Workshop 7: Panel Data

This workshop takes a final look at the wage equation using US data. It uses a panel data set drawn from the Current Population Survey of 595 persons who are observed for 7 periods, making a total of 4,165 observations.

1. **Data:**

   Use the data file wagecps.dta that you can download from CANVAS. You can find information on the variables in their labels by using the "describe" command. You could also summarize the data by using "summarize", but this may not be as helpful in this dataset as usual – why?

   - Get an overview of the organization of the data
     - Use "browse"
     - Have a further look at the data structure by typing "list id t exp wks in 1/21"
     - What is the role of the variables **id** and **t**?

   Set the unit and time identifiers as follows: "xtset id t"

   - Describe the data by "xtdescribe". Discuss the main patterns identified.
   - Summarize the variables lwage, ed, exp, and wks using "xtsum". Discuss the output.
   - Get a graphical representation of the dependent variable of interest:
     - Start by doing a "twoway scatter lwage t". What does it show? Why is that (not) helpful in this case?
     - Do something like an "xtline lwage if id < 10" or "xtline lwage if id < 10, overlay". Discuss the graphs. Do you see any potential indication of individual heterogeneity from these initial plots?

2. **Simple Pooled OLS**
   - Run an OLS regression of lwage on ed, exp, exp2, and wks. Discuss the potential weaknesses of this approach and what can be done to improve it in this context.
   - How would you recommend estimating the standard errors? Why?
   - Save the estimates for later use by something like "estimates store pooledOLS".

3. **Within-differences estimator**

- Transform your data following the steps described in slide 23. For each individual, calculate the means of the five variables that enter the model: lwage, ed, exp, exp2, and wks. Use something like “egen mlwage = mean(lwage), by(id)”.
- Then, for each individual, calculate the deviations from the means of the five variables that enter the model. Use something like “gen dlwage = lwage - mlwage”.
- Run the within-estimator as an OLS regression based on individual deviations-from-means.
- Run the within estimator using the Stata “xtreg” command with the “fe” option. Compare and save the estimates for later use.
- Why are the standard errors different for the two estimations? And how would you interpret the coefficient of the constant term in the FE regression? What is the purpose of the F-test at the bottom of the FE regression output, and what do you conclude from it?
- Another way to obtain the within estimates is to run an OLS regression with individual-specific constant terms, i.e. including a full set of individual dummies, the so-called least squares dummy variables (LSDV) estimator (see slide 22).
    - How many dummy variables are we talking about?
    - Before running this regression, you should first make sure that Stata allows you to handle that many variables. Do something like “set mat 800” (Note: Depending on the version of Stata you use, this may no longer be necessary). You can use the “i.” operator directly in the “reg” command to avoid constructing the dummies. What is the variable that you use for the “i.” operator?
    - Compare your results to the results of the within estimator. What do you conclude?

>>>>>> follow up in class <<<<<<<

4. **Random Effects estimator**

Recall that the random effects estimator combines the within and between dimensions to obtain an efficient estimator under the following assumptions:

- The individual-specific effect a_i is constant across time and drawn independently and identically across individuals.
- The idiosyncratic error component u_{it} is drawn independently and identically across individuals and time.
- The regressors are not correlated with the individual-specific effect a_i.

Run the random effects estimator using the Stata “xtreg” command with the “re” option. Compare and save the estimates for later use.

5. **Comparing RE and FE**

- Draw up a table in which you compare the estimates of the coefficients of the wage equation. (*Hint: use "estimates table", followed by the names of the stored estimates you saved before; if you want to produce a nicer table, explore the command "esttab"*)
- Run a Hausman test to compare the RE and FE estimates. What is the hypothesis being tested? What do you conclude – i.e., which model do you prefer according to the results?
- Why don't we get an estimate of the returns of education from the FE estimation?
- Re-estimate the FE equation adding interaction terms between education and time dummies. What can you conclude about the returns to education based on this specification?

6. **First-Differences estimator**

Let's do a final estimation using first-differences instead. For that, we need to do first-difference transformations of the data. Stata has the built-in "D." operator for this. You need to "sort id t" first to make it work.

- Transform the variables that enter the model (e.g., gen Dlwage=D.lwage) and run an OLS regression on the transformed data.
- Use the "D." operator directly in the "reg" command and check whether you obtain the same results.
- What causes the difference in the number of observations when using the FD estimator?
- Why don't we get estimates of the returns to education or the returns to experience from the first-differenced estimation?

>>>>>> follow up in class <<<<<<<