

Introductory Econometrics

Tutorial 4

PART A: To be done before you attend the tutorial. The tutors will ask you questions based on this part and that will be the basis for your participation point. The solutions will be made available at the end of the week.

In this exercise, you need to continue working with WDI.xlsx data set that you used in tutorial 2.

1. In the "A random sample of countries" sheet that you populated with GDP per capita and infant mortality rate in 2014, add two more columns and label them "healthpc2014" and "femsch2014". Then, for each country in the sample of 40 countries, use VLOOKUP that you learnt in tutorial 2 to place the health expenditure per capita and secondary school female enrollment rate in 2014 in these columns. Check a few of the entries to make sure that the formula has worked as expected. In order to make transferring these into Eviews easier, save this sheet separately in a new spreadsheet.
2. Read the data in "A random sample of countries" into Eviews. Get a table of summary statistics for GDPPC2014, INFMORT2014, HEALTHPC2014, FEMSCH2014. [In Eviews, you can CTL click on these variables, then open them as a Group and choose Descriptive Stats under the View button.]
3. The maximum of FEMSCH2014 is 175.38. What does that mean (see tutorial 2 for the definition of these variables)? Which country has the highest female secondary school enrollment rate?
4. Get the pairwise scatter plots of all 4 variables. You can do this in Eviews from the Group window of these four variables by choosing View/Graph and choosing Scatter, and then selecting the "Lower triangular matrix" from the "Multiple series" drop down menu. Based on these scatter plots, which two have closer to a linear relationship? Does it make sense that they should?
5. Get the pairwise sample correlation coefficients of the 4 variables. You can do this in Eviews from the Group window by choosing View/Covariance Analysis and checking the Correlation box and unchecking the Covariance box and then OK. The goal is for you to see that scatter plot can give you some insight about the functional form of the relationship (linear or nonlinear) between two variables, whereas the correlation coefficient quantifies the direction (positive or negative) and the strength of the linear relationship between two variables.
6. Suppose we want to estimate the following regression:

$$INFMORT2014_i = \beta_0 + \beta_1 GDPPC2014_i + u_i$$

using the data for this sample of countries. Write down the first 5 rows of the \mathbf{y} vector and the first 5 rows of the \mathbf{X} matrix.

Do not forget to bring your answers to PART A and a copy of the tutorial questions to your tutorial.

PART B: You do not need to hand this part in. It will be covered in the tutorial. It is still a good idea to attempt these questions before the tutorial.

1. (*Post-multiplying a matrix by a vector produces a linear combination of the columns of the matrix*): Let

$$\mathbf{X} = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}$$

and

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0.7 \\ 0.2 \end{bmatrix}.$$

Compute $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, and show that the result is 0.7 times the first column of \mathbf{X} plus 0.2 times the second column of \mathbf{X} . The learning objective of this question and its connection with multiple regression will be explained by your tutor.

2. Let's generalise the result in question 1. Suppose

$$\mathbf{X}_{n \times 3} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & & \vdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

and

$$\hat{\boldsymbol{\beta}}_{3 \times 1} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

Show that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is an $n \times 1$ vector which is a linear combination (a weighted sum) of columns of \mathbf{X} with weights given by the elements of $\hat{\boldsymbol{\beta}}$. That is:

$$\hat{\mathbf{y}} = \text{first column of } \mathbf{X} \times \hat{\beta}_1 + \text{second column of } \mathbf{X} \times \hat{\beta}_2 + \text{third column of } \mathbf{X} \times \hat{\beta}_3$$

In fact this is not specific to \mathbf{X} having 3 columns. It is true for any $n \times k$ matrix \mathbf{X} and $k \times 1$ vector $\hat{\boldsymbol{\beta}}$.

3. This question is based on question C4 in Chapter 2 of the textbook. It is based on data on monthly salary and other characteristics of a random sample of 935 individuals. These data are in the file wage2.wfl. We concentrate on *wage* as the dependent variable and the IQ as the independent variable.
 - (a) *Estimation, interpretation of the slope coefficient and R^2 of the regression*: Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in *wage* for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*? What is the relationship between the R^2 of this regression and the sample correlation coefficient between *wage* and *IQ*? Name your estimated equation **eq01**. Save the residuals of this regression in a variable called **uhat01**.
 - (b) *Interpretation of the intercept*: What does the intercept in eq01 mean? Now, run a regression of *wage* on a constant and (*IQ*-100), and name it **eq02**. Compare the results with your results in eq01 and note all similarities and differences. Save the residuals of this regression in a variable called **uhat02**.
Open uhat01 and uhat02 side by side and see if they are different. What is the interpretation of the intercept in eq02?
 - (c) Discuss what you learned from this exercise.
4. (*Regression on dummy variables*): Consider a data set in which each observation must belong to only one of two categories. For example, a data set on wage of random sample of n observations from the population of employed people. Each person in the data set is either male or female. The dummy variable *female* is a binary variable that is equal to 1 if the individual is female and 0 otherwise. The dummy variable *male* is a binary variable that is equal to 1 if the individual is male and zero otherwise. Consider regression of *wage* on these two dummy variables (no constant).
 - (a) Sketch the \mathbf{X} matrix for this regression (for ease of notation, you can assume that the first n_1 observations are female and the last n_2 observations are male. Obviously, $n_1 + n_2 = n$). Are the columns of \mathbf{X} linearly independent?
 - (b) Use the OLS formula $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ and derive the OLS estimator in this case. Comment on the result. Verify your result by creating the dummy variable *male* in wage1tute4 data set and running a regression of *wage* on *female* and *male* (no constant).
 - (c) Consider now that we had a constant in addition to these two dummy variables. Write down the \mathbf{X} matrix for this case. Are the columns of \mathbf{X} linearly independent? What is the dimension of the column space of \mathbf{X} ? Remember from first year ETC1000 that you were told that when a regression has a constant, you should only add one dummy variable for an attribute that has two categories (such as male, female). Explain where that rule came from.