Ec141, Spring 2019

Professor Bryan Graham

Problem Set 2

Due: February 26th, 2019

Problem sets are due at 5PM in the GSIs mailbox. You may work in groups, but each student should turn in their own write-up (including a narrated/commented and executed Jupyter Notebook). Please use markdown boxes within your Jupyter notebook for narrative answers to the questions below.

1 Using inverse probability weighting (IPW) to evaluate the returns to a college degree

The purpose of this problem set is to give you practice using the basic covariate adjustment methods introduced in lecture. You might find the following papers useful. The paper "Improving middle school quality in poor countries: evidence from the Honduran *Sistema de Aprendizaje Tutorial*" (McEwan et al., 2015) provides an example of inverse probability weighting in action. You should also review your lectures notes and read Holland (1986), Efron & Hastie (2016, Chapter 8) and Hirano & Imbens (2001).

Overview of dataset

This problem set uses the comma delimited dataset `nlsy79extract.csv`; available on the course GitHub page. The dataset includes information on 12,686 youth surveyed as part of the National Longitudinal Survey of Youth 1979 (NLSY79). In this problem set you will use the following variables:

`core_sample` – indicator for whether individual is part of the core NLSY70 sample

`year_born` – year in which individual was born

`live_with_mom_at_14` - dummy variable indicating whether individual resided with their mother at age 14

`live_with_dad_at_14` – dummy variable indicating whether individual resided with their father at age 14

`usborn` - dummy variable indicating whether individual was born in the United States

`male` - male/female dummy variable

`hispanic` – hispanic/non-hispanic dummy variable `black` – black/non-black dummy variable

`AFQT` – Armed Forces Qualification Test (AFQT) score percentile (0 to 100)

`HGC_Age28` – years of completed school at age 28

`HGC_Fath79` – father's years of completed schooling

`HGC_Moth79` – mother's years of completed schooling

`real_earnings_xxxx` – "xxxx" real earnings in 2010 prices (available for multiple years)

Preparing the dataset

1. Load the dataset into a pandas dataframe called `nlsy79`. Use `HHID_79` and `PID_79` as the multi-indices for the dataframe.

2. Drop any cases where `core_sample` equals zero.

3. Drop any units where `male` equals zero.

4. Drop any units where `year_born` *is not* 61, 62 or 63.

5. Create a dummy variable called `college` which equals 1 if `HGC_Age28` is *greater than or equal* to 16 and zero otherwise. Next drop any units where `HGC_Age28` is *less than* 12.

6. Create a variable called `earnings_in_2000` which is the average of `real_earnings_1997, real_earnings_1999, real_earnings_2001, real_earnings_2003.` When computing this variable average over all non-missing values; for example if earnings is observed in just two of the four years listed above, average over the two years it is observed.

7. Drop all variables except `live_with_mom_at_14`, `live_with_dad_at_14`, `usborn`, `hispanic`, `black`, `AFQT`, `HGC_Fath79`, `HGC_Moth79`, `college` and `earnings_in_2000`.

8. Finally retain only complete cases (you can use "dropna()" for this).

9. Those units that remain constitute your estimation sample. Use "describe()" to print out some basic summary statistics for your estimation sample. Write a short paragraph about your dataset.

Estimating the average earnings premium due to college attendance

A somewhat dated, although still useful, blog post on logistic regression using Python and statsmodels is available online at `http://blog.yhat.com/posts/logistic-regression-python-rodeo.html`. You might find this post useful for completing this portion of the problem set.

1. Compute the logistic regression of `college` onto a constant and the other variables in your dataset (except for `earnings_in_2000`).

2. Compute the fitted values ("propensity scores") associated with your regression fit, $\hat{e}\left(X_i\right)$ for $i = 1, \ldots, N$. Is the overlap condition satisfied? Why? Present *graphical* evidence for your answer.

3. Compute the IPW weights for average treatment effect (ATE) estimation as described in lecture and also Hirano & Imbens (2001). Compute the weighted least squares fit of `earnings_in_2000` onto a constant and `college` using these weights (this is a computational device which computes the IPW estimator described in lecture; you may use the WLS procedure in statsmodels for this step). Interpret the coefficient on `college.`

4. Use the bootstrap procedure described in lecture to construct a confidence interval for the ATE. Use at least 500 bootstrap samples.

5. Discuss your results. Is the selection on observables assumption reasonable? Why or why not? What additional data would you collect to improve your analysis?

References

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference*. Cambridge: Cambridge University Press.

Hirano, K. & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4), 259 – 278.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945 – 960.

McEwan, P. J., Murphy-Graham, E., Irribarra, D. T., Aguilar, C., & Rápalo, R. (2015). Improving middle school quality in poor countries: evidence from the honduras sistema de aprendizaje tutorial. *Educational Evaluation and Policy Analysis*, 37(1), 113 – 137.