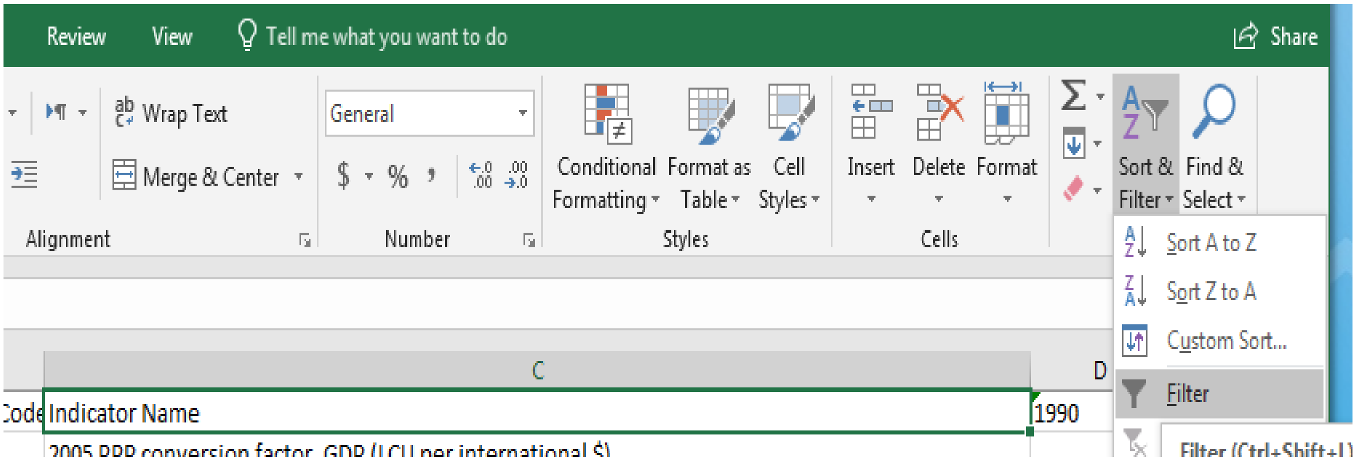
Introductory Econometrics

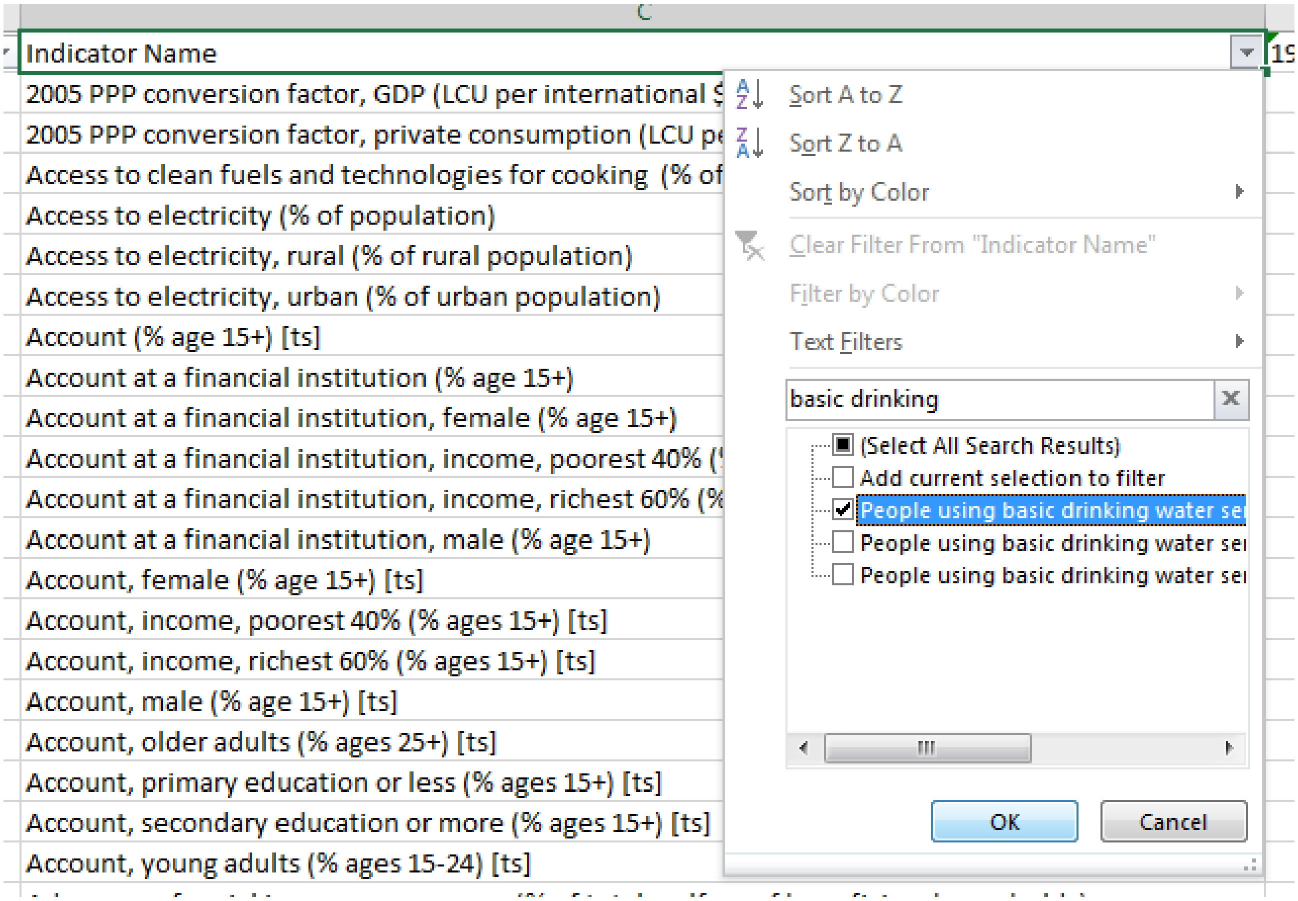
Tutorial 4

<u>PART A:</u> 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. 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 "People using basic drinking water services (as % of total population)". 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 'People using basic drinking water services (as % of total population)'. Since there are a large number of variables, an easy way to find this is to type 'basic drinking water' in the search window inside the filter window, which reduces the options to a small number of variables with 'basic drinking water' as part of their name. Make sure that you use the overall one (not the one that covers only the rural or the urban population). The screenshot below might help.



Indicator Name
2005 PPP conversion factor, GDP (LCU per international $
2005 PPP conversion factor, private consumption (LCU pe
Access to clean fuels and technologies for cooking (% of
Access to electricity (% of population)
Access to electricity, rural (% of rural population)
Access to electricity, urban (% of urban population)
Account (% age 15+) [ts]
Account at a financial institution (% age 15+)
Account at a financial institution, female (% age 15+)
Account at a financial institution, income, poorest 40% ('
Account at a financial institution, income, richest 60% (%
Account at a financial institution, male (% age 15+)
Account, female (% age 15+) [ts]
Account, income, poorest 40% (% ages 15+) [ts]
Account, income, richest 60% (% ages 15+) [ts]
Account, male (% age 15+) [ts]
Account, older adults (% ages 25+) [ts]
Account, primary education or less (% ages 15+) [ts]
Account, secondary education or more (% ages 15+) [ts]
Account, young adults (% ages 15-24) [ts]

basic drinking

- ■ (Select All Search Results)
- □ Add current selection to filter
- ☑ People using basic drinking water sei
- □ People using basic drinking water sei
- □ People using basic drinking water sei

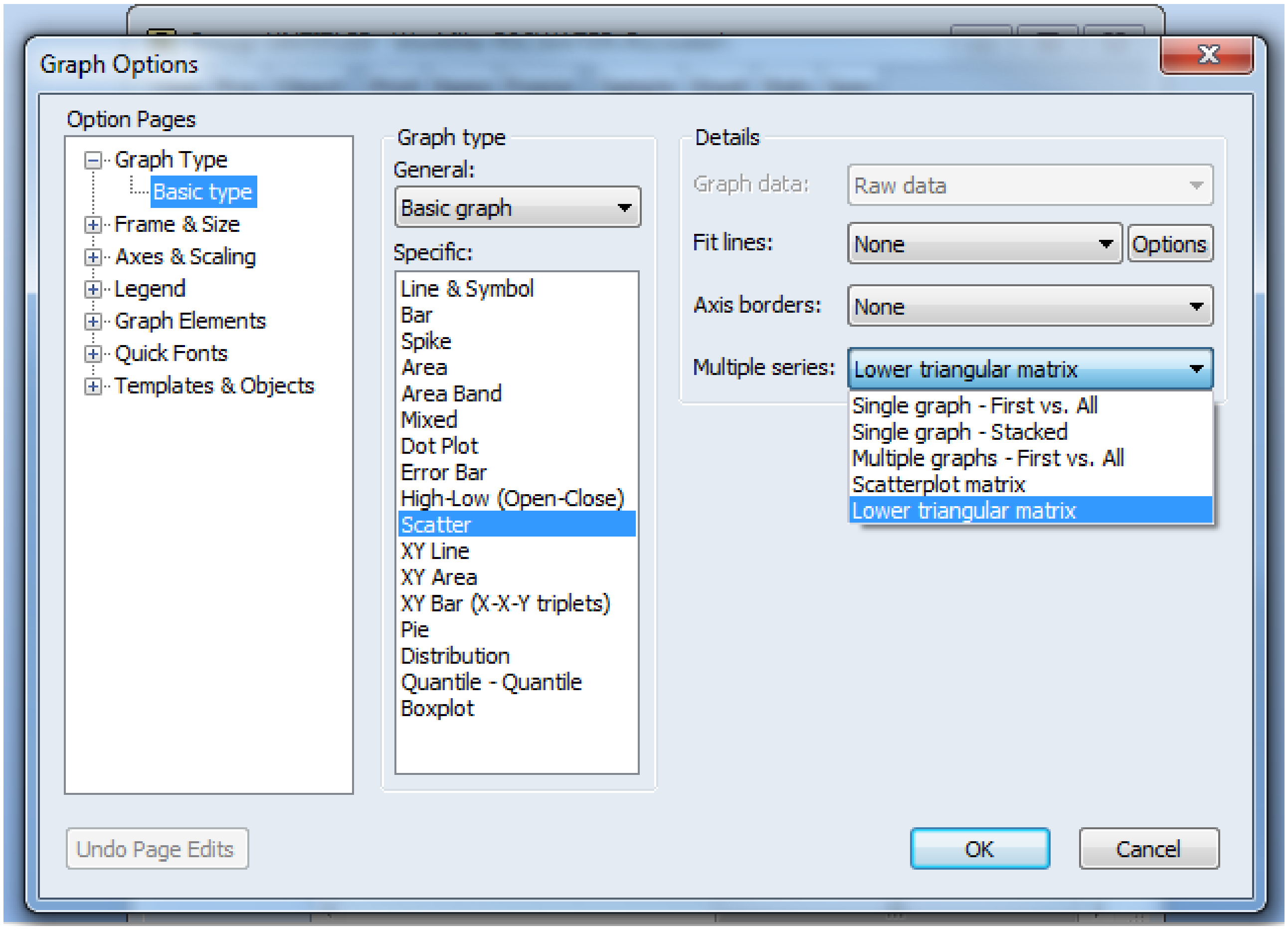
The filtered data will show only the % of population using basic drinking water services for each country in WDI database. Copy the entire area and paste into a new sheet in the WDI spreadsheet. Change the name of this new sheet to ‘water’. Save your copy of WDI.xlsx on a USB, so that you can use it later.

2. Sort the data in the ‘water’ sheet based on % population using basic drinking water services in 2015. Look and the top and the bottom of the distribution of sorted data. How many countries have missing values in 2015? What can you say about the countries at the lower end of the distribution? What can you say about those with 100% of population having basic drinking water services?

3. In the ‘A random sample of countries’ sheet in the WDI spreadsheet, add a new column labelled ‘water’ and populate it with data on % of population using basic drinking water services in 2015 for each of the 40 countries in this sample (VLOOKUP will help, but be careful that if a country has no data for 2015, 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 random sample of countries.). This sheet now should have data on GDP percapita, infant mortality and basic drinking water usage for a sample of 40 countries in 2015. Save this sheet separately and upload it into EViews.

4. Get the pairwise scatter plots of these 3 variables. 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. Based on these scatter plots, which two have closer to a linear relationship? Does it make sense that they

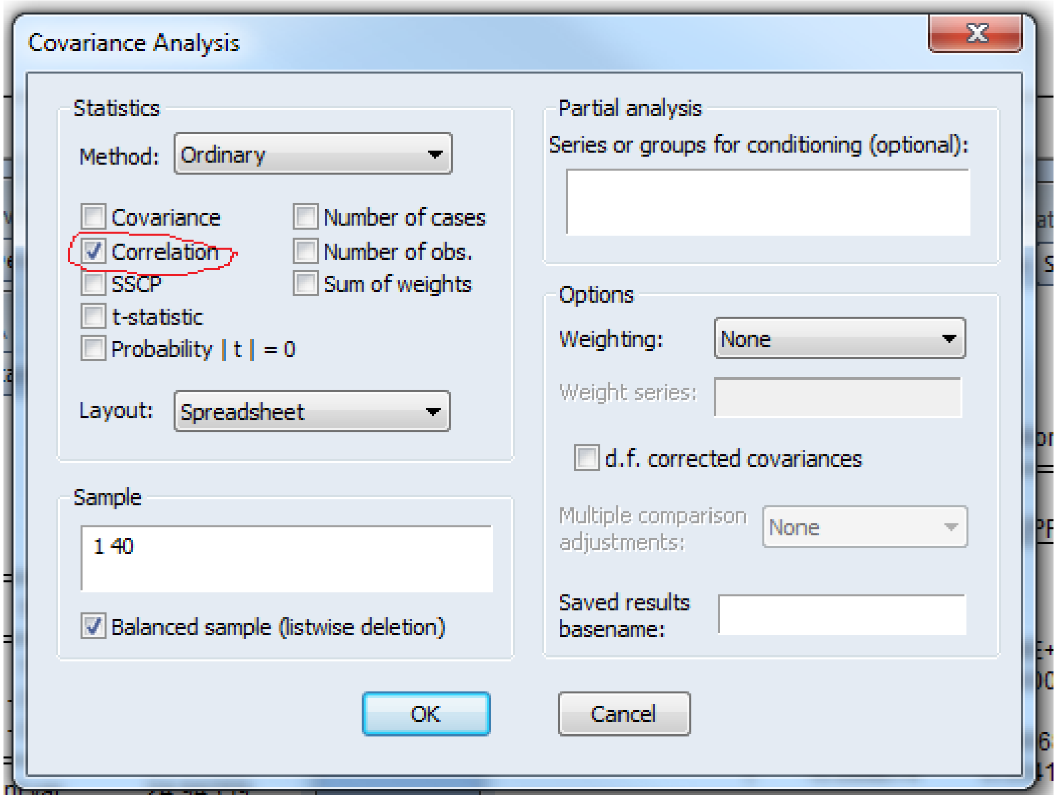
should?


Graph type
General: Basic graph
Specific: Scatter

Details
Graph data: Raw data
Fit lines: None
Axis borders: None
Multiple series: Lower triangular matrix

5. 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.



Covariance Analysis

Statistics

Method: Ordinary

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

Layout: Spreadsheet

Sample

1 40

☑ 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:

6. Neither the scatter plot nor the correlation coefficient can tell us if after taking into account the GDP per capita of a country, does access to basic drinking water have any additional effect on infant mortality. We can only investigate that with a multiple regression. Suppose we were to estimate
$$INFMORT_i = \beta_0 + \beta_1 GDPPC_i + \beta_2 WATER_i + u_i$$
using the data for this sample of countries. What would be the dimensions of the $\mathbf{y}$ vector, the $\mathbf{X}$ matrix and the $\boldsymbol{\beta}$ vector in this regression? Write down the first 3 rows of the $\mathbf{y}$ vector and 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

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

Compute $\widehat{\mathbf{y}}= \mathbf{X}\widehat{\boldsymbol{\beta}}$, 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

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

Show that $\widehat{\mathbf{y}}= \mathbf{X}\widehat{\boldsymbol{\beta}}$ 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 $\widehat{\boldsymbol{\beta}}$. That is:

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

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 $\widehat{\boldsymbol{\beta}}$.

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.wf1. 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 $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ 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.