Panel Data Analysis of Microeconomic Decisions: Assignment I

Fall 2024

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 answers of other group members. 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. At the beginning of the assignment, list the other members of your group. 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 assignment is October 21. The assignment should include the relevant Stata code and output, as well as your answers to the questions. The answers should be uploaded on the relevant Canvas assignment page; in case of technical problems, the assignment be sent to the following email address: `w.chen@uvt.nl`

1. The effects of age, cohort and time.

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.

Make use of the data, to answer the following questions.

(a) Explain why we can never simultaneously estimate the relationship between subjective well-being on the one hand, and calendar time, cohort (birth year) and age on the other hand.

(b) Produce a figure in which you plot life satisfaction against age, with separate lines for each cohort. For this, the command `collapse` will be helpful. Also produce a figure in which you plot changes in life satisfaction against age. Interpret the figure. How would the implied relationship between subjective life satisfaction and age look like?

(c) Directly plot life satisfaction against age. Does the figure contradict the implied relationship between subjective life satisfaction and age in Question (b)? How can they be reconciled?

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 whether you would expect the OLS estimate from a regression of `y_it` on `x_it` to be biased or not. Will they be upward or downward biased?

(c) Will the OLS estimate still be biased if we regress `y_it` on `x_it` and `alpha_i`? Explain. Are reported standard errors obtained from such a regression correct?

3. Many draws of simulated data.

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

(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)?

(c) Now, use clustered standard errors in the OLS regression. Moreover, estimate a random effects model. Compare the standard errors obtained from all three models. What do you observe? Explain.

4. Fixed Effects and First Differences Estimation.

Consider again the DGP in Question 2. Now, we are interested in the following two estimators: 1) the fixed effects (FE) estimator, and 2) the first-differences (FD) estimator.

(a) Estimate the model using the above two estimators. Explain under which conditions each of them yields consistent estimates (use relevant equations). Also explain which estimator you prefer for the given DGP. How would the errors have to be generated for the alternative estimator to be efficient.

(b) Again, estimate the model using the random effects estimator. What do you observe with the estimated coefficients? Explain.

5. Dynamic model.

The next thing we would like to try is what happens if we include `y_it` measured in the previous period as an additional explanatory variable. 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 Simulation with 100 draws such 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, 10, 20 and 50 time periods and for true value 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 simulation like in Question 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 consistently estimates the coefficient of the lagged dependent variable. Which lags of the dependent variable can be used as instruments? Why? (Hint: what conditions should an instrument fulfill?)