

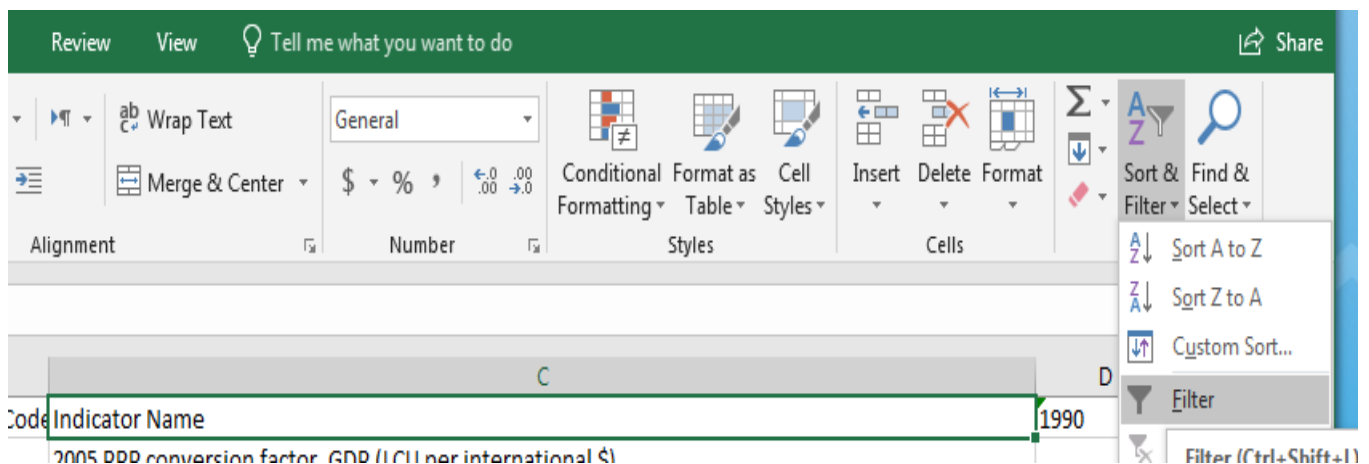
# Introductory Econometrics

## 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.

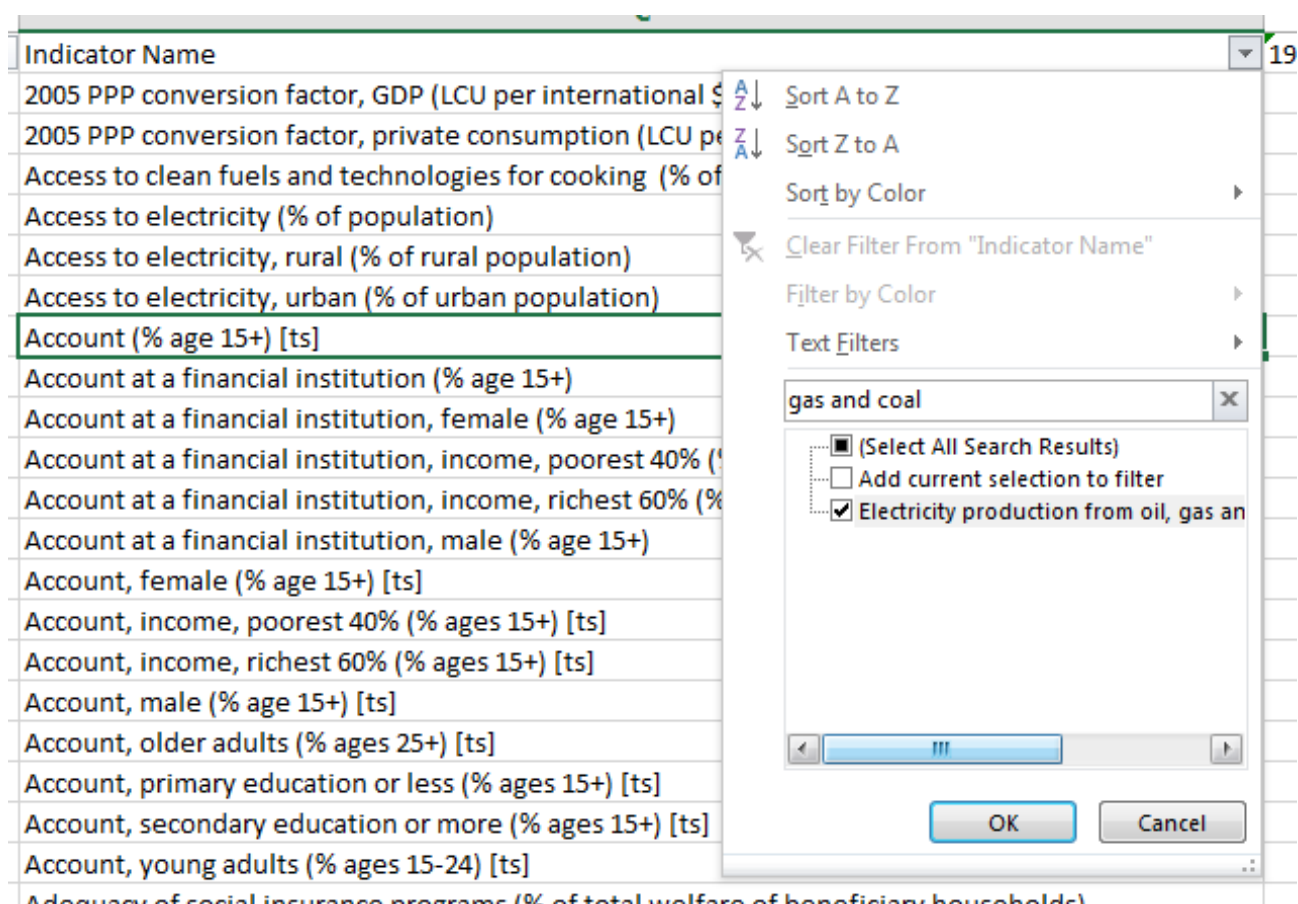
In this exercise, you need to continue working with WDI2019.xlsx data set that you used in tutorial 2. This is quite important because it gives you an idea of how to extract data that you want from a large data base and form it into a tidy format that can be easily read into any econometric software. The data set you form here will be the one that you will analyse in your first assignment, so pay attention and make sure that your data set is created correctly.

1. In the data sheet, filter the data and only keep “Electricity production from oil, gas and coal sources (% of total)”. To do that, first click on any of the cells in the first row and then click on “Sort & Filter”, and then choose “Filter”, as shown in the screenshot below:



This creates drop down menus for each of the cells on the first row. Click on the right corner of the “Indicator Name” cell to open the drop down menu, and choose ‘Electricity production from oil, gas and coal sources (% of total)’. Since there are a large number of variables, an easy way to find this is to type ‘gas and coal’ in the search window inside the filter window, which reduces the options to a small number of variables with ‘gas and coal’ as part of their name.

Then choose that variable that you want. The screenshot below might help.



The filtered data will show only the % of electricity generated by coal, gas and oil for each country in the WDI database. Copy the entire area and paste into a new sheet in the WDI2019 spreadsheet. Change the name of this new sheet to 'electricity from fossil fuels'.

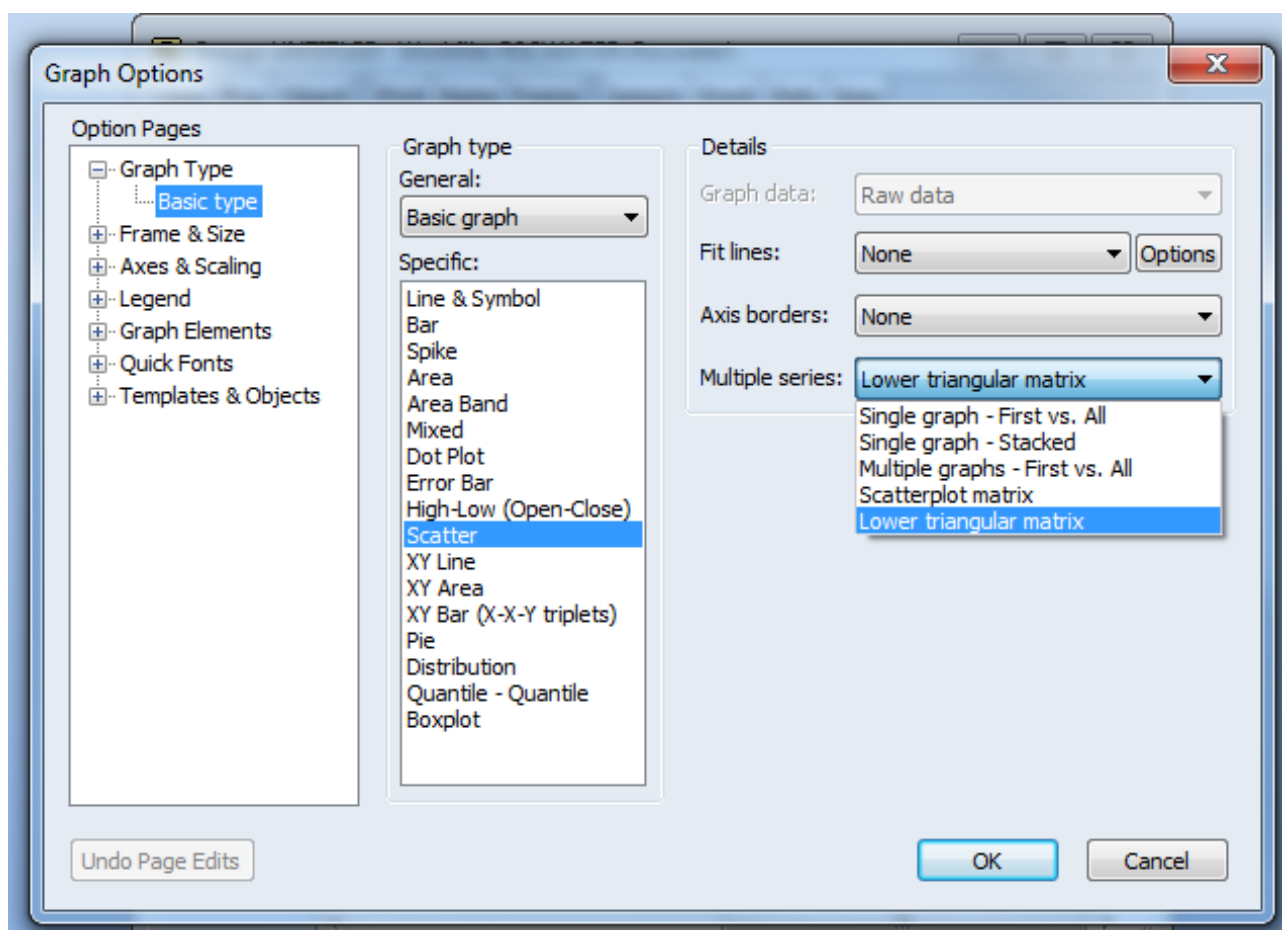
2. Repeat the above, but this time select 'Electric power consumption (kWh per capita)' in your filter. Copy the entire area and paste into a new sheet. Change the name of this new sheet to 'electricity per capita'. Save your copy of WDI2019.xlsx on a USB, so that you can use it later.
3. In the 'A sample of countries' sheet in the WDI spreadsheet, add two new columns labelled 'fosspect' and 'eleckwh' and populate them with data on % of electricity generated by fossil fuels (oil, gas and coal), and electric power consumption in 2014 for each of the countries in this sample (VLOOKUP will help, but be careful that if a country has no data for 2014, VLOOKUP returns 0. You need to delete the zero and leave the cell blank if one of the countries with missing data is in the sample of countries. There is only one country - Albania - which genuinely does not use any fossil fuels in electricity generation, so you should not delete the zero for Albania. If interested, read about electricity generation in Albania on the internet). This sheet now should have data on GDP per capita, CO<sub>2</sub> emissions per capita, % of electricity generated by fossil fuels in 2014, and electric power consumption per capita in kilo Watt hours in 2014 for 122 countries, with 17 countries having missing values. Save this sheet separately and upload it into EViews.

- You have electricity consumption per person, and you know the percentage of this electricity that was generated by fossil fuels. That is the part that contributes to CO<sub>2</sub> pollution. So, in EViews, generate a new variable called 'fosskwh' which shows electric power consumption per capita generated by fossil fuels. That is,

$$fosskwh = electkwh * foss pct / 100$$

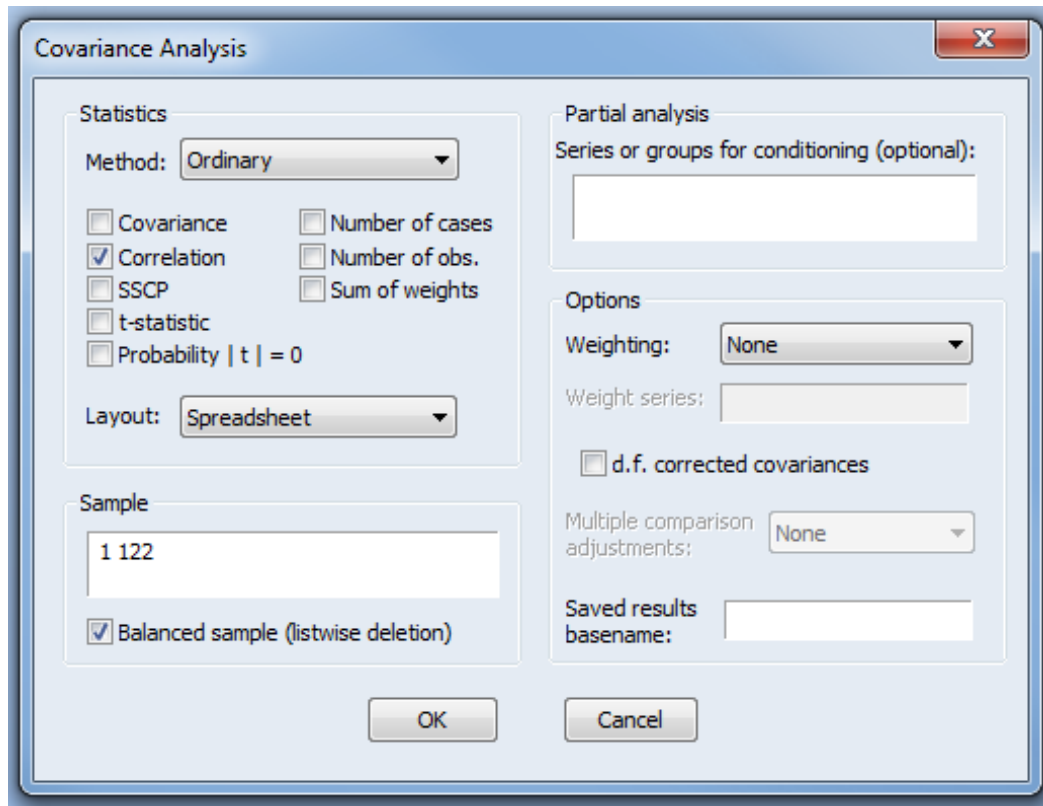
Find the top 5 countries in terms of electricity consumption per capita generated by fossil fuels.

- Get the pairwise scatter plots of co2pc, gdppc and fosskwh. You can do this in EViews from the Group window of these 3 variables by choosing View/Graph and choosing Scatter, and then selecting the "Lower triangular matrix" from the "Multiple series" drop down menu. No need to print the plots. Based on these scatter plots, comment on the pairwise relationship between these variables.



- Get the pairwise sample correlation coefficients of the 3 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. Nothing to print or

comment on here.



The image shows the 'Covariance Analysis' dialog box in SPSS. It is divided into several sections: 'Statistics', 'Partial analysis', 'Options', and 'Sample'. In the 'Statistics' section, the 'Method' is set to 'Ordinary'. Under the 'Statistics' group, 'Correlation' is checked, while 'Covariance', 'SSCP', 't-statistic', and 'Probability | t | = 0' are unchecked. In the 'Number of cases' group, 'Number of cases', 'Number of obs.', and 'Sum of weights' are all unchecked. The 'Layout' is set to 'Spreadsheet'. In the 'Sample' section, the 'Sample' list contains '1 122' and 'Balanced sample (listwise deletion)' is checked. The 'Partial analysis' section has an empty text box for 'Series or groups for conditioning (optional)'. The 'Options' section has 'Weighting' set to 'None', an empty 'Weight series' box, 'd.f. corrected covariances' unchecked, 'Multiple comparison adjustments' set to 'None', and an empty 'Saved results basename' box. 'OK' and 'Cancel' buttons are at the bottom.

**Covariance Analysis**

**Statistics**

Method: **Ordinary**

☐ Covariance ☐ Number of cases  
☒ Correlation ☐ Number of obs.  
☐ SSCP ☐ Sum of weights  
☐ t-statistic  
☐ Probability | t | = 0

Layout: **Spreadsheet**

**Sample**

1 122

☒ Balanced sample (listwise deletion)

**Partial analysis**

Series or groups for conditioning (optional):

**Options**

Weighting: **None**

Weight series:

☐ d.f. corrected covariances

Multiple comparison adjustments: **None**

Saved results basename:

OK Cancel

7. Print the summary statistics (for a common sample) for these three variables. You can do that by View -> Descriptive Stats -> Common Sample. "Common Sample" gives summary statistics for the subset of countries for which all three variables are available.

**Do not forget to bring your answers to 4, 6 and 7 above 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

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

and

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

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  $\hat{\boldsymbol{\beta}}$ . Because of this, in linear regression, the predicted value of the dependent variable  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$  is a linear combination of columns of the matrix of independent variables  $\mathbf{X}$ .

3. (*Regression on a constant only - one way to compute the sample mean and sample variance of any variable using regression*): Consider a regression model that has an intercept and no explanatory variables, i.e.

$$y_i = \beta_0 + u_i, \quad i = 1, \dots, n.$$

This means that the  $\mathbf{X}$  matrix is an  $n \times 1$  column of 1s.

$$\underset{n \times 1}{\mathbf{X}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- (a) Use the OLS formula  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$  and show that the OLS estimator of the constant term in this case is the sample average of the dependent variable.
- (b) What is  $\hat{u}_i$  (the residual for each observation) in this case? Show that  $\mathbf{X}'\hat{\mathbf{u}} = 0$  in this case is the same as  $\sum_{i=1}^n (y_i - \bar{y}) = 0$ , a result that we knew already.

- (c) What is  $\hat{\sigma}^2$  the estimator of the variance of  $u_i$  in this case? What is the standard error of regression in this case?
4. 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.wf1. We concentrate on *wage* as the dependent variable and the IQ as the independent variable.
- (a) *Verify question 3 with real data:* Run a regression of *wage* on a constant only. Verify that the OLS estimate of the intercept is the sample mean of *wage* and the standard error of regression is the sample standard deviation of *wage*.
- (b) *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**.
- (c) *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?
- (d) Discuss what you learned from this exercise.