Econometrics 1 - Problem Set 1

LMEC, Fall 2022

Prof. Matteo Barigozzi

TA: Filippo Pavanello

October 17, 2022

In groups (as they have been previously decided), use the software Stata (and your brain), to answer the questions below. Please return one zip-file (one for each group) to filippo.pavanello2@unibo.it by **October 17, 2022 (11:59pm, 23:59)**; write as object of the email PS1-Solutions'. The zip-file must contain (i) a document with answers to each specific question (pdf format); (ii) the Stata log-file; (iii) the Stata do-file. Name the zip-file as "surname1_surname2_surname3.zip", and in any case remember to write name, surname and id number (matricola) of each student in the document.

Throughout the problem set, after setting the number of observations (in Stata: set obs 1000), use the command ***set seed 1000 + G***, where G is the number of your group as indicated in the file attached to the email. For example, the members of Group 1 should specify: set seed 1001. Similarly, when running the *simulate* command, specify ***seed(1000+G)***.

Concise answers to all questions must be included in the pdf (either theoretical answers, or Stata output) but Stata commands and code can be contained in your do and log files. Good luck!

Question 1

Consider three random variables $(Y, \mathbf{X})$, which are jointly normally distributed in the population, and where $\mathbf{X} = (X_1, X_2)$.

The trivariate normal distribution of Y, X_1 and X_2 is characterised by $\mu_Y = 25$, $\mu_{X_1} = 18$, $\mu_{X_2} = 9$, and by the following variance/covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{X_1 Y} & \sigma_{X_2 Y} \\ \sigma_{X_1 Y} & \sigma_{X_1}^2 & \sigma_{X_1 X_2} \\ \sigma_{X_2 Y} & \sigma_{X_2 X_1} & \sigma_{X_2}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.3 & -0.2 \\ 0.3 & 1 & -0.4 \\ -0.2 & -0.4 & 1 \end{bmatrix}$$

Multivariate normal theory implies that in the population:

$$\mathbb{E}(Y \mid \mathbf{X}) = \beta_0 + \mathbf{X}\boldsymbol{\beta} \tag{1}$$

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \Sigma_X^{-1} \sigma_{XY} \tag{2}$$

σ_{XY} is the vector of covariances:

$$\begin{bmatrix} \sigma_{X_1 Y} \\ \sigma_{X_2 Y} \end{bmatrix}$$

and Σ_X is the variance-covariance matrix of the independent variables:

$$\begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1 X_2} \\ \sigma_{X_2 X_1} & \sigma_{X_2}^2 \end{bmatrix}$$

Moreover, the constant β_0 is defined as:

$$\beta_0 = \mu_Y - [\mu_{X_1}, \mu_{X_2}]\boldsymbol{\beta} \tag{3}$$

1. Generate a sample of $N = 1000$ observations on (Y, X_1, X_2) from the joint distribution above. Then, consider the multiple linear regression model:

$$Y = \beta_0 + \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{4}$$

   (a) Use the commands in MATA to compute the population values β_0, β_1 and β_2

   (b) Does the condition $\mathbb{E}(\epsilon|X_1, X_2) = 0$ hold in the above population model? Why?

   (c) Which interpretation can you attach to the population value β_1?

   (d) Define and compute the OLS estimator, by using the matrix formulation [*Start by using the command* `st_view` *or* `st_data` *in order to store* Y *and* X *as a vector and a matrix as we did in the class.*]

2. Generate 1,000 random samples of Y, X_1 and X_2 composed by 100 observations each ($N = 100$).

   (a) For each sample, estimate the parameters β_0, β_1 and β_2 in model 4 through the OLS method and return the estimates into a new dataset with 1000 observations, where each observation corresponds to one realisation of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

   (b) Use these replications to provide evidence about the biasedness/ unbiasedness of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.

   (c) Plot the distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$ and comment.

3. Now generate 10,000 random samples, of Y, X_1 and X_2 composed by 100 observations each (N = 100).

   (a) Replicate the exercise of part 3 and plot again the distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$. What do you notice?

Question 2

Dell, Jones and Olken (AER, 2012) empirically study the impact of temperature and precipitation on the economic output (GDP). You are provided with a part of the dataset ps1_group##.dta (sent to each group by email) the authors used for their study.

Consider the following linear model:

$$Y_i = \beta_0 + \beta_1 Temp_i + \beta_2 Prec_i + \epsilon_i \tag{5}$$

where the dependent variable Y_i is the real gross domestic product (GDP) of country i (in billions). $Temp_i$ is the average temperature (in Celsius degrees) in country i. The coefficient β_1 captures the impact of temperature on economic output. $Prec_i$ is the average precipitation level (in meters) in country i and β_2 coefficient capture its effect.

1. Load the data and Enter Mata. Use Mata's `st_view()` function to create matrices based on your Stata dataset.
   (a) Define and compute the OLS estimator
   (b) Define and compute the total sum of squares (SST), the explained sum of squares (SSE), and the residual sum of squares (SSR)
   (c) Define and compute the R-squared
   (d) Define and compute OLS residuals and fitted values $\widehat{Y}_i$
   (e) Compute the sample average of the OLS residuals and the sample covariance between regressors and residuals. Comment the results
   (f) Compute and compare the average fitted value and the average value of y [In MATA, to generate the mean of a vector write: `avrx=mean(x)`, where *avrx* is the name you choose for the new matrix/vector and x is the existing one.]
2. Now, exit MATA and use STATA command to obtain OLS estimation of model in equation 5
   (a) Compare these results with those obtained using the Matrix Linear Regression in Mata (the procedure run above)
   (b) Obtain the average fitted value and the average value of Y_i. Find an observation for which the GDP is overpredicted and one for which it is underpredicted
   (c) Show empirically whether the sample average of the OLS residuals is equal to zero
3. Now assume the actual population relationship between GDP and the set of independent variables is given by Equation 5, but suppose you accidentally specified the model as follows:

$$Y_i = \beta_0 + \beta_1 Temp_i + \epsilon_i \tag{6}$$

(a) Let denote with $\tilde{\beta}_1$ the OLS estimator of β_1. Do you expect it to be biased or unbiased for β_1? Why?

(b) Compare the estimated value $\hat{\beta}_1$ you obtain by estimating the correctly specified model with the estimated value $\tilde{\beta}_1$. When do you expect the two values to be very similar?

(c) Confirm the partialling out interpretation of the OLS estimates by explicitly doing the partialling out for Model 6

References

[1] Dell, M., Jones, B. F., & Olken, B. A. (2012). Temperature shocks and economic growth: Evidence from the last half century. *American Economic Journal: Macroeconomics*, 4(3), 66-95.