Panel Data Analysis of Microeconomic Decisions

Fall 2019

Notes about completing the assignment

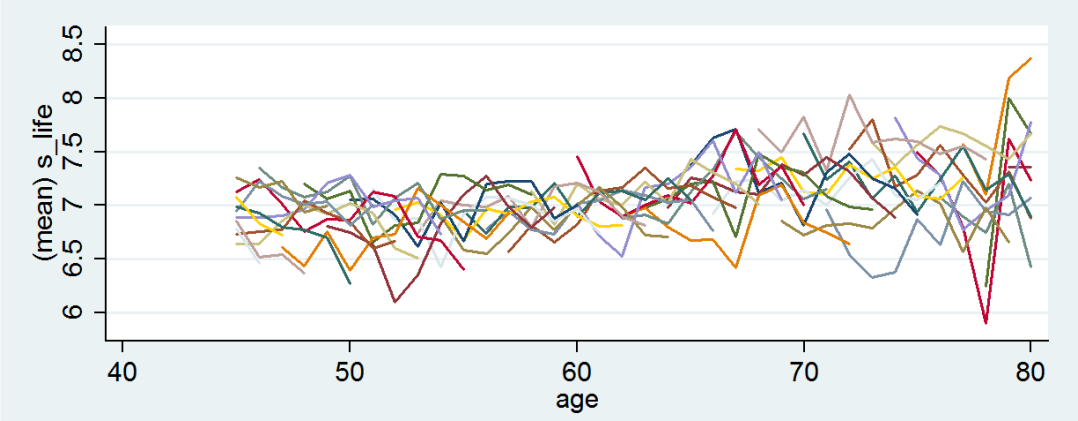
You can work on this exercise in groups of up to three people. However, each student is expected to submit an individual answer that differs from the answer the other group members hand in and to indicate who the other group members were. This means that explanations have to be given in your own words, while you can work on the code together and also discuss the results with your fellow group members before you write down your own answer. We will deduct points if formulations are too similar or if you don't name the other members of the group. The due date for this problem set is October 25th in class. The assignment should include the relevant Stata code and output, as well as your answers to the questions. The answers can be hand written/printed and handed in or sent to the email address: `m.husiatynski@uvt.nl`

Setup

In the first question, we use the same GSOEP data as in the first computer lab session. There, we have related subjective well-being to the variables age, couple, degreehandic, hhincome, and work using a linear specification and random as well as fixed effects estimation.

1. Profiles for different cohorts

The following figure plots the mean of life satisfaction for different birth cohorts against age. As you can see, it is a little bit messy.



a. Produce a similar figure in which you always plot life satisfaction against age, also for different cohorts, but always grouping 5 cohorts together. This gives one line for people born between 1951 and 1955, another one for 1956 until 1960 and so on. Also produce such a figure grouping always 10 cohorts together. Make an

effort to produce "nicer" figures, with proper axis labels and avoiding the blue background (use a different Stata color scheme). Hint: start with 1914-1919/1914-1924 cohort.

b. In the computer lab we talked about cohort effects. How does this discussion relate to what you now see in the figure you have just produced? Discuss this in detail.

c. A colleague of yours says that cohort effects are really interesting, and also that we need to take into account that subjective well-being changes with age. However, he wonders whether we have not forgotten to account for the fact that there is a time trend, so he wants to also control for that. That is, he wants to regress life satisfaction on age, calendar time and cohort dummies. Try it. What do you find? Explain.

2. One draw of simulated data.

We now generate artificial data using the following Stata code.

```
set seed 345398
drawnorm alpha_i, n(200)
expand 5
drawnorm nu_it e_it, n(1000)
g x_it=nu_it+alpha_i
drop nu_it
g y_it=3+alpha_i+2*x_it+e_it
```

a. Explain this data generating process (DGP) line by line. What is the underlying model (and equation) that has generated the data?

b. Use the command `pwcorr,sig` to display the correlation matrix between those variables. What is this command doing? Explain, based on this, whether you would expect OLS estimates from a regression of `y_it` on `x_it` to be biased or not. Will they be upward or downward biased?

c. Will they still be biased if we regress `y_it` on `x_it` and `alpha_i`? Explain. Are reported standard errors correct in that case?

3. Many draws of simulated data.

It is also possible to generate data many times and run a regression for each new data set. The following code performs such a Monte Carlo Study.

```
set seed 345398
capture program drop mcprog
program mcprog
  clear
  drawnorm alpha_i, n(200)
  expand 5
  drawnorm nu_it e_it, n(1000)
  g x_it=nu_it+alpha_i
  drop nu_it
  g y_it=3+alpha_i+2*x_it+e_it
  regress y_it x_it
end
```

```
simulate _b _se, reps(100):  mcprog
sum
```

a. Run the code and explain why the standard deviation of `_b_x_it` across simulated samples is substantially higher than the average estimate of the standard error, `_se_x_it` . Using clustered standard errors is an alternative way; why? Try it and discuss.

b. Explain why the mean of `_b_x_it` is not equal to 2. When does this matter and when does it not matter (catchword: prediction vs. causal effect)?

4. Fixed Effects and First Differences Estimation.

Go back to the Monte Carlo in 2, but now use two different estimators: 1) the fixed effects (FE) estimator, and 2) the first-differences (FD) estimator. Explain under which conditions each of them yields consistent estimates (use relevant equations). Also explain which estimator you prefer for the given DGP. How could you obtain the correct standard errors for both of these estimators?

5. Dynamic model.

The next thing we would like to try is what happens if we include an additional explanatory variable, `y_it`, in the previous period. For this, use `l.y_it` to refer to the lagged value of `y_it` . Generate (pragmatically) the data for the first year of each person with a coefficient on `l.y_it` equal to zero (so that it actually does not depend on it). That is, only let `y_it` depend on `l.y_it` for periods after the first (you can do this with a combination of generate for the first year and replace for all other years in Stata). Then use the fixed effects estimator and run a Monte Carlo as in question 3. Produce a table in which you show how the bias in your estimate depends on the number of time periods for each individual and on the true value of the coefficient on the lagged value `l.y_it` . Do this for 5 time periods, 10, 20 and 50 and for true values of the coefficient of 0.5. Explain why the fixed effects estimator is biased in the first place.

6. Instrumental variables estimation.

Do a Monte Carlo as in 5 for `t=5` time periods in which you use the Arellano-Bond estimator to estimate the effect of lagged dependent variable. Explain why this estimator provides a consistent estimator for coefficient of the lagged dependent variable. Which lags of the dependent variables can be used as instruments? Why? Hint: what conditions should an instrument fulfill?