Homework 4: Censoring and Panel Data

Due on April 20th 2022 at 11 PM

Your final output should consist of raw code and a pdf file with answers. Output should be uploaded on Github before the due date/hours. For this assignment,

April 6, 2022

Data

This assignment is based on data from the NLSY97.[1]

Variable Descriptions

"dat_A4.csv" is a dataset extracted from the NLSY97 with particular columns listed as follows. For further information about the variables, please refer to the documentation file "dat_A4/variables_doc.pdf".

[1]The NLSY97 consists of a nationally representative sample of 8,984 men and women born during the years 1980 through 1984 and living in the United States at the time of the initial survey in 1997. Participants were ages 12 to 16 as of December 31, 1996. Interviews were conducted annually from 1997 to 2011 and biennially since then. The ongoing cohort has been surveyed 18 times as of date. Data are available from Round 1 (1997-98) through Round 19 (2019-20). The NLSY97 collects extensive information on respondents' labor market behavior and educational experiences. The survey also includes data on the youths' family and community backgrounds to help researchers assess the impact of schooling and other environmental factors on these labor market entrants. Data from the NLSY97 also aid in determining how youths' experiences relate to establishing careers, participating in government programs, and forming families. Finally, information from the NLSY97 allows researchers to compare the progress of this cohort with that of other NLS cohorts (See Cross-Cohort NLSY79/97 Data Harmonization). Please see https://www.bls.gov/nls/nlsy97.htm for more information.

Exercise 1 Preparing the Data

- Create additional variable for the age of the agent "age", total work experience measured in years "work_exp". *Hint:* "CV_WKSWK_JOB_DLI.01" denotes the number of weeks a person ever worked at JOB 01.
- Create additional education variables indicating total years of schooling from all variables related to education (eg, "BIOLOGICAL FATHERS HIGHEST GRADE COMPLETED") in our dataset.
- Provide the following visualizations.
  - Plot the income data (where income is positive) by i) age groups, ii) gender groups and iii) number of children
  - Table the share of "0" in the income data by i) age groups, ii) gender groups, iii) number of children and marital status
  - interpret the visualizations from above

Exercise 2 Heckman Selection Model

Using the variables created above, estimate the following models.

- Specify and estimate an OLS model to explain the income variable (where income is positive).
  - Interpret the estimation results
  - Explain why there might be a selection problem when estimating an OLS this way
- Explain why the Heckman model can deal with the selection problem.
- Estimate a Heckman selection model (Note: You can not use a pre-programmed Heckman selection package. Please write down the likelihood and optimize the two-stage Heckman model). Interpret the results from the Heckman selection model and compare the results to OLS results. Why does there exist a difference?

Exercise 3 Censoring

Note that the "YINC-1700" variable is censored because of privacy issues. In other words, high wages are top-coded in this data set.

- Plot a histogram to check whether the distribution of the income variable. What might be the censored value here?
- Propose a model to deal with the censoring problem.

- Estimate the appropriate model with the censored data (please write down the likelihood function and optimize yourself without using the pre-programmed package)
- Interpret the results above and compare to those when not correcting for the censored data

In the second part, we use the panel dimension of NLSY97 data.

Exercise 4 Panel Data

Some variables used in previous exercises such as marital status, how many weeks of experience in each job, highest degree ever received, total income are selected in the new "dat_A4_panel.csv" dataset. We now have the panel structure that includes 19 rounds of survey with these variables. Variable descriptions can be found in the "dat_A4_panel/variables_doc.pdf". We are interested in the effect of education, marital status, experience and education on wages.

- Explain the potential ability bias when trying to explain to understand the determinants of wages
- Exploit the panel dimension of the data to propose a model to correct for the ability bias. Estimate the model using the following strategy.
  - Within Estimator.
  - Between Estimator
  - Difference (any) Estimator
- Interpret the results from each model and explain why different models yield different parameter estimates