

## Tutorial 2

**keywords:** cross sectional data, time series data, histogram, summary statistics, scatter plot, confidence interval of population mean, logarithmic transformation, trends, seasonality

**estimated reading time:** 34 minutes

Quang Bui

July 31, 2018

# Question 1

## Background

### Types of data sets

- A *cross-sectional* data set consists of a sample of individuals, households, firms, cities, countries, or a variety of other units, taken at a given point in time. For example, 40 countries' wine consumption level, GDP per capita, and infant mortality in 2015.
- A *time-series* data set consists of observations on a variable or several variables over time. For example, Australia's wine consumption level, GDP per capita, and infant mortality every year from 2005 to 2015.
- A *panel* data set consists of a time series of *each* cross-sectional member in the data set. For example, 40 countries' wine consumption level, GDP per capita, and infant mortality every year from 2005 to 2015.

EViews workfile: *A random sample of countries.wf1*

*A random sample of countries.wf1* contains data on 40 randomly selected countries from the population of all countries. Since this data set was capture at a single time period, we call this a cross-sectional data set. The variables in this data set include:

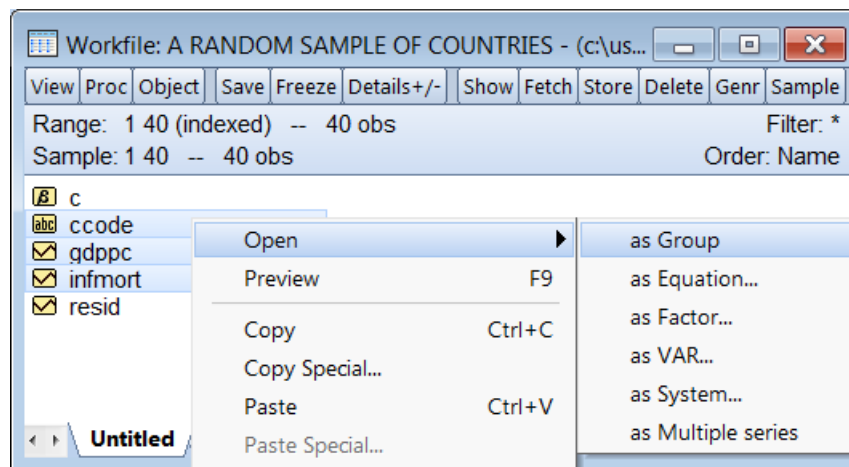
*ccode* – country code

*gdppc* – 2015 GDP per capita measured in USD & adjusted for purchasing power difference

*infmort* – 2015 infant mortality measured in number of deaths per 1000 lives

To view the data set, select and highlight *ccode*, *gdppc*, and *infmort* then,

*Right click* → *Open* → *as Group*



The screenshot shows the EViews Group: UNTITLED window titled "Group: UNTITLED Workfile: A RANDOM SAM...". The menu bar includes View, Proc, Object, Print, Name, Freeze, Default, Sort, Edit+/-, and SmpI. The table displays data for 10 countries, with columns for CCODE, GDPPC, and INFmort.

	CCODE	GDPPC	INFmort
DZA	DZA	13724.72	21.9
ARG	ARG	19101.30	10.3
ARM	ARM	8195.934	12.5
AUS	AUS	43832.43	3.2
BGD	BGD	3132.568	29.7
BOL	BOL	6531.519	30.5
BRA	BRA	14666.02	14.0
CAN	CAN	42983.10	4.5
CHL	CHL	22536.62	7.3
CHN	CHN	13569.89	9.2
CIV			

1. Obtain the summary statistics 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?

## Background

### Population parameters and sample statistics

Population parameters describe the true behaviour of the population considered. For example, the population mean of 2015 GDP per capita is a value that describes the average GDP per capita in 2015 in the population of all countries. Without population data (and this is often the case in practice), we cannot calculate population parameters and the true behaviour of the population is unknown.

Greek letters are used to denote population parameters,

- Population mean,  $\mu$
- Population standard deviation,  $\sigma$
- Population variance,  $\sigma^2$
- Population intercept coefficient,  $\beta_0$
- Population slope coefficient of variable  $j$ ,  $\beta_j$
- etc.

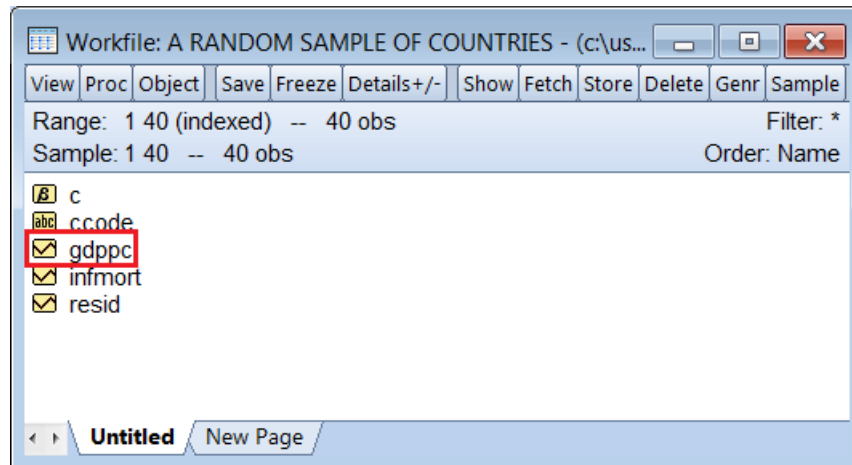
Since population data is often unavailable in practice, we cannot calculate population parameters and the true behaviour of the population is unknown. In practice, we draw a random sample from the population (we hope this is representative of the population) and use this sample to make inferences and learn about the true behaviour of the population. For example, suppose it were not feasible to gather data on the 2015 GDP per capita of every country, so instead, we obtain a sample of 40 randomly selected countries,

#	Country Code	GDP Per Capita
1	DZA	13724.72
2	ARG	19101.30
3	ARM	8195.934
4	AUS	43832.43
5	BGD	3132.568
6	BOL	6531.519
7	BRA	14666.02
8	CAN	42983.10
9	CHL	22536.62
$\vdots$	$\vdots$	$\vdots$
40	VNM	5667.409

With this sample, we can obtain summary statistics of the 2015 GDP per capita (among other things). Summary statistics, which summarise the central tendency, variability, and shape of a variable, include sample statistics like the sample mean, sample median, sample standard deviation, sample maximum & sample minimum and are estimates of the population mean, population median, population standard deviation, population maximum & population maximum.

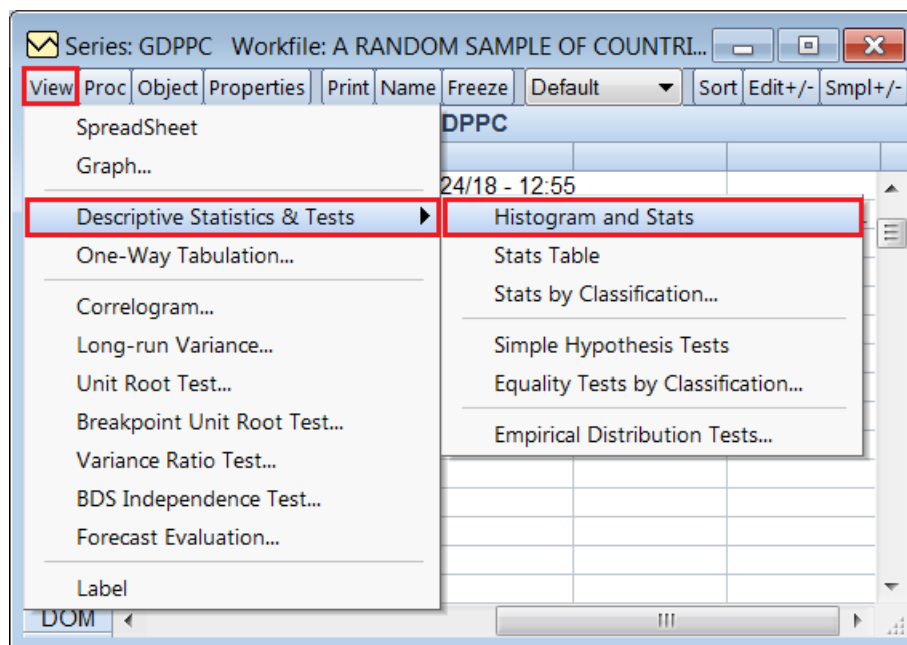
Sample statistics are random variables because they vary from sample to sample e.g. with a different sample of 40 countries, the sample mean changes. In fact, if we applied repeated sampling, we would have a ‘sample of sample means’ and this too would have its own set of summary statistics.

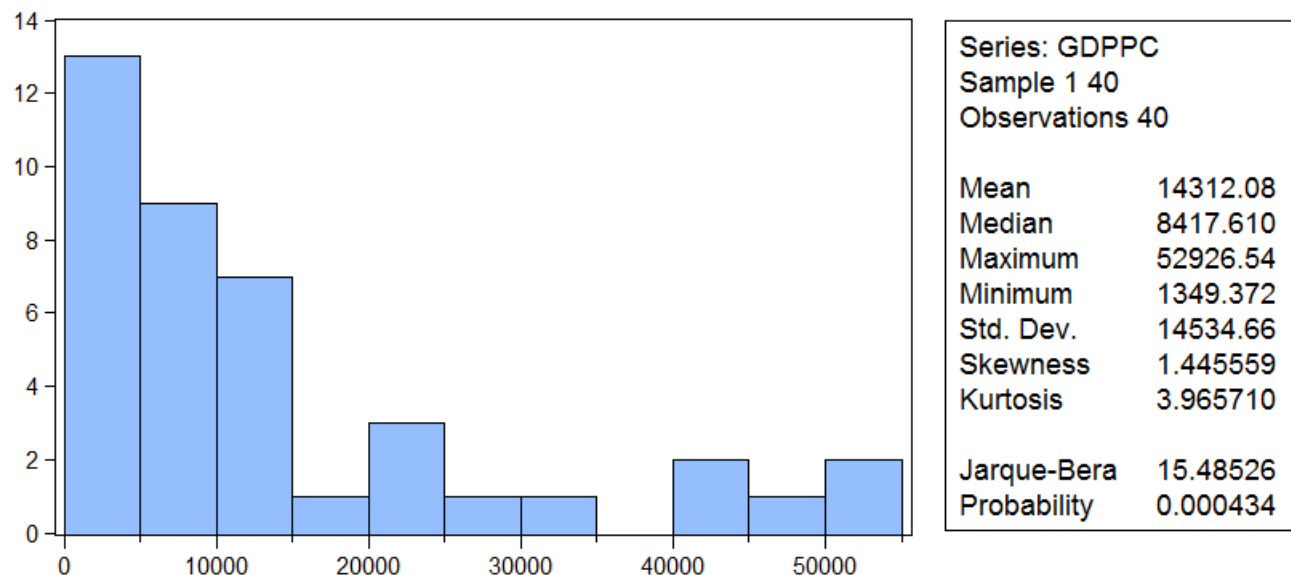
To obtain the summary statistics and histogram of *gdppc*,



double-click *gdppc* then,

*View* → *Descriptive Statistics & Tests* → *Histogram and Stats*





- GDP per capita has a positively skewed (right-tailed) distribution.
- The sample mean of GDP per capita is much higher than the sample median

$$\text{sample mean} = \$14312.08 > \text{sample median} = \$8417.61$$

which suggests huge inequality among these 40 countries i.e. some countries in our random sample have a particularly high GDP per capita relative to the rest, which causes the mean to be greater than the median.

- The minimum GDP per capita is \$1349.37, which is less than 1/6 of the median in this sample and translates to \$3.70 per day per person. (The more descriptive your explanation, the better. Since GDP per capita is skewed, the median is a better measure of the central tendency of GDP per capita than the mean, so we compare the minimum against the median instead of the mean.)
- The maximum GDP per capita is \$52926.54, which is more than 6 times the median in this sample and translates to \$145 per day per person.

If we had 40 different countries, the summary statistics would change. This means that each sample statistic e.g. sample mean, sample median, sample maximum, sample minimum, sample standard deviation, etc. are random variables i.e. their values are uncertain and differ from sample to sample.

2. Using the sample average (mean) and sample standard deviation, compute the 95% confidence interval for the population mean of *gdppc*. The confidence interval is given by equation [C.23] in Appendix C of the textbook. It is okay to use the rule of thumb given by equation [C.26] in the textbook. Explain what the confidence interval shows. The average *gdppc* for all countries in the WDI data base in 2015 is \$17406. Does your confidence interval contain this value?

## Background

### Point Estimate of Population Mean

The sample mean is a point estimate of the population mean and of course, there are many estimators of the population mean, e.g. randomly picking a number from the sample and using this as an estimate of the population mean, but this is clearly a poor estimator of the population mean. Point estimates alone do not inform us about the precision with which we have estimated the population parameter.

$$\overline{gdppc} = 14312.08$$

$$\mu = ???$$

### Confidence Interval of Population Mean

Unlike a point estimator, a confidence interval provides a range of values

$$[\dots, \dots]$$

so that we have a sense of what the population parameter might be, and the precision with which we have estimated it.

For example, a 95% confidence interval of the population mean tells us that we are 95% confident that the population mean will lie between this range of values. If you can be 95% confident that the population mean lies between a narrow range of values,

$$[14302.18, 14322.08]$$

then you would be very confident about your point estimate of the population mean (think about the formula of the confident interval, and what causes the width of the interval to increase/decrease).

Using the simple rule of thumb for a 95% confidence interval of the population mean,

$$[\overline{gdppc} \pm 2 \times se(\overline{gdppc})]$$

$$[\overline{gdppc} - 2 \times se(\overline{gdppc}), \overline{gdppc} + 2 \times se(\overline{gdppc})]$$

$$\begin{aligned}
& [\overline{gdppc} - 2 \times \frac{\text{sample std.dev}_{gdppc}}{\sqrt{n}}, \overline{gdppc} + 2 \times \frac{\text{sample std.dev}_{gdppc}}{\sqrt{n}}] \\
& [14312.08 - 2 \times \frac{14534.66}{\sqrt{40}}, 14312.08 + 2 \times \frac{14534.66}{\sqrt{40}}] \\
& [9716, 18908]
\end{aligned}$$

To interpret this confidence interval, we say that there is a 95% chance that the 2015 population mean of GDP per capita lies between \$9716 and \$18908. It is incorrect to say that “95% of the time, the 2015 population mean of GDP per capita will lie between \$9716 and \$18908”.

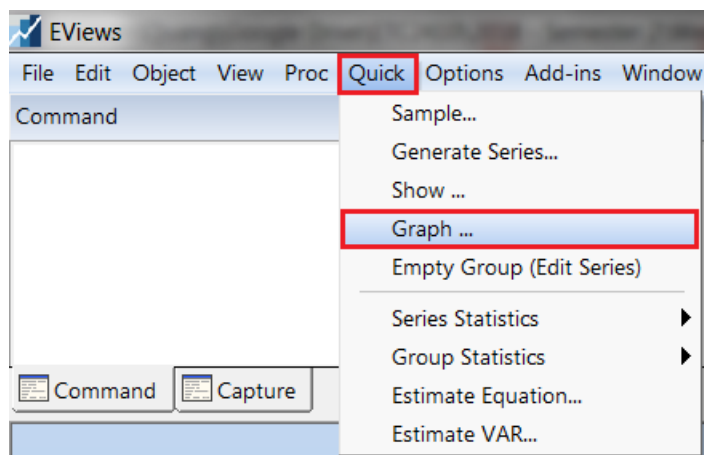
This confidence interval contains \$17406, which is the mean GDP per capita per annum in 2015 for all countries (this is a population mean so it is constant and does not change). Since the confidence interval depends on the sample mean and sample standard deviation, which differ from sample to sample, a different sample of 40 countries results in a slightly different confidence interval.



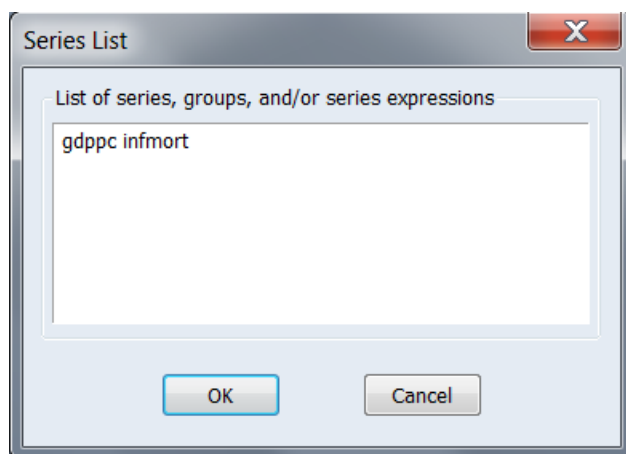
3. We want to explore the association of infant mortality and GDP per capita. What kind of graph can give us an insight into the nature of this relationship? Based on this graph, are infant mortality and GDP per capita positively or negatively correlated? Is their relationship linear?

Scatter plots can give us an insight into the nature of the relationship between two variables. To obtain a scatter plot of infant mortality (y-axis) against GDP per capita (x-axis),

*Quick*  $\rightarrow$  *Graph* . . .

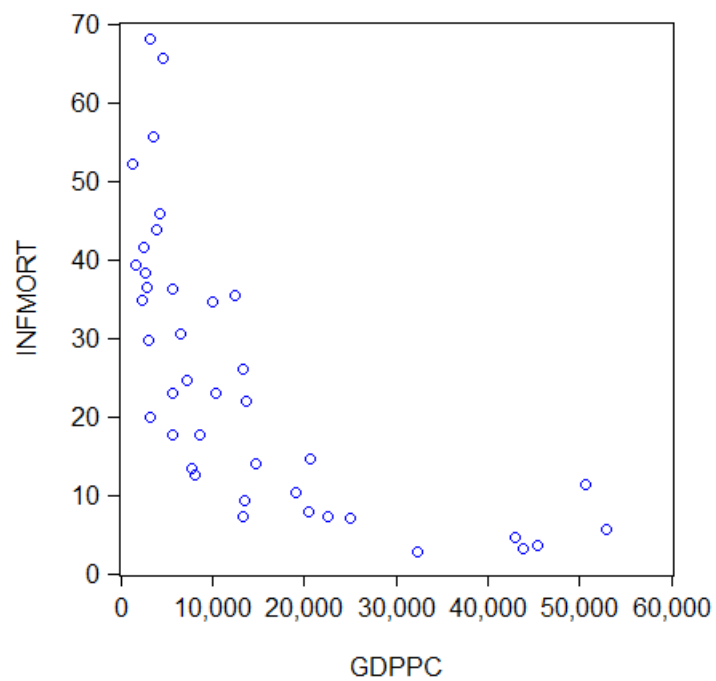
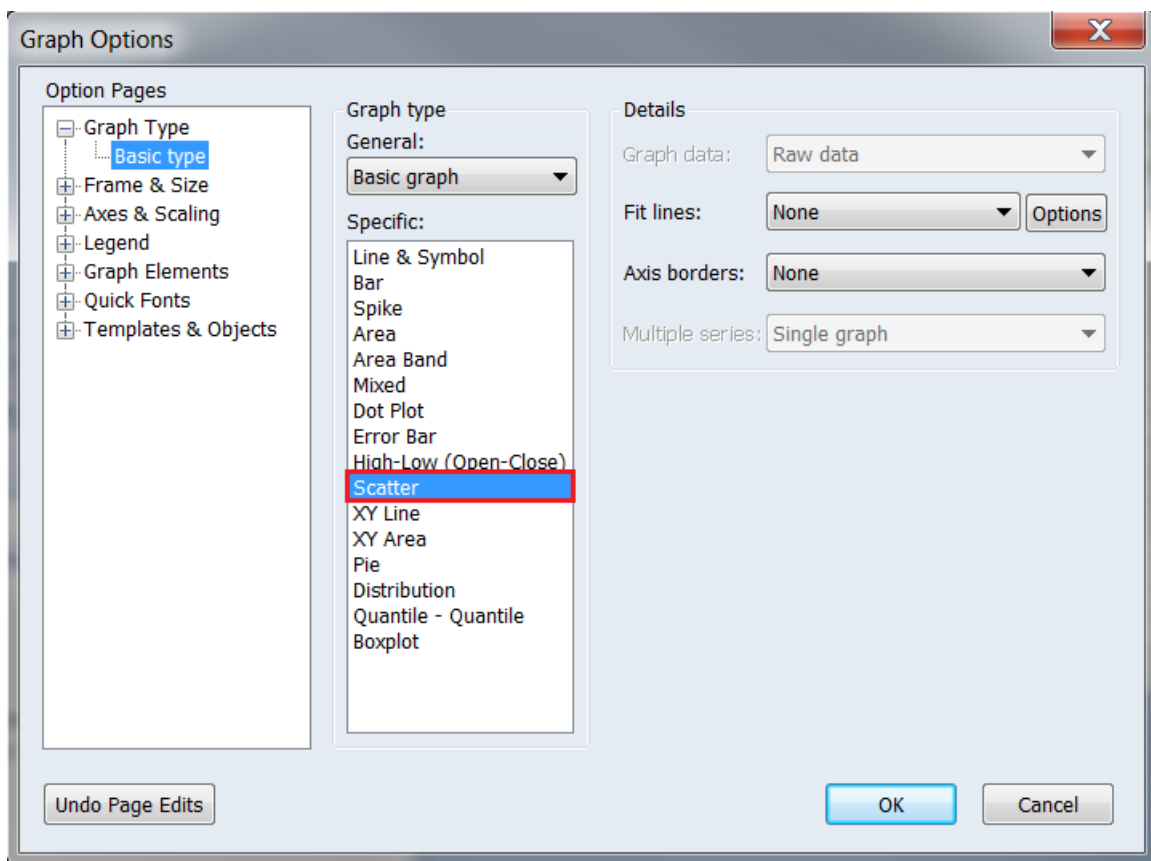


in the *Series List* window, type the variable to go on the x-axis followed by the variable to go on the y-axis,



Under *Specific*, select *Scatter* and press *OK*,

*Specific* : *Scatter*  $\rightarrow$  *OK*

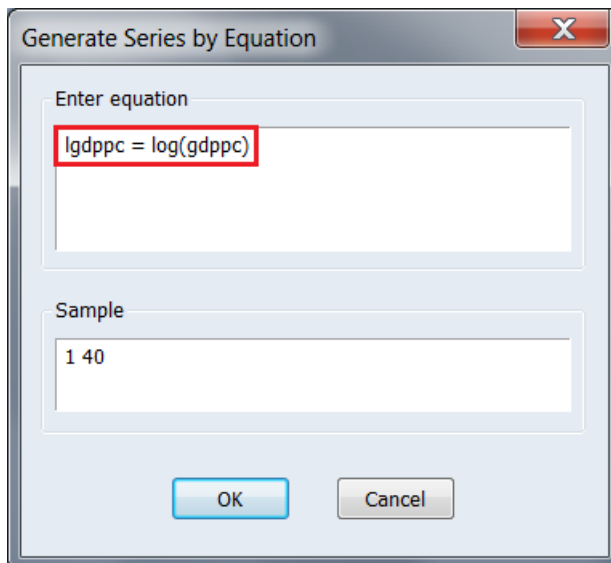
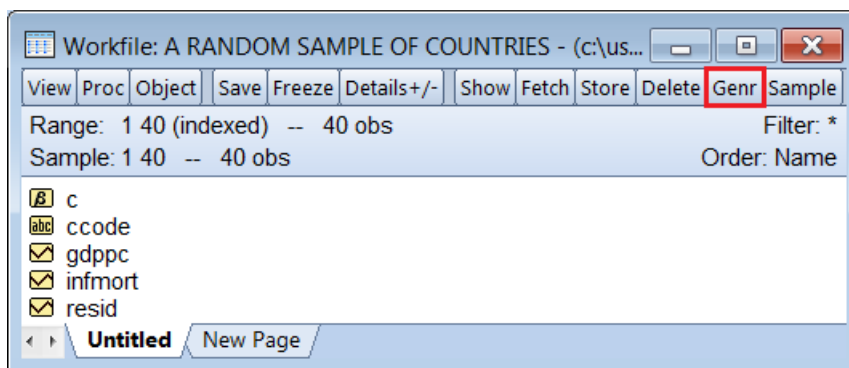


The scatter plot shows that infant mortality and GDP per capita are negatively related but this relationship is non-linear.

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.

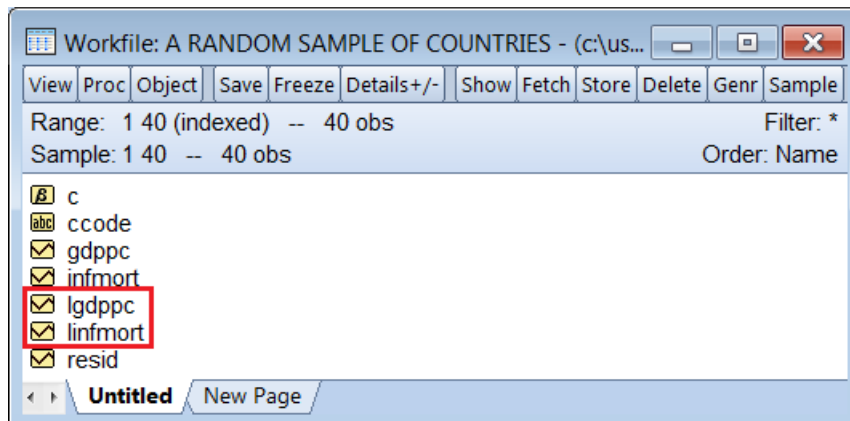
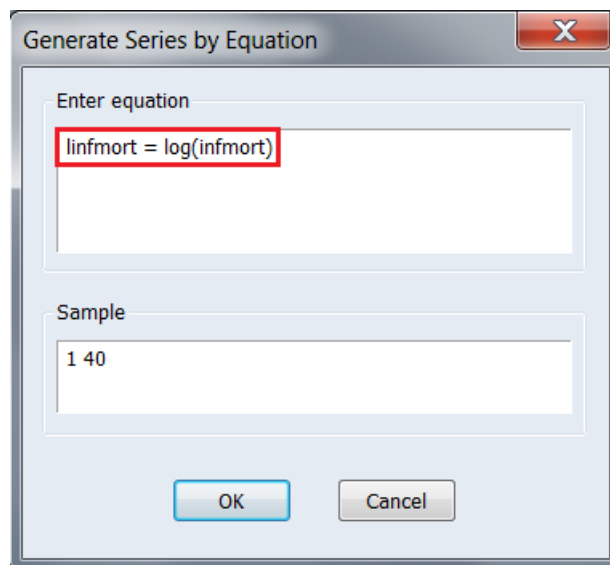
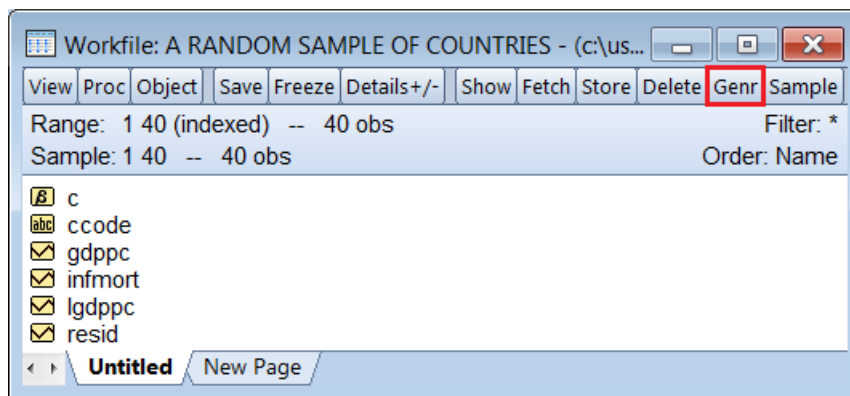
To generate the variables *lgdppc*, which are the natural logarithm of *gdppc*, in EViews,

*Genr* → Enter equation :  $\text{lgdppc} = \log(\text{gdppc})$  → OK



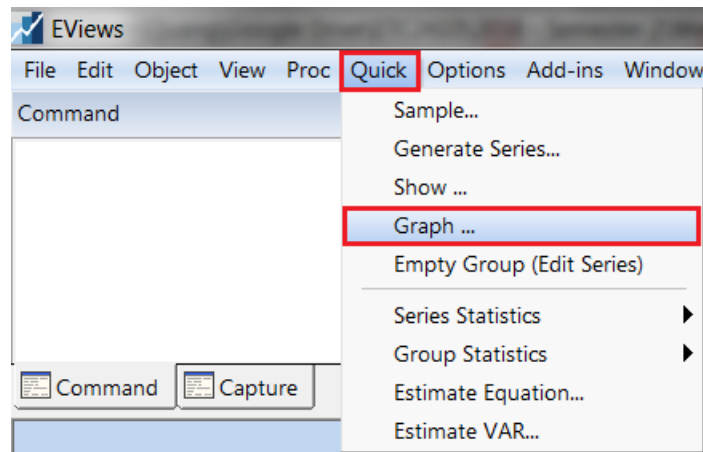
To generate the variables *linfmort*, which are the natural logarithm of *infmort*, in EViews,

*Genr* → Enter equation :  $\text{linfmort} = \log(\text{infmort})$  → OK

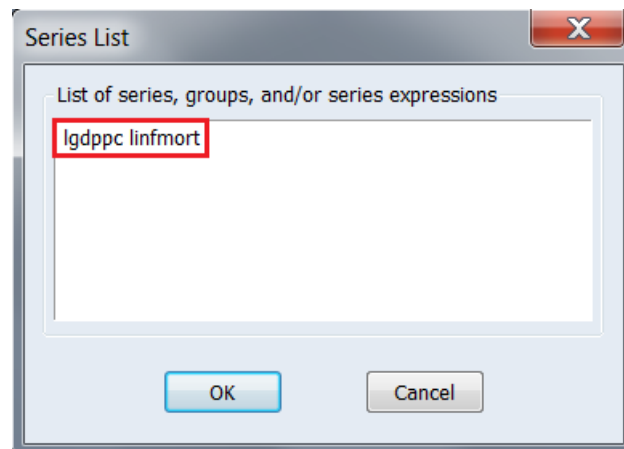


To obtain a scatter plot of  $linfmort$  (y-axis) against  $lgdppc$  (x-axis),

*Quick*  $\rightarrow$  *Graph ...*

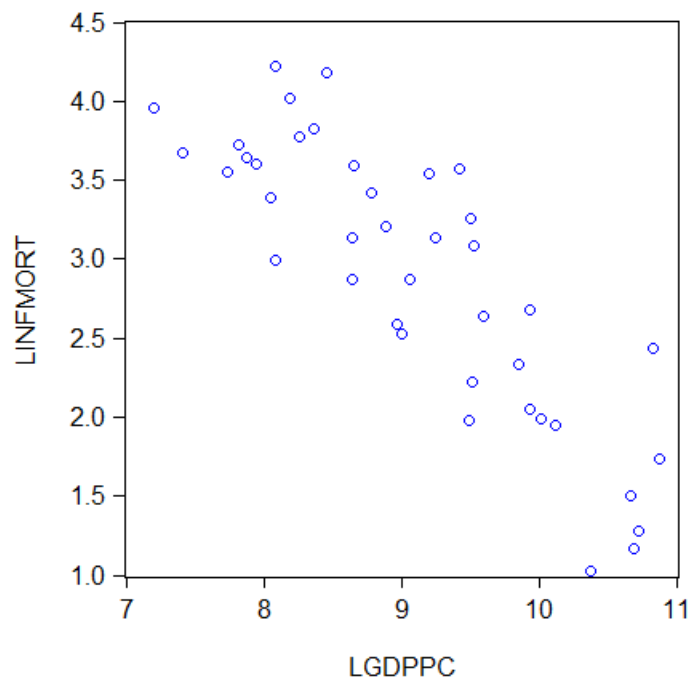
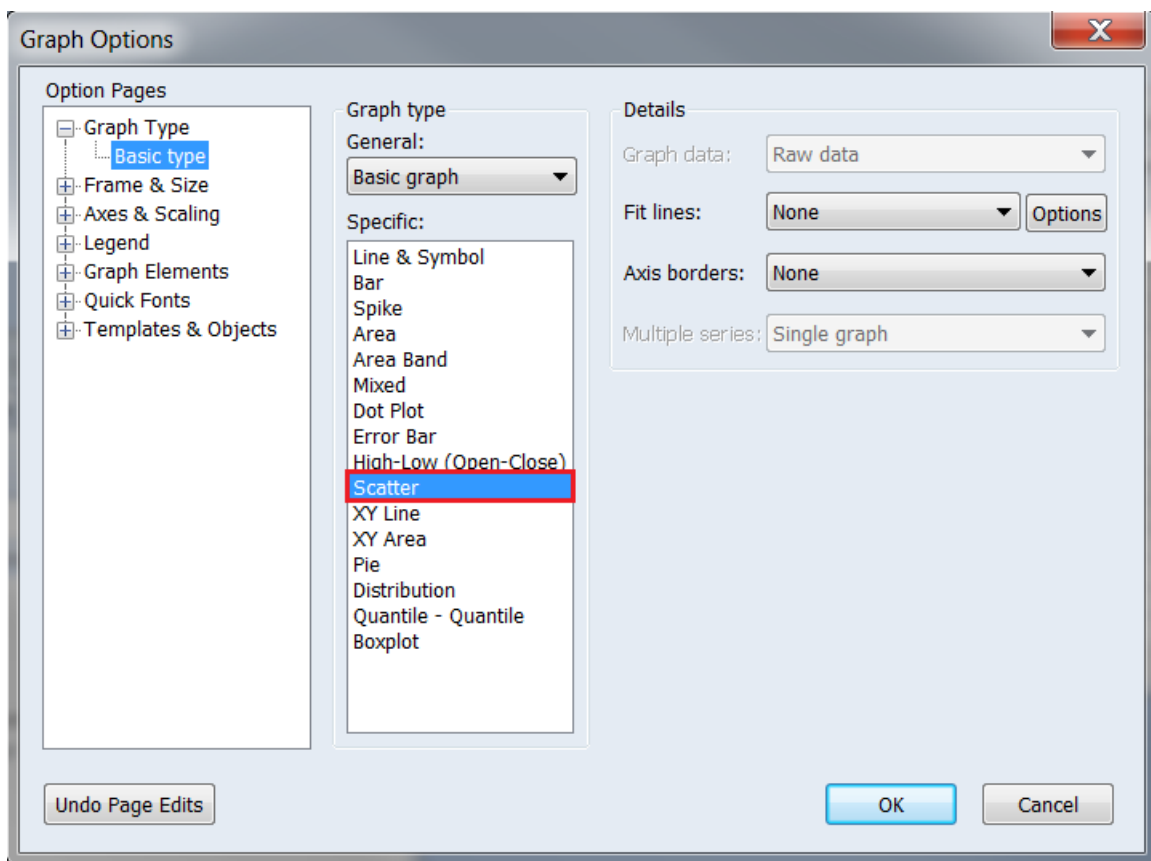


*Series List* :  $lgdppc$   $linfmort$



Under *Specific*, select *Scatter* and press *OK*,

*Specific* : *Scatter*  $\rightarrow$  *OK*



Unlike *infmort* & *gdppc*, which have a non-linear relationship, the log of infant mortality & log of GDP per capita have a stronger linear relationship. We learn from this exercise that we may have to transform the raw data before we model their relationship in a model with linear parameters.

Consider the following simple linear regression models,

$$l\text{infmort} = \beta_0 + \beta_1 l\text{gdppc} + u \quad (1)$$

$$\text{infmort} = \alpha_0 + \alpha_1 \text{gdppc} + v \quad (2)$$

Both models are linear in the parameters, but (1) specifies a linear relationship between log of infant mortality and log of GDP per capita, while (2) specifies a linear relationship between infant mortality and GDP per capita. Clearly, (1) is more appropriate than (2).

This exercise also tells us that for every dollar change in GDP per capita, the change in infant mortality is not constant, rather, for every percentage change in GDP per capita, the subsequent percentage change in infant mortality is constant (more on this next time).

## Question 2

Download hourly wage from

*www.ausmacrodata.org* → *Categories* → *wage price index* → *first series that shows up*  
→ *Download CSV*

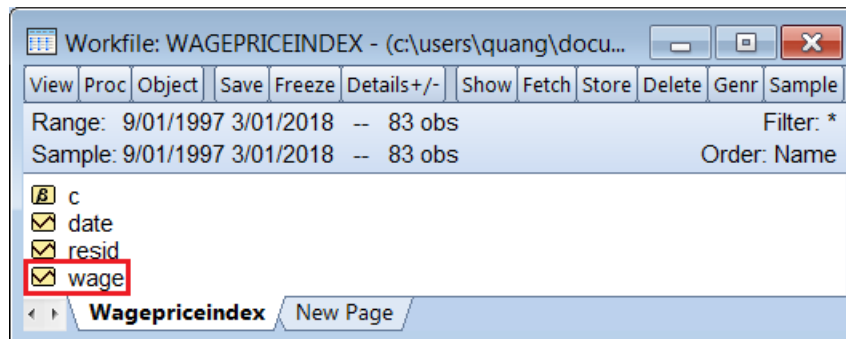
Open the CSV file and tidy up the data set i.e. only keep the first two columns and deleted everything else, and give a better name for the second column, such as *wage*.

	A	B
1	date	wage
2	Sep-97	67.4
3	Dec-97	67.9
4	Mar-98	68.5
5	Jun-98	68.8
6	Sep-98	69.6
7	Dec-98	70
8	Mar-99	70.4
9	Jun-99	70.8
10	Sep-99	71.5

Save the CSV file and read it in EViews. Note that EViews automatically realises that you have quarterly data.

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 values of the series in a spreadsheet. This window has a menu bar. Under *View* is *Graph* and the default graph is a line plot.) What can we learn from this plot?

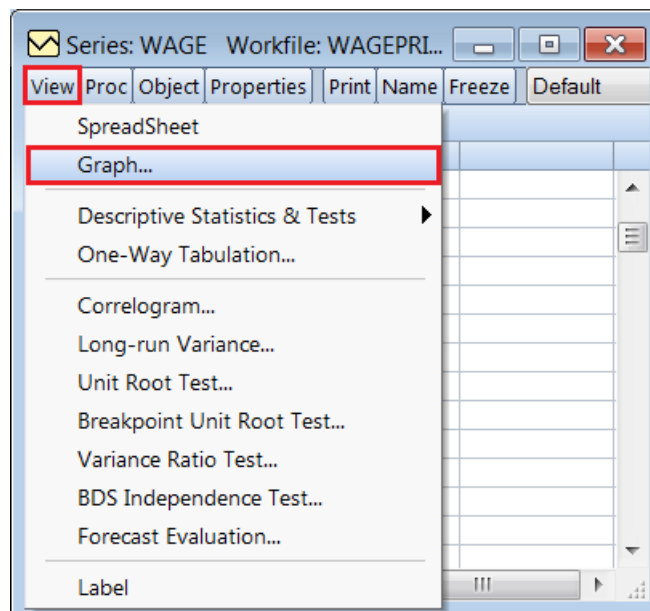
Double click on *wage*,



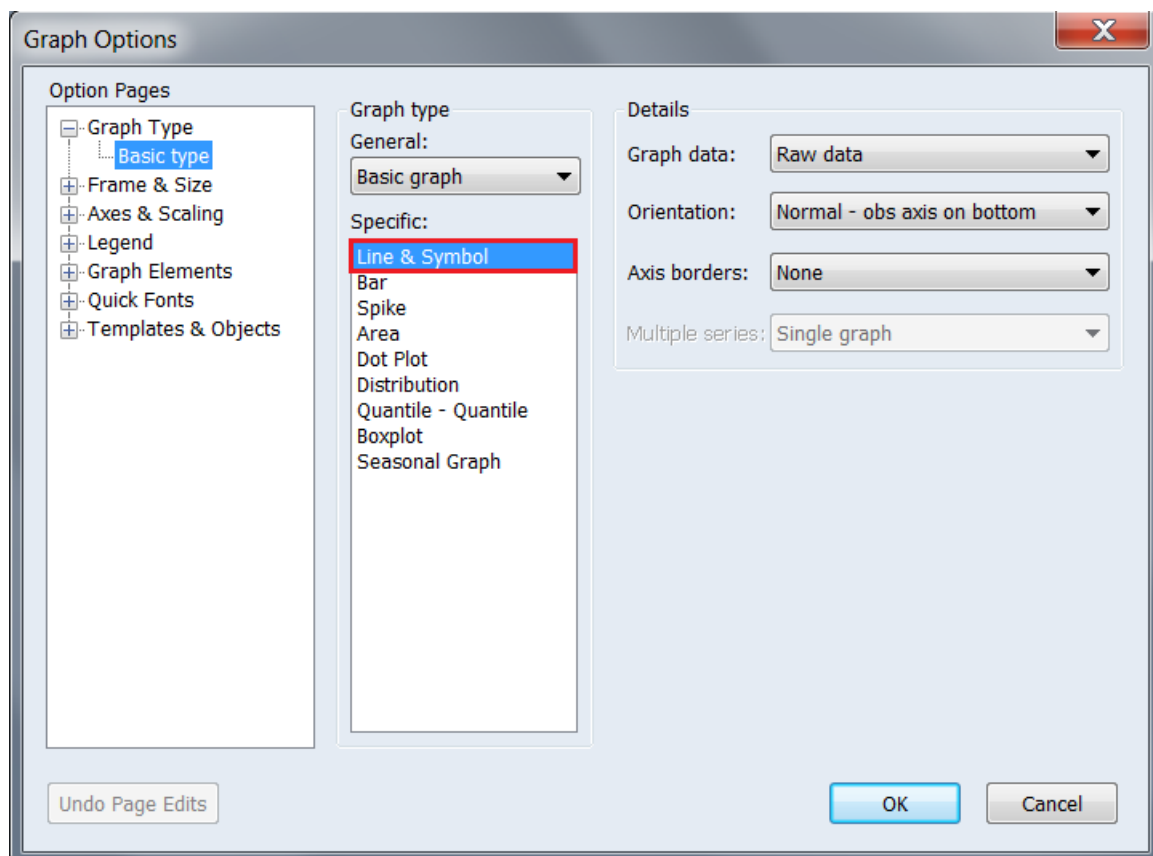
then,

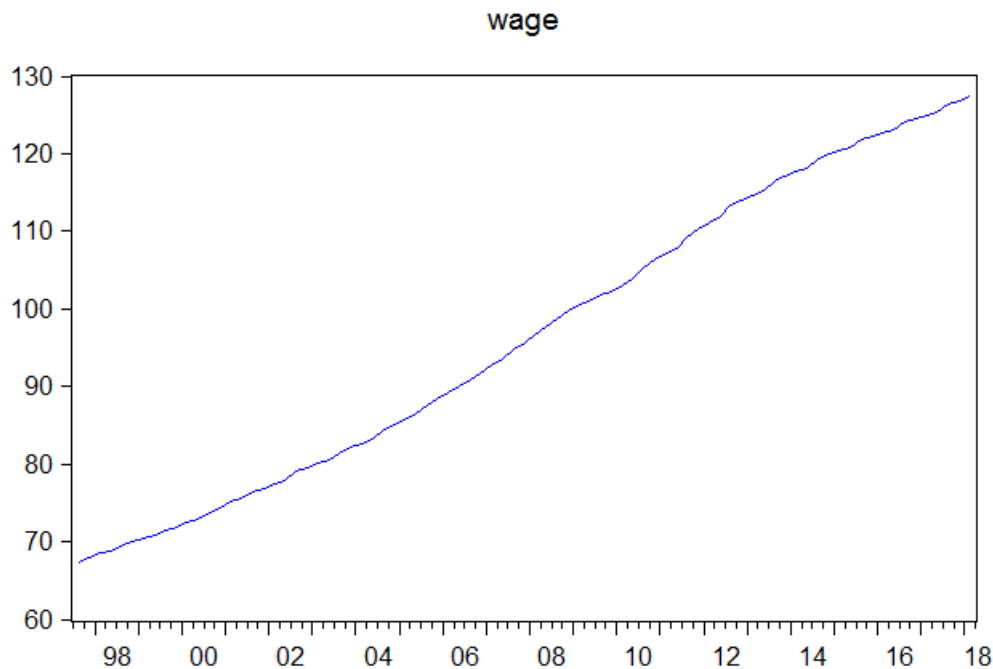
*View* → *Graph*





*Specific : Line & Symbol*





## Background




### Trends and seasonality

A persistent downward or upward movement, often over 10+ years, indicates that there is a *trend* in the time series. If we observe fluctuations within a one-year period that repeats in the same fashion year-after-year, then the time series exhibits *seasonality*.





From the line graph of hourly wage, we observe an upward trend i.e. hourly wage increases with time. Hourly wage is also highly persistent as each observation is very close to observations before it.

2. If we were interested in forecasting hourly wage in the next period, would sample average be a good forecast? Suggest more appropriate forecast.

No, the sample mean of hourly wage would not be a good forecast of hourly wage in the next period because the mean increases with time. For example, the mean hourly wage index in 2017 is higher than the mean in 2016, 2015, etc. and the mean hourly wage index in 2016 is higher than the mean in 2015, 2014, etc.

Series: WAGE Work...   

View	Proc	Object	Properties	Print	Name	Freez
wage						
12/01/2013		117.0				
3/01/2014		117.7				
6/01/2014		118.2				
9/01/2014		119.3				
12/01/2014		119.9				
3/01/2015		120.3				
6/01/2015		120.8				
9/01/2015		121.8				
12/01/2015		122.3				
3/01/2016		122.7				
6/01/2016		123.1				
9/01/2016		124.1				
12/01/2016		124.5				
3/01/2017		124.9				
6/01/2017		125.4				
9/01/2017		126.4				
12/01/2017		126.9				
3/01/2018		127.4				

We could use the previous observation of hourly wage as a forecast of hourly wage in the coming period but there are many better approaches. Another example is to base the next quarter's growth on an average quarter on quarter growth of hourly wage of the same quarters.

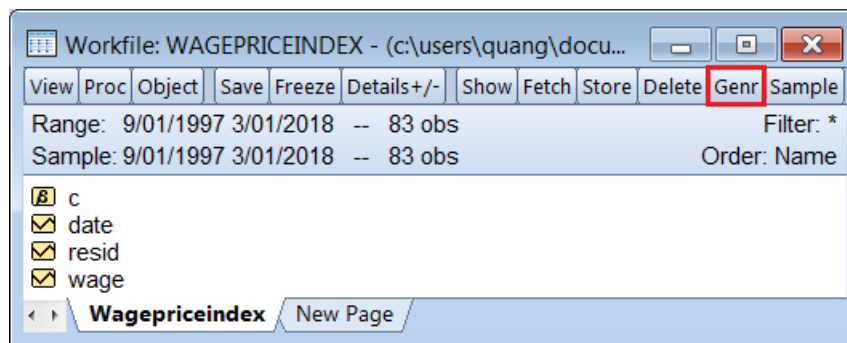
- Forecasting hourly wage in 2018 quarter 2.
- Compute average growth from quarter 1 to quarter 2.
- Assume quarterly growth from 2018 quarter 1 to 2018 quarter 2 is the same as the average growth from quarter 1 to quarter 2.
- $forecast\_2018\_quarter\_2 = 2018\_quarter\_1 \times (1 + average\_growth\_Q1\_to\_Q2)$

3. Generate the growth rate of hourly wage using the log-returns formula.

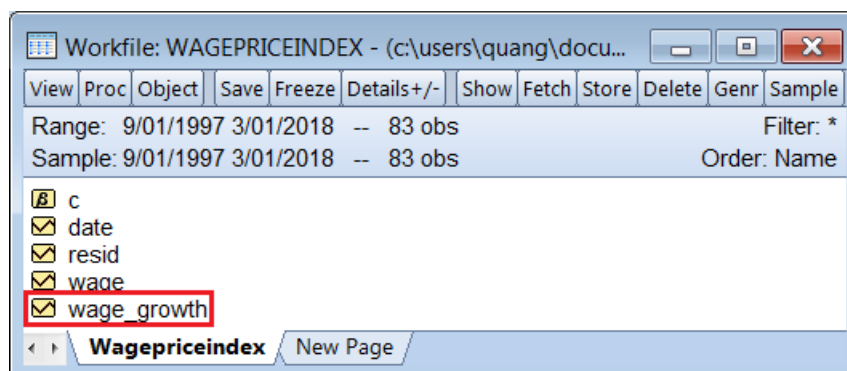
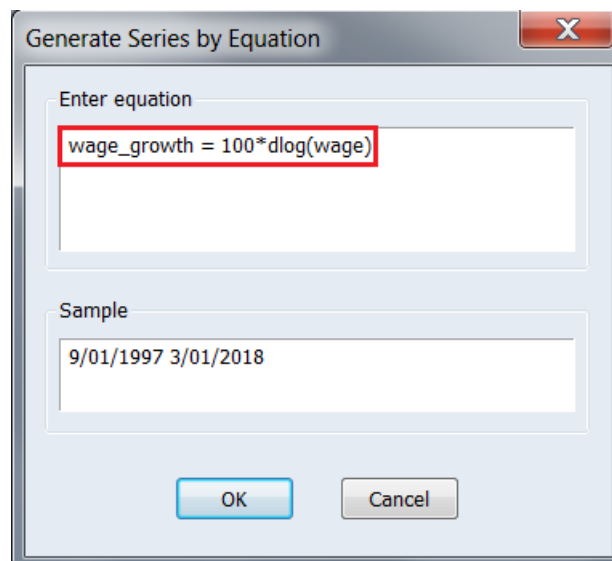
$$\begin{aligned}
 g_t &= 100 \times \Delta \log(wage_t) \\
 &= 100 \times (\log(wage_t) - \log(wage_{t-1}))
 \end{aligned}$$

EViews has a built in function 'dlog(X)' which computes the difference of logarithm of X. Open this series. Why is the first value of this series NA?

To generate the growth rate of hourly wage using the log-returns formula in EViews click *Genr*,



*Enter Equation :  $wage\_growth = 100 * dlog(wage)$*



Series: WAGE\_GROWTH    Workfile: WAGEPRICEINDEX...

View Proc Object Properties Print Name Freeze Default Sort Edit+/- Sm

**WAGE\_GROWTH**

Last updated: 07/30/18 - 19:50  
Modified: 9/01/1997 3/01/2018 // wage\_growth = 100\*dlog(wage)

9/01/1997	NA
12/01/1997	0.739102
3/01/1998	0.879771
6/01/1998	0.437000
9/01/1998	1.156082
12/01/1998	0.573067
3/01/1999	

The first value in this series is the growth rate of hourly wage in 1997 quarter 3 and is *NA* because we require hourly wage in 1997 quarter 2 to compute this value, which we do not have.

$$\begin{aligned}
 wage\_growth_t &= 100 \times (\log(wage_t) - \log(wage_{t-1})) \\
 wage\_growth_{1997Q3} &= 100 \times (\log(wage_{1997Q3}) - \log(wage_{1997Q2})) \\
 &= 100 \times (\log(67.4) - \log(wage_{1997Q2}))
 \end{aligned}$$

Series: **WAGE**    Wor...

View Proc Object Properties Print Name Freez

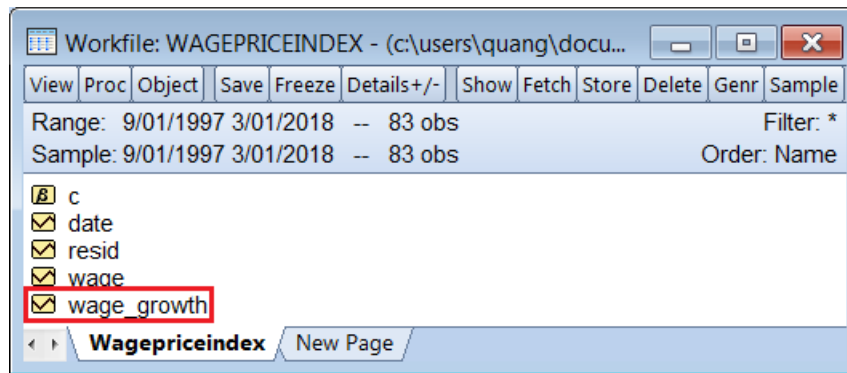
**wage**

Last updated: 07/30/18 - 18...  
Imported from 'C:\Users\Q...

9/01/1997	67.4
12/01/1997	67.9
3/01/1998	68.5
6/01/1998	68.8
9/01/1998	69.6
12/01/1998	70.0
3/01/1999	

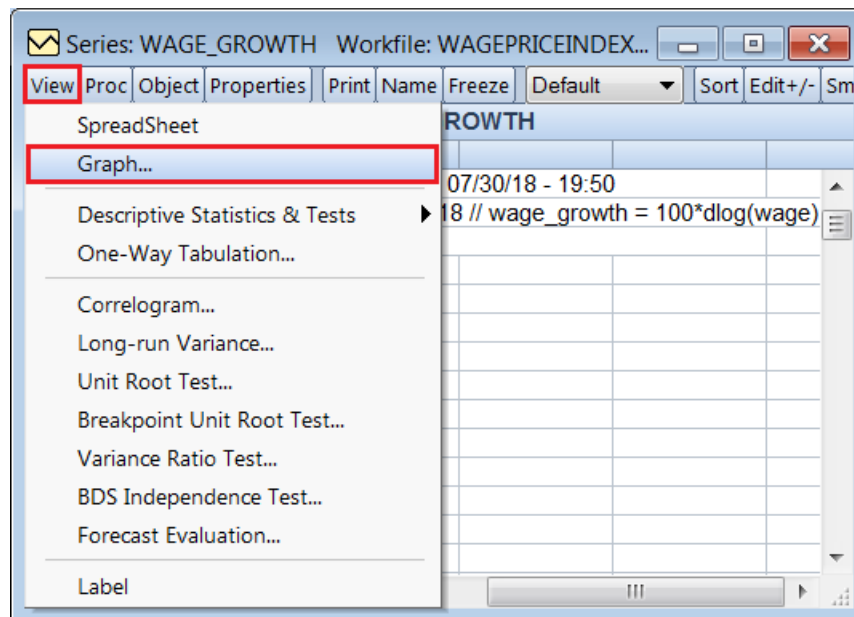
4. Plot the growth rate of wage. What does this plot tell you?

Open *wage\_growth*,

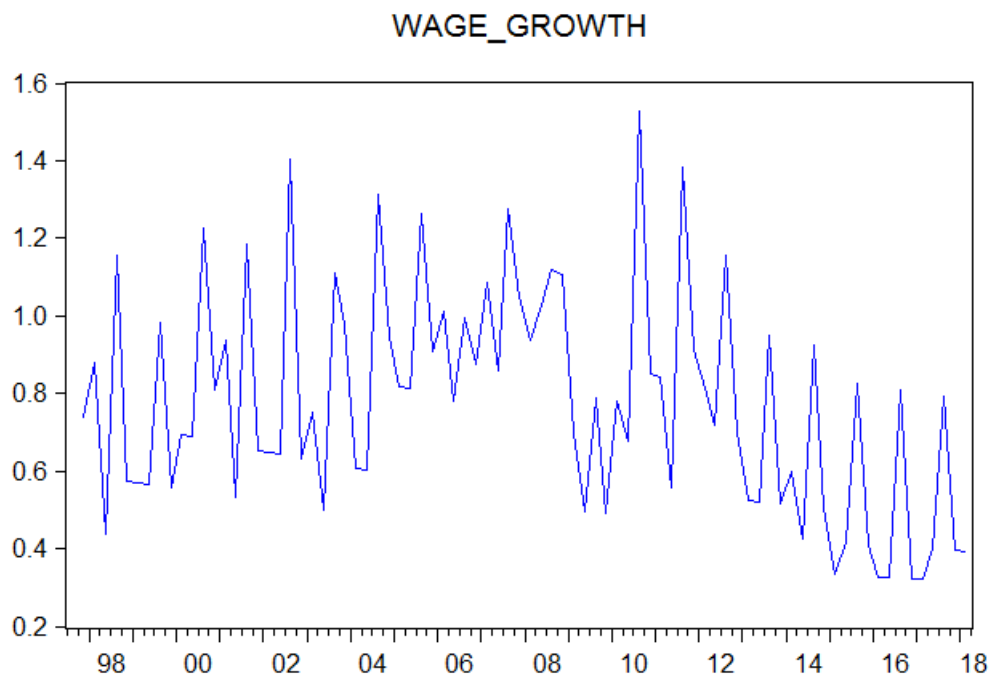
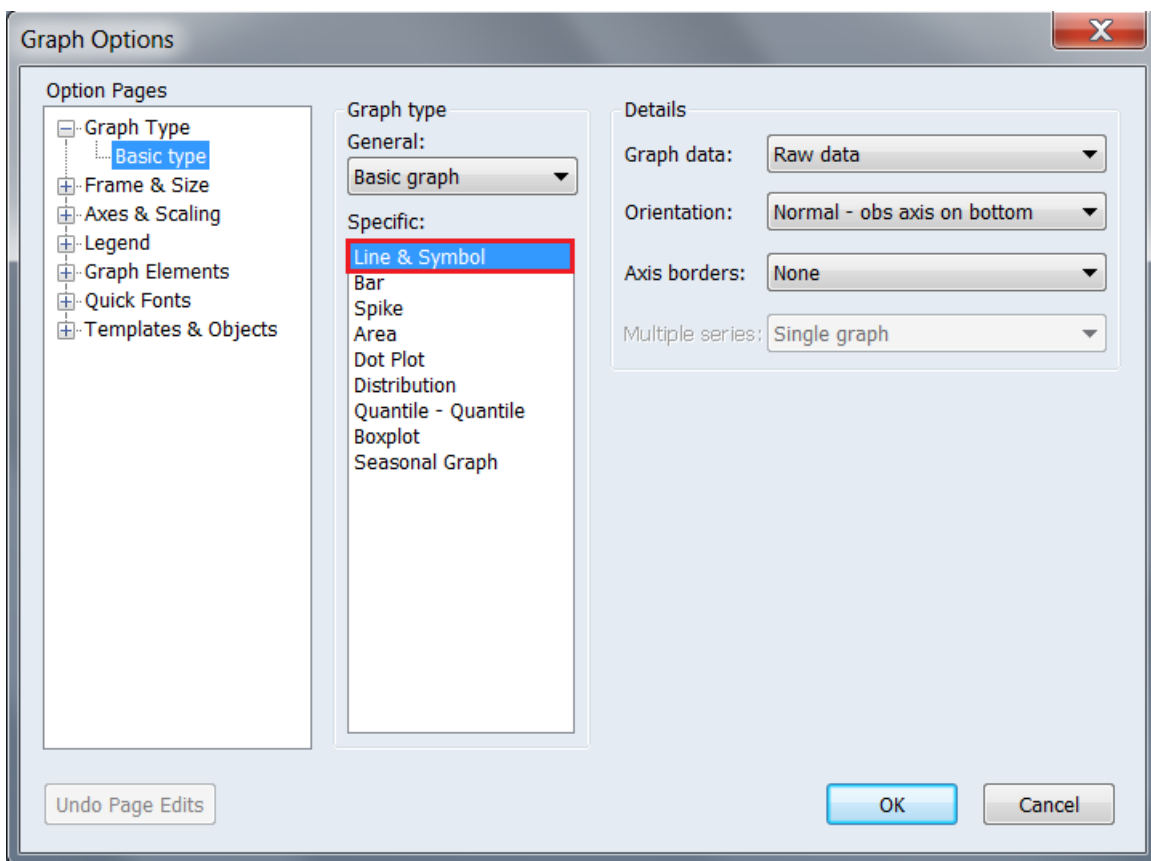


then

*View → Graph*

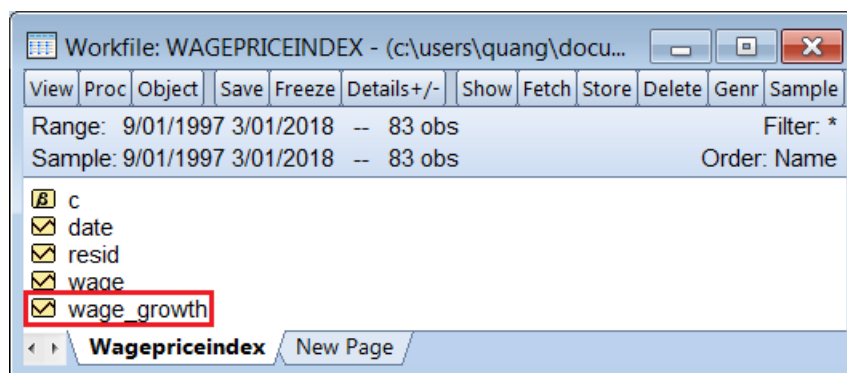


*Specific : Line & Symbol*



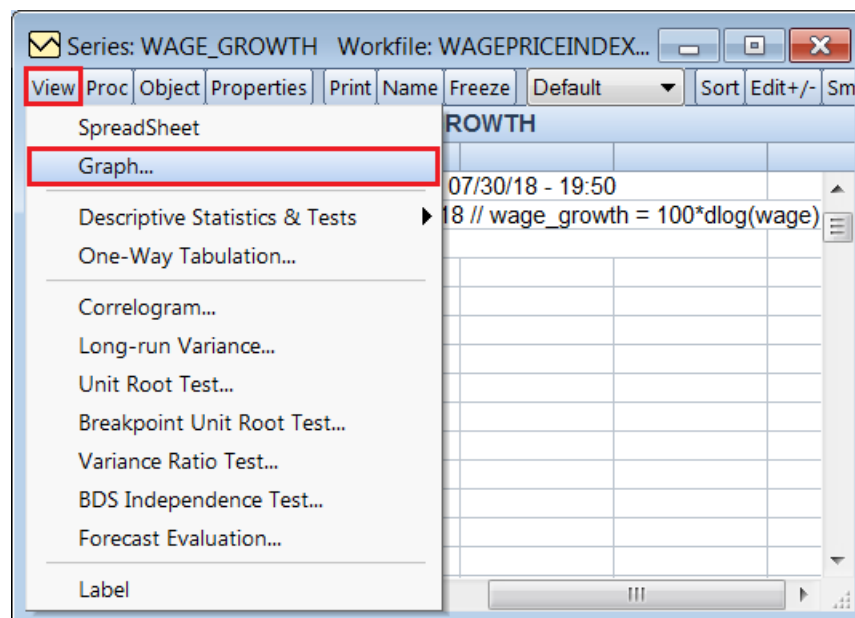
There is clear seasonality in the growth rate of hourly wage i.e growth rate of hourly wage is highest in the 3rd quarter of each year. This may be due to the fact that most people's wage increases become effective at the beginning of the financial year.

5. Look at the seasonal plots of the growth rate of wage. To plot seasonal plots of *wage\_growth*, open *wage\_growth*, Open *wage\_growth*,



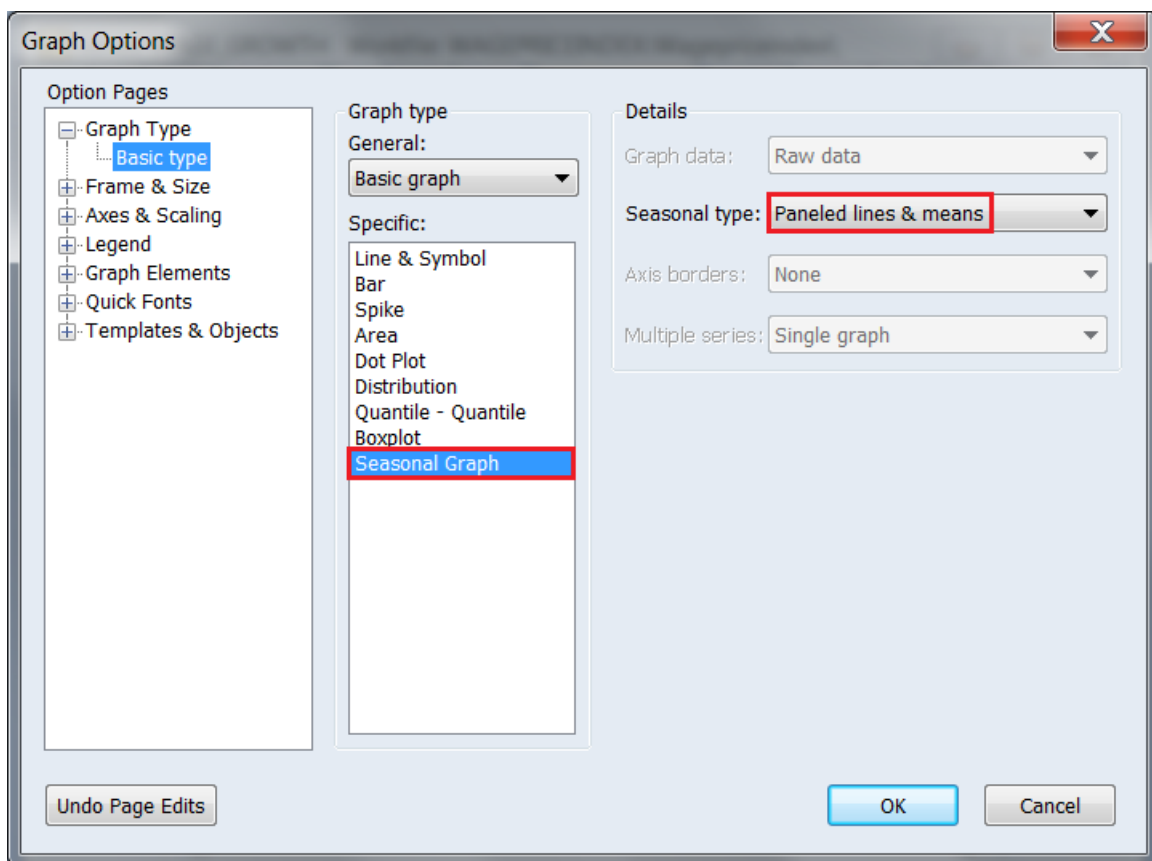
then

*View → Graph*

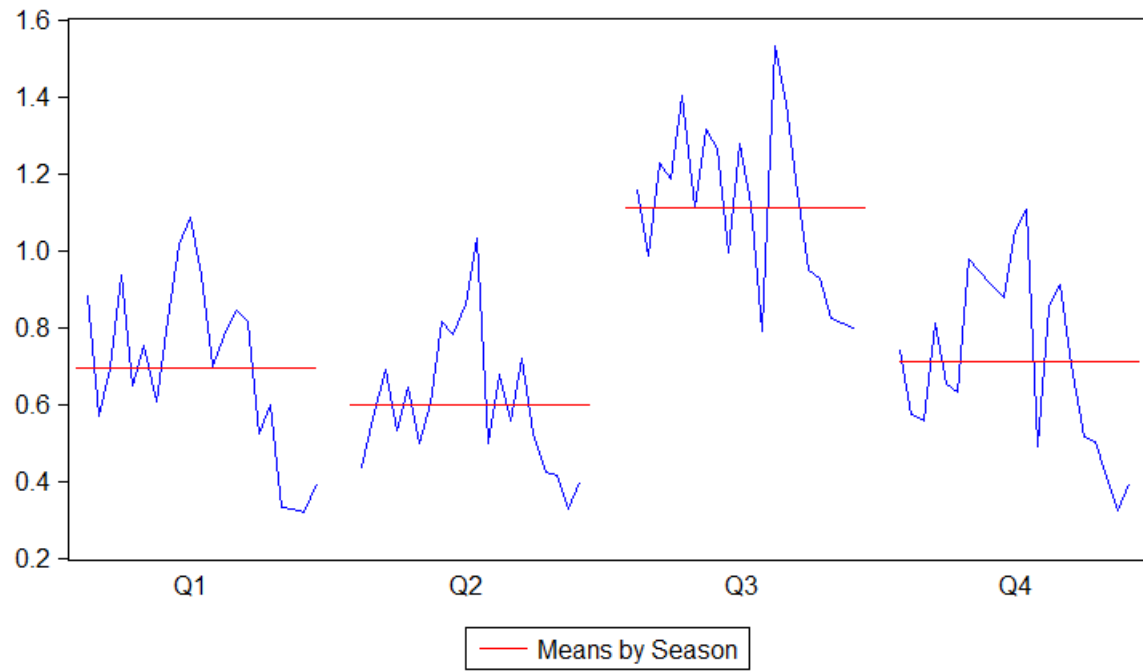


*Specific : Seasonal Graph Seasonal type : Paneled lines & means*

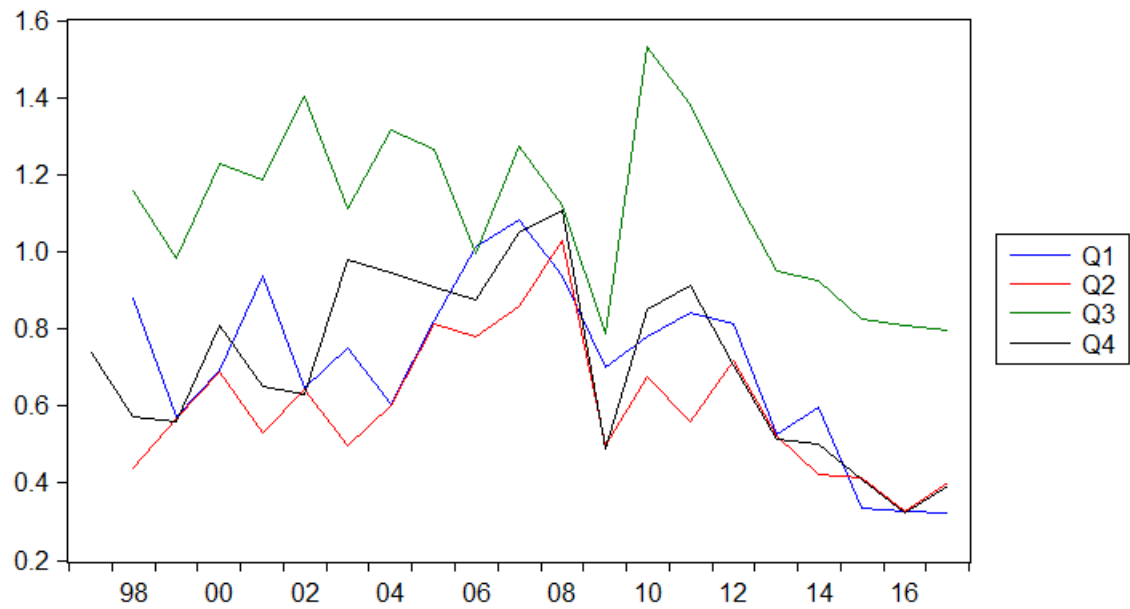




WAGE\_GROWTH by Season



WAGE\_GROWTH by Season



Wage growth in the 3rd quarter is higher than in other quarters.

6. Look at the time series plot of the growth rate of wage again. Mentally adjust for seasonal variation. Do you see that 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?

- These wages are nominal wages which have not been adjusted for inflation.
- Real wage is adjusted for inflation and represents how much we can buy with our wage.
- If nominal wage growth is lower but inflation is lower as well, then real wage growth could be lower, steady, or higher, depending on the relative magnitude of nominal growth rate and inflation rate.
- Suppose your nominal wage increases, but inflation increases at a rate that is greater than your wage increase. In this instance, your purchasing power has weakened because although your wage has increased the price of goods & services has increased proportional greater than your wage.
- We need to look at the inflation rate to determine what happened to real growth rate.