Final Exam

Ph.D. Applied Microeconometrics
KDI School Fall 2023

2023-12-04

Due date: Friday, December 1st at 11:59pm

Please work by yourself. As before, please submit the following files on eKDIS:

- Your R Markdown file
- Your knitted PDF file
- Any other scripts you used to complete the assignment

I would like all of your answers (including the data section) to be in a single markdown file. However, you are welcome to use another script for any of the analyses, if you would prefer. If you do, please include the script in your submission.

1 Part 1: Short answer

Question 1

We spent quite a bit of time discussing recent research on two-way fixed effects (TWFE). Discuss the following:

- What is the main problem with two-way fixed effects?
  - When does this problem arise?
- What are some of the solutions that have been proposed?
- Do you agree with the following statement? Why or why not? "Recent advancements in this area are the most important advancements in applied microeconometrics in the last 10 years."

Question 2

Consider the following production function:

$$log(y_{it}) = \alpha_i + \beta_t + \gamma log(L_{it}) + \delta log(K_{it}) + \phi \left(log(L_{it}) \times log(K_{it})\right) + \epsilon_{it} \tag{1}$$

where y_{it} is output, L_{it} is labor, K_{it} is capital, and α_i and β_t are firm and time fixed effects, respectively. I want to estimate the marginal product of labor, which is given as:

$$\frac{\partial y_{it}}{\partial L_{it}} = \frac{y_{it}}{L_{it}} \left(\gamma + \phi log(K_{it})\right) \tag{2}$$

- What are the identification assumptions for estimating γ? (You can ignore the recent literature on TWFE.)
- For estimating the production function itself, how would you deal with standard errors? Why?
- Suppose I want to test the hypothesis that $\gamma = \delta = \phi = 0$.
  - What is the appropriate test statistic?
  - How would you calculate the test statistic? (I mean by hand, not by using a canned function.)
- How would you calculate a confidence interval for the marginal product?

Question 3

A commonly taught assumption of OLS is that the error term is normally distributed. However, this is not strictly necessary for OLS. Discuss when this assumption is particularly important and when it is not. (Hint: think about the CLT.)

2 Part 2: Getting your hands dirty with data

The dataset "pollution.csv" is subset of data from my pollution paper. The variables are:

- shrid: village identifier (this is panel data)
- year: year
- distfe: district identifier
- pm25: pollution concentration (PM 2.5) during the growing season, logged
- wind: the number of days (in 10s) in the season in which the village is downwind from a pollution source
- yield: agricultural yield (in kilograms per hectare)
- rain_z: rainfall (z score) during the growing season
- temp_mean: temperature during the growing season

I do not specify how to deal with standard errors. I will leave that up to you.

Question 4

Consider the following regression:

$$log(yield_{it}) = \alpha_i + \gamma_t + \beta_1 pm25_{it} + \beta_2 rain_{it} + \beta_3 temp_{it} + \epsilon_{it}, \tag{3}$$

where α_i is village fixed effects and γ_t is year fixed effects.

- Estimate the regression and output the results in a table.
- Interpret the coefficient on $pm25$.
- What are the requirements for the coefficient on $pm25$ to be interpreted as the causal effect of pollution on yield?
  - Do you think these requirements are satisfied? Why or why not?
- How did you specify the calculation of the variance-covariance matrix (standard errors)? Why?

Question 5

Consider the following two-stage least squares (2SLS) set up:

$$pm25_{it} = \alpha_i + \gamma_t + \beta_1 wind_{it} + \beta_2 rain_{it} + \beta_3 temp_{it} + \epsilon_{it} \tag{4}$$

$$log(yield_{it}) = \phi_i + \psi_t + \delta_1 pm25_{it} + \delta_2 rain_{it} + \delta_3 temp_{it} + \eta_{it} \tag{5}$$

- Estimate the *reduced form* regression and output the results in a table. Interpret the coefficient on $wind$.
- Estimate the first and second stages of the 2SLS regression and output the results in a table. Interpret the coefficient on $pm25$.
- What are the requirements for the coefficient on $pm25$ to be interpreted as the causal effect of pollution on yield?
  - Do you think these requirements are satisfied? Why or why not? (You can refer to my paper if you'd like.)

2.1 Question 6

Suppose you want to use *randomization inference* to test the hypothesis that the coefficient on $pm25$ in equation (5) is equal to zero. To do this, you need to give each village a *random* treatment assignment from the rest of the villages. In other words, you need to take a randomly selected combination of {wind, pm} and assign it to each village (the randomly selected combination should come from the same village). The steps are as follows:

1. Set a seed to ensure that your results are replicable.
2. Randomly select a combination of {wind, pm} and assign it to each village. However, you need to randomly select *from the same year*.
3. Estimate the 2SLS regression using the randomly assigned IV and treatment variable.

4. Save the coefficient on $pm25$ from the second stage (δ_1).
5. Repeat this process 500 times.

- Do steps 1 through 5.
- Create a figure that shows the distribution of δ_1 from the 500 iterations.
- Compare the distribution of treatment effects to the actual treatment effect from the 2SLS regression. What do you conclude?
  - In other words, interpret the output and what this teaches us about the estimated treatment effect.