Econometrics 1 - Problem Set 1

LMEC, Fall 2021

Prof. Matteo Barigozzi

TA: Nektaria Glynia

October 8, 2021

In groups (as they have been decided), use the software Stata (and your brain), to answer the questions below. Please return one zip-file (one for each group) to nektaria.glynia2@unibo.it by **October 15, 2021 (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 four random variables $(Y, \mathbf{X})$, where $\mathbf{X} = (X_1, X_2, X_3)$ which are jointly normally distributed in the population. The joint normal distribution of Y, X_1, X_2 and X_3 is characterized by $\mu_Y = 10$, $\mu_{X_1} = 15$, $\mu_{X_2} = 15$, $\mu_{X_3} = 10$ and by the following variance/covariance matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{X_1Y} & \sigma_{X_2Y} & \sigma_{X_3Y} \\ \sigma_{X_1Y} & \sigma_{X_1}^2 & \sigma_{X_1X_2} & \sigma_{X_1X_3} \\ \sigma_{X_2Y} & \sigma_{X_2X_1} & \sigma_{X_2}^2 & \sigma_{X_2X_3} \\ \sigma_{X_3Y} & \sigma_{X_3X_1} & \sigma_{X_3X_2} & \sigma_{X_3}^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 & 0 & 0.2 \\ 0.6 & 1 & 0 & 0.3 \\ 0 & 0 & 1 & 0 \\ 0.2 & 0.3 & 0 & 1 \end{bmatrix}$$

Multivariate normal theory implies

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

where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \Sigma_X^{-1} \sigma_{XY} \tag{1}$$

σ_{XY} is the vector of covariances:

$$\begin{bmatrix} \sigma_{X_1Y} \\ \sigma_{X_2Y} \\ \sigma_{X_3Y} \end{bmatrix}$$

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

$$\begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1X_2} & \sigma_{X_1X_3} \\ \sigma_{X_2X_1} & \sigma_{X_2}^2 & \sigma_{X_2X_3} \\ \sigma_{X_3X_1} & \sigma_{X_3X_2} & \sigma_{X_3}^2 \end{bmatrix}$$

The constant β_0 is defined as $\beta_0 = \mu_Y - [\mu_{X_1}, \mu_{X_2}, \mu_{X_3}]\boldsymbol{\beta}$.

1. Generate a sample of $N = 100$ observation on (Y, X_1, X_2, X_3) from the joint distribution above.

   Consider the multiple linear regression model:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{2}$$

   (a) Use the commands in MATA to compute the theoretical/population values β_0, β_1, β_2 and β_3.

   (b) which interpretation can you attach to the population value β_1?

   (c) define and compute the OLS estimator, by using the matrix formulation *[Start by using the command st_view in order to store y and X as a vector and a matrix.]*

   (d) define and compute the total sum of squares (SST), the explained sum of squares (SSE), and the residual sum of squares (SSR)

(e) define and compute the R-squared and the adjusted R-squared

(f) define and compute OLS residuals and fitted values $\widehat{y}$

(g) compute the sample average of the OLS residuals and the sample covariance between regressors and residuals. Comment the results

(h) compute and compare the average fitted value and the average value of y *[In MATA, to generate the mean of a vector write: avr_x=mean(x), where avr_x is the name you choose for the new matrix/vector and x is the exisisting one.]*

2. Now, exit MATA and use STATA command to obtain OLS estimation of model in equation 2.

   Compare these results with those obtained using the Matrix Linear Regression in Mata (the procedure run above).

3. Generate 1000 random samples of $N = 100$ observation on (Y, X_1, X_2, X_3) from the joint distribution above. For each sample:

   (a) estimate the parameters $\beta_0, \beta_1, \beta_2, \beta_3$ through the OLS regression in equation (2), and store these results in a new dataset where each observation corresponds to one realization of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$. *[Whithin the 'program' you write to run the simulation, DO NOT use the command set seed, use instead seed(number) after the command simulate.]*

   (b) Are the OLS estimators $\widehat{\boldsymbol{\beta}}' = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \hat{\beta}_3)$ unbiased for parameters $\boldsymbol{\beta}$ in equation 2? Supplement your theoretical answer by computing their averages over the 1000 repeated samples.

   (c) Plot the sampling distributions of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, , $\hat{\beta}_3$ and comment about them.

Question 2

You are asked to study the relationship between the family and the labor market, by focusing in particular on the link between fertility and labor supply. This is a long-standing research question in economics and there are many studies reporting estimates of this relationship. You are provided with the dataset ps1_group##.dta (sent to each group by email), which contains a sample of women aged 21 to 35 with two or more children, that is randomly drawn from the 1980 US Census .
Consider the following model:

$$labor\_supply_i = \beta_0 + \beta_1 more\_kids_i + \beta_2 X_i + \epsilon_i \tag{3}$$

when $labor\_supply_i$ can be:
1. the number of weeks worked in the year prior to the census
2. the average hours worked per week
3. labor earnings in year prior to census, in 1995 dollars;
$more\_kids_i$ is a binary variable equal to 1 if the woman has more than two children; and X_i contains individual covariates such as age, age at first birth, indicator variables for Black, Hispanic and Other race and years of education. ϵ_i is the idiosyncratic component.

1. Load the data and Enter Mata. Use Mata's `st_view()` function to create matrices based on your Stata dataset. Choose **one** of the three variables that measure the labor supply and compute the OLS estimator of model in equation 3.

2. Now, exit MATA and use STATA command to obtain OLS estimation of model in equation 3 and compare these results with those obtained using the Matrix Linear Regression in Mata.

3. Now assume the actual population relationship between labor supply and the set of independent variables is given by equation 3, but suppose you accidentally specified the model as follows:

$$labor\_supply_i = \beta_0 + \beta_1 more\_kids_i + \beta_2 education_i + u_i \tag{4}$$

   (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 $\widehat{\beta}_1$ 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 4