Panel Data Seminar Exercise

October 5, 2018

Introduction

In this seminar exercise, you will be working from the "National Longitudinal Study of Youth" 1997 (NLSY97), a major longitudinal study tracking about 9,000 youths who were between 12 and 15 on December 31, 1996. NLSY97 is intended to follow young adults both as they complete their education and over their working lives.

In this dataset, you will be predicting the determinants of income for individuals, using:

- demographic characteristics: race/ethnicity, sex, & age,
- geographical characteristics: region and urbanity of residence,
- ability measures: high school GPA and ASVAB (Ability) Test Score,
- educational completion: highest grade completed in school, and
- work experience: years working at current main job,

Part 1: Prepare Variables

Before proceeding to the analysis, you will first need to do a small amount of data preparation.

(a). Create *"female"* and *"urban"* indicator (dummy) variables

- First, look at the structure of the variables *sex* and *rural_urban* using the **codebook** command.
- Then, use **gen** and **replace** to create indicators for whether the individual is female and whethey they live in an urban area.
- Provide descriptive labels for the new variables using the **label variable** command.
- Then **drop** the *sex* and *rural_urban* variables.

(b) Create an indicator for whether an individual has completed high school, using the *highest_grade_compl* variable.

- First, look at summary information for *highest_grade_compl* using the **summarize** command.
- Then create an indicicator for whether someone has completed high school, using the condition that highest grade completed is greater than or equal to 12.
- Once again, label the variable.

(c) Create an *age* variable, equal to *year* - *birthyr*.

(d) Create a variable for the log of earnings.

Part 2: Exploratory Analysis

With the data prepared, now let's do some exploratory analysis.

(a) Plot the average earnings by *age*.
(b) Create a scatterplot of GPA and log earnings for individuals who are 32 and living in the Northeast in 2015.

- Add titles to the graph and save it in your working directory.

Part 3: Regression Analysis

(a) First, run pooled OLS regression, with the log of earnings as the dependent variable and suitable choices of variables corresponding to each of the different explanatory variables mentioned in the introduction.

(b) Repeat the regression for each of random and fixed effects.

- Why are some variables omitted from the fixed effects regression?
- How (and why) do the standard errors of estimates change between estimates?
- After the fixed effects regression, save the residuals using the following command "predict earnings_resid, residuals"
  - Create a scatterplot of earnings and earnings residuals for the individual with person_id 49. Do the residuals look autocorrelated? (Feel free to also check out the residuals plot for other people too. Here's some IDs with several observations: 2,13,21,28,31,35,41,53)