Econometrics 1

TD 7: linear regressions and causality

(Chapter 4, Section 1)

In a famous article,[1] Robert LaLonde investigated the effect of the National Supported Work Demonstration (NSWD) program, which took place in 1976 and 1977, using experimental data (that is, observations generated by some randomized experiment).

The purpose of this exercise is to replicate some of his results and to test the validity of the experiment using, among other things, simple linear regressions. The file nsw.dta (available on Pamplemousse) will be used to this end.
The names of the variables are generally explicit but, for all practical purposes :

- `treat` = 1 if the individual receives the treatment, 0 otherwise;
- `nodegree` = 1 if the individual does not have a degree, 0 otherwise;
- `re75` and `re78` are the annual incomes in 1975 and 1978, respectively.

We denote D_i the treatment variable of the individual i. We are interested in the effect of the treatment on the 1978 income: Y_i is the *observed* 1978 income of an individual i while $Y_i(d)$ corresponds to his *potential* income in 1978 in the situation where i receives the treatment $d \in \{0, 1\}$; remember that $Y_i = Y_i(D_i)$.

1. How should we interpret $Y(0)$ and $Y(1)$ in this context? To which variable in the database does Y correspond to?
2. Describe the sample and perform some descriptive statistics.
3. Compute $\overline{Y}_1 - \overline{Y}_0$, where we let $\overline{Y}_d = \sum_{i:D_i=d} Y_i/n_d$ and n_d denotes the size of the subsample $\{i : D_i = d\}$. Show, under an assumption to be specified, that this difference estimates without bias a causal parameter to be also specified.
4. Perform a regression of Y on D. Was the result on the coefficient expected? Has the treatment a significant effect (at the 5% level) on income?

[1] R. J. LaLonde (1986). "Evaluating the econometric evaluations of training programs with experimental data". *The American Economic Review*, 76 (4), 604-620.

5. How can we test that the assignment to treatment was indeed made randomly? Perform these tests and comment them.

6. Run a regression of Y on 1975 income and then on `nodegree`. Comment on the results. Do the results challenge the assumption posited in Question 3 ?

We now append data from the whole US population using the PSID data[2] (psid_controls.dta). The idea is to see what we would obtain if we used individuals from the PSID as a "control group", instead of using the control group from the NSWD experiment.

7. Merge the two datasets, the one from LaLonde experimental sample (nsw.dta) and the one from the whole US population (psid_controls.dta). Describe the variables and compute descriptive statistics according to the sample source. What do you remark? Could you argue the LaLonde sample is representative of the whole US population? If not, how would you describe that sample's characteristics?

We now drop the control observations from the experimental data (that is, the observations from LaLonde sample that were not treated: the individuals who did not follow the NSWD program).

8. With the remaining observations, compute an estimate of the treatment effect using a simple linear regression. What do you find? Explain the result.

9. How can you improve the previous estimate using a multiple linear regression? Under which assumptions would the regression identify a causal effect?

10. Perform such a regression. Comment on the obtained estimates and standard errors. Draw conclusions about your previous assumptions and the relevance of an experimental approach.

[2] The Panel Study of Income Dynamics (PSID) is the longest-running longitudinal household survey in the world. The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics. The PSID is directed by faculty at the University of Michigan, and the data are available on this website without cost to researchers and analysts.