Data Task for Beshears, Choi, and Laibson Research Assistantship

Updated September 2021

1: Instructions

Time limit: 6 hours

Please answer the questions in a separate document that clearly indicates:

- Question number
- Briefly, the steps taken to arrive at the answer
- Any figures or tables should be inserted directly in the document near corresponding text

Also, include (separately) any code used to generate the responses, where applicable.

As you work, imagine that you will need to share your findings with the faculty, such as during a team meeting. Be sure to take note of any methods, results, or questions you would want to discuss with the team.

2: Introduction

Imagine that we wish to assess the impact of the 2008 financial crisis on the labor supply choices of recent graduates. We think that attending college during a recession may cause students to enter work at different rates compared to students who did not experience a recession during university. In other words, students who graduated in the years before 2008 may have different labor force participation trajectories compared to students who graduated after 2008. We decide to compare these two cohorts to identify the effect that the 2008 financial crisis had on the labor trajectories of new graduates.

We will use panel data that allow us to trace an individual's employment status across time. In addition to comparisons across cohorts, we will also consider an individual's age as well as secular trends in employment that are unrelated to the financial crisis.

3: Empirical approach

In our dataset (described below), we observe employment status (in several broad categories, including working, student, unemployed, etc.) for each individual and survey year. We have assumed that individuals leave the education system when their marginal benefit of human capital falls below their opportunity cost of education. Thus, as individuals age, and potentially acquire more schooling, they are more likely to be in the workforce in the following year. We model this by including an individual's age.

This model suggests the following empirical specification:

$$Y_{it} = \beta_0 + \beta_1 cohort_i + \beta_2 age_{it} + \beta_3 grad_{it} + \beta_4 age_{it} \times grad_{it} + \beta_5 age_{it} \times cohort_i + \beta_6 cohort_{it} \times grad_{it} + \beta_7 age_{it} \times grad_{it} \times cohort_i + \gamma_t + \varepsilon_{it}, \quad (1)$$

where Y_{it} is a labor-supply related outcome variable for individual i in year t, cohort is the individual's

cohort (pre or post), $grad$ takes value 1 if an individual has graduated (age > 22), and 0 otherwise.[1] γ_t denote year fixed effects.

Under some assumptions, we can simplify the above model to:

$$Y_{it} = \delta_0 + \delta_1 cohort_i + \delta_2 age_{it} + \delta_3 age_{it} \times grad_{it} \times cohort_i + \gamma_t + \mu_{it} \quad (2)$$

QUESTION 1: What are the assumptions made to simplify the model? Discuss in terms of the interpretation of the original coefficients. Is there an assumption that you expect may not hold? If so, choose one assumption and explain why it may not hold.

QUESTION 2: What is the coefficient of interest? What is its interpretation? Give an expression for the cumulative net effect of being in the post-recession cohort, comparing the effect 2 years after graduation to 4 years after graduation.

QUESTION 3: What assumption(s) are necessary to interpret the coefficient of interest as a causal effect?

Next, we will review and process the data to estimate the above model.

4: Data

`PSID_subset.csv`

The data used is a subset of the Panel Study of Income Dynamics (PSID) cross-year individual file, accessed from: https://simba.isr.umich.edu/Zips/ZipMain.aspx. Further details about the structure and coding of this file can be found here: https://psidonline.isr.umich.edu/Guide/FileStructure.pdf.

We have already limited the dataset to 2001-2015 and variables of interest. The variables included (using year 2001 as an example) are:

Variable	Original variable name	Description
`FID_68`	`ER30001`	Unique family identifier ("1968 Interview Number")
`IID_68`	`ER30002`	Unique individual identifier ("Person Number")
`age1`	`ER33604`	Age of the individual in year 2001
`EA_01`	`ER33616`	Years Completed Education, where values 0, 98, 99 are coded to missing
`EMPL_1`	`ER33612`	Employment status, where values 0, 8, 9 have been coded to missing. Original variable coding: https://simba.isr.umich.edu/cb.aspx?vList=ER33612

- *A note about the data:* Because the PSID tracks the same group of individuals over time since 1968, the rows of the data are unique individuals, and a separate set of variables is used for each year. For example, an individual's age or work status is assessed each survey year. Some individuals may enter or leave the dataset after 1968, due to death, marriage, childbirth, etc. Thus, certain variables may be missing in some

[1] The choice of defining the graduation age as 22 is somewhat arbitrary; one could think about how this assumption changes the interpretation of the results.

years for some individuals. Also note that we may not have given you the data in the ideal format for the analysis below. Feel free to restructure the data as you desire (e.g., with STATA `collapse` or `reshape`).

First, identify the individuals in each cohort. There is no right way to do this, but we want to see a logical way of generating cohorts that is well-documented in your code. Options include but are not limited to:

- Identify individuals who are approximately of college-graduation age (21 OR 22 years old) in 2007 and 2009 for the pre and post cohorts respectively.
- Identify individuals who completed their education before 2008 (e.g. individuals who are 22 or older in 2007). For the post cohort, identify individuals who experienced the financial crisis at some point during their education (e.g. individuals who are between 18 and 22 in 2009).

Regardless of strategy, the individuals in each cohort should not overlap. Limit the dataset to these individuals.

QUESTION 4: Choose a method for defining the pre and post cohorts. If choosing one of the above, explain the rationale behind the choice, as well as the specific coding and implementation used. If devising your own method, explain the method and the rationale behind it.

Some things to consider are the total sample size after filtering, availability of variation with which to identify the model, and what ages constitute the sample at each year. We recommend dropping observations where individuals are younger than 17, and thus not likely to be making labor market vs. schooling decisions. We also suggest generating indicator variables for work status. Some categories could include "working," "student," and "unemployed."

- *Follow-up question:* Imagine that you decide to identify individuals who are exactly age 22 in 2007 and 2009 for the pre and post cohorts, respectively. This makes it statistically impossible to estimate the model. What are some potential reasons why?

QUESTION 5: Before estimating the model, it is useful to observe the data visually, to check modelling assumptions and aid in the later interpretation of results. Suppose we are interested in "working" and "unemployed" statuses, create one figure (of the same type) for each that shows an insightful cut of the data and demonstrates our empirical approach, and describe what conclusions can (or cannot) be drawn from it.

QUESTION 6: Estimate the model in equation (2) using your prepared dataset using OLS and report a table of regression estimates. Interpret and discuss the results.

[continued next page]

5: Discussion

QUESTION 7: Imagine you are presenting these results to the team in a meeting. In addition to the interpretation of results above, what are some potential issues, problems, or questions you would raise in the meeting? If you made any important judgement calls with regards to any of the results, how would you explain them? No need to repeat anything already discussed in previous answers.

QUESTION 8: The PSID includes a "transition to adulthood" supplement with additional variables. Take a quick look through the documentation for this dataset and discuss a variable or two that would be interesting to examine in follow-up analyses. Mention any potential challenges or concerns from using these variables. There is no need to download or merge additional data.