}_i$) to the mixed dataset. Is there overlap? Explain.
>
> *Hint:* You can access predictions from a model using `$fitted.values`.
> *Another hint:* Try histograms grouped/filled by treatment status.
>
> **13C.** Enforce overlap using the minimum p_i observed in the treated group and the maximum p_i observed in the control group.
>
> **13D.** Estimate the treatment effect using a regression that conditions on p_i. What happens if you also include p_i interacted with treatment?
>
> **13E.** Now estimate the treatment effect by blocking on p_i.
>
> **Extra credit** Use the *doubly robust method* that combines regression and blocking.

14. Compare the various treatment effects that you've estimated in **10–13**. How do they compare to the effects you estimated **03.**? Which estimates should we trust? Why?

Data description

Variable	Description
data_id	Dataset identifier.
treat	Treatment indicator (select to be part of NSW).
age	Age (years).
education	Education (years).
black	Indicator for whether the individual is black.
hispanic	Indicator for whether the individual is Hispanic.
married	Indicator for whether the individual is married.
nodegree	Indicator for individuals without a high-school diploma.
re74	Real earnings in 1974 (1982 dollars).
re75	Real earnings in 1975 (1982 dollars).
re78	Real earnings in 1978 (1982 dollars).

Note: The NSW dataset does not include re74.