Econometrics I (Second Half)
Problem Set 1

Instructions (Read Carefully)

Please submit all of the files related to your solution to this problem set via email to me (andrea.flores@fgv.br) and your TA Taric (tariclatif@gmail.com) by **Thursday, December 4th at 11:00am**.

Your solution to this problem set should consist of (1) a pdf with your responses to the items in each of the questions as if it was a report (for items that require coding with no discussion, include a reference to a particular part of your code), and (2) the files containing the code used in each question that requires estimation.

Go as far as you can. Even if you don't manage to complete all items in a question, partial credit will be applied generously if you clearly describe the issues you faced in approaching that particular question and intuitively explain how you think these issues could be addressed.

Good Luck!

Question 1: Identification and Covariation Restrictions [20 points]

Throughout this question, we will consider a structure S describing a relationship between variables Y_1 and Y_2.

For parts **(a)-(c)**, suppose that the structure describes the relationship between Y_1 and Y_2 through the following set of equations

$$Y_1 = \beta X + \Lambda + \eta_1 \tag{1}$$
$$Y_2 = \beta Z + \Lambda + \eta_2 \tag{2}$$

Assume that the latent variables ϵ_1 and ϵ_2 satisfy the following covariation restrictions:

$$\mathbb{E}[\eta_1 | X, Z] = \alpha \tag{3}$$
$$\mathbb{E}[\eta_2 | X, Z] = \alpha \tag{4}$$

for some constant α. Further assume that Λ satisfies the following condition:

$$\mathbb{E}[\Lambda | X, Z] = \gamma_0 + \gamma_1 (X + Z) \tag{5}$$

(a) Suppose that we **only** have data on Y_1 and X and that we assume that $\mathbb{E}[Z|X] = 0$. Show that β is not identified in this case. What **exclusion restriction** could we impose to identify β?

(b) Suppose we obtain access to a better data set that contains information on Y_1, Y_2, X, and Z so that all these variables are observable to us. Show that we can now identify β without imposing any exclusion restriction.

Question 2: Measurement Error and Identification [10 points]

Consider the following model implying a linear relationship between Y and a covariate X^*, which we unfortunately capture with error in the data as X:

$$\begin{aligned} Y &= \alpha + \beta X^* + \epsilon \\ X &= X^* + \eta \end{aligned}$$

Suppose that there exists some variable Z such that $\mathbb{E}[X^*|Z = z]$ exists and that:

$$\mathbb{E}[\eta|Z = z] = 0$$

(a) Write down and intuitively describe the restriction(s) on $\mathbb{E}[\epsilon|Z = z]$ – if any – needed to identify β. If no such restriction is needed, explain why.

(b) Derive an expression that allows us to identify β in terms of observed variation in Y,X, and Z. Intuitively explain

Question 3: Identification in Practice [25 points]

Go through the paper *Using Geographic Variation in College Proximity to Estimate the Return to Schooling* by David Card, which I am including in the files for this problem set. Intuitively work through the identification analysis of the paper by breaking it down following the Hurwicz (1950) framework to determine the identifying power of a model. Throughout your analysis, make sure to discuss the following within the context of the paper:

1. Model (restrictions imposed to define admissible structures) and structure

2. Structural characteristics/features relevant to addressing the research question of the paper

3. Identifying assumptions made to identify the structure characteristic of interest. That is, describe which of the identification strategies mentioned during lectures was implemented in the paper.

4. Describe the mapping between the data and the model that ultimately yields identification under the assumptions made

Question 4: Panel Data [45 points]

Use the dataset called `cornwell.dta` for this question. The data consist of a panel of 90 North Carolina counties captured between 1981 and 1987 and containing information on county-level crime rates (crimes per person) and other characteristics.

Consider the following econometric model:

$$Y_{it} = X_{it}'\beta + \alpha_i + \epsilon_{it}$$

where X_{it} denotes a vector of "deterrent" variables which could affect the likelihood that people would commit a crime. These include the following: *lprbarr* (log of probability of arrest), *lprbconv* (log of probability of conviction), *lprbpris* (log of probability of prison sentence), *lavgsen* (log of average sentence length) and *lpoplc* (log of police per capita).

(a) Estimate the model implementing a fixed effects approach by using the within estimator for β. Report your estimates and explain what does $(\widehat{\beta}_{FE})$ imply regarding the relationship between county-level crime rates and deterrance measures.

(b) Estimate the model implementing a random effects approach by using a GLS estimator for β Report your estimates and explain what does $(\widehat{\beta}_{GLS})$ imply regarding the relationship between county-level crime rates and deterrance measures?

(c) Implement the Hausman test manually (only using the coefficients that you can test with both FE and GLS estimators) and report the test statistic. Do we reject the null hypothesis that there is no correlation between county-specific unobserved heterogeneity and counties' deterrence characteristics?

(d) Estimate your model using first differences, compare your point estimates and standard errors with those obtained in part (a). Intuitively explain if there are noticeable differences in the estimates obtained from both models.

(e) Suppose that you want to control for variables that capture the returns to legal opportunities. For this, you want to extend your model to control also for weekly (log) wages in different sectors (*lwcom-lwser*). Extend your FE and RE estimators to include these controls and re-do parts (a)-(c) (FE and RE estimators along with the Hausman test) with the new specification.