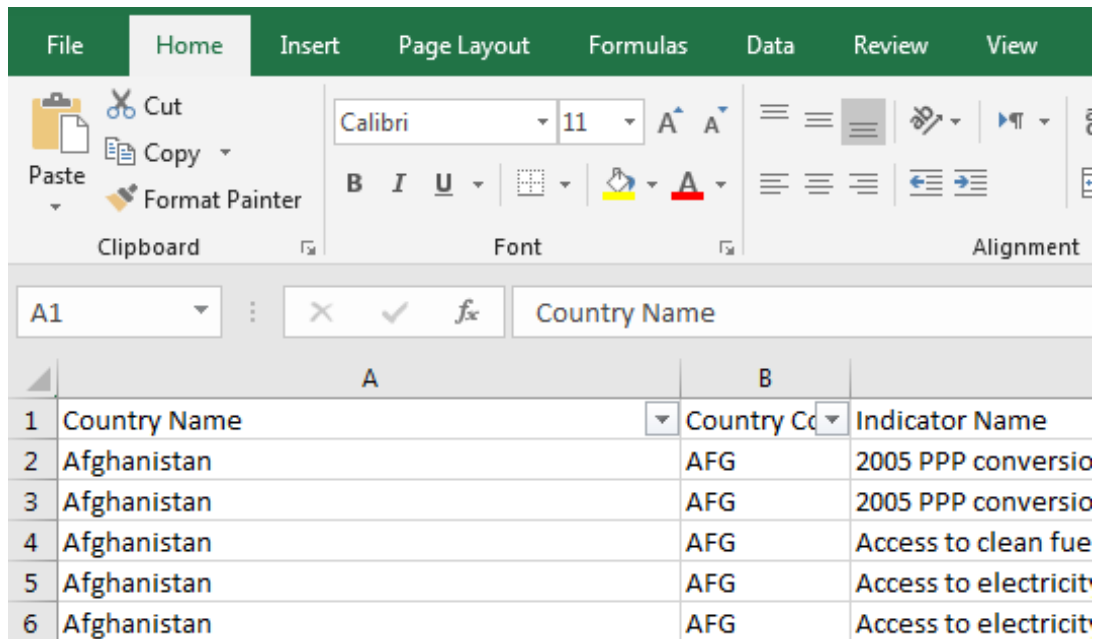


# Introductory Econometrics

## Tutorial 2

**PART A:** To be done before you attend the tutorial, and handed in to your tutor at the beginning of the tutorial to earn 1 point. You may write the answers by hand. The solutions will be made available at the end of the week.

This tutorial is about different kinds of data (cross section, time series and panel) that we often work with in business and economics. Download the data set WDI.xlsx from Moodle to your local directory (this is a huge data set with more than 340000 rows, so it may take a while to download. Do not try to download it on your phone!). When you open the data sheet, you will see something like this:



	A	B	
1	Country Name	Country Code	Indicator Name
2	Afghanistan	AFG	2005 PPP conversion factor
3	Afghanistan	AFG	2005 PPP conversion factor
4	Afghanistan	AFG	Access to clean fuel
5	Afghanistan	AFG	Access to electricity
6	Afghanistan	AFG	Access to electricity

This is a panel data set because it contains data for a set of countries from 1960 to present (I have deleted all years before 1990 to make the file smaller). Of course not all countries have data available for each year, so there are many missing values. The source of this data set is the World Bank. The data set is called the "World Development Indicators" and it is publicly available from <https://datacatalog.worldbank.org/dataset/world-development-indicators>. As you can see in the sheet labeled "Data", it is quite messy and we need to do quite a bit of work to make it ready for analysis. I have done some pre-processing and created 2 sheets labeled "gdppc" and "infmort", which contain data on GDP per capita and infant mortality respectively.

- GDP per capita (gdppc) is measured in US dollars and has been adjusted for purchasing power differences among countries (this is referred to as "purchasing power parity (PPP) dollars". For example, one Swiss franc is equal in value to one US dollar. However, because Switzerland is an expensive country, with one Swiss franc in Switzerland you cannot buy as much as you can buy with one US dollar in the USA. A 1 Swiss franc in Switzerland may only have the purchasing power equivalent to 90 US cents in USA. These considerations have been taken into account and adjusted for, so the dollar values have the same purchasing power and hence are directly comparable across countries.
- Infant mortality (infmort) is measured in number of deaths per 1000 live births.

The purpose of this pre-tutorial exercise is for you to learn to use the VLOOKUP function of Excel to make a *tidy data set* for a group of countries from your messy WDI data set. A *tidy data set* is a data set that has the variables that we want to use in our data analysis in columns, and each row of

it contains one observation on those variables. The following figure shows a snap shot of the top 10 rows of a tidy data set containing GDP per capita and infant mortality for a cross section of countries in 2015.

	A	B	C
1	CCODE	GDPPC	INFMORT
2	DZA	13724.72	21.9
3	ARG	19101.3	10.3
4	ARM	8195.934	12.5
5	AUS	43832.43	3.2
6	BGD	3132.568	29.7
7	BOL	6531.519	30.5
8	BRA	14666.02	14
9	CAN	42983.1	4.5
10	CHL	22536.62	7.3

The first row contains variable names. The first column often contains an observation identifier. For cross section data an identifier is some unique tag for that observation. In the above figure, the identifier is three letter country code that the United Nations has assigned to each country. If we were dealing with data for students in a class for example, student ID would be a good identifier. While these identifiers are not used in the analysis, they are useful for merging data from different sources or for descriptive analysis (e.g. identifying which country has the highest GDP per capita in our sample). For time series data, the first column is usually the date. Fortunately, statistical packages these days are intelligent enough to read dates written in various formats.

Remember that while having data in a tidy format is important for transferring data into a statistical package and analysing it, it is more important to know what our data measure. That is, we need to know that GDPPC and INFMORT stand for, we need to know their unit of measurement, and we need to know the time period that the data correspond to. Also, the country codes are not always easy to guess. For example, it is very difficult to guess that DZA is the code for Algeria. So, we need a table of country codes, or we can add a column to our data containing country names, although since some country names have spaces and strange characters in them, they may cause problems when reading the data in some statistical packages (fortunately EViews is intelligent enough to not get confused). Now the exercise that you need to do:

1. There is a sheet in WDI.xlsx named “A random sample of countries”. This sheet contains 40 country codes that were selected at random from the list of all countries. Using Excel’s VLOOKUP function, get the country’s GDP per capita in 2015 and infant mortality rate in 2015 in columns B and C in front of each country code. The syntax of VLOOKUP function is:

= VLOOKUP(lookup\_value, table\_array, col\_index\_num, type\_of\_match)

lookup\_value is the cell that contains the country code(e.g. A2)

table\_array is the range of data to search (e.g. gdppc!\$A\$2:\$AD\$160)

col\_index\_num is the column that contains the data that you want (relative to the first column in table\_array)

type\_of\_match is the accuracy of match. You always want FALSE here to find exact matches

Note the use of \$ in the examples so that the formula can be repeated for all countries. Note that the infmort sheet has the country code in its second column, and it contains many more countries than the gdppc sheet. This is deliberate to give you some practice to learn the VLOOKUP

function. Do some random checks to make sure that you have transferred the data correctly and then save your spreadsheet. Since this spreadsheet is very large, you may want to save the “A random sample of countries” sheet as a new spreadsheet on its own, so that it can be opened faster.

2. To check if you have done this correctly, unhide a hidden sheet in the WDI.xlsx and check the data with what you have created (if you don't know how to unhide a hidden sheet in Excel, google it!). If they don't match, click on the first GDPPC and INFMORT cells and compare the VLOOKUP formulae with yours. If you have any questions, come to the consultation session of the lecturer or any of the tutors for explanation.
3. Import the data in “A random sample of countries” into EViews. An easy way is to open EViews, then choose File->Open->Foreign Data as Workfile and then find and click on the file where you saved the “A random sample of countries” sheet. Note that if you are using EViews on MoVE, the file has to be in your directory on the Monash server or on your Monash google drive (it cannot be in a file on your laptop or a USB device or google drive related to your personal google account).
4. Obtain the summary statistics of GDPPC and INFMORT in EViews, print those and hand them to your tutor in your tutorial to obtain the participation point for week 2. Save the EViews workfile.

**END OF PART A. YOU MAY WANT TO REFER TO THE EViews LESSON ON MOODLE FOR A REMINDER ON HOW TO IMPORT DATA INTO EViews.**



## EViews

EViews is available on all PCs in the computer labs in the Menzies building. You can also access it anywhere, any time using your own personal electronic device (Windows, Mac, iOS, or Android) via the Monash Virtual Environment (MoVE). The video in this link

<http://guides.lib.monash.edu/learning-tools/move>

tells you all the information you need to get to EViews. The following lesson takes you through the process of accessing EViews on MoVE and also doing some elementary data analysis.

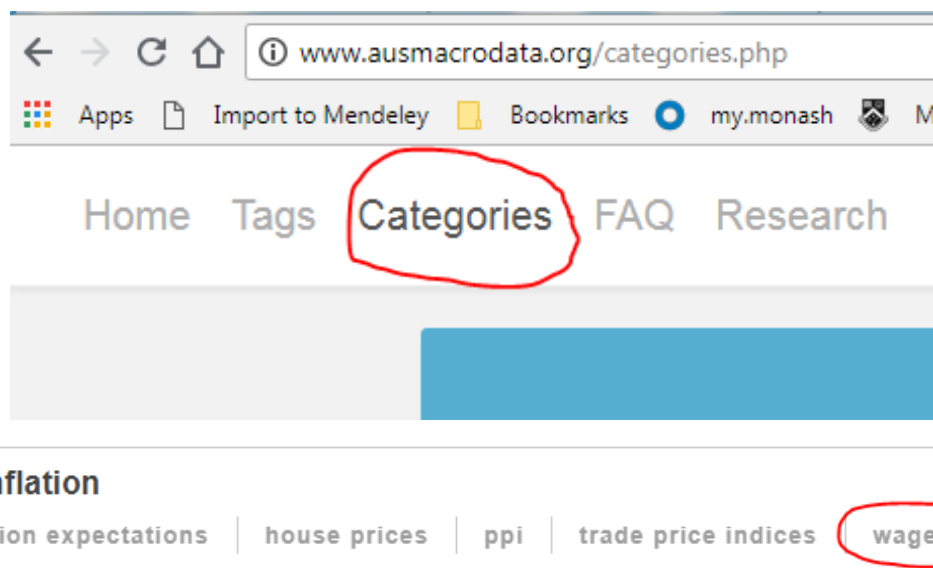
 **Lesson 1: EViews Software**

## **PART B: To be done in the tutorial.**

**Question 1:** *Important topics of the first year statistics: Histograms, scatter plots, confidence intervals + remembering logarithms:* Open the EViews workfile that you generated in Part A. If you have not saved it or have difficulty finding it, download “A random sample of countries.wf1” from the moodle site and use that.

1. Obtain the summary statistic and histogram of GDPPC. Discuss what you can learn from the histogram and summary statistics. If you had a different set of 40 countries, would the summary statistics be the same?
2. Using the sample average and the sample standard deviation, compute the 95% confidence interval for the population mean of GDPPC (if you have forgotten the formula for the confidence interval for the population mean, refer to equation (11.1) on page 66 of ETC1000 lecture notes that are on moodle (under Extra Study Materials related to tutorial 1), or equation [C.23] in Appendix C of the textbook. It is OK to use the rule of thumb given in equation [C.26] of the textbook. Explain what this confidence interval shows. The average GDPPC of *all* countries in WDI data base in 2015 is \$17406. Does your confidence interval contain this value?
3. We want to explore the association between infant mortality and GDP per capita. What kind of a graph can give us an insight into the nature of this relationship? Based on this graph, are infant mortality and GDP per capita positively or negatively correlated? Is their relationship linear?
4. Generate the logarithmic transformation of GDPPC and INFMORT. Use the label LGDPPC for the natural logarithm of GDPPC and LINFMORT for the natural logarithm of INFMORT. Note: In EViews,  $\log(X)$  calculates the natural logarithm of  $X$ . EViews does not recognise  $\ln(X)$ . Explore the association between LINFMORT and LGDPPC with the aid of an appropriate graph. Compare and contrast it with what you got in the previous part.

**Question 2:** *Working with time series: plots, trends, seasonality, growth rate (log-returns):* Download hourly wages from [www.ausmacrodata.org](http://www.ausmacrodata.org) -> Categories -> wage price index -> the first series that shows up -> Download CSV. Open the CSV file, and tidy up the data set, i.e. only keep the first two columns and delete everything else, and give a better name for the second column, such as WAGE. Save the CSV file and read it in EViews. Note that EViews automatically realises that you have quarterly data.



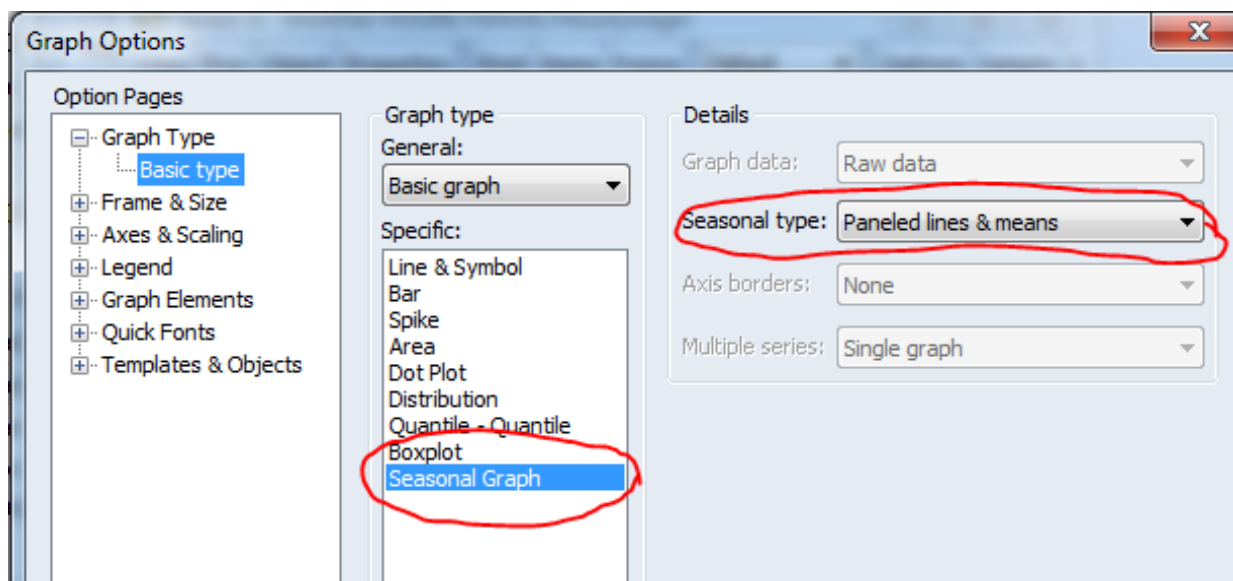
Other

1. Plot the WAGE series (plotting a time series means producing a line plot of the series in which the x-axis is time. In EViews, clicking on a series opens a window that shows the value of the series in a spreadsheet. This window has a menu bar. Under View is Graph. And the default for Graph is a line plot). What can we learn from this plot?
2. If we were interested in forecasting hourly wage in the next period, would sample average be a good forecast? Suggest more appropriate forecasts.
3. In most financial or economic time series, trend is so dominant that it is the only thing that we can immediately see and all other aspects of the time series are dwarfed by its trend. To see other aspects of the series we have to remove its trend. One way is to make a model with time as an explanatory variable (we will do this later in the course). Another way is to compute the growth rate, in this case  $g_t = 100 \times \frac{wage_t - wage_{t-1}}{wage_{t-1}}$  (multiplication by 100 is just to express it in percentage points). A more prevalent way of calculating the growth rate, in particular in finance, is to use what is known as “log-returns”

$$g_t = 100 \times \Delta \log(wage_t) = 100 \times (\log(wage_t) - \log(wage_{t-1}))$$

These two methods of calculating the growth rate produce values that are close to each other as long as growth rate is less than 10% in absolute value. EViews has a built in function ‘dlog(X)’ that computes the difference of logarithm of X. Generate the growth rate of hourly wage using the log-returns formula. Open this series. Why is the first value of this series NA?

4. Plot the growth rate of wage. What does this plot tell you?
5. Look at the seasonal plots of the growth rate of wage. To produce seasonal plots in EViews, in the series window, View -> Graph, and choose Seasonal Graph (the last option under Graph type). There are two seasonal plots: one that plots each season in a different panel side by side and shows average growth rate for each season. Another type shows four line plots, one for each season, overlayed on one graph. Look at both plots and discuss what you learn from these plots.



6. Look at the time series plot of the growth rate of wage again. Mentally adjust for seasonal variation. Do you see that the wage growth has been declining since 2010? Should that by itself worry us? What else do we need if we are worried about the value of one hour of work?