MOEC.0021
Empirical Methods
University of Zurich — HS 2019
PROF: Greg Crawford / TA: E. Dicarlo, M.R. Greco, S. Bagagli, A. Jenni — Handout 4

Exercise 4

1 Theory

1. **Motivating Linear Panel Data.**
Suppose you are interested in estimating the production function for agricultural output (just like a seminal article in 1961 by Mundlak in the Journal of Farm Economics). You have access to data for a large number of farms i for $T \geq 1$ time periods. The production function you want to estimate is

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}$$

where y_{it} is log(output), x_{it} is log(labour) - a variable input, α_i is log(soil quality) - a fixed input, and ε_{it} is rainfall - a random input. Each farmer knows the price of output P_t, the wage rate W_t, and the soil quality of his farm α_i. However, as the econometrician you only observe (y_{it}, x_{it}). Assume that ε_{it} is iid and independent of everything in the model

(a) **Solve the farmer's profit maximization problem assuming he sells output at a common (across farmers) market price P_t and pays common wages W_t.** (Hint: It may help to write down the production function in levels instead of logs.) For notational convenience, assume that $Ee^{\epsilon_{it}} = \lambda$. Does the labor demand depend on α_i? Explain the economic intuition behind the result.

(b) Under what assumption can you recover a consistent estimate for β by running (pooled) OLS? Based on what you found in (1a) do you think this assumption is violated in this case? (no proof required)

(c) Suppose you only had $T = 1$ period of data. Propose an estimation strategy that would consistently estimate β. Be careful to explain what assumptions need to hold. Could the variables P_1 and W_1 possibly help? If you would need access to another variable besides $(y_{i1}, x_{i1}, P_1, W_1)$ provide an example of what might work. Discuss what properties your variable must satisfy.

(d) Now suppose that you have access to $T = 10$ periods of data. Which of the following estimators would consistently estimate β

i. Random Effects

ii. Fixed Effects

iii. First Differences

Explain your answers by (briefly) discussing which assumptions are needed for consistent estimates from each of the three methods, and whether they are likely to hold in this example.

(e) Would you prefer your estimation strategy in (c) or your preferred estimator identified in (d)? Why?

(f) **[Extra]** Farmers take their harvesting decisions (i.e. how many workers to hire) also based on rain *forecasts* for the season. This is another variable that the farmers observe. Suppose you are not able to observe this information. Obviously, as for the case of soil quality, rain forecasts are correlated with labor decisions, but they end up in the error term. This means your ε_{it} also contains a time-varying shock. You still have $t = 10$ periods of data. How would this affect your preferred estimator? How could you fix this?

2 Empirical Applications

1. **Basic Panel Data Models**
This question uses a panel data set taken from Baltagi and Griffin (1983) "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures" in the European Economic Review. The data set contains infomation on gasoline consumption in 18 OECD countries (i) over the 19 years from 1960-1978 (t). This is the dataset used by the authors in their publication and was submitted along with their article. The dataset is available on to download here: `http://bit.ly/1VMDlpi`. The appendix contains information describing each variable.

   (a) Consider the following specification for a gasoline consumption equation:

$$\ln\left(\frac{Gas}{Car}\right)_{it} = \beta_0 + \beta_1 \ln\left(\frac{Y}{N}\right)_{it} + \beta_2 \ln\left(\frac{P_{MG}}{P_{GDP}}\right)_{it} + \beta_3 \ln\left(\frac{Car}{N}\right)_{it} + u_{it}$$

   where $u_{it} = \alpha_i + \varepsilon_{it}$.

   Provide some economic rationale for the regression specification. That is, explain why each variable is included, and the likely sign of the coefficients.

   (b) Estimate this specification by a Pooled OLS regression.

   (c) What are the necessary assumptions for this OLS model to be consistent? Are they likely to be satisfied here? In particular, are there any potential sources of endogeneity that we should worry about? Given your answer to these questions, is OLS the best linear unbiased estimator for this model?

   (d) Let's focus on β_2, the coefficient on the real price of gasoline. Given your answer to part (c), what is the likely direction of the bias in the Pooled OLS coefficients? Be careful in explaining the economic mechanism driving the bias.

   (e) Estimate the regression specification using:

      i. The LSDV Estimator. How do the coefficients compare to what you find in (b)? Is this what you expected?

      ii. The Within Groups ("Fixed Effect") Estimator. Are the β's different to what you obtained in (**??**)? Are they different from what you obtained in (**??**) ? Explain.

      iii. The Generalised Least Squares ("Random Effect") Estimator

   (f) Focusing on your estimates of β_2

      i. What does the OLS estimate of $\hat{\beta}_2$ imply about the relationship between gas prices and gas consumption? Does this estimate pass the "smell test", i.e. do you think it's magnitude is likely to be too big, about right, or too small?

      ii. Do the results suggest your concerns about endogeneity bias may be reasonable?

(g) Perform the Classical- and Regression Based Hausman Tests to investigate whether the Random Effects assumption is viable in this setting. What do you find? Does this conclusion surprise you?

3 Appendix: Data Description for gasoline.csv

Variable	Description
co	Factor indicating country.
year	Year
c	Logarithm of motor gasoline consumption per car
y	Logarithm of real per-capita income
p	Logarithm of real motor gasoline price
car	Logarithm of the stock of cars per-capita