Introduction to EViews 6.0/7.0

Authors:

Anders Thomsen

Rune Sandager

Andreas Vig Logerman

Jannick Severin Johanson

Steffen Haldrup Andersen

Last updated: **Jan 2013**

Table of contents

Preface

Before reading this manual there are a few things you need to be aware of. First of all, this manual is made by the Analytics Group (www.asb.dk/AG) to support BSS's 5th semester economic students and Can.merc Finance students in their use of Eviews. It is far from a complete guide on how to use the software, but only meant to support the students with their specific needs. The manual is not a statistics guide or a textbook, and should not be read as a substitute for either. To make this point crystal clear, we will be making references to the two text books set by the professors:

- Keller. Statistics for management and economics, 8th Edition 2009. Thomson..
- Wooldridge. Introductory Econometrics - A Modern Approach. 4th Edition 2009. Thomson.
- Marno Verbeek – A guide to modern Econometrics. 2nd edition 2004. John Wiley & Sons, Ltd.

The aim of this manual is to show you how you use a specific software application to make statistical analysis. It is not intended to teach you statistical theory. We do believe that to be able to use this kind of software, for making valid and meaningful analysis, one must have a sufficient understanding of the underlying statistical theory.

Throughout this manual we will be using eight different work files to illustrate the use of Eviews. The first file is based on a survey made among ASB's students back in 1998. It contains around 20 variables all of which can be found in appendix A at the very end of this manual. The second work-file is from a research on wage differences between the sexes. The variables and their names in this work-file is considered easy to interpret and should not require any further notice. Both of these work files are available on www.asb.dk/AG

1 Introduction to Eviews

1.1 What is Eviews?

E-views is a spreadsheet software used for various types of data analysis. It has some similarity to the commonly used Microsoft Excel and does support this type of files. According to its creators E-views is characterized as: *"EViews provides sophisticated data analysis, regression, and forecasting tools on Windows based computers".* While you are able to conduct some data analysis in Excel, E-views enables you to do traditional Excel analysis, like descriptive statistics, but also more advanced calculations, regressions and simulations, which you won't find in Excel. In addition to its increased functionality, it also operates at a much faster pace, both in terms of calculation time and in terms of ease of use. Especially Eviews data series analysis functions are superior to many of its competitors.

1.2 Installing Eviews

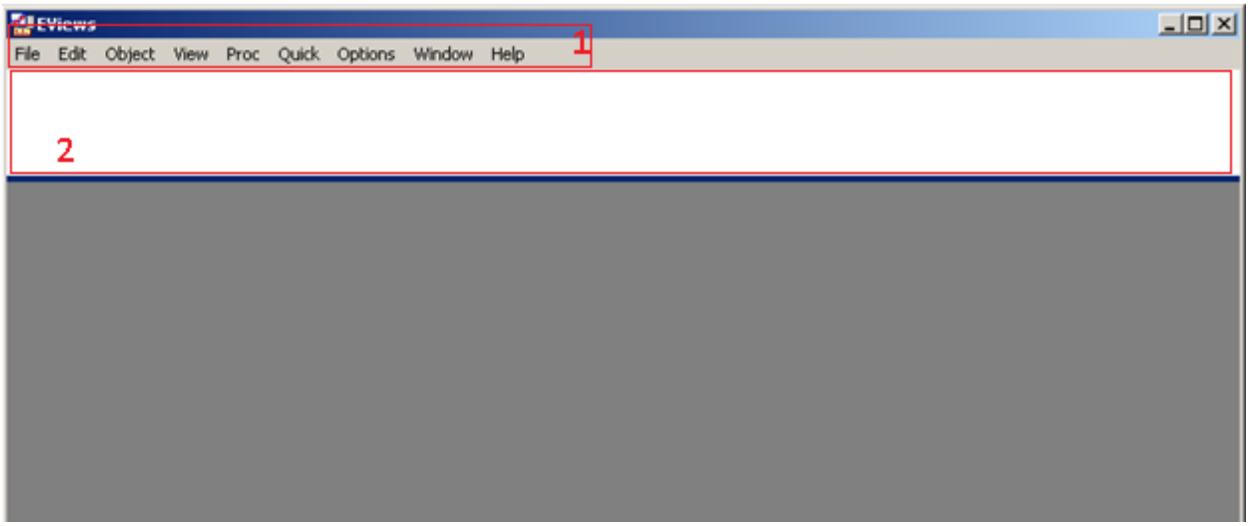
At the moment Eviews only exists for Windows operation system. Mac and Linux users need to install a version of Windows (XP, Vista, 7 all work) to be able to run the application. The system requirements are quite modest and all computers bought in the last five years should have no problems running it smoothly. A full version of Eviews 6.0/7.0 is currently installed on the student computer labs PCs in the H.

This manual is based on version 7.0 of Eviews. There might be minor differences from the student version of the application, but these differences will not be touched upon in this manual.

1.3 The EViews Interface

1.3.1 The empty interface

At a first glance, Eviews doesn't look like much. But its power lies not in its appearance, but in its ease of use, which despite the simple user interface, is very accesible.

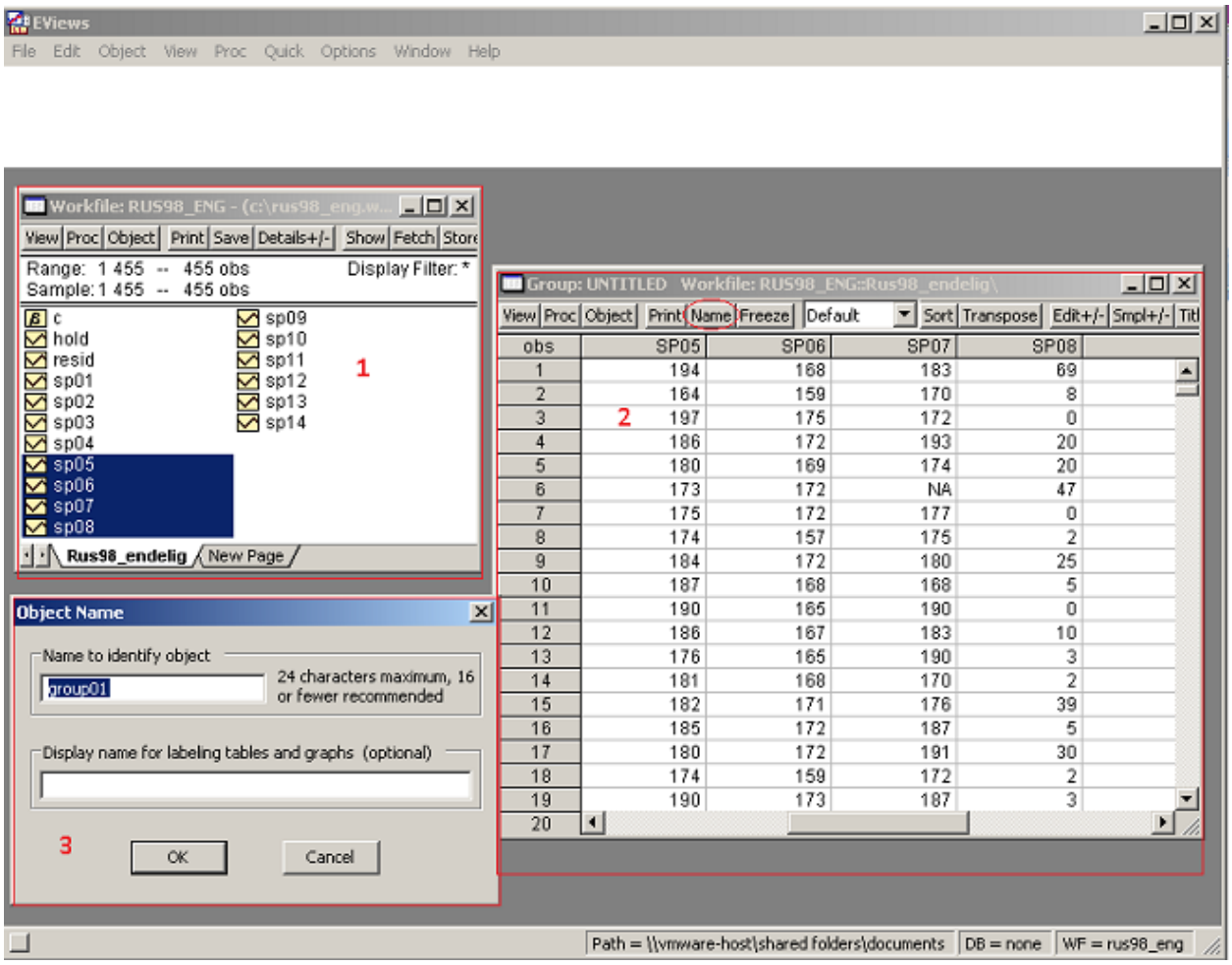


At this point the interface only includes areas of importance: At 1. Is the traditional tool bar, which includes the different tools, used later in this manual. It is important to notice that the content of these different dropdown menus depends on which Eviews window you select beforehand. E.g. not selecting a data set and clicking the *proc* bottom gives you no options at all, while the same click gives multiple different opportunities after selecting a window containing data.

At 2. is the coding area/prompt. This area allows you to apply different text based commands, which is used for both data manipulation and as a potential shortcut for making different regressions. The grey area below the coding line is somewhat similar to the desktop of your PC. It can include numerous windows, including data spreadsheets, regression results, graphs and several different outputs.

1.3.2 Objects and variables in the interface

After importing data, making some calculations, graphs etc.(see the following parts of this manual) the interface could look something similar to the following:



After opening an existing work file, you will see the window at 1. The windows contains a list of all the variables in the work-file, this list is somewhat similar to the columns of an Excel sheet. To view every single observation and its number, one must select the variables of interest by holding down the ctrl bottom and clicking the variables of interest. To the spread sheet window similar to 2, you must either right click the group or click the view bottom, then click *open /as group*. It is often a good idea to save groups, equations, graphs (called objects) by a specific name. This is done by clicking the name button, which is circled in the picture above. After assigning a name and clicking OK, the object will appear along the variables in the first window. The object will appear with a symbol matching its kind of object (graph, group, equation etc). The order in which you select the different variables, is the same order you get when you open the "group." The above window is achieved, by first pressing variable sp05, then holding ctrl and press sp06 and so forth. When you have selected the 4 different variables, right click, and open as group.

2 Data Import

Importing data is straight forward as long as the structure of the data file is correct. In general you need to make sure that the data is structured with variable names in the top row of your spreadsheet and then having the observation following below (see the below illustration from Excel)

A	B	C	D	E	F	G	H	I
hold	sp01	sp02	sp03	sp04	sp05	sp06	sp07	sp08
5	2	1	0	90	194	168	183	69
1	1	1	3	74	164	159	170	8
2	2	1	3	75	197	175	172	0
5	2	1	0	76	186	172	193	20
1	2	2	7	64	180	169	174	20
5	1	2	0	62	173	172	NA	47

It should be noted that besides the following ways of importing, Eviews also support several other file types and application for importing, but we will focus on the most common ones.

2.1 Importing from Excel

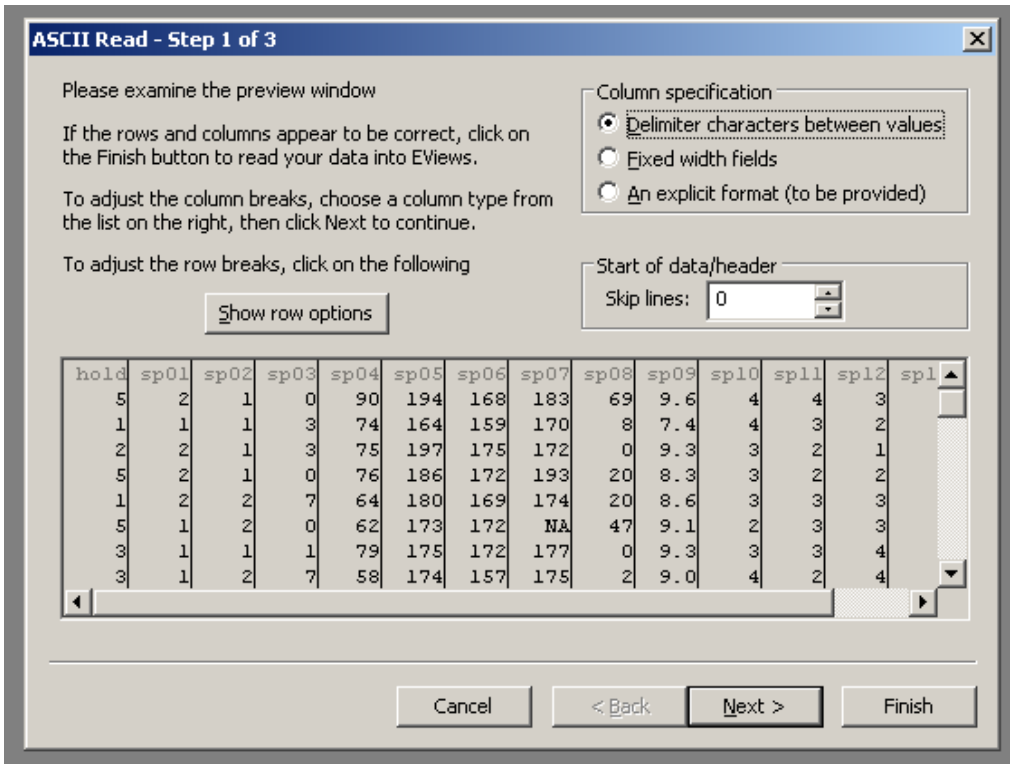
Importing Excel files can be as easy as 1-2-3, if the structure is as described above. One can simply drag-and-drop the Excel file to the Eviews window, and it will automatically open the file and show the included variables. If you on the other hand have an Excel file which does not have the support structure, you must manually adjust the structure. Remove graphs and all none observation within the Excel file, save the file and try to import it again. The alternative to the drag and drop option is going: *file/open/Foreign data or work file* and then browsing your way to the Excel file. When you save in Excel, it is important that you choose "save as ..." and then "Excel 97-2003 Workbook." Eviews will have problems if you import a 2007 file, so remember this.

2.2 Importing from SPSS

Importing data sets of the SPSS file format .SAV will result in problems from time to time. One common problem is that Eviews reads all the variables within the SPSS file to be nominal instead of ratio scaled. This can be solved within Eviews, but takes a very long time, and is beyond the scope of this text. In general you must make the necessary adjustments within SPSS before trying to import the file to Eviews (read the Analytics Group SPSS manual for Bachelor Students – www.asb.dk/AG). To import the SPSS files: *file /open /Foreign data or work file*

2.3 Importing from text-files

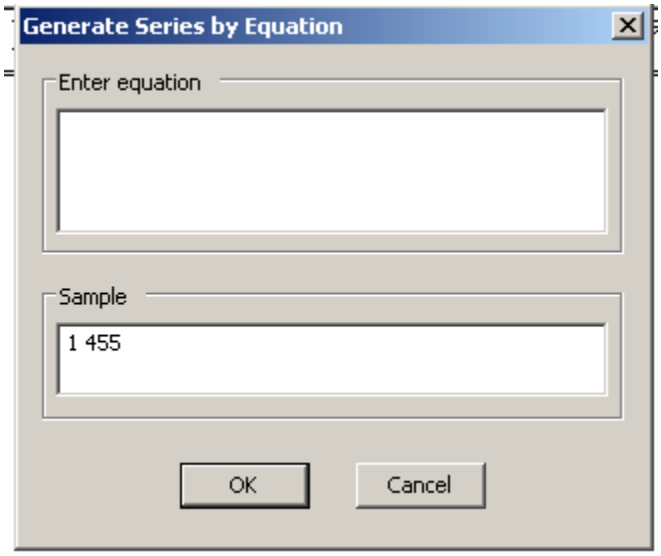
Like Excel, Eviews can import from different types of text files. The process is very similar to the one used in Excel. It is however very important to be aware that to import text files (.txt), you must still use the : *file/open/Foreign data or work file* process and *not* the *file/open/text file,* since this will not lead to Eviews treating the content of the file as data, but as plain text. When opening comma or tap separated text files, Eviews automatically detects the structure of the file, but will let you preview the result before the final import. We found that in some more advanced cases, the text file importer of Excel may be considered superior – and thus you might want to import the text file first in Excel, and then import the resulting Excel file in Eviews.


ASCII Read - Step 1 of 3
Please examine the preview window
If the rows and columns appear to be correct, click on the Finish button to read your data into EViews.
To adjust the column breaks, choose a column type from the list on the right, then click Next to continue.
To adjust the row breaks, click on the following
Show row options
Column specification
Delimiter characters between values
Fixed width fields
An explicit format (to be provided)
Start of data/header
Skip lines: 0
hold sp01 sp02 sp03 sp04 sp05 sp06 sp07 sp08 sp09 sp10 sp11 sp12 sp1
5 2 1 0 90 194 168 183 69 9.6 4 4 3
1 1 1 3 74 164 159 170 8 7.4 4 3 2
2 2 1 3 75 197 175 172 0 9.3 3 2 1
5 2 1 0 76 186 172 193 20 8.3 3 2 2
1 2 2 7 64 180 169 174 20 8.6 3 3 3
5 1 2 0 62 173 172 NA 47 9.1 2 3 3
3 1 1 1 79 175 172 177 0 9.3 3 3 4
3 1 2 7 58 174 157 175 2 9.0 4 2 4
Cancel
< Back
Next >
Finish

3 Creating new variables

3.1 General creation

The general process of creating new variables within Eviews can be done by either using the coding area or by going through the tool bar: *Object/Generate series* which leads to the following window:



Alternately you may use the coding area command: *genr* followed by the new variable name and the new variable equation – GENR 'NEW_VARIABLE_NAME' 'VARIABLE EQUATION'.

3.2 The variable equation and operators

Using the different sign and functions for making new variables: LN – LOG.. e + EXP etc. addition etc. the log(1 plus var hint)

The challenge of creating variables in Eviews all comes down to the use of the variable equations. Eviews allows you to use all of the traditional mathematical operations in the following way:

"+" – addition

"-" – subtraction

"*" – multiplication

"/" – dividing

The use of parentheses can be used as on any normal calculator. To illustrate the use of these operators, an example of these could be the following:

```
genr VAR_NEW = (VAR1 + VAR2)*100 – VAR2/20.
```

Besides these common mathematical operators, you may also ask Eviews to apply logarithmic and exponential function on the variables. Using the Eviews command "Log(VAR)" will result in Eviews using the natural logarithmic function on the

variable VAR, the function also known as LN() on must calculators. Using the natural exponential function in Eviews is done with the command 'exp(VAR)" – similar to using e^x on must calculators. An example using these functions:

```
genr VAR_NEW2 = log(VAR1+1) + exp(VAR2)
```

Note how we add 1 to VAR1 before applying the log function. This is done to ensure that we do not take the logarithm of zero. In general, making operations which are not mathematical possible e.g. dividing by zero or taking LN(0) will result in error pop-up showing, in the middle of the screen.

3.3 Creating dummies

An important extension of the variable equation is how dummy variables are made in Eviews. A dummy variable has a value of either 0 or 1 for any observation, e.g. having 1 for observations with an age above 20 and 0 if not. Creating dummies like this, by using existing variable like age, can be valuable in many different analyses. To create dummies like this we need to use the following operations:

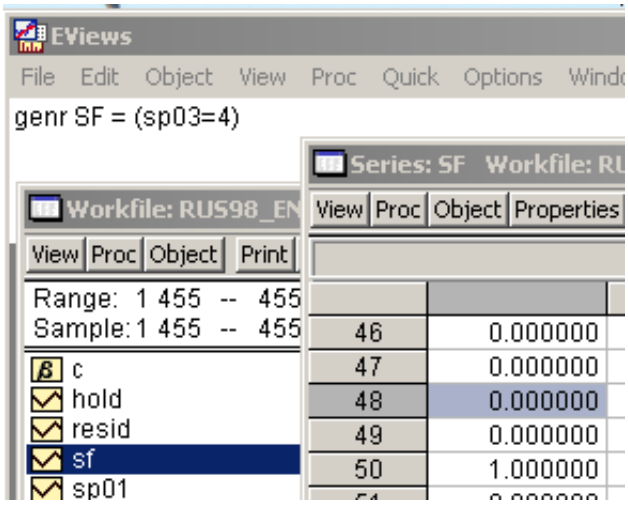
```
(VAR1=value) – will equal one if VAR1 is equal to 'value'

(VAR1>value) – will equal one if VAR1 is greater than 'value'

(VAR1<value) – will equal one if VAR1 is less than 'value'
```

To illustrate making dummies like this, consider the example mentioned above. Let's say I want to create an 'adult' variable using an already existing 'age' variable. I simply type in:

```
genr adult = (age>20).
```

Using our ASB student survey work-file, we can create a dummy variable based on the political party variable as shown below: We want to call this variable SF. To make it using the command line we simply have to write:

```
genr SF = (sp03=4)
```



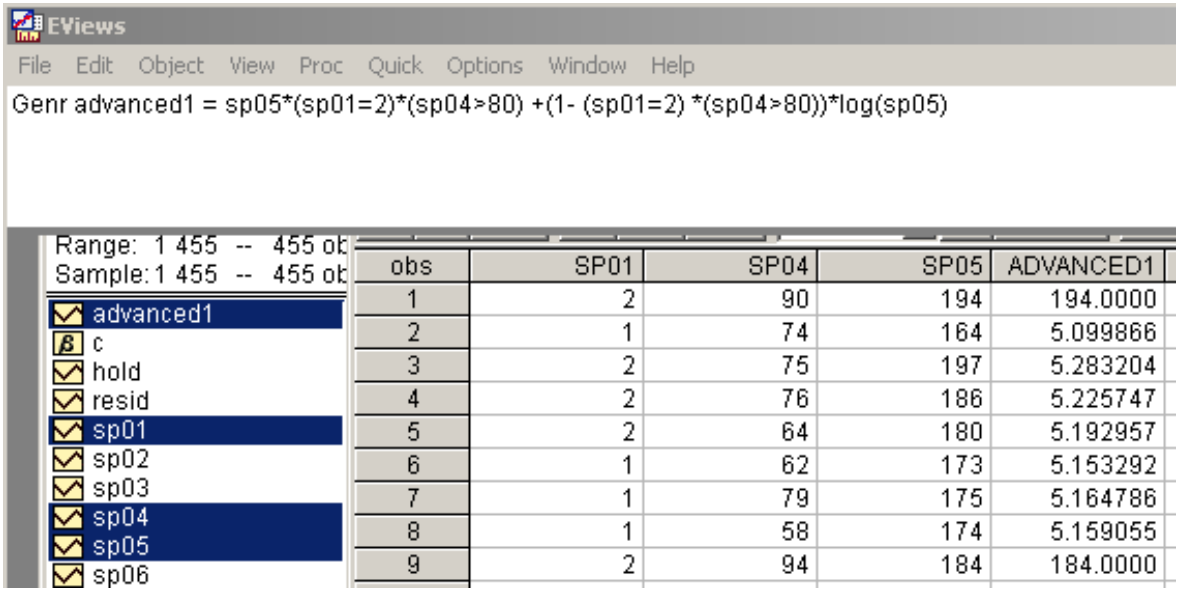
This use of dummies can freely be combined with the previous shown operations, allowing you to create more advanced resulting variables.

Also the use of these constraints (=, <,>), combined with the operators 'AND' and 'OR' ,can be used in numerous ways when creating new variables. To illustrate this claim here is a more advanced example, which uses multiple of these constraints.

We want a variable which states the height (sp05) of the respondent, but only if it's a male, (sp01=2) with a weight of more than 80kg (sp04>80). If not we want the variable to equal the natural logarithm of the height[1]:

```
Genr advanced1 = sp05*(sp01=2)*(sp04>80) +(1- (sp01=2) *(sp04>80))*log(sp05)
```

Note that *(1- (sp01=2) *(sp04>80))* will equal 0 if, and only if, both the constraints are true and 1 if not.



obs	SP01	SP04	SP05	ADVANCED1
1	2	90	194	194.0000
2	1	74	164	5.099866
3	2	75	197	5.283204
4	2	76	186	5.225747
5	2	64	180	5.192957
6	1	62	173	5.153292
7	1	79	175	5.164786
8	1	58	174	5.159055
9	2	94	184	184.0000

3.4 Creating group based on existing variables

The concept used to make the dummy variables above, can be expanded when creating grouping variables with more than two outcomes. Say we have a discrete variable, *var*, which takes the values: 0,1,2,4,5,6,7,8,9,10 and want to make a grouping variable, *group,* taking the value 1,2 or 3 depending on values of the existing variable, *var*. We want the groups to be the following: *group*=1 if var is 1 or below, *group* = 2 if var equal 2 or 3 and *group*=3 if var equal 4 or above:

[1] Don't try to make any sense of this variable; it's nothing but an example of a more advanced way of using the dummy variables to create more advanced equations.

Grouping variable, group:	Existing variable, Var values:
Group=1	0, 1
Group =2	2, 3
Group =3	4, 5 , 6 , 7, 8, 9, 10

Generating this grouping variable can be done in different ways. First we will illustrate how to create this *group* variable by first creating tree dummy variables (gr1, gr2 and gr3) and next we will show you how to create the same variable without using these dummies.

The dummy method is made by using the command line:

```
genr gr3 = (var >= 4)

genr gr2 = (var = 2 or var = 3)

genr gr1 = (var <= 1 )
```

This takes care of the tree dummies. To create the final grouping variable, *group*, we use these 3 dummies in the following way:

```
genr group = gr1*1 + gr2*2 + gr3*3
```

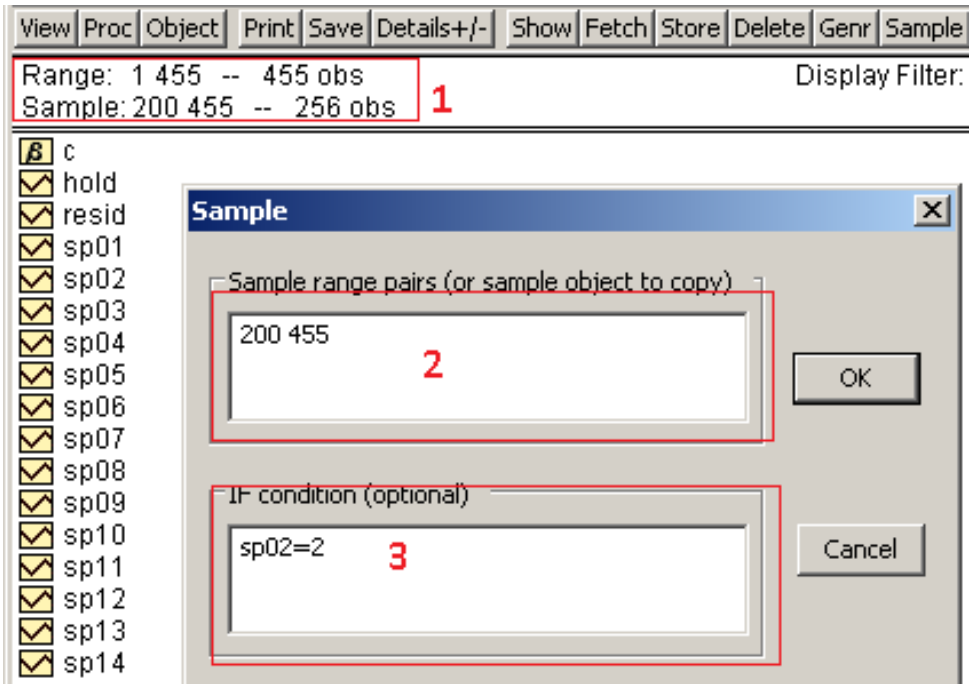
An alternative method is to combine the dummy creation with the above code in one line of code:

```
genr group = (var>=4)*3 + (var=2 or var=3)*2 + (var<=1)*1
```

Both of these methods yields the same result.

3.5 Sample range

Using the constraints from the above section and the range tool can be used to focus on specific parts of the sample. This specification can be in regard of both the sample number and based upon a characteristic of the respondents. To access this 'sample' tool you simply double click the area marked as 1:



Within this sample range tool, area 2 concerns the observation number of the sample. To include all observation you need to write @All. To include a specific range you simply write the starting point followed by the ending observation – in this case the starting point is observation number 200 and the ending point is observation 455.

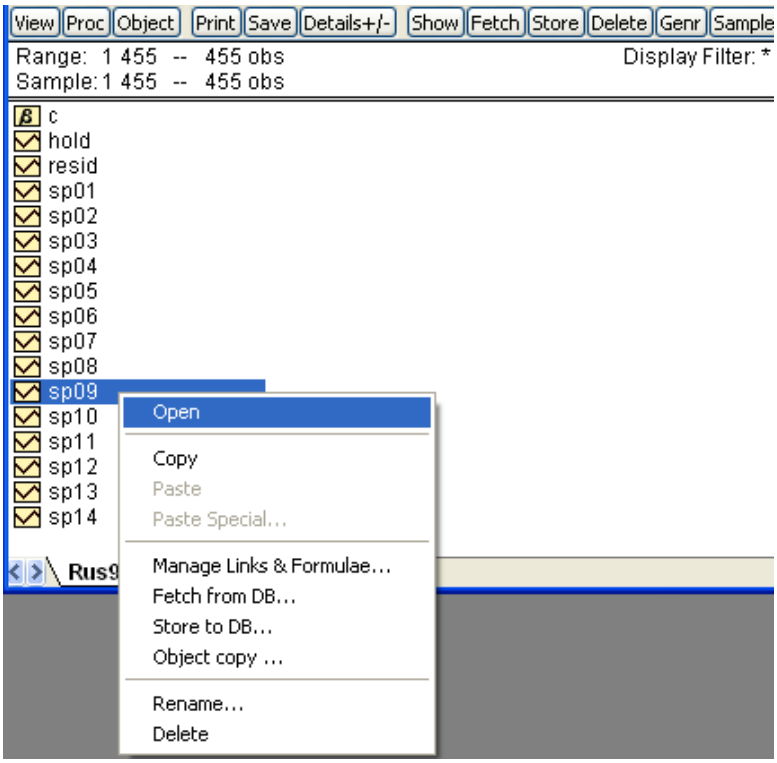
Area 3 allows you to constrain your analysis of the sample by using the previous mentioned constraints. In this case the analysis is reduced to only including observations which has sp02=2, that is all respondents whom expect an annual income above 300.000 Dkr. You can make these constraints more advanced by using the words 'and' & 'or' while adding more constraints. E.g. to only analyse male respondents who expect an annual income above 300.000 Dkr would be:

```
sh01=2 and sh02=2
```

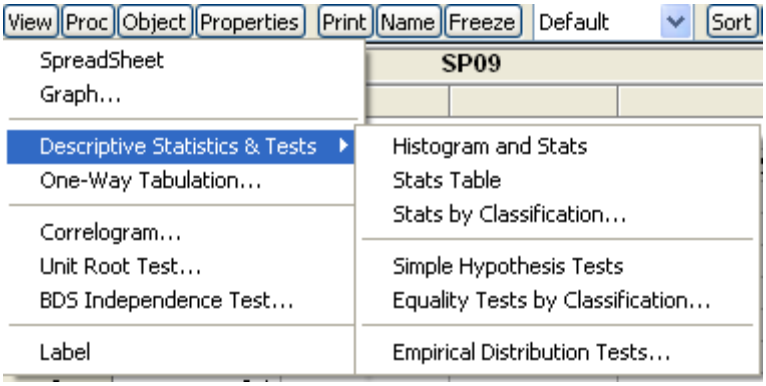
4 Descriptive Statistics

4.1 The Basics

In the following we use the data set called rus98_eng in use.wf1, which contains information concerning 455 students at ASB, such as grade, age, gender etc. Getting the most basic descriptive statistics in Eviews is very straight forward. First you need to select the variable of interest (i.e. sp09, which is the average grade of the students) by double clicking it, or right clicking and choosing open as group.



Next you need to click *View* within the new window and select *Descriptive statistics & Tests*. Doing so gives you a list of different options, as shown below.



"Stats Table" reports the following:

	SP09
Mean	8.476180
Median	8.500000
Maximum	10.40000
Minimum	6.300000
Std. Dev.	0.738161
Skewness	-0.040541
Kurtosis	2.611562
Jarque-Bera	2.919541
Probability	0.232290
Sum	3771.900
Sum Sq. Dev.	241.9275
Observations	445

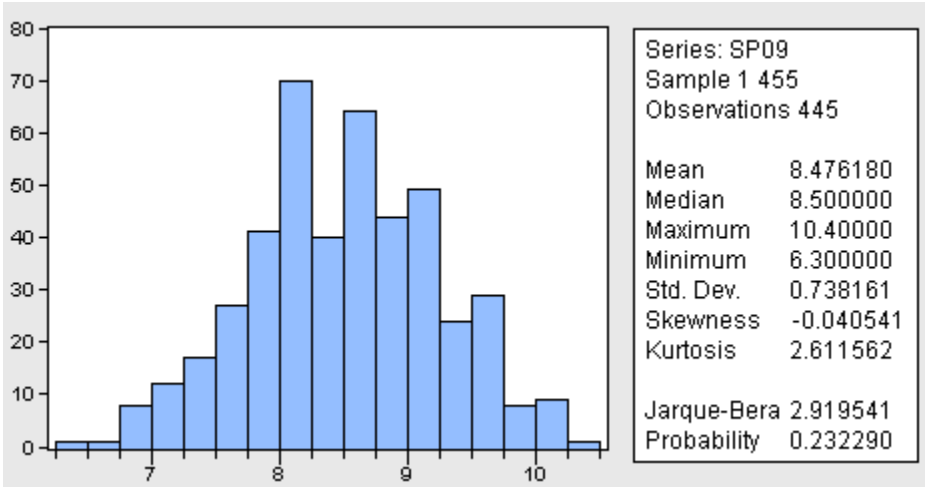
That is, the average of sp09 (average grade) is 8,47, and the std. dev. is 0,738. The variance is not directly reported, but can be obtained by:

Var(x) = (std dev (x))^2

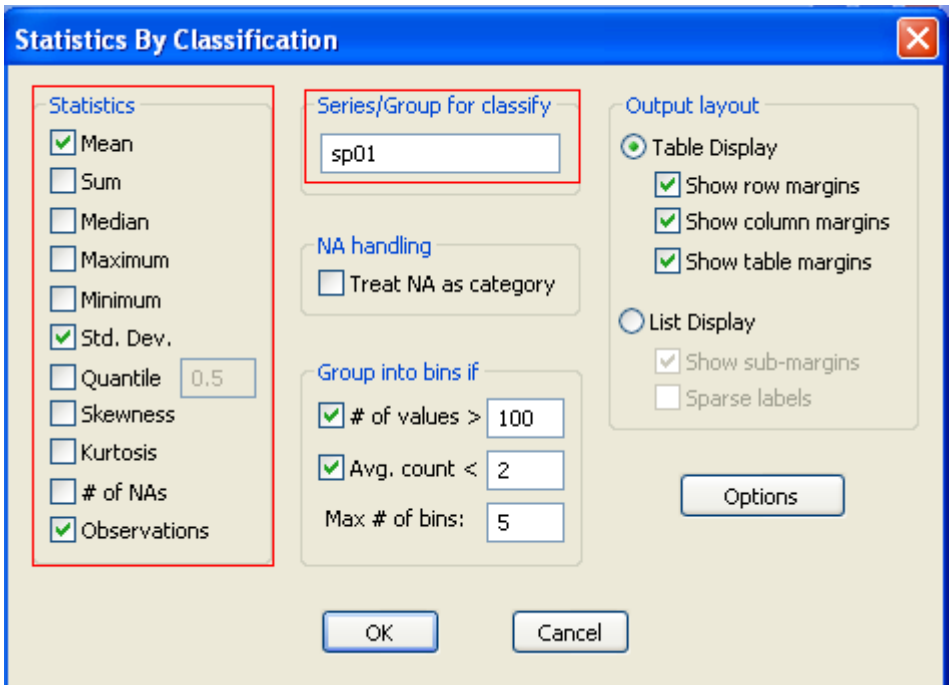
Var(sp09) = $0{,}738^2$

The number of observations is 445. Be aware that this means that 10 of our original respondents have not answered the question "sex."

"Histogram and stats" reports similar stats, but includes a distribution histogram:



The "stats by classification" will report statistics grouped by another variable, sp01, sex in this case (sex equals 1 if it's a female, and equals 2 if it's a male):



This results in the following output:

Descriptive Statistics for SP09
Categorized by values of SP01
Date: 03/30/10 Time: 14:13
Sample: 1 455
Included observations: 445

SP01	Mean	Std. Dev.	Obs.
1	8.534118	0.712285	170
2	8.440364	0.752761	275
All	8.476180	0.738161	445

The output above shows that in our sample the 170 females have a mean average grade of 8,53, while the 275 males in the sample have a mean average grade of 8,44.

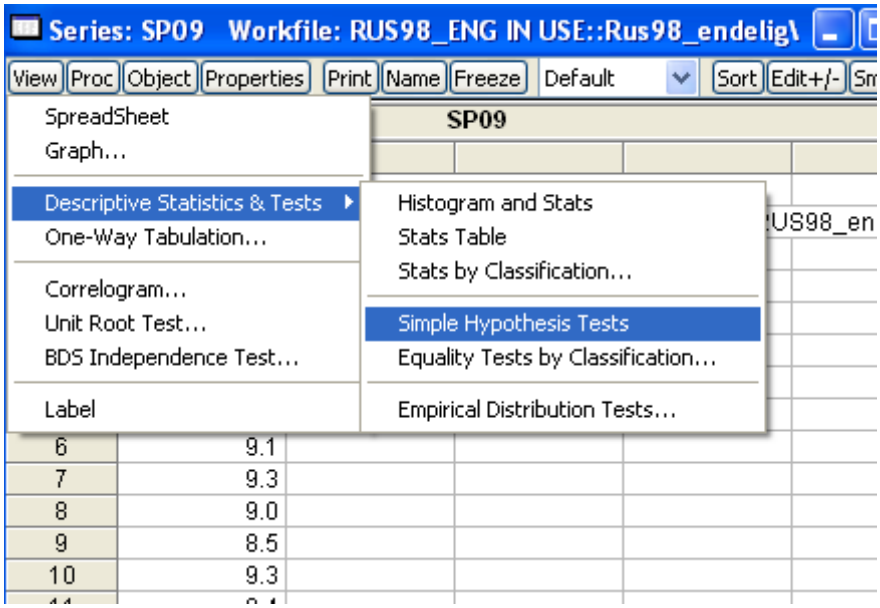
4.2 One sample t-test – two sided

Let's continue the prior example of the average grade. A simple t-test is used, when you want to test whether the average of a variable is equal to a given value. In this example we want to determine if the average grade for students at ASB is 7. First we conduct a two-sided test, and afterwards we make it a one-sided test. The H_0 hypothesis should look like this:

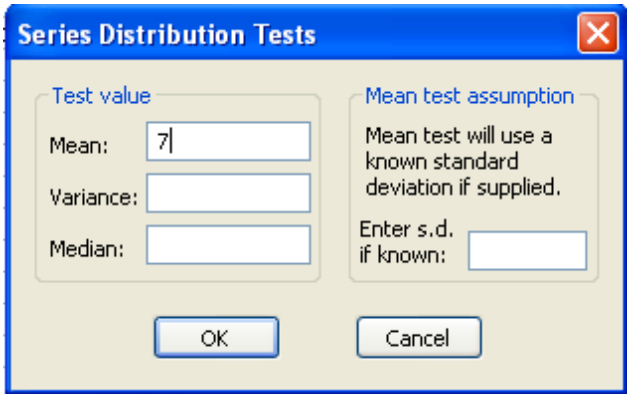
H_0: $\mu_{ave.grade} = 7$

H_1: $\mu_{ave.grade} \neq 7$

To make this test in Eviews we first select the variable called sp09 (average grade) by double clicking it, and then choose *"view – descriptive statistics and tests – simple hypothesis tests:"*



Then we have to type in the value from H_0. In our example our H_0 hypothesis is that the average grade equals 7. Thus we write the following, and click OK.



Then we get the following output:

Hypothesis Testing for SP09
Date: 04/21/10 Time: 16:31
Sample: 1 455
Included observations: 445
Test of Hypothesis: Mean = 7.000000

Sample Mean = 8.476180
Sample Std. Dev. = 0.738161

Method	Value	Probability
t-statistic	42.18598	0.0000

This gives us our t-statistic on 42,19, which we have to compare to the critical value of the t-distribution with 444 degrees of freedom at a significance level of 5%. Because we are applying a two-sided test, the critical values are -1,96 and 1,96.

Our conclusion is that we reject our null hypothesis because the test statistic falls in the critical region. The very low p-value indicates that our conclusion is not sensitive to changes in the significance level.

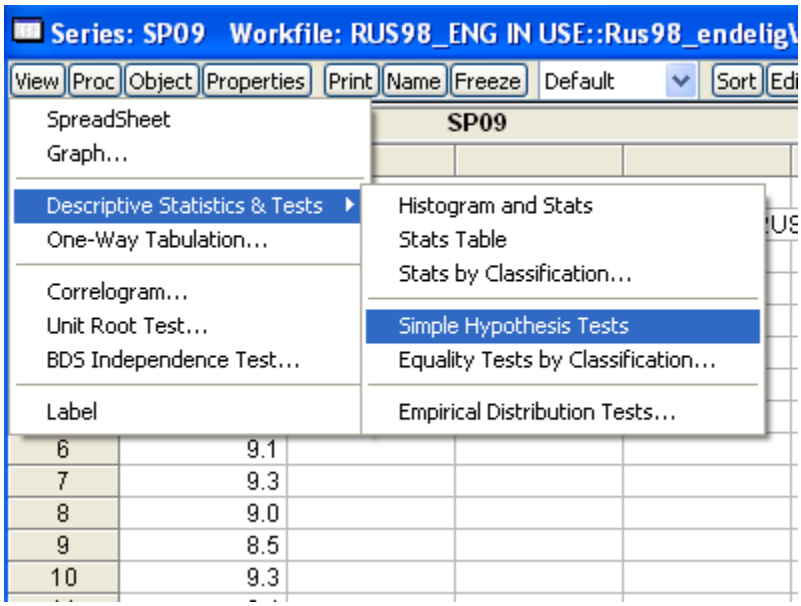
4.3 One sample t-test – one sided

Now we want to make a one-sided t-test, and we want to test, whether the average grade of ASB-students is higher than 8,3. The hypothesis should then look like this:

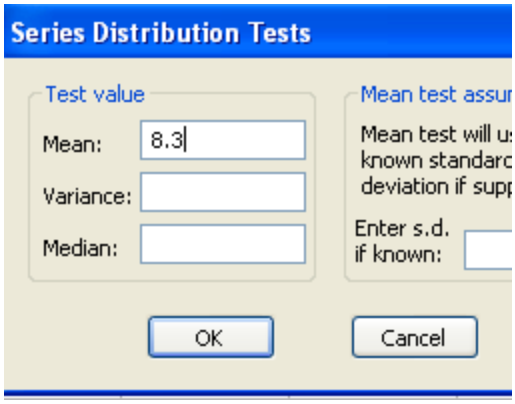
H_0: $\mu_{ave.grade} = 8,3$

H_1: $\mu_{ave.grade} > 8,3$

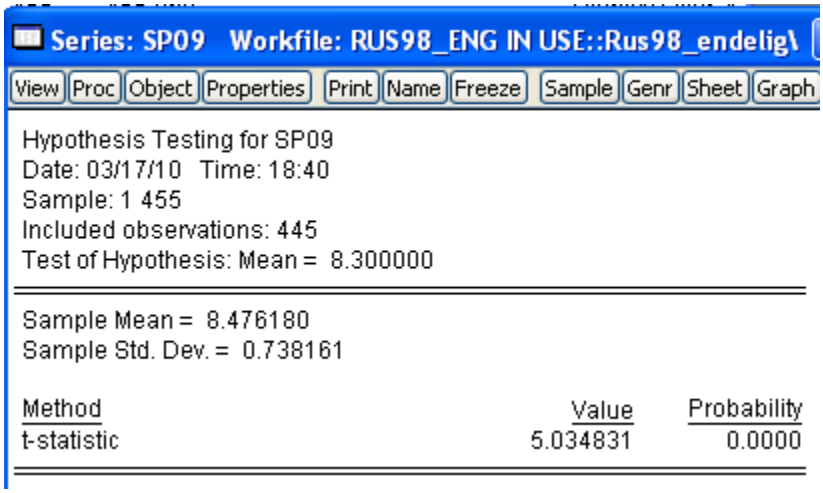
In Eviews we first have to open the variable we are interested in, sp09. We select the test the same place as the two sided test: "View – Descriptive Statistics and Tests – Simple hypothesis test"



Our test-value is 8,3, so in Eviews we simply write: 8.3 (We use . instead of ,)



We then get the following output:



This gives us a t-statistic of 5,03, which we have to compare to the critical value of the t-distribution with 444 degrees of freedom and a significance level of 5%. Note that this time we apply a one-sided test and the critical value is changed to 1,645. Thus we reject the null hypothesis. Our conclusion is not sensitive to changes in the significance level. Note that the p-value in Eviews refers to a two-sided test!

4.4 Testing for differences in mean – based on two groups

If you want to compare two means based on two independent samples you have to make an independent sample t-test. E.g. you want to compare the average grade for students at ASB for women versus men. The hypothesis looks as follows:

H_0: $\mu_{grade\_men} = \mu_{grade\_women} \Leftrightarrow \mu_{grade\_men} - \mu_{grade\_women} = 0$

H_1: $\mu_{grade\_men} \neq \mu_{grade\_women} \Leftrightarrow \mu_{grade\_men} - \mu_{grade\_women} \neq 0$

We want to test if the average grade for females is different from the average grade for males. **The first step** is to determine if the variances are equal. To determine this we could either use the False F-test or Levene's test. (To test for equality in variance using Levene's test please see the section about ANOVA).

4.4.1 False F-Test

To conduct the false F-test, Eviews is used for calculating the sample variance for each group. This is done as shown in section 4 – descriptive statistics. Eviews does not contain the test by default. The hypothesis looks like this:

H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

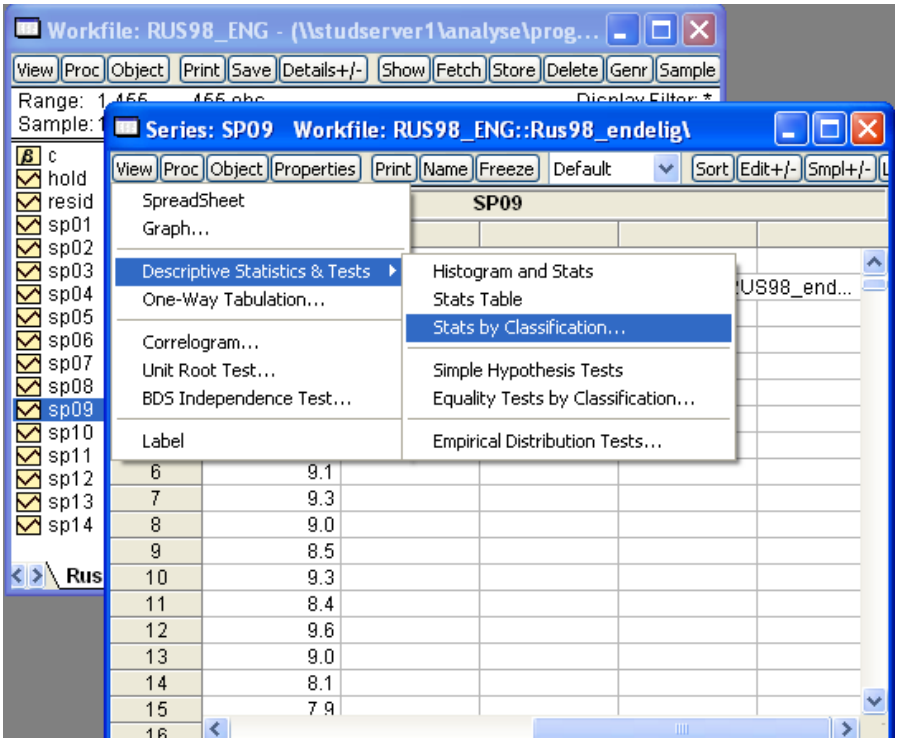
H_1: At least two σ^2 differ'

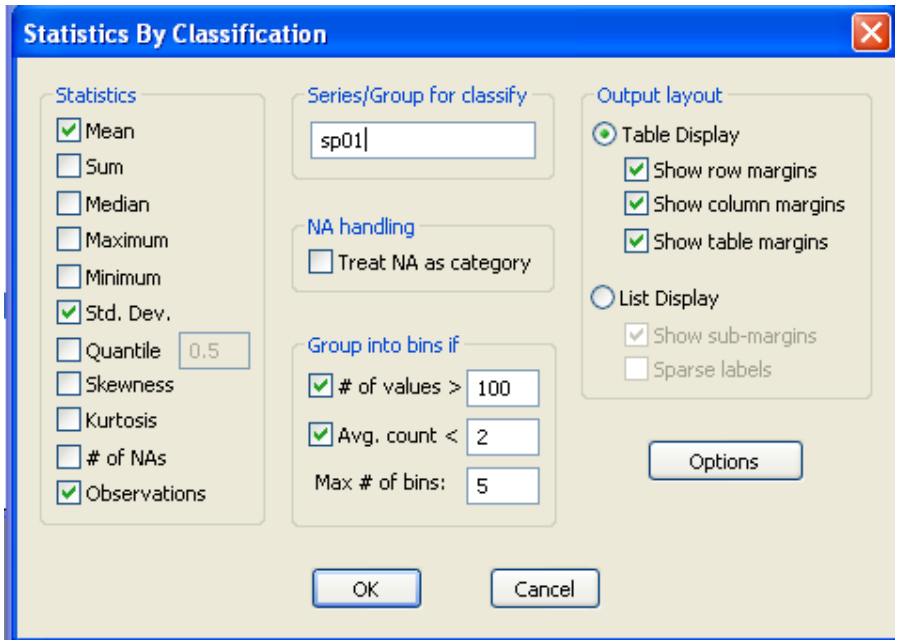
We then want to know if the variance in variable sp09 (average grade) is the same for the two groups (male, females).

The False F-test looks like this if we only have 2 groups:

$$\frac{S_{max}^2}{S_{min}^2} \sim F_{n_1-1;n_2-1}$$

We then need to find the largest (S_{max}^2) and smallest (S_{min}^2) variance in our sample. To do this, open the variable of interest, sp09 (average grade) and do the following:





Our group variable was sex (sp01)

The resulting output contains the mean, standard deviation and number of observations by group (Recall that σ^2 = Variance):

Descriptive Statistics for SP09
Categorized by values of SP01
Date: 04/21/10 Time: 16:53
Sample: 1 455
Included observations: 445

SP01	Mean	Std. Dev.	Obs.
1	8.534118	0.712285	170
2	8.440364	0.752761	275
All	8.476180	0.738161	445

Then we have to calculate our F-statistic by dividing the largest sample variance ($0{,}753^2$ = 0,567) with the smallest sample variance ($0{,}712^2$ = 0,507), and compare this to a critical value: $\frac{S_1^2}{S_2^2} \sim F_{n_1-1;n_2-1}$ ➔ 0,567 / 0,507 = 1,119

The critical value of the F-distribution for a two sided test with 169 and 274 degrees of freedom at a significance level of 5% can be found to be: $f_{169;\,274;0,025}$ = 1,31. This indicates that we cannot reject the null-hypothesis, and therefore we conclude that the variance is equal across the two groups.

If you have more than two groups, the critical value is calculated a bit different.[2] The test is still $\frac{S^2_{max}}{S^2_{min}}$ but the significance level has to be adjusted. We use $\alpha^* = \frac{2\alpha}{k(k-1)}$

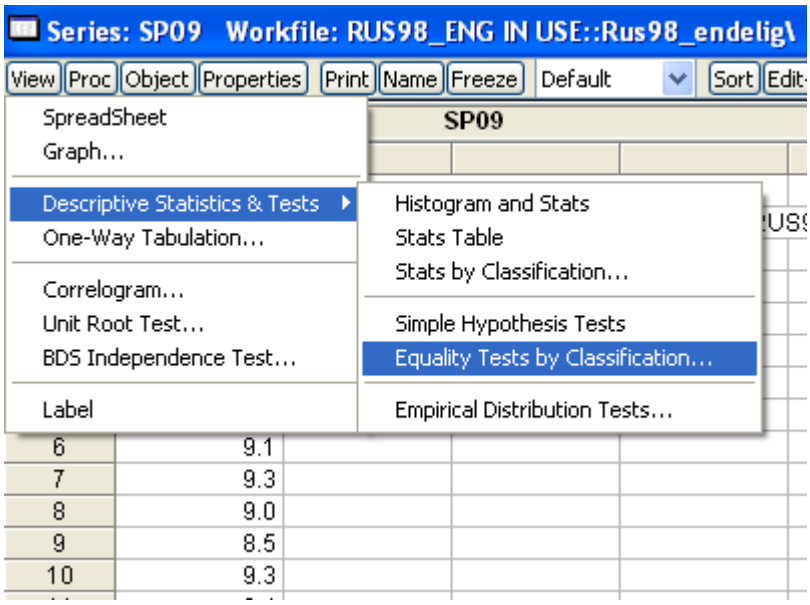
Let's say we have 5 groups instead of only 2. If this is the case, the critical value of the f-distribution can be found as:

$f_{169,274,\,\alpha^*/2}$ where $\alpha^*/2 = \frac{2\alpha}{k(k-1)*2} = \frac{0,05}{5(5-1)} = 0,0025$

And therefore:

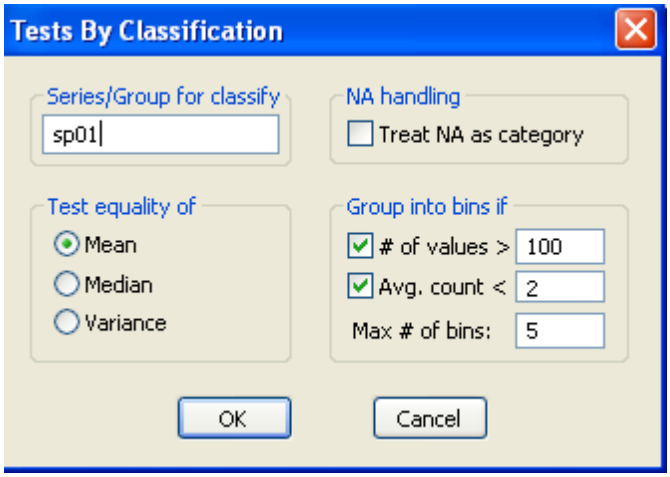
$f_{169;\,274;\,0,0025}$ = 1,47.

Now let's continue with our test for the difference in means based on two groups. If you need to test more than two groups, you need to use another test, such as ANOVA. Given equality of the variance, **the second step** is to determine if the means are equal. The test is performed in Eviews by choosing the variable of interest (sp09 = grade) and then "view – Descriptive Statistics and Tests – Equality Test by classification"

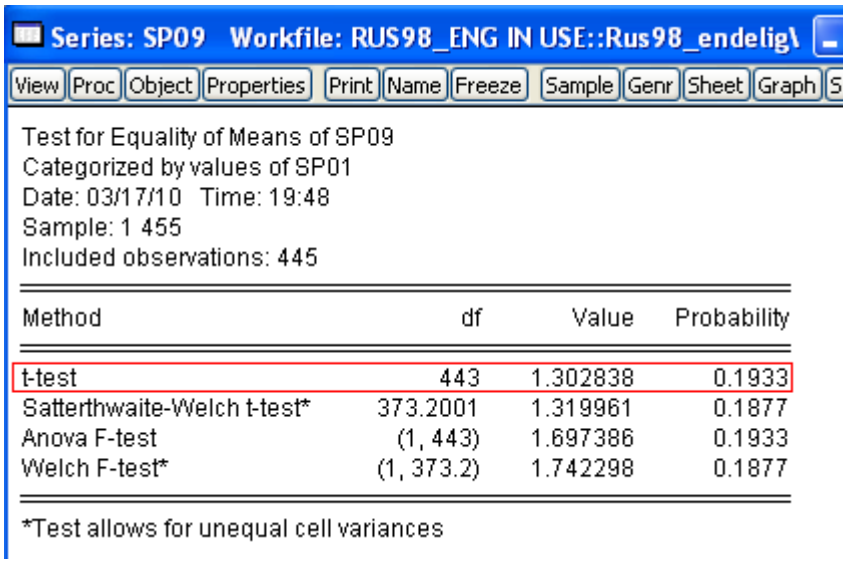


[2] Lecture Notes in Business Statistics page 45

Our grouping-variable in this example is sex (sp01). We want to test the null hypothesis, that the 2 means are equal, which is why we choose "mean" in the window:



Which results in the following output:



This gives us a t-statistic of 1,30, which we have to compare to the two-sided critical values in the t-distribution with 443 degrees of freedom, and a significance level of 5%. This critical values can be found to be approximately equal to -1,96 and 1,96. This indicates that we cannot reject the H_0 hypothesis. Therefore we conclude that there is no difference between the average grade of males and females. The p-value is relative high, which indicates that the conclusion is not sensitive to changes in the significance level.

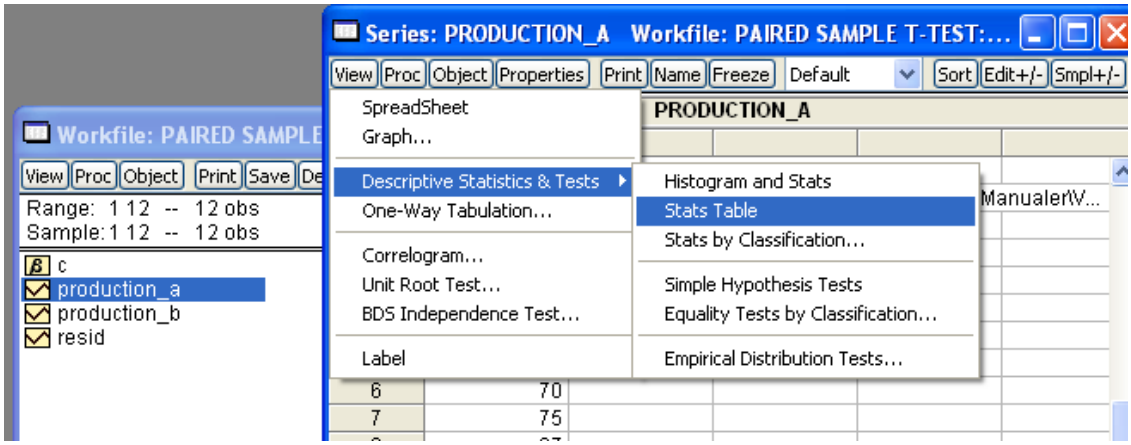
4.5 Paired Sample T-tests[3]

The use of the paired sample t-test will be shown using the following example:

During the summer, a farmer compared two harvesters for the past 12 days. They were tested on the same field, right next to each other. This means that they were exposed to the same weather and the same topsoil conditions. The farmer wants to test which harvester is the most effective. In this example we have two samples: One sample for the first combine harvester, and one sample for the second harvester. Because of the experiment conditions (same weather, and same topsoil) the two samples are dependent, and therefore we will make a Paired Sample t-test.

The production is measured as production_a for harvester a, and production_b for harvester b. The dataset for the following test is named "Paired Sample t-test"

First let's look at some stats about the two different harvesters. This is done by choosing one of the variables and then "view – descriptive statistics and tests – stats table"



[3] Keller (2008) chapter 13.1

We then get the following:

Series: PRODUCTION_A Workfile: P

View | Proc | Object | Properties | Print | Name | Free

	PRODUCTION_A
Mean	61.83333
Median	61.50000
Maximum	75.00000
Minimum	40.00000
Std. Dev.	9.768533
Skewness	-0.596415
Kurtosis	3.236097
Jarque-Bera	0.739293
Probability	0.690979
Sum	742.0000
Sum Sq. Dev.	1049.667
Observations	12

The same can be done with the other variable:

Series: PRODUCTION_B Workfile:

View | Proc | Object | Properties | Print | Name | Fr

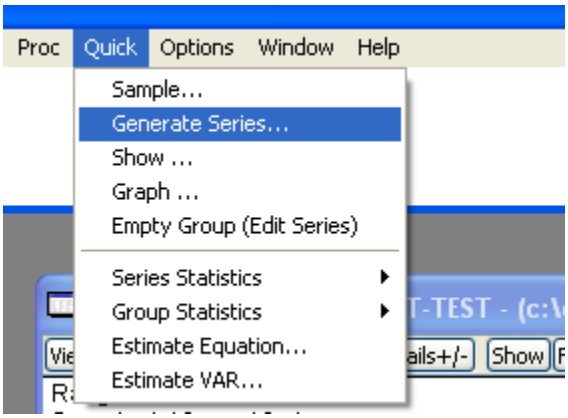
	PRODUCTION_B
Mean	63.83333
Median	65.00000
Maximum	76.00000
Minimum	45.00000
Std. Dev.	7.941071
Skewness	-0.993774
Kurtosis	3.908453
Jarque-Bera	2.387818
Probability	0.303034
Sum	766.0000
Sum Sq. Dev.	693.6667
Observations	12

These output tells us that production_b has a larger sample mean than production_a. When we have a paired sample t-test, the hypothesis looks as follows:

H_0: $\mu_{produktion\_a} = \mu_{produktion\_b} \Leftrightarrow \mu_{produktion\_a} - \mu_{produktion\_b} = 0$

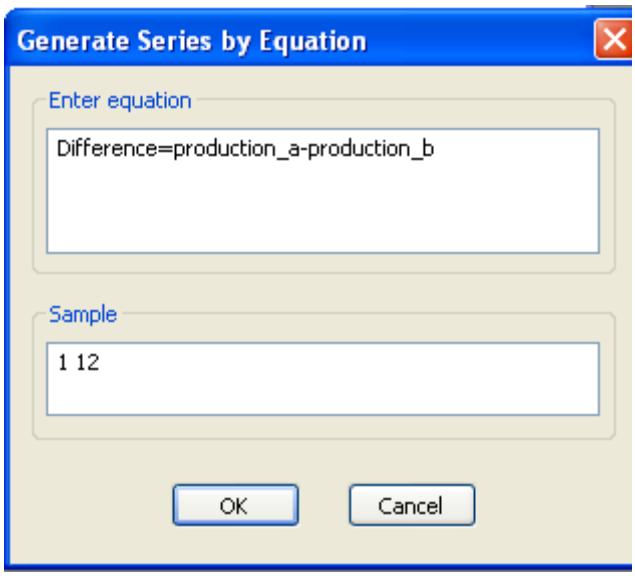
H_1: $\mu_{produktion\_a} \neq \mu_{produktion\_b} \Leftrightarrow \mu_{produktion\_a} - \mu_{produktion\_b} \neq 0$

First of all we have to construct a new variable, which measures the difference between the two daily productions. In Eviews we make this variable in "quick" – "Generate Series":

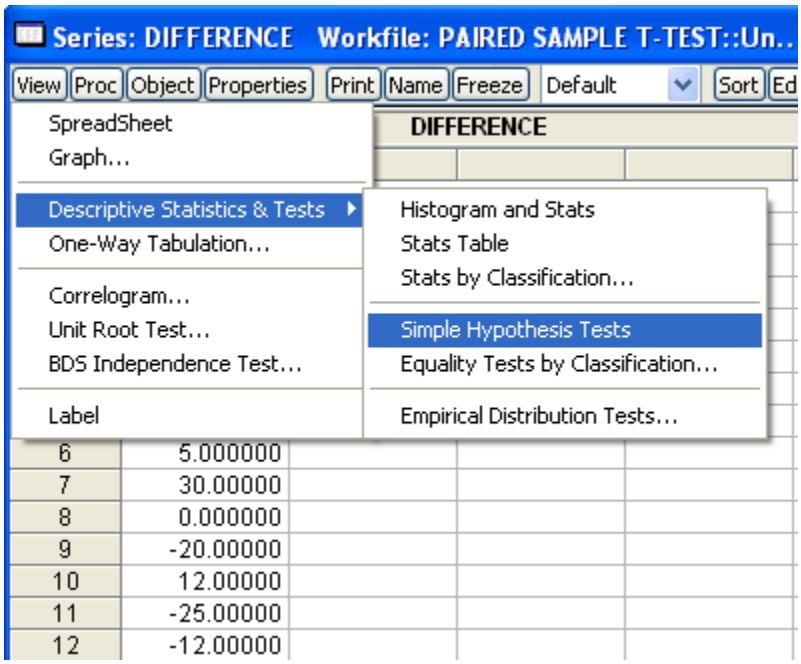


First we have to name our new variable. In this example we just name it "Difference". We then have to explain to Eviews how this new variable should be calculated, so we write:

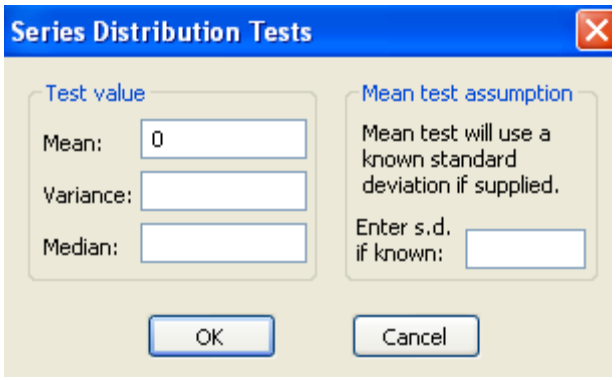
"Difference=production_a-production_b"



The new variable is constructed, and now we can use the simple hypothesis test to do paired sample t-test. Start by opening the new variable "Difference" – and then choose *"view – Descriptive statistics and Tests – Simple hypothesis test"*



Please recall our problem. We were interested in testing the hypothesis if there was a difference between the two different harvesters. Our test-value in our example will therefore be 0, just like the hypothesis.



Then we get the following output:

Hypothesis Testing for DIFFERENCE
Date: 03/17/10 Time: 17:47
Sample: 1 12
Included observations: 12
Test of Hypothesis: Mean = 0.000000

Sample Mean = -2.000000
Sample Std. Dev. = 15.05747

Method	Value	Probability
t-statistic	-0.460117	0.6544

This gives us a t-value of -0,46, which we have to compare to a critical value of t-distribution with 11 degrees of freedom at a significance level of 5% for a two sided test. This critical value is +/-2.228, which tells us that we cannot reject the null hypothesis. Thus, we cannot conclude that there is a difference between the two harvesters. The very high p-value indicates that our conclusion is not sensitive to changes in the significance level.

5 Analysis of Variance (ANOVA)

5.1 The basics

An analysis of variance (=ANOVA) is a statistical method, to detect if there is a statistical difference between the means of the populations. To get a proper understanding of the ANOVA theory see Keller section 14.1-6 page 513-579.

The null hypothesis in the simple ANOVA test is the following:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

Against the alternative

H_1: at least two μ's differ

Where k is the different groups of interest and μ_j is the mean within that group. In the following example we will try to determine if there exist a statistical significant difference between the students weight (the dependent variable of interest) and their choice of political party (the grouping variable). Since the sample only contains very few observations for some of the political parties, we have used the sample constraints showed earlier to focus the analysis on the political parties[4]:

0 – Undecided(0) , 1 – Soc.Dem.(S), 3 – Kons.(K), 7 – Venstre (V) – (the number is the observation number in party variable sp03 – see appendix A)

So the resulting hypothesis for our test becomes:

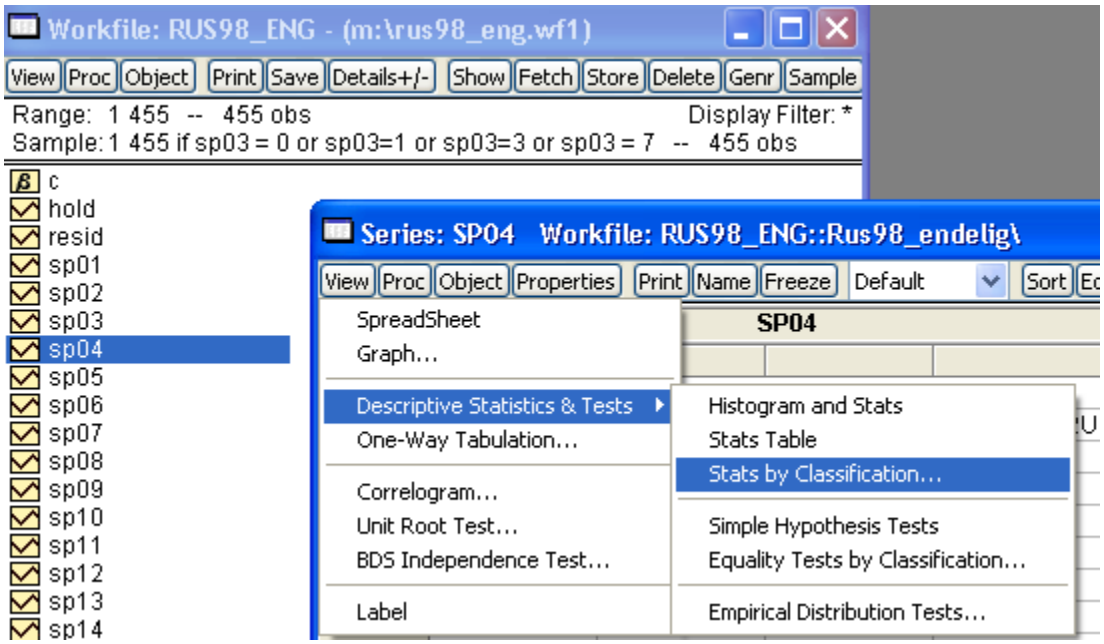
[4] "sp03 = 0 or sp03=1 or sp03=3 or sp03 = 7" in the sample range window

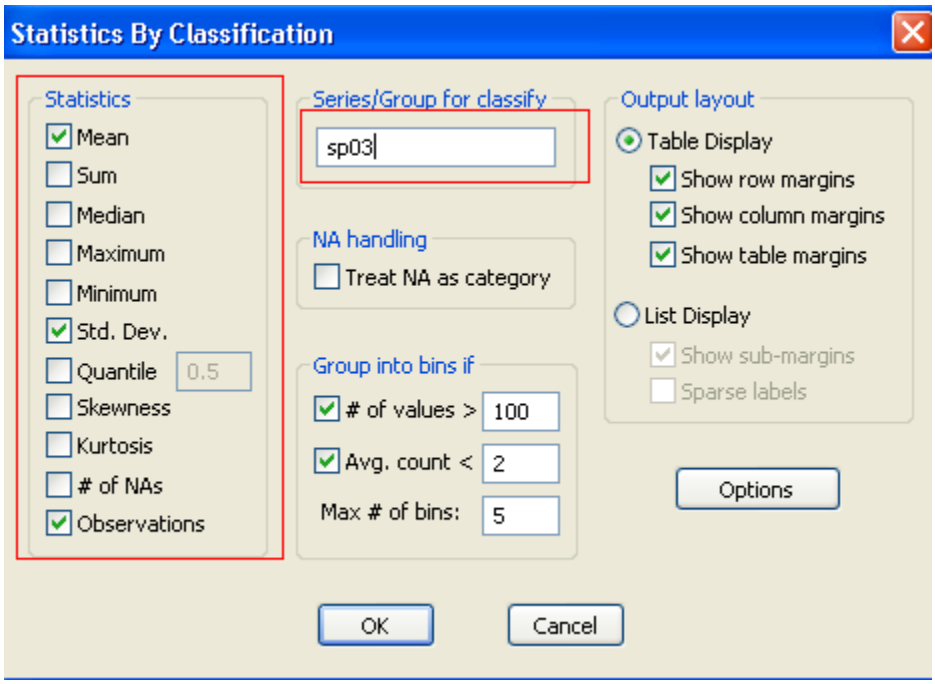
$H_0: \mu_0 = \mu_S = \mu_K = \mu_V$

$H_A: at\ least\ two\ \mu's\ differ$

The sample mean and standard deviation is calculated, like shown in the descriptive statistics section, by selecting the variable by double clicking it and going *View/ Descriptive Statistics/Tests/Stats by classification..*

In the resulting window, the parameters of interest, mean and standard deviation in this case, and name the grouping variable – political party, sp03.





The resulting output contains the mean, standard deviation and number of observations by group:

Descriptive Statistics for SP04
Categorized by values of SP03
Date: 02/13/10 Time: 12:07
Sample: 1 455 IF SP03 = 0 OR SP03=1 OR SP03=3 OR SP03 = 7
Included observations: 394

SP03	Mean	Std. Dev.	Obs.
0	67.87500	12.97563	48
1	69.36538	11.28426	52
3	71.38158	11.96324	76
7	72.67431	12.01418	218
All	71.40355	12.10943	394

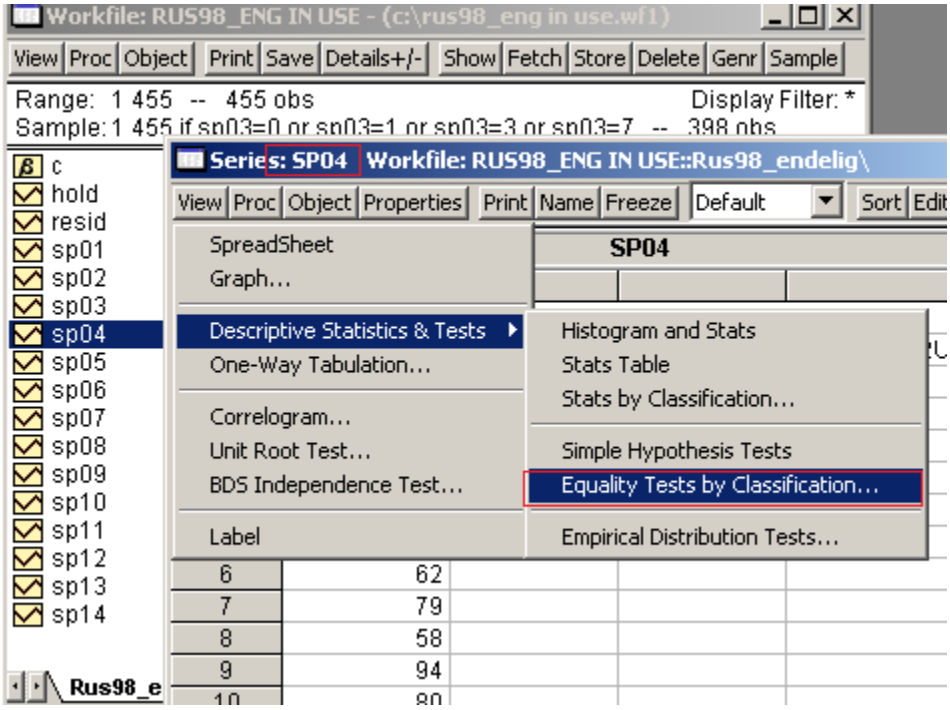
To determine if there is a statistically significant difference, we need to run an ANOVA test, which is shown in the following section.

5.2 The ANOVA test in Eviews

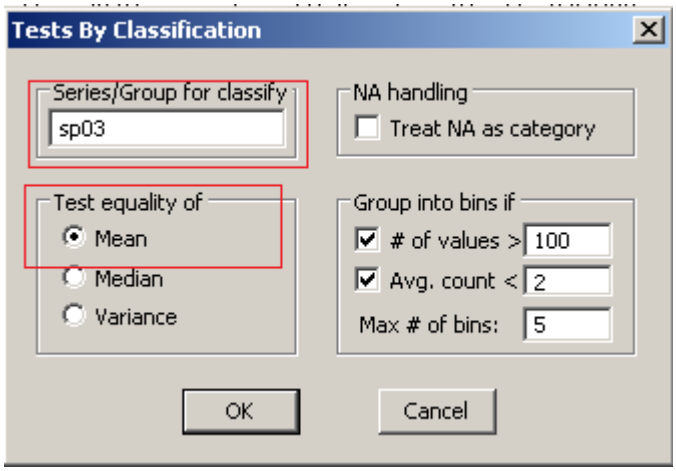
Before we run the test, remember to change the data range, so it fits what we want. Double click on the sample range, as shown in section 3.4. In this example we write "sp03 = 0 or sp03=1 or sp03=3 or sp03 = 7" in the IF condition. This makes Eviews conduct the test, only on the observations needed for filling the restriction.

To test the hypothesis in Eviews, you first need to select the variable of interest. In this case the variable of interest is the weight of the students, sp04. Selecting the variable is done simply by double clicking it, which opens the *Series: SP04* window. To make Eviews perform the ANOVA test you need to go:

View > Descriptive Statistics & Tests> Equality Tests by Classification… (see the picture below)



As a result you will see the following window appearing:



To let Eviews know that we want to group the variable based on the political party, which is variable sp03, we type in sp03 as shown above. The '*Test quality of*' is set to "mean" as default, so we simply leave this setting. Clicking *OK* will give us the desired ANOVA output:

Test for Equality of Means of SP04
Categorized by values of SP03
Date: 01/03/10 Time: 10:41
Sample: 1 455 IF SP03=0 OR SP03=1 OR SP03=3 OR SP03=7
Included observations: 394

Method	df	Value	Probability
Anova F-test	(3, 390)	2.683933	0.0464
Welch F-test*	(3, 122.42)	2.533548	0.0601

*Test allows for unequal cell variances

Analysis of Variance

Source of Variation	df	Sum of Sq.	Mean Sq.
Between	3	1165.717	388.5723
Within	390	56463.12	144.7772
Total	393	57628.84	146.6383

To determine whether to reject the null hypothesis or not we focus on the highlighted *ANOVA F-test* output. The column named *Probability* contains the p-value of interest. Since the p-value is below 5% we reject the null hypothesis and conclude that there is a statistical significant difference in weight between the groups.

5.3 Testing assumptions

A number of assumptions must be met to ensure the validity of the above analysis of variance.

The following three assumptions will be checked in this section

1) Homogeneity of variance

2) Normally distributed errors

3) Independent error terms

5.3.1 Homogeneity of variance (1)

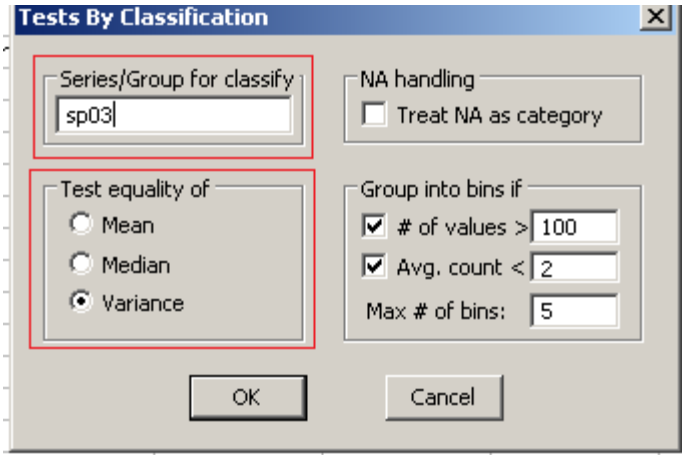
To test for homogeneity of variance between the different groups in the analysis, we use Levene's test for equality of variance. The hypothesis for the test, in our case, is:

$H_0: \sigma_O^2 = \sigma_S^2 = \sigma_K^2 = \sigma_V^2$

$H_1: at\ least\ two\ \sigma's\ differ$

To have EViews run Levene's test, is somewhat similar to running the ANOVA test in the first place. Once again you need to select the variable of interest, sp04, and then go:

View /Descriptive Statistics /Tests/Equality Tests by Classification...



In the resulting window you once again put in the grouping variable, sp03, but this time you ask Eviews to *Test equality of* **Variance** and not mean.

Test for Equality of Variances of SP04
Categorized by values of SP03
Date: 01/03/10 Time: 11:02
Sample: 1 455 IF SP03=0 OR SP03=1 OR SP03=3 OR SP03=7
Included observations: 394

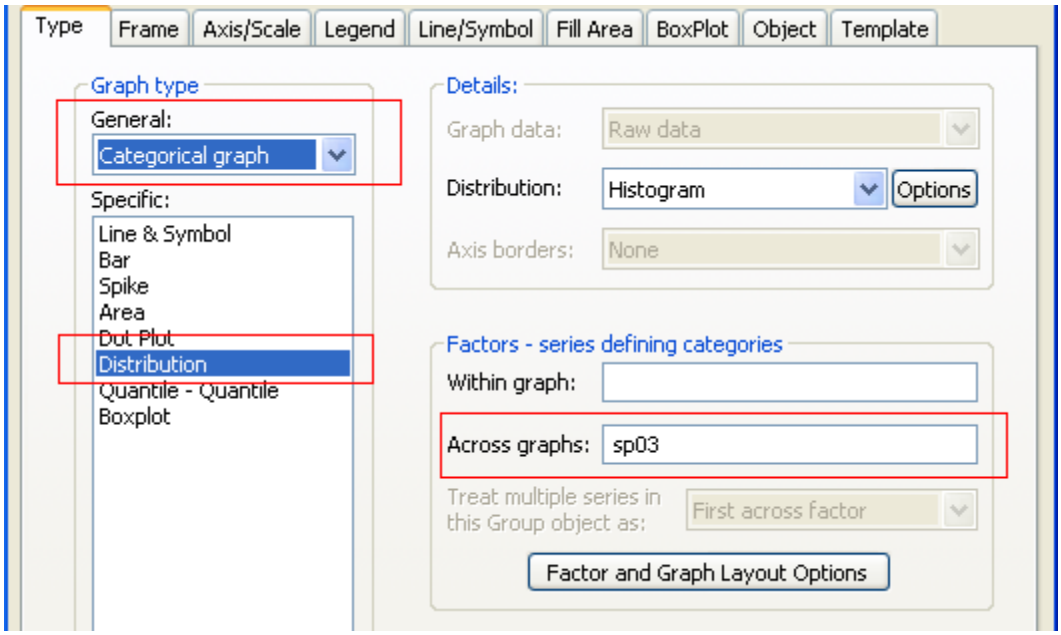
Method	df	Value	Probability
Bartlett	3	0.966214	0.8094
Levene	(3, 390)	0.516296	0.6713
Brown-Forsythe	(3, 390)	0.457641	0.7120

Just like in the ANOVA case we base our conclusion on the resulting p-value. But unlike in the ANOVA case we get a p-value of 0.67, which is way above any reasonable level of significance. Therefore we cannot reject the null hypothesis and assumption of homogeneity of variance is considered satisfied.

To test for the homogeneity of variance you could also use the False F-test. This is done in section 4.4.1 .

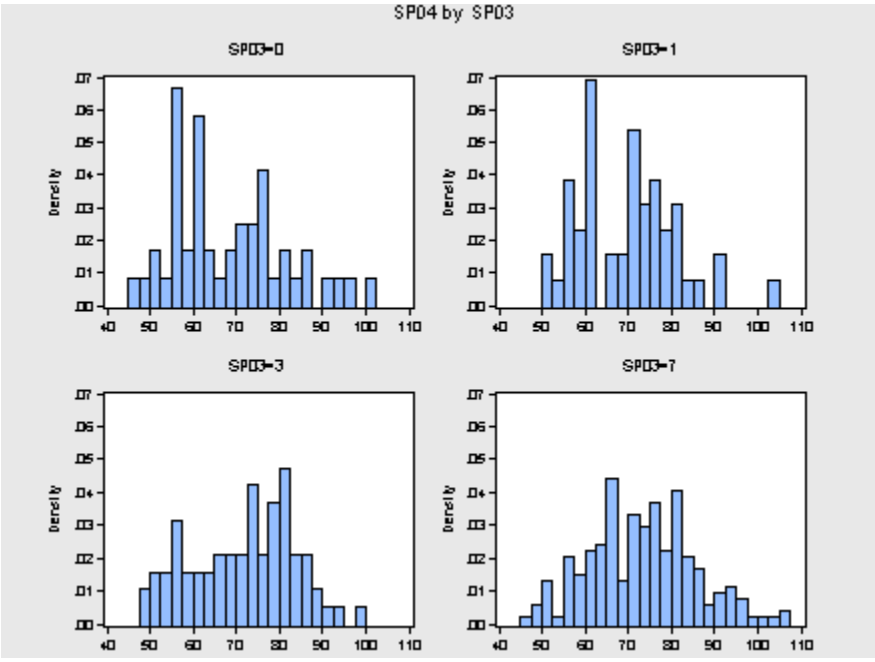
5.3.2 Normally distributed errors

We address the issue of normality within each group. Doing so can be done in different ways. First we address the assumption by creating distribution histograms for each group. Doing so is done by first selecting the dependent variable, the weight, sp04, by double clicking it. Then clicking *Graph* will result in the following option window:



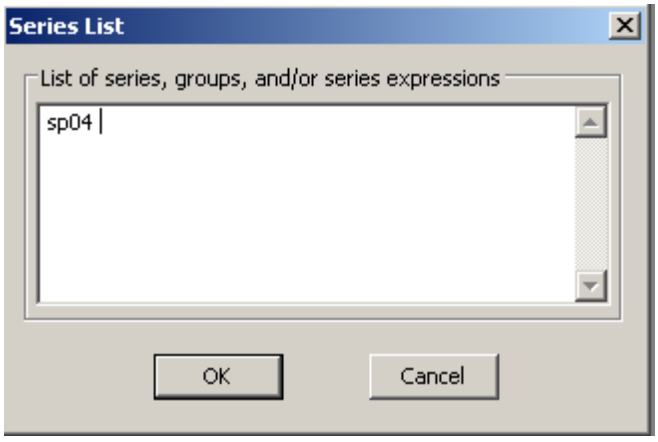
To make Eviews group the observation, first select *Categorical Graph* which gives you the additional option to the right. Then select *Distribution* and make Eviews do the actual grouping by writing the variable name sp03, in the *Across graphs* window.

The result should look similar to this:

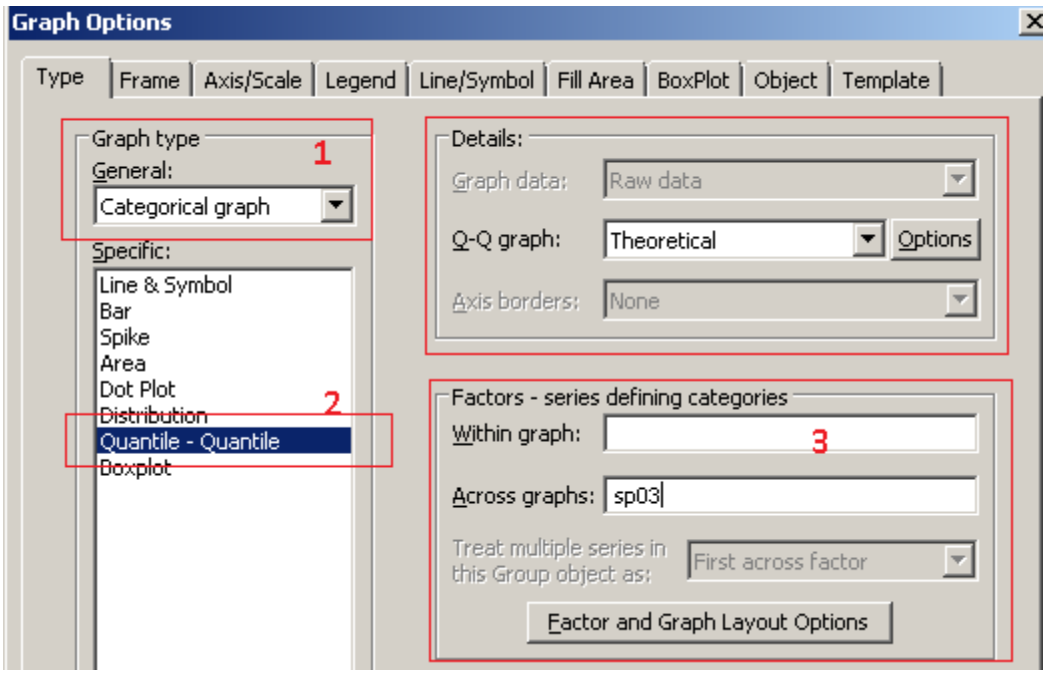


An alternative way of checking for normality is doing so across the different groups. Making this cross group analysis is done by using Q-Q plots to determine whether or not the observations follow a normal distribution when analyzed within their group. To make this analysis in EViews do the following:

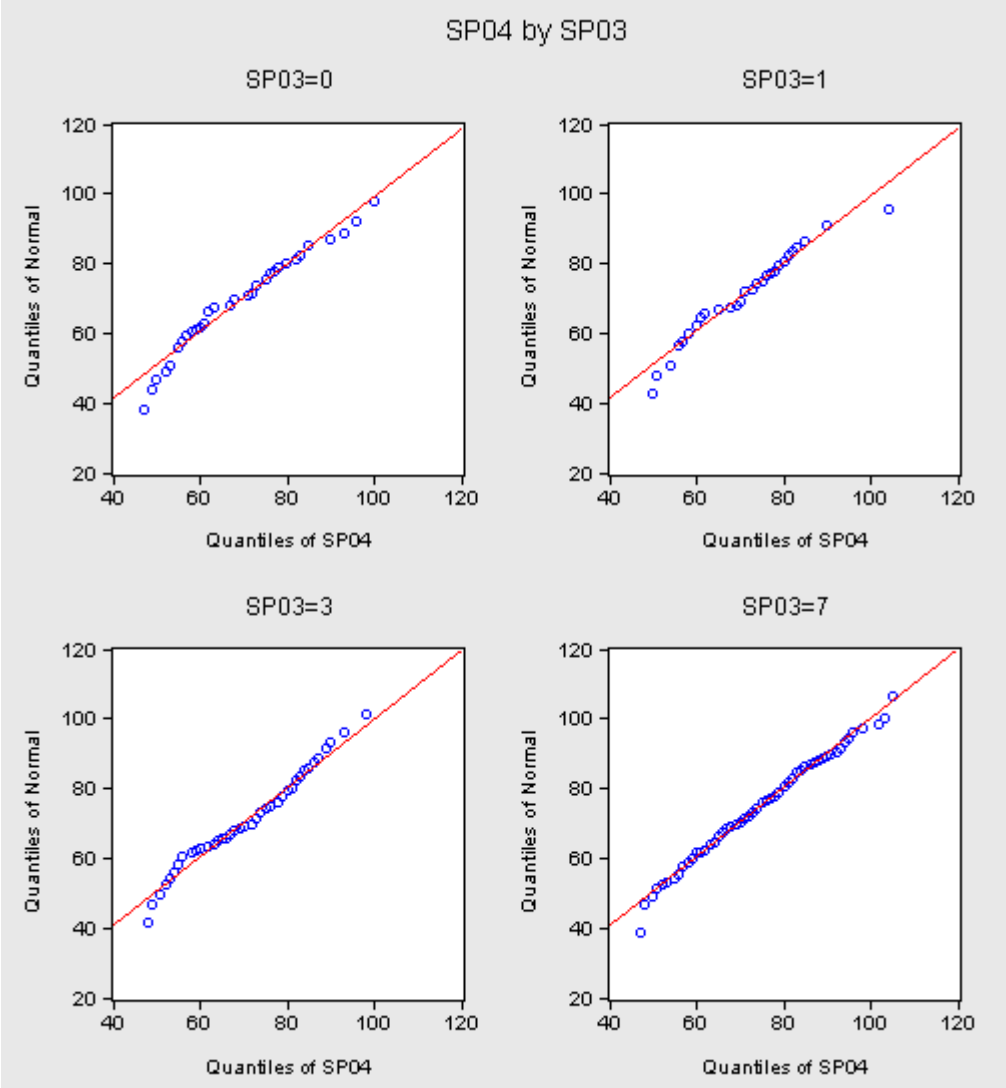
Select *Quick > Graph* from the top tool, which should result in the following windows:



Within this windows type in the variable of interest, the weight sp04 (in this case), and click *OK* – and you will face the following option window:



First you need to choose *Categorical Graph* from the dropdown menu (1). Then select the specific graph *Quantile – Quantile* (2), which is also known as the Q-Q plot. To make Eviews create a separate graph for each outcome in the grouping variable, you need to type in the grouping variable in the *Across Graphs* window. If you for some reason want Eviews to test for another distribution then the normal distribution you can change the options of the test in the Details window, but this is not of interest in our example. After clicking *OK* the resulting output presents itself:



The output displays the perfect normal distribution, the red line, and the actual observations, the blue dots, within each group. As we can see there exists only minor deviations from the red line and therefore we conclude that the assumption concerning normal distributed errors is satisfied.

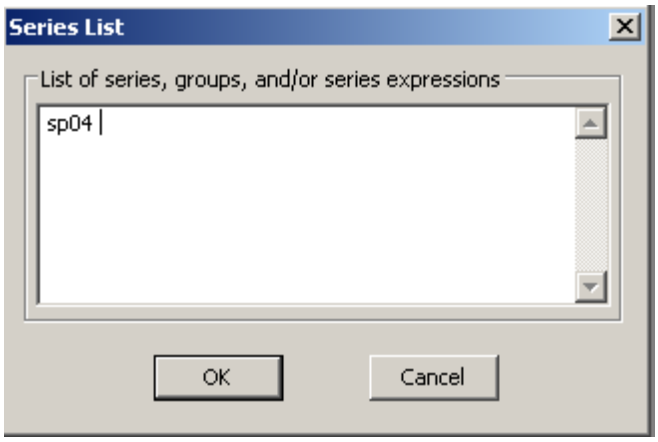
Before rejecting the assumption of normality one should always consider the properties of the central limit theorem – see Keller p. 300 for further.

5.3.3 Independent error terms (3)

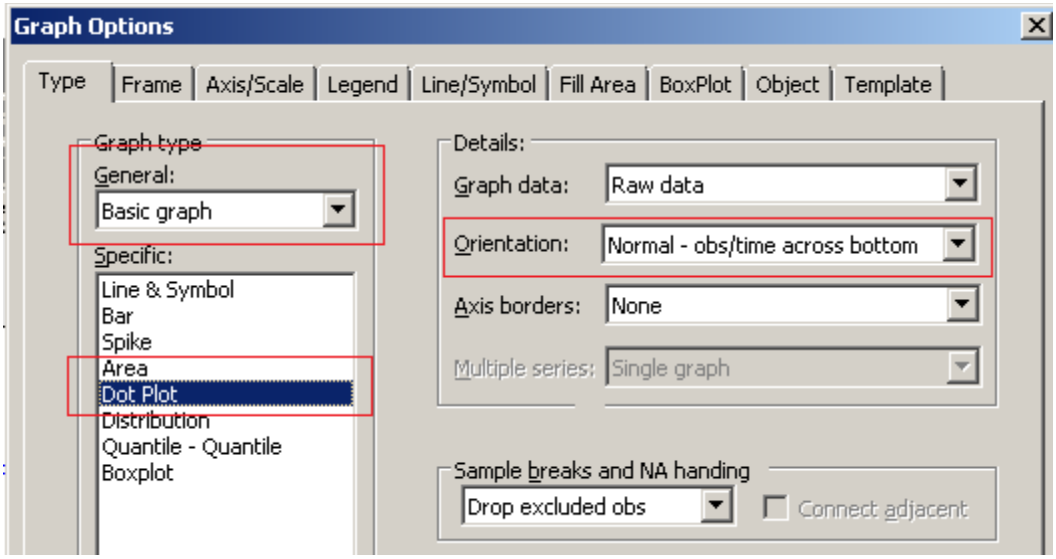
Problems concerning this assumption is by construction rarely a problem when analysing cross-sectional data, but is still mentioned in this manual to illustrate how the assumption is treaded in Eviews.

Assumptions concerning independent error terms is simply done, by making scatter plots of the variable of interest and the observation numbers. This is done to ensure that a pattern related to order in which the sample is collected, doesn't exist. Making a scatter plot diagram like this is somewhat similar to the graphs made above:

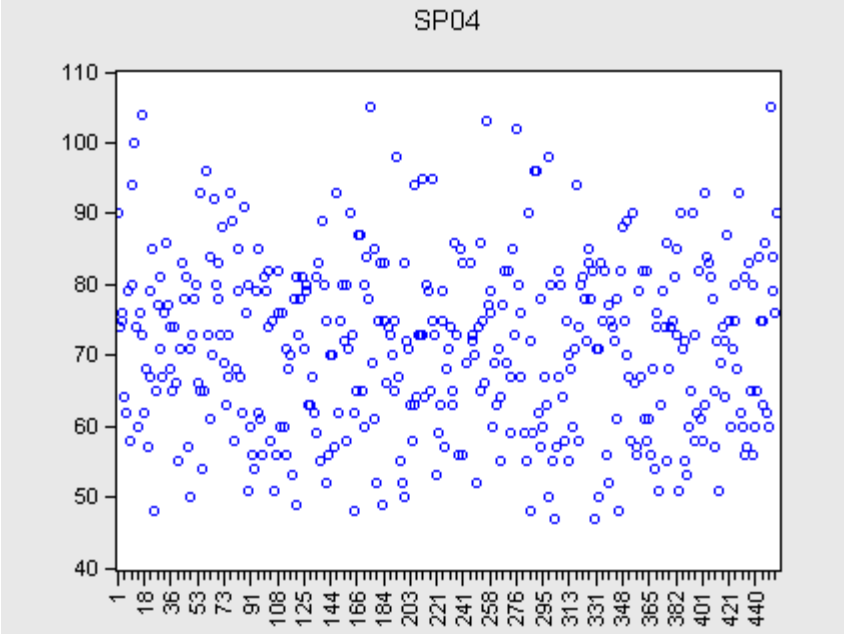
Select *Quick/Graph* and type in the variable of interest in the resulting window and click *OK*.



Once again you will face the graph option window:



This time you need to leave the option at *Basic Graph* and select *Dot Plot* from the specific window. Before clicking *OK* make sure that you window match the one shown in the picture above.



The resulting output is shown above. Since there is no evidence of a pattern cross the observation numbers, we conclude that the assumption concerning independent error terms is satisfied. Once again note that Eviews doesn't report the errors and we used the actual observations. In this case this using the actual observations should not make any difference.

6 Simple linear regression (SLR)

6.1 The basics

The basic idea of simple linear regression, is analyzing the relationship between two interval/ratio scaled variables. More specifically we want to determine (1) if the variable x causes y (2) and how large is the economical effect of variable x on variable y. That is, how much does y change when x changes by one unit. Put in mathematical terms:

In relation to the linear model relating x and y

$y = \beta_0 + \beta_1 \cdot x + u$

1) Is β_1 significant different from zero and

2) What is the magnitude of β_1

In this and the following section we will be using the work-file "FEMALEPRIVATWAGE.wf1" to determine the relation between a person's hourly wage and the co-variants such as education, experience, marriage and children. To illustrate how SLR could be used in this framework, consider the following example:

We are interested in determine how education (*educatio* in the work-file) is related to hourly wage (*hourwage*) and thus the relationship:

$hourwage = \beta_0 + \beta_1 \cdot educatio + u$

That is determining how large, if any, an effect education has on the hourly wage.

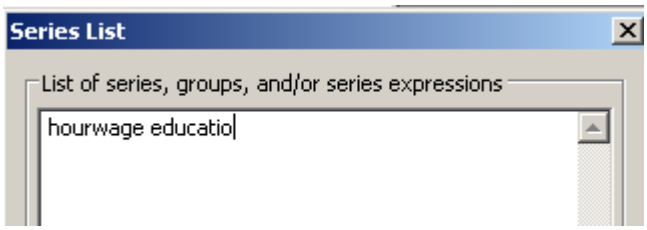
Please note:

The implications, theory and challenges concerning simple linear regression analysis are the main topics of Wooldridge chap. 2. In the following we will assume that the content of this chapter is somewhat known theory and we will not go into detail with more advanced implications of SLR such as reverse causality or omitted variable bias.

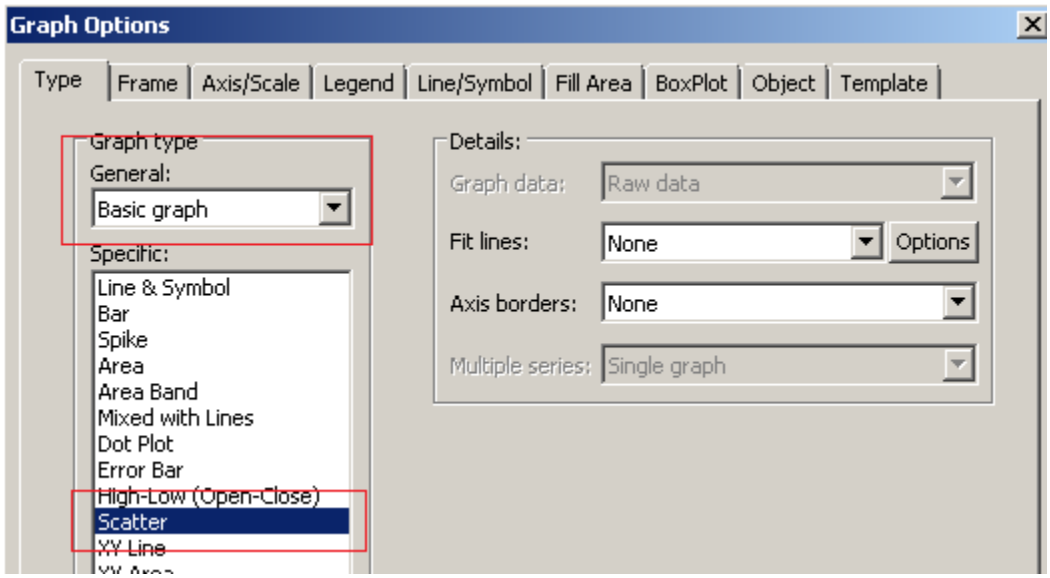
6.2 Scatter dot graphs

Descriptive statistics such as scatter dot graphs of the variables of interest is a very essential part of the regression analysis, since it allows you to explore the observed relation between the two variables and allows you to adjust your model accordingly[5]. Thus the use of these scatter dot graphs can be used to avoid of model misspecification.

To create scatter dot graphs in Eviews clicking *Quick > Graph* in the top tool bar will get you the following window:

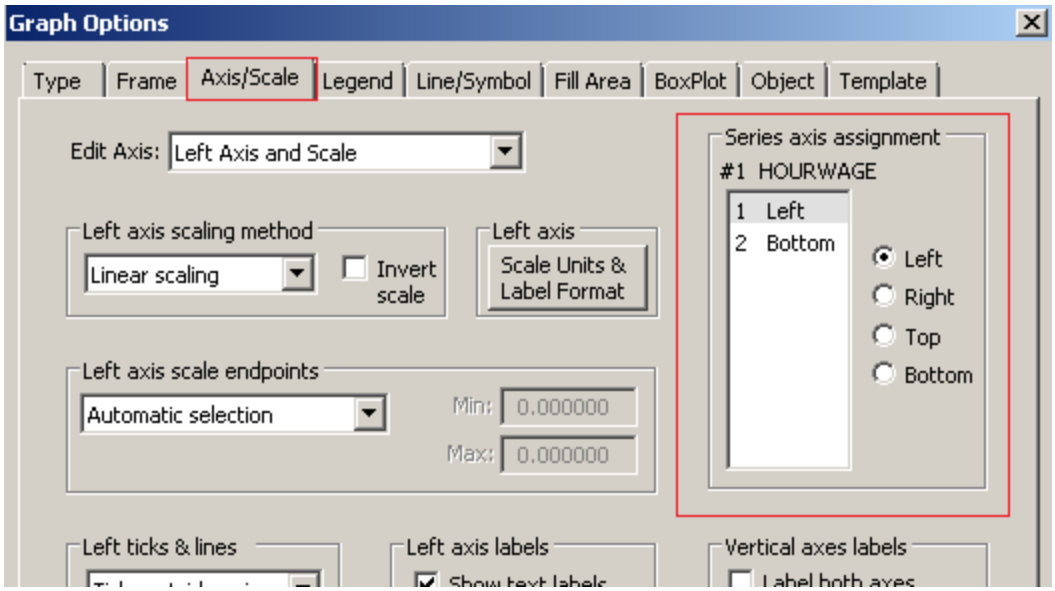


Here you simply type in the names of the two variables of interest, the order is not important, and click *OK*.



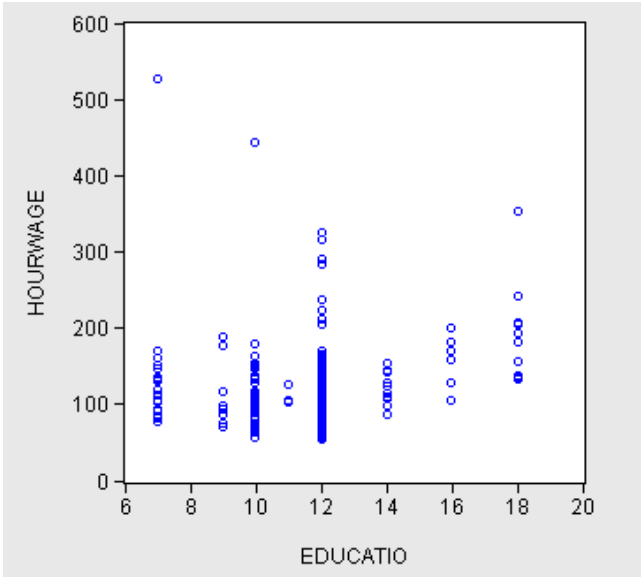
In graph type, select the default *Basic Graph* and *Scatter* from the specific list. To make sure that your variables are on the right axis, click the *Axis/Scale* tab and adjust the axis like shown below.

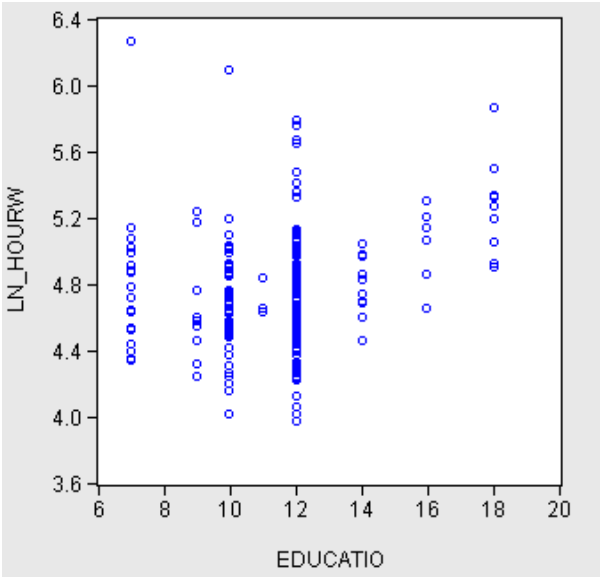
[5] E.g. let's say you plot y against both x, ln(x), x^2 and discover a better fitting relationship.



The resulting output should look somewhat similar to the one shown here. Note how we made the scatter plot for both hourly wage and LN(hourly wage) against education. This is done to show how scatter plots can be used to explore different relationships between variables. In this case we find that the LN(wage) vs. education show evidence of a better fitting relationship than wage vs. education . Based on this observation we could consider rewriting our model to the form:

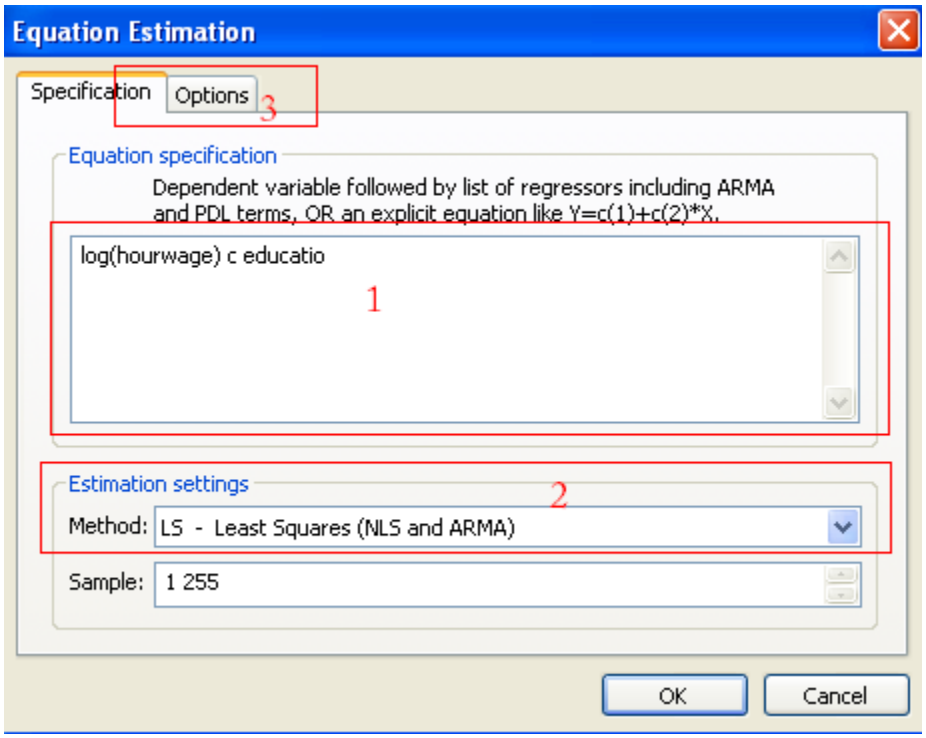
$$LN(hourwage) = \beta_0 + \beta_1 \cdot educatio + u$$





6.3 Model estimation in Eviews

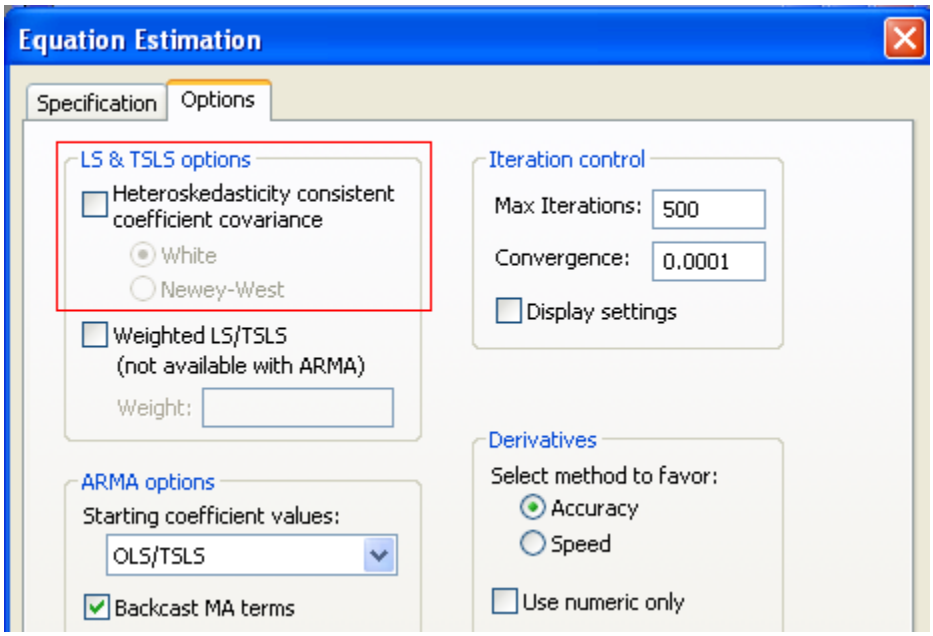
Running model estimation in Eviews, that is, determine the coefficient and their standard deviation in our model, is one of the notable strengths of the software. Doing so can like the variable creation, be done by using the command line or the estimation tool. In this manual we will focus on the use of the estimation tool. Opening this tool is done by selecting: *Quick > Estimate Equation...*



The Equation Estimation tool is shown in the previous picture. Most importantly is the *Equation specification* in which you have to specify which variables you want to include and their internal relations. It is really important to name your dependent variable (often called y) first. Following your dependent variable you should type the constant (β_0) written in EViews as "c" for constant (forgetting this constant has huge implications on the resulting estimation). Your regressior/explanatory/independent variable (also known as x) should follow the constant c (the β_0 from our model)[6]. In the section marked as 2 in the above picture is where you tell EViews which statistical method it should use to estimate the equation. In this manual we will not cover other than the default setting called least squares (see Wooldridge p. 27 for how to *"Derive the Ordinary Least Squares"* - the use of more advanced methods such as maximum likelyhood and two state least square is covered in Wooldridge*).*

Clicking the *options* tab (marked as 3) results in the following equation option window:

[6] In the MLR section we expand this part by using more variables.



Especially the marked area concerning heteroskedasticity might be of relevance somewhere down the line, since it's used to adjust for the problems concerning heteroskedasticity (Wooldridge chap. 8).

When the equation has been specified and all options are in place, clicking *OK* will result in an output similar to the one shown and described in the following section.

6.4 Model output

Equation: UNTITLED Workfile: FEMALEPRIVATEWAGE E...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: LOG(HOURWAGE) 1
Method: Least Squares
Date: 01/04/10 Time: 18:10
Sample: 1 255
Included observations: 255

Variable	Coefficient	Std. Error	t-Statistic	Prob.
2 C	4.355741	0.109058	39.93975	0.0000
EDUCATIO	0.033528	0.009343	3.588477	0.0004

R-squared 3	0.048433	Mean dependent var	4.739935
Adjusted R-squared	0.044672	S.D. dependent var	0.339227
S.E. of regression 4	0.331564	Akaike info criterion	0.637819
Sum squared resid	27.81344	Schwarz criterion	0.665594
Log likelihood	-79.32195	Hannan-Quinn criter.	0.648991
F-statistic 5	12.87717	Durbin-Watson stat	1.913342
Prob(F-statistic)	0.000399		

The output produced by Eviews (see the above picture) might seem overwhelming at first, but far from all of the data has relevance at this stage[7]. We have highlighted the most important aspects of the output to make the interpretation easier accessible:

1. Section shows the dependent variable (your y) – always make sure this part of the output is as intended. – The output reports that LOG(HOURWAGE) is the dependent variable just as intended.
2. Section contains the most critical information. The table shows the estimated constant, $\hat{c}$, and the estimated coefficient, $\widehat{\beta_1}$, their standard deviations, t statistics and resulting p-values.

   In this case the output reports that:

   $$\widehat{\beta_0} = 4.3557 \; \sigma_{\beta_0} = 0{,}10905 \; and \; \widehat{\beta_1} = 0{,}03352 \;\; \sigma_{\beta_1} = 0{,}009343$$

   Resulting in a final model estimate:

   $$\widehat{LN(hourly\ wage)} = 4{,}3557 + 0{,}03352 \cdot education$$
   $$R^2 = 0{,}048433 \qquad (0{,}10905) \;\; (0{,}009343)$$
3. Shows R-squared and R-squared adjusted. Both of which are so call goodness-of fit indicators. To understand the difference between the two see Wooldridge p. 199-203.
4. S.E. regression is useful when the model is used for forecasting/prediction – see. Wooldridge p. 206-9 on "*Prediction and Residual Analysis*".
5. The F-statistic is somewhat similar to the one used in the ANOVA section. Also see Wooldridge p. 143-154.

6.5 Testing SLR assumptions

Not all the relevant assumptions can be tested using statistical software. We will in this section focus on the ones that are testable. To make the use of SLR valid we must satisfy all of the assumptions referred to as *The Guass-Markov Assumptions for Simple Regression* in Wooldridge and additionally; we need to assume that the errors are independent and normally distributed with mean 0 and variance σ^2 - that is $u|x \;\; i.i.d. \sim N(0, \sigma^2)$ - to be able to run the above used hypothesis test.[8]

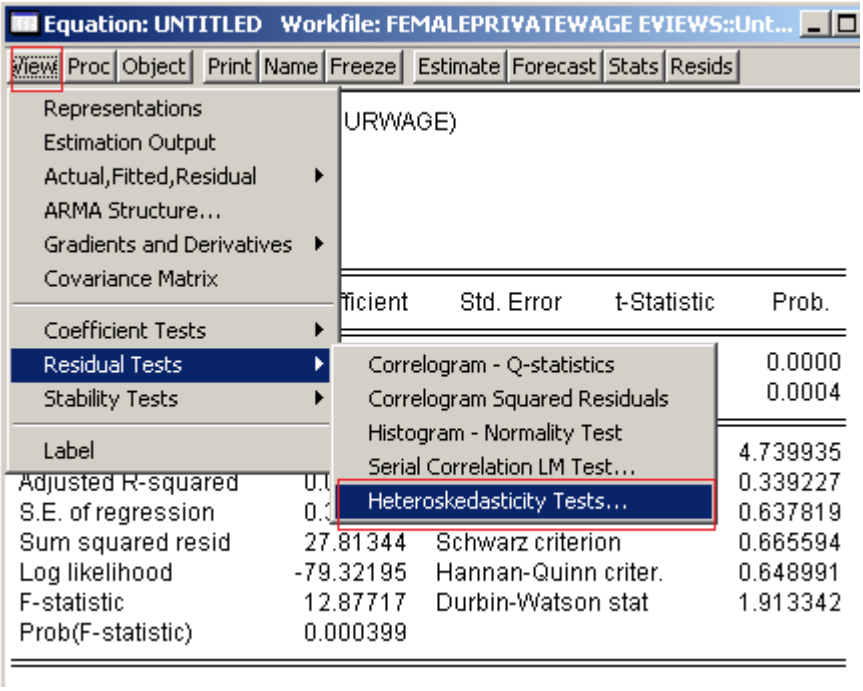
6.5.1 Testing for heteroskedacity – SLR.5

To understand the meaning of homoskedasticity, see Wooldridge p. 52-58. Besides this introduction to the phenomena, Wooldridge has dedicated the entire chapter 8 to explain, how to test for this assumption and how to adjust the method of estimation accordingly. We will only focus on how to run these tests and how to interpret the resulting output, not the underlying theory.

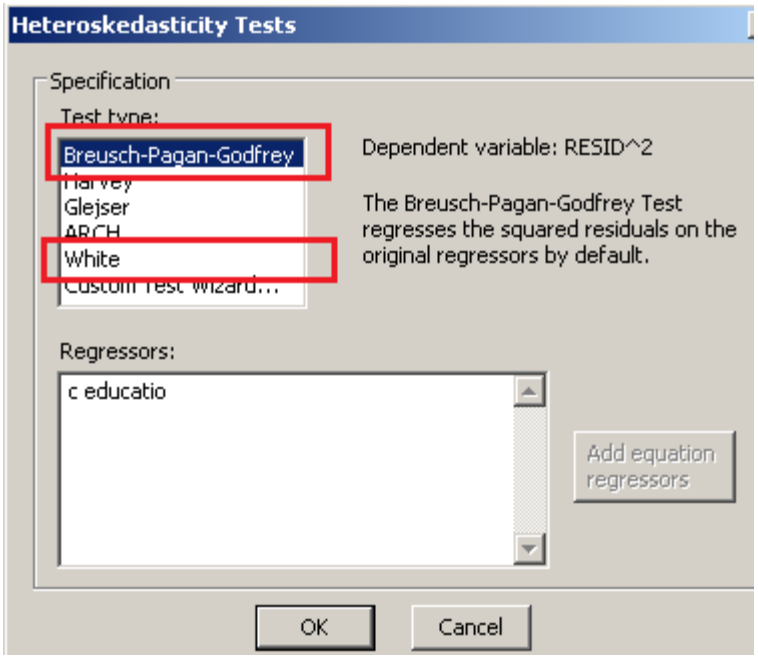
To run test related to any estimated model, you must first of all estimate the model as shown in the previous section. Then, within the resulting equation window, click *View/Residual Tests/Heteroskedasticity Tests..* – as shown below.

[7] In this case stage is not only a reference to the indended user but also the SLR.

[8] In relation to validity the most importance of the assumptions is SLR.4. – that is $E(u|x) = 0$ – see Wool. Chap 2.



Doing so will result in the following window:



The window shows a list of possible tests, all testing for heteroskedasticity. The tests covered in Wooldridge are the Breusch-Pagan-Godfrey [Wooldridge p. 273] and White [Wooldridge p. 274].

No matter which test we use for testing heteroskedacity, the null hypothesis is identical:

$H_0: Var(u|x) = \sigma^2$ – There's homoskedacity

$H_1: Var(u|x) \neq \sigma^2$- There's heteroskedacity

The resulting outputs when running these test is shown below:

Heteroskedasticity Test: White

F-statistic	1.528063	Prob. F(2,252)	0.2190
Obs*R-squared	3.055453	Prob. Chi-Square(2)	0.2170
Scaled explained SS	8.445976	Prob. Chi-Square(2)	0.0147

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 01/06/10 Time: 12:26
Sample: 1 255
Included observations: 255

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.533200	0.250592	2.127765	0.0343
EDUCATIO	-0.066562	0.041491	-1.604239	0.1099
EDUCATIO^2	0.002485	0.001705	1.457393	0.1463

Heteroskedasticity Test: Breusch-Pagan-Godfrey

F-statistic	1.173079	Prob. F(1,253)	0.2798
Obs*R-squared	1.176895	Prob. Chi-Square(1)	0.2780
Scaled explained SS	13.47424	Prob. Chi-Square(1)	0.0002

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 01/06/10 Time: 18:51
Sample: 1 255
Included observations: 255

	Coefficient	Std. Error	t-Statistic	Prob.
C	7698.450	4548.246	1.692620	0.0918
EDUCATIO	-422.0367	389.6607	-1.083088	0.2798

What both tests does is using the squared residuals (RESID^2) as the dependent variable and try to determine whether these can be explained using different forms of the original independent variables [see. Wooldridge chap. 8 for further detail]. To conclude whether we have to reject the null hypothesis or not, using the resulting F statistic is enough. The F-test tests for the joint significant of all the included independent variables (see the future sections on this topic and Wooldridge chap. 4). If these are not jointly significant, then we cannot reject the null hypothesis and assume homoskedadicity. To reject the null hypothesis we would need a prob. value (or p-value) less than 0.05. None of the two tests reports p-values anywhere close to 5% so cannot reject the null hypothesis – in other words, heteroskedacity does not seem to be a problem.

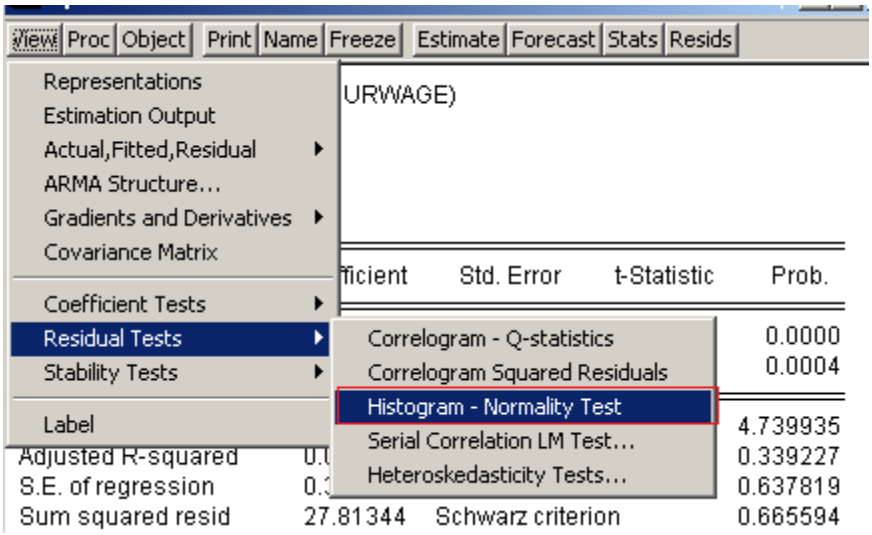
6.5.2 Testing for normally distributed errors

To test for normal distributed errors we use the Jarque-Bara test for normality. The hypothesis of the Jarque-Bera test is a follows:

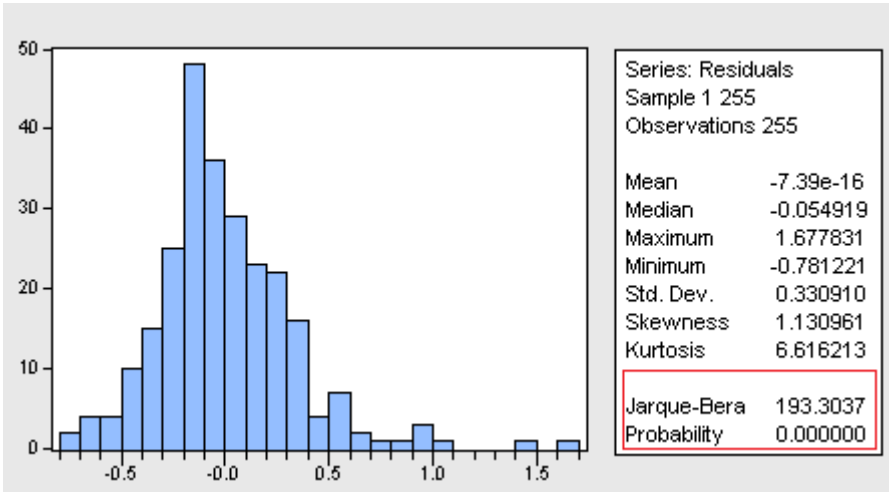
H_0: *errors are normally distributed*

H_1: *errors are not normally distributed*

Running the test in Eviews is somewhat similar to running the tests for heteroskedacity. First you must estimate the model (which creates the residuals on which the test is based), then simply go: *View/Residual Tests/Histogram /Normality test*



Doing so will result an output similar to the following:



To determine whether the assumption of normal distributed errors are satisfied or not, we once again turn our attention to the highlighted test statistic and p-value. The p-value in this case turns out to be 0, and as a result, we reject the null hypothesis

7 7. Multiple linear regression (MLR)

7.1 The basics

Multiple linear regression is quite similar to simple linear regression but with more than one independent variables [see. Wooldridge p. 72]. MLR determines $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ in a model of the following kind:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + u$$

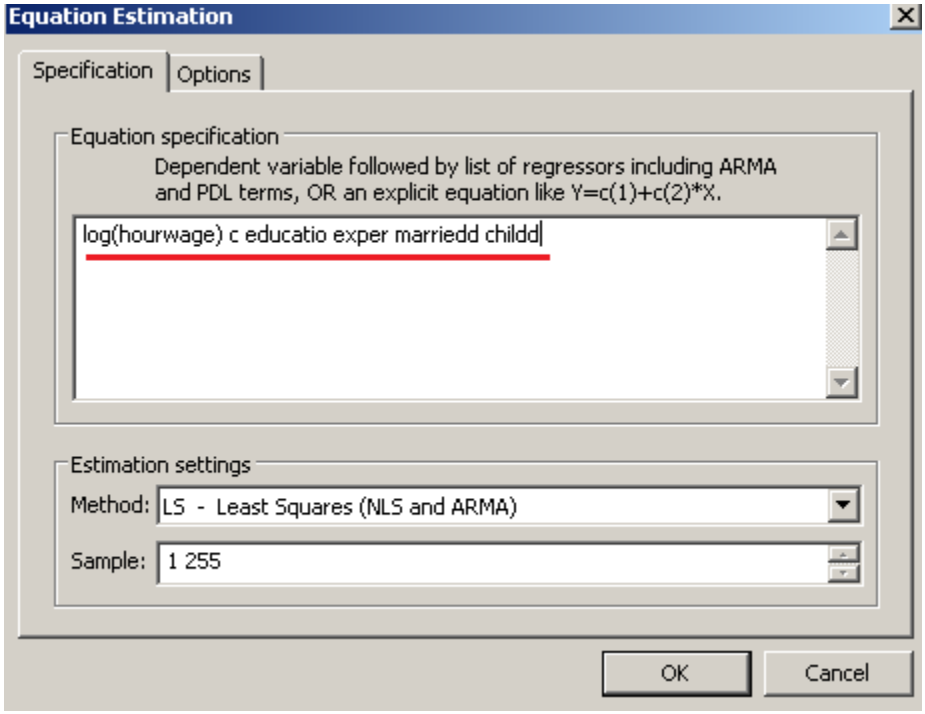
The purpose of including more than one variable, to explain the variance in y, are as follow:

1) Deal with the problems of omitted variable bias – that is to make the "everything else equal" assumption SLR.4 valid – see. Wooldridge Section 3.1 on motivation for multiple regression.
2) Including more variables, and thereby increase R^2 i.e. the explanation power of the model, which can increase the precision of forecasting.

Wooldridge dedicates a large part of his book to this subject. To gain a basic understanding of the concept reading chapter 3 will get you started. But to gain a sufficient understanding reading chapter 3-9 is strongly recommended.

7.2 Model estimation in EViews

Running the MLR model estimation in Eviews is similar to running the model estimation of SLR. Simply choose *Quick/Estimate Equation* which result in the familiar estimation window:



The model estimated in the above case is:

$$LN(hourly\ wage) = \beta_0 + \beta_1 \cdot education + \beta_2 \cdot experience + \beta_3 \cdot married + \beta_4 \cdot child + u$$

We are still trying to determine why some people make more money than others (the variance in *hourwage)*, but this time we include more potential explanations (independent variables). Note that married and child are both dummy. Married is 1 if a person is married and 0 otherwise. Child is 1 if a person has at least one child has at least one child and 0 otherwise. Running the above equation results in the following output:

Dependent Variable: LOG(HOURWAGE)
Method: Least Squares
Date: 01/07/10 Time: 12:57
Sample: 1 255
Included observations: 255

	Coefficient	Std. Error	t-Statistic	Prob.
C	4.067641	0.120822	33.66642	0.0000
MARRIEDD	-0.068956	0.047681	-1.446175	0.1494
CHILDD	0.011722	0.048919	0.239615	0.8108
EDUCATIO	0.044374	0.008909	4.980669	0.0000
EXPER	0.017511	0.002965	5.905731	0.0000

R-squared	0.177484	Mean dependent var	4.739935
Adjusted R-squared	0.164324	S.D. dependent var	0.339227
S.E. of regression	0.310106	Akaike info criterion	0.515606
Sum squared resid	24.04138	Schwarz criterion	0.585042
Log likelihood	-60.73972	Hannan-Quinn criter.	0.543536
F-statistic	13.48641	Durbin-Watson stat	2.082515
Prob(F-statistic)	0.000000		

When running MLR model estimation the first place to look in the output, is at the F-statistic and its p-value (underlined in the above figure). As described in Wool. Section 4.5 – *Testing Multiple Linear Restriction: The F Test* – the F-test tests multiple linear restrictions. In Eviews the hypothesis tested by the F-test in the basic MLR estimation output is:

$$H_0: \beta_1 = 0\ , \beta_1 = 0\ , \beta_2 = 0\ , \ldots\ , \beta_k = 0$$

$$H_1: H_0 is\ not\ true$$

A simple interpretation of the null hypothesis is that the union of all used regressors do not have a significant effect on y. In our example we find that the regressors used to have a significant effect on y (the p-value is 0 thus we reject the null hypothesis).

The analysis of each specific variable, their significance and effect on y is somewhat similar to the analyse SLR – see the above section.

7.3 Models with interaction terms

To really gain a understand of some of the many other possibilities available when using MLR for data analysis we refer to Wooldridge chap. 6 which describes the use of *Models with Interaction Terms.* Estimating interaction models in Eviews is no difficult task. To illustrate how it's done, consider the following example: We want to determine if the effect of an extra

year of education has a different effect on hourly wage when employed in the province vs. working in the capital. To determine whether or not this is the case, we need to estimate the following model:

$$\mathrm{LN(hourly\ wage)} = \beta_0 + \beta_1 \cdot \mathrm{education} + \beta_2 \cdot (\mathrm{education} \cdot \mathrm{province}) + \beta_3 \cdot \mathrm{province} + \mathrm{u}$$

If $\beta_2$is significant different from zero we must conclude that the effect of education does differ depending on your location of work.

To run this form of model in Eviews we can either construct a new variable, like shown in a previous section, and then run the model estimation or we can do the following:

In the estimate equation window type:

```
log(hourwage) c education province educatio*province
```

Thus giving the output:

Dependent Variable: LOG(HOURWAGE)
Method: Least Squares
Date: 02/13/10 Time: 13:19
Sample: 1 255
Included observations: 255

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.386599	0.134926	32.51118	0.0000
EDUCATIO	0.035530	0.011376	3.123112	0.0020
PROVINCE	0.008689	0.226919	0.038290	0.9695
EDUCATIO*PROVINCE	-0.010709	0.019563	-0.547402	0.5846

R-squared	0.077449	Mean dependent var	4.739935
Adjusted R-squared	0.066422	S.D. dependent var	0.339227
S.E. of regression	0.327768	Akaike info criterion	0.622538
Sum squared resid	26.96534	Schwarz criterion	0.678087
Log likelihood	-75.37363	Hannan-Quinn criter.	0.644882
F-statistic	7.023857	Durbin-Watson stat	1.924037
Prob(F-statistic)	0.000149		

Thus we conclude that the effect of education does not differ depending on whether one lives in a province or not.

7.4 The assumptions of MLR

The assumptions of any test is always or great importance when considering its validity. The importance and implications of each specific assumption is discussed in great detail in Wooldridge. Not all of these assumptions (MLR1-6 – Wooldridge p. 157-158) can be formally tested, just like in the SLR case. In the SLR section we show how to use Eviews to test for normal distributed errors and heteroskedacity. Even though the underlying theory does differ when running these tests in the MLR case, there are no differences in the way these are run in Eviews. For this reason we will not repeat that part of the manual.

For normality: Section 6.5.2 page 30, and for homoskedacity: Section 6.5.1 page 28

7.5 Testing multiple linear restrictions – the Wald test

Assume we have the following model:

$$\text{LN(hourly wage)} = \beta_0 + \beta_1 \cdot \text{education} + \beta_2 \cdot \text{education}^2 + \beta_3 \cdot \text{experience} + \beta_4 \cdot \text{age} + u$$

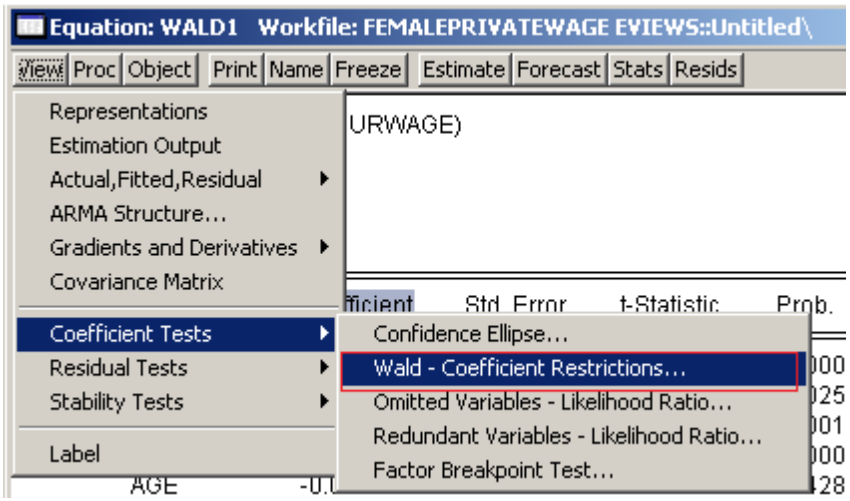
We suspect that almost all of these variables are somewhat positively correlated with each other. To test for joint significance one option would be to use the F-test as described in Wooldridge p. 143, another option would be to use the Wald test in Eviews.

The Wald test tests one or more linear restriction on the model. Let say we want to test the join significance of age and experience in the above example (note that just because one variable is significant does not necessary mean that the group including the variable is significant). The hypothesis in that case would be:

$$H_0: \beta_3 = 0 \text{ and } \beta_4 = 0$$

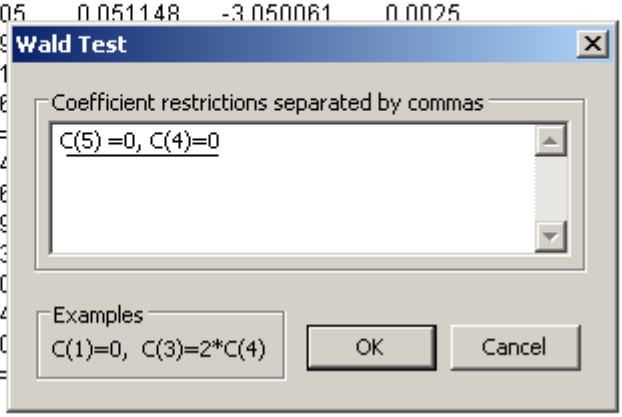
$$H_1: H_0 \text{ is not true}$$

To run the Wald test in Eviews is done by first estimating the model including all reggressors of interest (see the former section on estimation), within the resulting output window go: *View > Coefficient Tests >Wald - Coefficient Restrictions* like shown in the picture below.



The resulting window is where the restrictions from our hypothesis are written. Doing so can be a little misleading since Eviews names the variables a little different than what is normally done. Eviews names the variables according to the number by which they appear in the output (or in the estimated equation for that matter). So the constant β_0(called C in Eviews) is in this case *not* C(0) but C(1) since it appear as the first variable on the list.

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)= C	5.296199	0.341635	15.50252	0.0000
C(2)= EDUCATIO	-0.156005	0.051148	-3.050061	0.0025
C(3)=EDUCATIO^2	0.00829			
C(4)= EXPER	0.02181			
C(5)= AGE	-0.00456			
R-squared	0.21994			
Adjusted R-squared	0.20746			
S.E. of regression	0.30199			
Sum squared resid	22.8003			
Log likelihood	-53.9820			
F-statistic	17.6224			
Prob(F-statistic)	0.00000			



Writing the restrictions is done in the white field like shown above. Note that we write more than one restriction by separating each with a comma like in the above example. An alternative to the above used way of writing our restriction would be to write:

```
C(5)=C(4)=0
```

The resulting output when running the test (clicking *OK*) is shown below:

Wald Test:
Equation: WALD1

Test Statistic	Value	df	Probability
F-statistic	17.52742	(2, 250)	0.0000
Chi-square	35.05485	2	0.0000

Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(5)	-0.004566	0.003106
C(4)	0.021817	0.004642

Restrictions are linear in coefficients.

Eviews reports both the Chi-square and the F-statistic statistics. The choice between the two should not make that big of a difference, since the resulting p-value will not differ by any significant amount. Like in any other hypothesis test we reject the null hypothesis if the resulting p-value falls below our predetermined level of significance (we use 0.05). In this specific case we get a resulting p-value equalling zero and thus reject the null hypothesis and conclude that the variables does have a jointly significant effect on our dependent variable.

8 General ARMA processes

8.1 Univariate time series: Linear models

In this introduction to univariate time series, the ARMA process will be discussed. In general we consider a time series of observations on some variable, e.g. dividends $Y_1 \ldots Y_t$. These observations will be considered realizations of a random variable that can be described by some stochastic process.

A simple way to model dependence between consecutive observations states that Y_t depends linearly upon its previous values $Y_{t-1} \ldots Y_{t-p}$, that is:

$$Y_t = \delta + \rho Y_{t-1} \ldots \rho Y_{t-p} + \varepsilon_t$$

The above equation states that the current value Y_t equals to a constant δ plus ρ times its previous value plus an unpredictable component ε_t. The equation above is called an auto-regressive process of order p, or in short, AR(ρ).

In this case we are interested in determining the best ARMA representation of D, which is defined as dividends. This gives us the following equation:

$$div_t = \delta + \rho div_{t-1} + \varepsilon_t$$

8.2 Testing for unit root in a first order autoregressive model

Before determining the appropriate ARMA representation of D, a unit root test is advantageous. In the equation above ρ = 1, which corresponds to a unit root. The consequence on a unit root is among other things non-stationarity. Non-stationarity implies that the distribution of the variable of interest does depend on time, therefore we must exercise caution in using them directly in regression models.

A convenient equation for carrying out the unit root test is to subtract y_{t-1} from both sides of the equation above and to define $\theta = \rho - 1$:

$$\Delta div_t = \delta + \theta div_{t-1} + \varepsilon_t$$

If the null hypothesis H_0: θ=0 (or the first autocorrelation ρ = 1) is true, then a unit root is obtained, which indicates that the time series is non-stationary. To test the null hypothesis that θ=1, it is possible to use the standard t-statistic, but with different critical values calculated using Dickey-Fuller. The DF test is estimated by using three different equations, as presented in E-Views. The three test equations are:

$$\Delta y_t = \delta + \theta y_{t-1} + \varepsilon_t$$

$$\Delta y_t = \delta + \theta y_{t-1} + \gamma t + \varepsilon_t$$

$$\Delta y_t = \theta y_{t-1} + \varepsilon_t$$

The first equation has an intercept, indicated by the parameter δ, represents a random walk model with drift. The second equation with a trend (γ) and an intercept represents a random walk model with drift around a stochastic trend, and the last one represents a random walk model.

The data analysis for testing the unit root can be done as follows.

- Click the *div* variable, and the values will appear on the screen.
- Click *View/Unit root test* and the options for the unit root test is as follows:

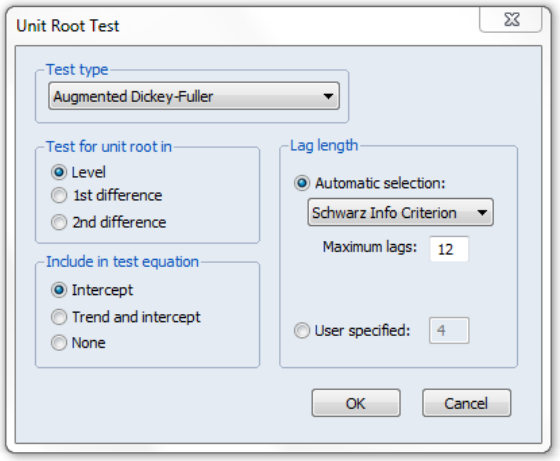


Figure 1

Test type: Different types of statistical options are available. The most common tests are the Augmented Dickey-Fuller, Philips-Perron and Kwiatkowski-Philips-Schmidt-Shin (KPSS).

Test for unit root in: Level, 1st difference and 2nd difference. These options are related to the amount of times that the series have to be differenced before it gets stationary. The hypothesis goes as follows:

Level: I(1) vs. I(0) for H0 vs. H1
1st: I(2) vs. I(1) for H0 vs. H1
2nd: I(3) vs. I(2) for H0 vs. H1

In general, an I(*d*) process is a series that is stationary after differencing *d* times.

Lag length: According to Wooldrigde J. M. the number of lags included is a trade-off between losing power and a wrong test statistic. There are no general rules to follow in any case, but for annual data, one or two lags usually suffice. For monthly data, we might include 12 lags.

To choose at which level the test should be executed, an initial visual inspection is required.

Figure 2 illustrates *div* displayed in level, 1st difference and 2nd difference. The visual inspection tells us that the dividends are trended, either by a random walk with drift or a deterministic trend. Furthermore it seems stationary by taking the 1st difference.

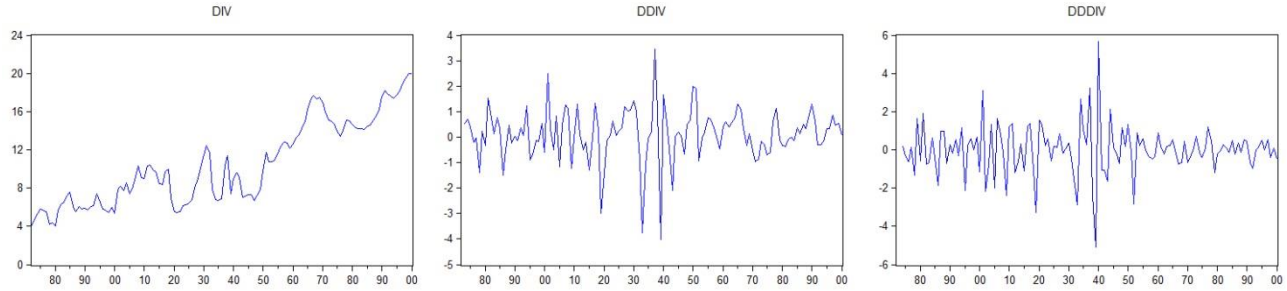


Figure 2

By choosing the following two setups, we intend to do a test on the two auxiliary equations below, which represents the equation with zero lag (standard Dickey-Füller) and one lag respectively (Augmented Dickey-Fuller). We could of course extend the test including more lags, but remember the above-mentioned conclusion.

$$\Delta y_t = \delta + \theta y_{t-1} + \gamma t + \varepsilon_t$$

Null Hypothesis: D(DIV) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 0 (Automatic - based on SIC, maxlag=0)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.380237	0.0000
Test critical values:	1% level	-4.031899	
	5% level	-3.445590	
	10% level	-3.147710	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DIV,2)
Method: Least Squares
Date: 01/24/13 Time: 15:31
Sample (adjusted): 1874 2000
Included observations: 127 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(DIV(-1))	-0.829242	0.088403	-9.380237	0.0000
C	0.022164	0.179555	0.123441	0.9020
@TREND(1872)	0.001188	0.002408	0.493370	0.6226

R-squared	0.415071	Mean dependent var	-0.003767
Adjusted R-squared	0.405637	S.D. dependent var	1.289123
S.E. of regression	0.993848	Akaike info criterion	2.848875
Sum squared resid	122.4791	Schwarz criterion	2.916060
Log likelihood	-177.9035	Hannan-Quinn criter.	2.876171
F-statistic	43.99585	Durbin-Watson stat	1.923809
Prob(F-statistic)	0.000000		

$$\Delta y_t = \delta + \theta y_{t-1} + \gamma t + \beta \Delta y_{t-1} + \varepsilon_t$$

Null Hypothesis: D(DIV) has a unit root
Exogenous: Constant, Linear Trend
Lag Length: 1 (Automatic - based on SIC, maxlag=1)

		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		-9.001326	0.0000
Test critical values:	1% level	-4.032498	
	5% level	-3.445877	
	10% level	-3.147878	

*MacKinnon (1996) one-sided p-values.

Augmented Dickey-Fuller Test Equation
Dependent Variable: D(DIV,2)
Method: Least Squares
Date: 01/25/13 Time: 14:55
Sample (adjusted): 1875 2000
Included observations: 126 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(DIV(-1))	-1.019931	0.113309	-9.001326	0.0000
D(DIV(-1),2)	0.227302	0.087998	2.583038	0.0110
C	0.012187	0.178822	0.068150	0.9458
@TREND(1872)	0.001621	0.002391	0.678103	0.4990

R-squared	0.446958	Mean dependent var	-0.005175
Adjusted R-squared	0.433359	S.D. dependent var	1.294171
S.E. of regression	0.974195	Akaike info criterion	2.816820
Sum squared resid	115.7847	Schwarz criterion	2.906860
Log likelihood	-173.4596	Hannan-Quinn criter.	2.853400
F-statistic	32.86608	Durbin-Watson stat	2.006178
Prob(F-statistic)	0.000000		

The Durbin-Watson statistics is close to two, including one lag, which indicates that the errors are serially uncorrelated against the alternative that they follow a first order autoregressive process.

By choosing different test types according to Figure 1, we can increase the robustness of the test. As mentioned above, the Phillips-Perron and the KPSS test is preferred. The hypothesis for the ADF and the Phillips-Perron is the same, which means that the null hypothesis claims that a unit root is present, i.e. the more negative the test statistic is, the stronger is the probability of rejecting the null hypothesis; that there is a unit root at the given level of confidence. The KPSS test works the other way around, which means that it tests the null hypothesis of stationarity against the alternative of a unit root. The first table test whether div is I(2) or I(1) (i.e. first difference with intercept), and the second table test whether div is I(1) or I(0) (i.e. in levels with trend and intercept).

Test	Test statistics	5 % critical limit*	Conclusion
ADF	-9,079	-3.45	Reject
Philips-Perron	-70,111	-3.45	Reject
KPSS	0,500	0,146	Reject

The test statistics is all above the critical limit, therefore, we reject I(2) in favor of I(1).

*The limits viewed in E-views is not correct, instead selected percentiles of the appropriate distribution are developed and published in several works by Dickey and Fuller.

Test	Test statistics	5 % critical limit*	Conclusion
ADF	-3,497	-3.45	Reject
Philips-Perron	-2,931	-3.45	Fail to reject
KPSS	0,174	0,146	Fail to reject

The unit root tests in levels show some different results. The Philips-Perron and KPSS tests suggest a unit root at the 5 % significance level, while the ADF test rejects a unit. As we have two tests pointing I(1) and the ADF test is only marginally rejecting a unit root, it could be sign of dividends having a unit root in levels.

8.3 Formulating ARMA processes

As mentioned in the beginning of the previous subchapter, a simple way to model dependence between consecutive observations states that Y_t depends linearly upon its previous values $Y_{t-1}…Y_{t-p}$, that is:

$$Y_t = \delta + \rho Y_{t-1} \dots \rho Y_{t-p} + \varepsilon_t$$

Or even simpler:

$$Y_t = \delta + \rho Y_{t-1} + \varepsilon_t$$

This corresponds to the most simple first-order autoregressive growth model. It says that the current value Y_t equals to a constant δ plus ρ times its previous value plus an unpredictable component ε_t. The first equation above is called an auto-regressive process of order p, or in short, AR(p), and the second equation above is called an auto-regressive process of order 1, or in short, AR(*1*).

To estimate the equation above, which actually could be specified as *y c y(-1),* either by *Quick -> Estimate Equation* and then type the equation or type *ls y c y(-1)* through the command field. But this option do not allow for the time-series part of Eviews. Instead one can use the following possibilities:

y c ar(1), y c ar(1) ar(2) … y c ar(1) ar(2) ar(p)
y c ma(1), y c ma(1) ma(2) … y c ma(1) ma(2) ma(p)
y c ar(1) ma(1), y c ar(1) ar(2) ma(1), y c ar(1) ar(1) ma(1) ma(2) … y c ar(1) ar(p) ma(1) ma(q)

Let's take the previous discussed data into account, and estimate the best *ARMA(p,q)* representation of *dividends (D_t).* To determine the best representation, it's important to regress our model upon a stationary process, and as proven we should use ΔD_t instead.

To estimate the equation $\Delta D_t = \delta + \rho \Delta D_{t-1} + \varepsilon_t$ equation type *ls d(div) c ar(1)* in the command field. The d in front of div is a Eviews command for taking the first difference. The second difference is easily done by adding another d, as in d(*d(div)).* The obtained statistical results are as follows:

Dependent Variable: D(DIV)
Method: Least Squares
Date: 02/03/13 Time: 15:50
Sample (adjusted): 1874 2000
Included observations: 127 after adjustments
Convergence achieved after 3 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.119850	0.106291	1.127564	0.2617
AR(1)	0.172805	0.088038	1.962841	0.0519

R-squared	0.029900	Mean dependent var	0.120637
Adjusted R-squared	0.022140	S.D. dependent var	1.001990
S.E. of regression	0.990836	Akaike info criterion	2.835088
Sum squared resid	122.7195	Schwarz criterion	2.879878
Log likelihood	-178.0281	Hannan-Quinn criter.	2.853285
F-statistic	3.852745	Durbin-Watson stat	1.923194
Prob(F-statistic)	0.051886		

Inverted AR Roots	.17

Figure 3

1. We fail to reject the null hypothesis of no first-order autocorrelation, H_0: $\rho = 0$, with a p-value of >5%. The point estimator is $\hat{\rho} = 0{,}172$ with std. Error 0.088.
2. The DW statistic is 1.92, which indicates that the model only has a little problem towards a positive autocorrelation in the residuals. This can be tested further by a Breush-Godfrey serial correlation LM test.
    a. In the interpretation window *click view/Residual Diagnostics/Serial Correlation LM test*
    b. *Enter 1 lag*, for testing H_0: $\rho = 0$ (no AR(1) in the error terms). The number of lag is as previous discussed a trade-off, but because of the fact that the data is annual, we are using one lag.
    c. The p-value is p=0,0174 indicating first-order serial correlation of order 1.
3. The statistical results in the figure below can be obtained by selecting *View/Representations.* This figure shows the estimation command and equation, as well as the regression function.

```
Estimation Command:
==========================
LS D(DIV) C AR(1)

Estimation Equation:
==========================
D(DIV) = C(1) + [AR(1)=C(2)]

Substituted Coefficients:
==========================
D(DIV) = 0.119849933931 + [AR(1)=0.172804701395]
```

4. The next step is to repeat the previous steps, and through trial and error finding the best representation of the dependent variable, in this case *dividends.*
5. As a help, or an indicator, one can use the correlogram Q-statistics as shown below. Furthermore some residual tests are appropriate through *view/Residual Diagnostics.*

Correlogram Q statistic

The figures 4 - 7 shows three statistics.

10.6 The AC (autocorrelation coefficient)
10.6 The PAC (partial autocorrelation coefficient)
10.6 A box-pierce Q-statistic with its probability

The lines in the graph of AC and PAC approximate two standard error bounds. Figure 4 represents an AR(1) model with Durbin-Watson statistic 1.923. The graph shows that at lagged *k=2,* the hypothesis of no autocorrelation is rejected. The Q-statistic is a test statistics for the joint hypothesis that all of the autocorrelation coefficients ρ_k up to certain lagged values are simultaneously equal to zero. The results in figure 4 show that $H_0: \rho_1 = \rho_k = 0$ is rejected up to and including 4 lags. If the mean equation is correctly specified, all Q-statistics should not be significant. However, there remains the practical problem of choosing the order of the lagged variables to be utilized for the test.

As you can see the model in figure 5, AR(2), has a Durbin-Watson statistic close to 2, which is recommended, and according to the correlogram the model seems correctly specified. The ARMA(2,1) model in figure 6 does have some problems similar to the discussed above. Figure 7, the ARMA(2,2) could also be correctly specified, but the model offers more lags than necessary.

Date: 02/04/13 Time: 13:24
Sample: 1874 2000
Included observations: 127
Q-statistic probabilities adjusted for 1 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.037	0.037	0.1803	
		2	-0.217	-0.219	6.3577	0.012
		3	-0.041	-0.024	6.5758	0.037
		4	-0.050	-0.100	6.9141	0.075

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.119850	0.106291	1.127564	0.2617
AR(1)	0.172805	0.088038	1.962841	0.0519

Figure 4: AR(1), DW: 1,923

Date: 02/04/13 Time: 13:25
Sample: 1875 2000
Included observations: 126
Q-statistic probabilities adjusted for 2 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.002	-0.002	0.0007	
		2	-0.023	-0.023	0.0711	
		3	-0.001	-0.001	0.0714	0.789
		4	-0.079	-0.080	0.9028	0.637

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.116040	0.085305	1.360290	0.1762
AR(1)	0.210159	0.087754	2.394870	0.0181
AR(2)	-0.225330	0.087757	-2.567667	0.0114

Figure 5: AR(2), DW: 2,004

Date: 02/04/13 Time: 13:27
Sample: 1875 2000
Included observations: 126
Q-statistic probabilities adjusted for 3 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.024	0.024	0.0720	
		2	-0.146	-0.146	2.8391	
		3	0.049	0.057	3.1507	
		4	0.024	-0.001	3.2250	0.073
		5	-0.021	-0.007	3.2853	0.193
		6	0.084	0.088	4.2374	0.237

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.116551	0.015518	7.510759	0.0000
AR(1)	1.104360	0.088662	12.45579	0.0000
AR(2)	-0.252085	0.088470	-2.849378	0.0051
MA(1)	-0.992776	0.013257	-74.88844	0.0000

Figure 6: ARMA(2,1), DW: 1,944

Date: 02/04/13 Time: 13:28
Sample: 1875 2000
Included observations: 126
Q-statistic probabilities adjusted for 4 ARMA term(s)

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.008	-0.008	0.0082	
		2	-0.016	-0.016	0.0403	
		3	-0.008	-0.008	0.0483	
		4	0.017	0.016	0.0856	
		5	-0.031	-0.031	0.2178	0.641
		6	0.059	0.059	0.6799	0.712

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.115353	0.020409	5.652145	0.0000
AR(1)	0.631729	0.258642	2.442481	0.0160
AR(2)	0.155542	0.239898	0.648365	0.5180
MA(1)	-0.473680	0.230911	-2.051354	0.0424
MA(2)	-0.507980	0.227479	-2.233083	0.0274

Figure 7: ARMA(2,2), DW: 2,008

9 Endogeneity

9.1 The basics

If one of our explanatory variables is correlated with our error term the assumption of $E(\varepsilon_i x_i) = 0$ does not hold. The problem of endogeneity causes our OLS estimator to be inconsistent and biased. Endogeneity can be present in models with measurement error, missing variables, and or simultaneity/reverse causality. To solve this problem we need another estimator which is consistent and unbiased even though we have endogeneity. The problem of endogeneity can be solved by using instrumental variables.

Instruments are variables that are uncorrelated with the error term, and correlated with the endogenous variable. Consider the model without endogeneity

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$$

With the moment conditions

$$E(\varepsilon_i x_{1i}) = 0$$
$$E(\varepsilon_i x_{2i}) = 0$$

Isolating the error term and inserting yields

$$E\{(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)x_{1i}\} = 0$$
$$E\{(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)x_{2i}\} = 0$$

We would be able to solve for β_1 and β_2 by replacing the expectation with the sample moments.

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)x_{1i} = 0$$
$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)x_{2i} = 0$$

But these moment conditions does not hold anymore because of endogeneity, namely $E(\varepsilon_i x_{1i}) = 0$ is violated. The solution is to find an instrument z_i (or vector of instruments), to replace the endogenous variable above. The error term and the instrument should be uncorrelated and the instrument and the instrumental variable should be correlated. If x_{1i} is the endogenous variable we have

$E(\varepsilon_i z_i) = 0$, validity criterion
$E(x_{1i} z_i) \neq 0$, relevance criterion

With the following sample moment conditions

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)z_i = 0$$
$$\frac{1}{N}\sum_{i=1}^{N}(y_i - x_{1i}\beta_1 - x_{2i}\beta_2)x_{2i} = 0$$

It should now be possible to solve for the consistent estimate of β_1 and β_2, and thus we have the IV-estimator instead of the OLS-estimator. Our IV-estimator can generally be written (in matrix notation) as

$$\beta_{IV} = (Z'X)^{-1}Z'y$$

Where Z is an NxK matrix with i'th row z_i'. It will only be possible to use the IV-estimator in the exactly identified case, where number of moment conditions (R) equals number of parameters to be estimated (K) $R = K$. If we have more valid and relevant instruments, it will be inefficient to throw some of them away and thereby only use a subset. To solve this problem we use the two-step least square (2SLS) estimator. The 2SLS estimator is an IV estimator. In a just-identified model it simplifies to the IV estimator given above. The 2SLS estimator is given by

$$\hat{\beta}_{2SLS} = [X'Z(Z'Z)^{-1}Z'X]^{-1}[X'Z(Z'Z)^{-1}Z'y]$$

The 2SLS estimator gets its name from the result that it can be obtained by two consecutive OLS regressions: OLS regression of x on z to get $\hat{x}$ followed by OLS of y on $\hat{x}$ which provides $\hat{\beta}_{2SLS}$.

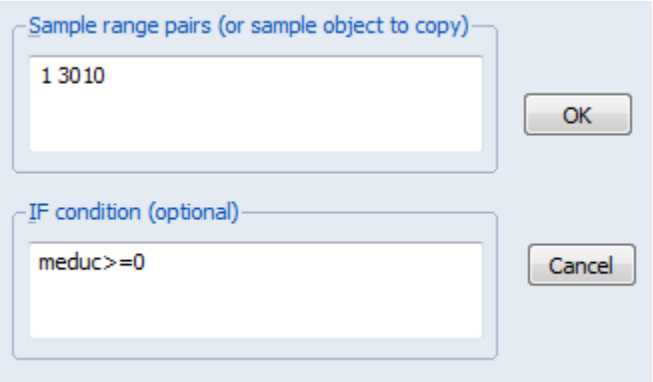
9.2 IV estimation using EViews

We use the dataset wage.wf1, that contains wage information for 3010 individuals, but we only consider the restricted data set, for which we have data for meduc (mother education) N=2657. We will be following the lines of section 5.4 in Verbeek, except for the reduction in sample size.

We want to examine what determines wage, and therefore we estimate the following regression equation:

$$lwage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 black_i + \beta_5 smsa_i + \beta_6 south_i$$

We start by restricting the sample size, by clicking Quick/Sample.



Estimating the wage equation by OLS (Quick/Estimate equation) (potentially inconsistent and biased), we get the following output

Dependent Variable: LWAGE
Method: Least Squares
Date: 12/27/12 Time: 14:18
Sample: 1 3010 IF MEDUC>=0
Included observations: 2657

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.550755	0.691967	5.131395	0.0000
EDUC	0.033423	0.002916	11.46107	0.0000
AGE	0.117315	0.048674	2.410244	0.0160
AGE^2	-0.001312	0.000846	-1.551060	0.1210
BLACK	-0.193553	0.019789	-9.780655	0.0000
SMSA	0.160955	0.016809	9.575475	0.0000
SOUTH	-0.128241	0.016250	-7.891824	0.0000

R-squared	0.277500	Mean dependent var	6.270610
Adjusted R-squared	0.275864	S.D. dependent var	0.444069
S.E. of regression	0.377886	Akaike info criterion	0.894183
Sum squared resid	378.4143	Schwarz criterion	0.909687
Log likelihood	-1180.922	Hannan-Quinn criter.	0.899794
F-statistic	169.6366	Durbin-Watson stat	1.847614
Prob(F-statistic)	0.000000		

We suspect that we have endogenous variable educ. To check the relevance of a potential instrument we make an auxiliary regression. By regressing *educ* on all the explanatory variables and the potential instrument *nearc4*, we can examine the relevance of the instrument *nearc4*. *nearc4* is a dummy variable equal to one if the individual lived close to a college, and is therefore assumed to be positively correlated with *educ* but uncorrelated with the error term, making it a good candidate for instrument.

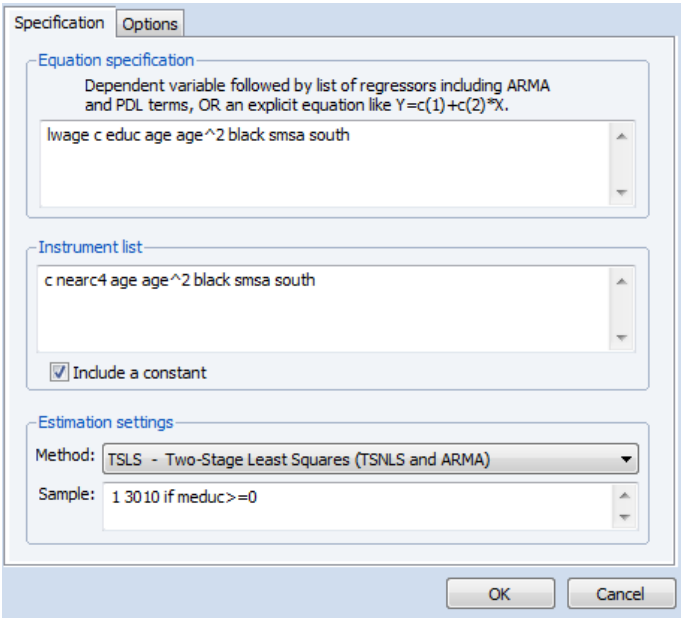
The auxiliary regression is performed by regressing *educ* with a constant and all explanatory variables, including the potential instrument.

Dependent Variable: EDUC
Method: Least Squares
Date: 12/27/12 Time: 14:02
Sample: 1 3010 IF MEDUC>=0
Included observations: 2657

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.764438	4.601598	-0.166124	0.8681
AGE	0.970301	0.323077	3.003313	0.0027
AGE^2	-0.016786	0.005617	-2.988684	0.0028
BLACK	-1.426201	0.128653	-11.08567	0.0000
SMSA	0.740329	0.116882	6.333963	0.0000
SOUTH	-0.403924	0.109241	-3.697550	0.0002
NEARC4	0.378655	0.114322	3.312172	0.0009

R-squared	0.101772	Mean dependent var	13.42868
Adjusted R-squared	0.099738	S.D. dependent var	2.647511
S.E. of regression	2.512015	Akaike info criterion	4.682678
Sum squared resid	16722.08	Schwarz criterion	4.698183
Log likelihood	-6213.938	Hannan-Quinn criter.	4.688290
F-statistic	50.04204	Durbin-Watson stat	1.774649
Prob(F-statistic)	0.000000		

We see that the coefficient on *nearc4* is significant, making it a relevant instrument. We'll now estimate the regression equation by 2SLS, using the exogenous explanatory variables as instruments for themselves and *nearc4* as instrument for *educ*. In *Quick/Estimate equation the following is entered:*



Obtaining the given results, with 2SLS:

Dependent Variable: LWAGE
Method: Two-Stage Least Squares
Date: 12/27/12 Time: 14:18
Sample: 1 3010 IF MEDUC>=0
Included observations: 2657
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.558878	0.700485	5.080588	0.0000
EDUC	0.056797	0.045965	1.235652	0.2167
AGE	0.095049	0.065848	1.443445	0.1490
AGE^2	-0.000928	0.001141	-0.813160	0.4162
BLACK	-0.160364	0.068141	-2.353432	0.0187
SMSA	0.140694	0.043247	3.253253	0.0012
SOUTH	-0.117359	0.026954	-4.354059	0.0000

R-squared	0.259985	Mean dependent var	6.270610
Adjusted R-squared	0.258309	S.D. dependent var	0.444069
S.E. of regression	0.382439	Sum squared resid	387.5880
F-statistic	144.5015	Durbin-Watson stat	1.851742
Prob(F-statistic)	0.000000	Second-Stage SSR	396.9483
J-statistic	0.000000	Instrument rank	7

Standard errors for 2SLS are bigger than for OLS. It could indicate a weak instruments problem. And our coefficient estimates differ in the OLS and the 2SLS case, which indicates a problem with endogeneity, since they both should be consistent without endogeneity.

Testing the exogeneity (making it a valid instrument) of the instrument *educ* by *clicking View/IV Diagnostic & Tests/Regressor Endogeneity Test*, type the endogenous variable *educ* and click *OK*.

Endogeneity Test
Equation: TWO_SLS
Specification: LWAGE C EDUC AGE AGE^2 BLACK SMSA SOUTH
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH
Endogenous variables to treat as exogenous: EDUC

	Value	df	Probability
Difference in J-stats	0.266638	1	0.6056

J-statistic summary:

	Value
Restricted J-statistic	0.266638
Unrestricted J-statistic	0.000000

Where H_0 claims that there are no differences between the model in which *educ* is treated as endogenous and the model where it is treated as exogenous. We cannot reject H_0 and therefore *educ* can be considered exogenous.

We have options in checking the validity of instruments if we have an over identified model. To construct an over identified model we'll use meduc as additional instrument for educ. Estimation results:

Dependent Variable: LWAGE
Method: Two-Stage Least Squares
Date: 12/27/12 Time: 15:05
Sample: 1 3010 IF MEDUC>=0
Included observations: 2657
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH
MEDUC

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.555910	0.695341	5.113905	0.0000
EDUC	0.048258	0.007523	6.414653	0.0000
AGE	0.103183	0.049354	2.090664	0.0367
AGE^2	-0.001068	0.000858	-1.245457	0.2131
BLACK	-0.172488	0.022186	-7.774518	0.0000
SMSA	0.148095	0.017927	8.261038	0.0000
SOUTH	-0.121334	0.016645	-7.289711	0.0000

R-squared	0.270444	Mean dependent var	6.270610
Adjusted R-squared	0.268792	S.D. dependent var	0.444069
S.E. of regression	0.379727	Sum squared resid	382.1098
F-statistic	153.1730	Durbin-Watson stat	1.850689
Prob(F-statistic)	0.000000	Second-Stage SSR	391.2385
J-statistic	0.035981	Instrument rank	8
Prob(J-statistic)	0.849555		

Again we start by testing for exogeneity, by clicking *View/IV Diagnostic & Tests/Regressor Endogeneity Test*

Endogeneity Test
Equation: TWO_SLS
Specification: LWAGE C EDUC AGE AGE^2 BLACK SMSA SOUTH
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH
MEDUC
Endogenous variables to treat as exogenous: EDUC

	Value	df	Probability
Difference in J-stats	3.849734	1	0.0498

J-statistic summary:

	Value
Restricted J-statistic	3.885856
Unrestricted J-statistic	0.036122

There is no clear answer to this test. We cannot say by any certainty that our *educ* can be treated as exogenous.

In our over identified model we will now check for validity of instruments using the C-orthogonality test, found under **IV Diagnostics & Tests**, and we need to test the instrument one by one. Starting with *meduc* we get:

Instrument Orthogonality C-test Test
Equation: TWO_SLS
Specification: LWAGE C EDUC AGE AGE^2 BLACK SMSA SOUTH
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH MEDUC
Test instruments: MEDUC

	Value	df	Probability
Difference in J-stats	0.035981	1	0.8496

J-statistic summary:

	Value
Restricted J-statistic	0.035981
Unrestricted J-statistic	0.000000

Instrument Orthogonality C-test Test
Equation: TWO_SLS
Specification: LWAGE C EDUC AGE AGE^2 BLACK SMSA SOUTH
Instrument specification: C NEARC4 AGE AGE^2 BLACK SMSA SOUTH MEDUC
Test instruments: NEARC4

	Value	df	Probability
Difference in J-stats	0.035981	1	0.8496

J-statistic summary:

	Value
Restricted J-statistic	0.035981
Unrestricted J-statistic	1.21E-32

We get the same J-statistic in both tests; this is a general result when we only have two instruments and one endogenous variable. Both tests strongly indicate validity of our instruments.

Another way of testing the validity of the instruments is the LM-test. To calculate the LM-statistic we need to run an auxiliary regression with the residuals from the 2SLS estimation as dependent variable. To save the residuals from our 2SLS estimation we press *Proc/Make residual series*, type a name for the residuals and click ok. We regress the saved residuals with regards to all explanatory variables and instruments (excluding the endogenous variable). This provides the following output:

Dependent Variable: E_IV
Method: Least Squares
Date: 12/27/12 Time: 15:28
Sample: 1 3010 IF MEDUC>=0
Included observations: 2657

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.003598	0.695869	-0.005170	0.9959
AGE	0.000157	0.048844	0.003207	0.9974
AGE^2	-3.02E-06	0.000849	-0.003556	0.9972
BLACK	-7.08E-05	0.020201	-0.003507	0.9972
SMSA	-0.001103	0.017750	-0.062115	0.9505
SOUTH	0.000535	0.016601	0.032217	0.9743
NEARC4	0.003316	0.017284	0.191853	0.8479
MEDUC	-7.10E-06	0.002481	-0.002863	0.9977

R-squared	0.000014	Mean dependent var	-2.64E-15
Adjusted R-squared	-0.002629	S.D. dependent var	0.379244
S.E. of regression	0.379742	Akaike info criterion	0.904359
Sum squared resid	381.9973	Schwarz criterion	0.922079
Log likelihood	-1193.442	Hannan-Quinn criter.	0.910772
F-statistic	0.005258	Durbin-Watson stat	1.850615
Prob(F-statistic)	1.000000		

To calculate the LM-statistic we multiply the number of observations with the R^2

$$n \cdot R^2 = 2657 \cdot 0{,}000014 = 0{,}037$$

P-value

$$P(\chi_1^2 > 0{,}037) = 0{,}8475$$

We fail to reject H_0 of there being valid instruments.

10VAR (Vector Autoregressive models)

10.1 The basics

In Vector Autoregressive models we have generalized the univariate autoregressive model to the multivariate case. This gives us some beneficial features like:

- Take into account the data generating process for all included variables
- Estimate and model the dynamic interrelation between variables
- Perform cointegration test without the shortcommings of the Engle-Granger cointegration procedure
- Be indifferent about the choice of dependent variable

Our VAR model is defined as

$$Y_t = \delta + \Theta Y_{t-1} + \epsilon_t$$

Where $\epsilon_t \sim iid(0, \Sigma)$, In this Σ represents the covariance matrix, Θ the coefficient matrix, Y_t vector of our variables, δ is the vector containing the intercepts, ϵ_t is a vector containing all our error terms.

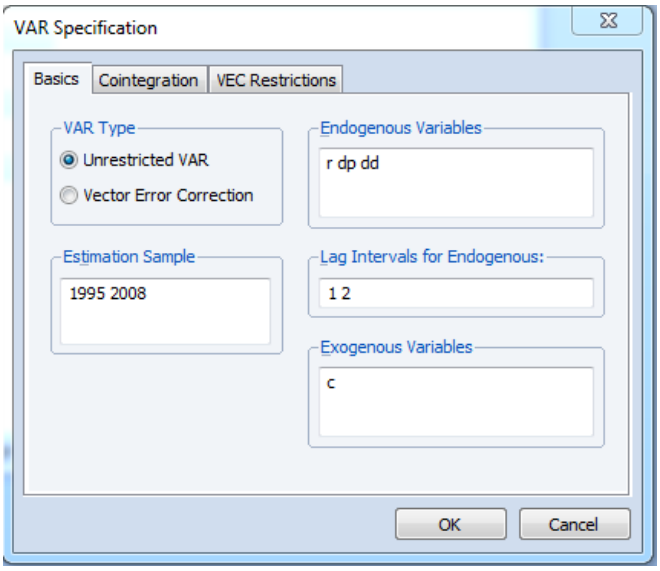
10.2 Estimating a model

The dataset we will use here is uk1.wf1. It is suited for exploration of the dynamic relationship between stock returns, dividend growth and the dividend-price ratio.

Variable	Description
dp	Log dividend-price ratio
dd	Log dividend growth
r	Log stock return

Before making the VAR model we should check if our variables are stationary, and thereby not having a stochastic or deterministic trend, so we avoid spurious regression.

Open up the workfile. Mark all the variables you want to include in your VAR model. Right click and press Open/as VAR. The following dialogue box will appear:



It will be possible to estimate both an Unrestricted VAR and a Vector Error Correction model. In this model we choose to include two lags of our dependent variable. We could also include more lag intervals. If we instead had written "1 2 5 6" in lag intervals, we would include lag 1-2 and 5-6.

The determination of lag length is a trade-off between the curse of dimensionality[9] and reduced models, which are not appropriate to indicate the dynamic adjustment.

If the lag length is to short, autocorrelation of the error terms could lead to apparently significant and inefficient estimators. Therefore, one would receive wrong results.

The results we get by estimating the model is (with lag interval 1 1, and sample set to 1901 to 2000):

[9] Even with inclusion of a small lag length interval we will have to estimate many parameters. Increasing number of parameters causes the degrees of freedom to decrease.

Vector Autoregression Estimates
Date: 02/02/13 Time: 13:48
Sample: 1901 2000
Included observations: 100
Standard errors in ( ) & t-statistics in [ ]

	R	DP	DD
R(-1)	0.139648	-0.130547	0.019183
	(0.10256)	(0.12675)	(0.08849)
	[1.36156]	[-1.02996]	[0.21677]
DP(-1)	0.354888	0.504163	-0.159394
	(0.08531)	(0.10543)	(0.07361)
	[4.15981]	[4.78198]	[-2.16543]
DD(-1)	0.043930	-0.147483	-0.097302
	(0.12158)	(0.15025)	(0.10490)
	[0.36132]	[-0.98159]	[-0.92757]
C	1.143143	-1.537161	-0.496843
	(0.26200)	(0.32378)	(0.22606)
	[4.36308]	[-4.74751]	[-2.19786]
R-squared	0.183411	0.267148	0.090060
Adj. R-squared	0.157893	0.244246	0.061624
Sum sq. resids	3.080476	4.704444	2.293195
S.E. equation	0.179132	0.221370	0.154556
F-statistic	7.187400	11.66500	3.167140
Log likelihood	32.11045	10.93927	46.86736
Akaike AIC	-0.562209	-0.138785	-0.857347
Schwarz SC	-0.458002	-0.034579	-0.753140
Mean dependent	0.053141	-3.103365	-0.002613
S.D. dependent	0.195204	0.254641	0.159550

Determinant resid covariance (dof adj.)	5.77E-09
Determinant resid covariance	5.11E-09
Log likelihood	528.9670
Akaike information criterion	-10.33934
Schwarz criterion	-10.02672

The coefficient estimate is presented without brackets, the number in brackets is the p-value from the significance test of the parameter estimate, and the number in square brackets is the critical value.

In this VAR model the data generating process for R is estimated to

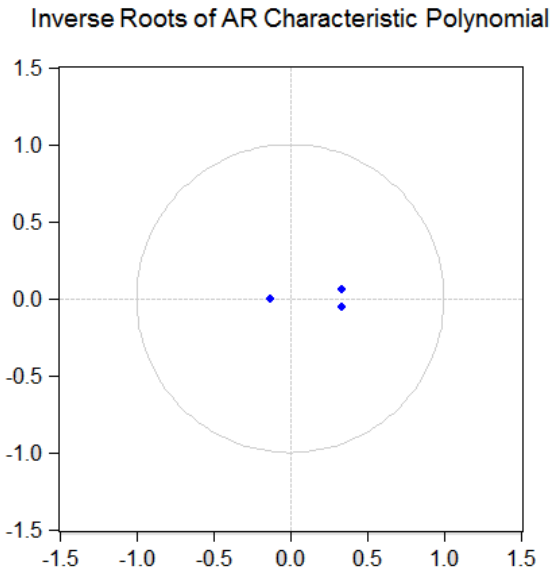
$$R = 0{,}143 + 0{,}140 \cdot R_{t-1} + 0{,}355 \cdot DP_{t-1} + 0{,}044 \cdot DD_{t-1}$$

Note: none of our variables are significantly different from zero at a 5% significance level in the above example.

In the bottom of the estimation output we have the Akaike and Schwartz information criterion. One way to find "the right" specification and the right number of lags in our VAR model, is to minimize the information criteria.

10.3 Stationary

To check if our VAR(1) model is stationary, we choose the option *View/Lag structure/AR root graph*



All inverse roots smaller than 1 indicates that our VAR model is stationary.

10.4 Granger causality

To test if one of our variables Granger-causes one of the other variables, choose *View/Lag structure/Granger causality test.*

VAR Granger Causality/Block Exogeneity Wald Tests
Date: 02/02/13 Time: 15:38
Sample: 1901 2000
Included observations: 100

Dependent variable: R

Excluded	Chi-sq	df	Prob.
DP	17.30404	1	0.0000
DD	0.130555	1	0.7179
All	21.42719	2	0.0000

Dependent variable: DP

Excluded	Chi-sq	df	Prob.
R	1.060825	1	0.3030
DD	0.963521	1	0.3263
All	2.538495	2	0.2810

Dependent variable: DD

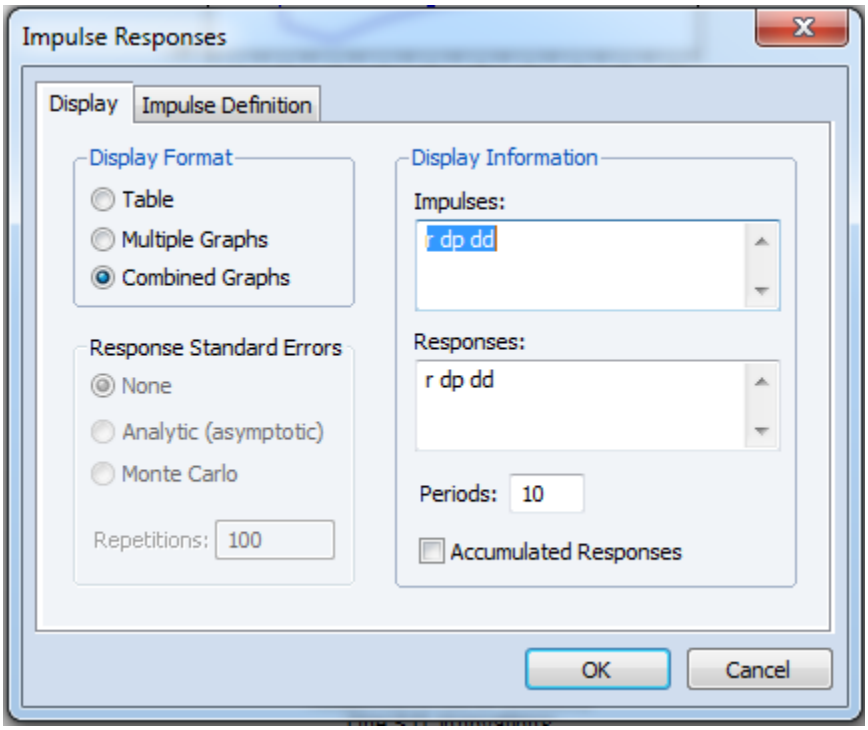
Excluded	Chi-sq	df	Prob.
R	0.046989	1	0.8284
DP	4.689078	1	0.0304
All	6.308589	2	0.0427

Where our H_0 is that we do not have Granger-causaility. In our sample we get that DP causes R, and DP causes DD

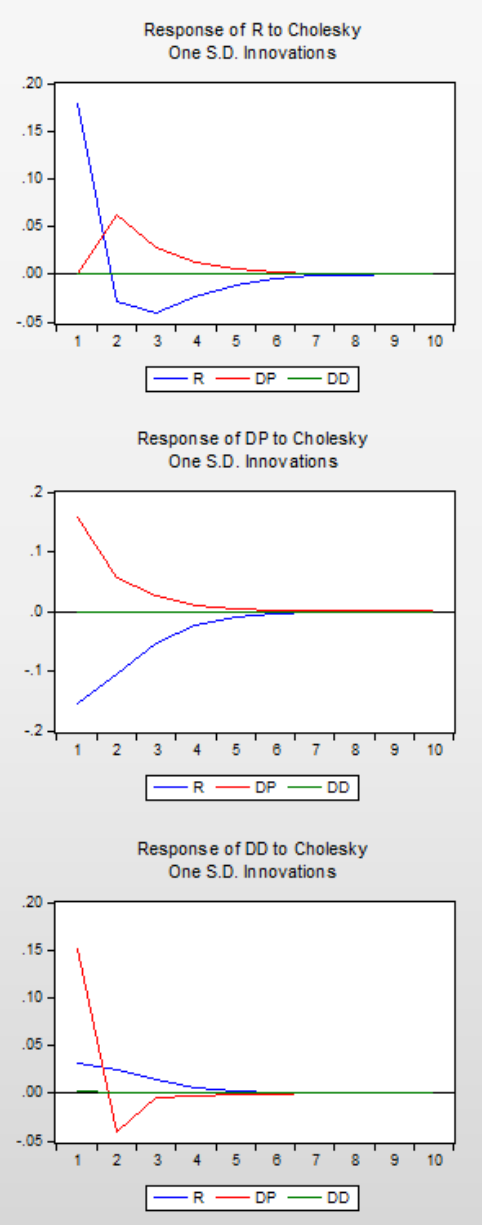
10.5 Impulse/response functions

A shock to the *i*-th variable not only directly affects the *i*-th variable but is also transmitted to all of the other endogenous variables through the dynamic (lag) structure of the VAR. An impulse response function traces the effect of a one-time shock to one of the innovations on current and future values of the endogenous variables.

We obtain the impulse response function plot by selecting *View/Impulse response*



It is possible to specify which variables we want to give an impulse, and which variables we want to see the response from. In our model, we will get a combined 3x3 graph with the above selection.



The first graph shows how the variable *R* react to a shock to the variables *R, DD, DP*. The response of a shock to *R* is a big positive change in *R*, the following effect the first period the response is negative. For a shock to *DP* the response of *R* is positive, and the effect dies out over time, meaning that a positive shock to log dividend-price ratio will cause a positive shock to the log stock return.

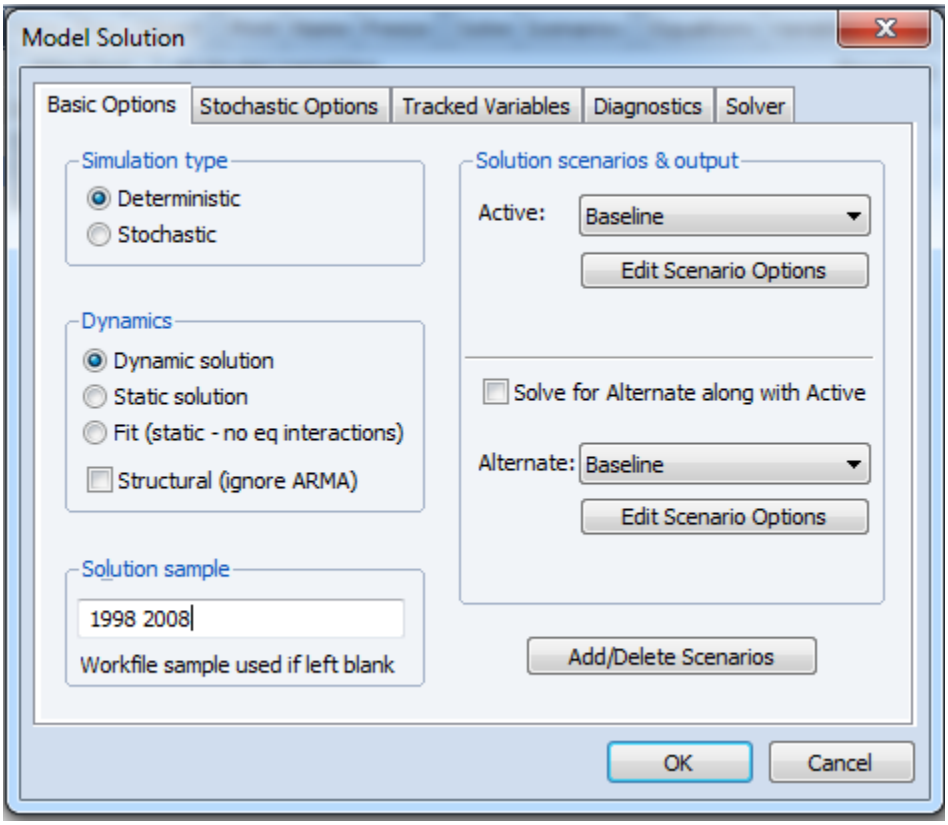
10.6 Forecasting

There are different forms of forecasting. In-sample forecasting uses a sample within another sample which means, that we only use a subset of the data. Then we use the rest of the data to compare our forecasts to the real data. This method is used if we want to evaluate our forecast.

Out-of-sample forecasting uses all the data from the sample, and forecasts in periods ahead of the sample, this is considered "real forecasting".

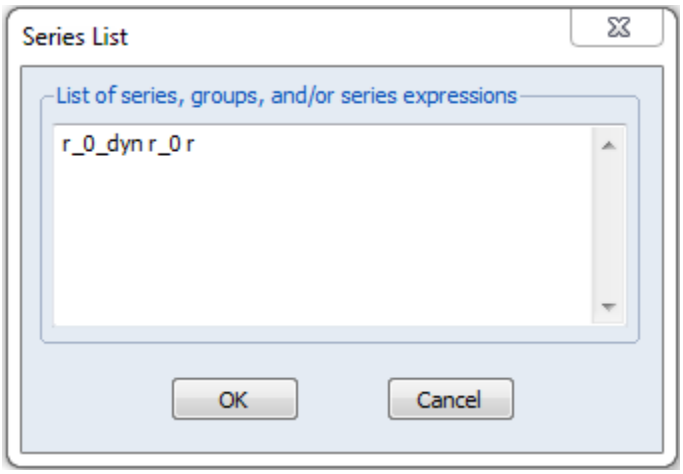
We would like to evaluate our VAR-models ability to forecast, therefore we use in-sample forecasting. Choose *proc/make model/solve*

It is possible to use two different types of simulation dynamics, dynamic solution and static solution. Static dynamics forecasts one period and updates the information in the forecast, and thereby uses a 'rolling window' of data. Dynamic solution constructs an *h*-period forecast based on data in the sample, and thus does not update the information after each forecast.

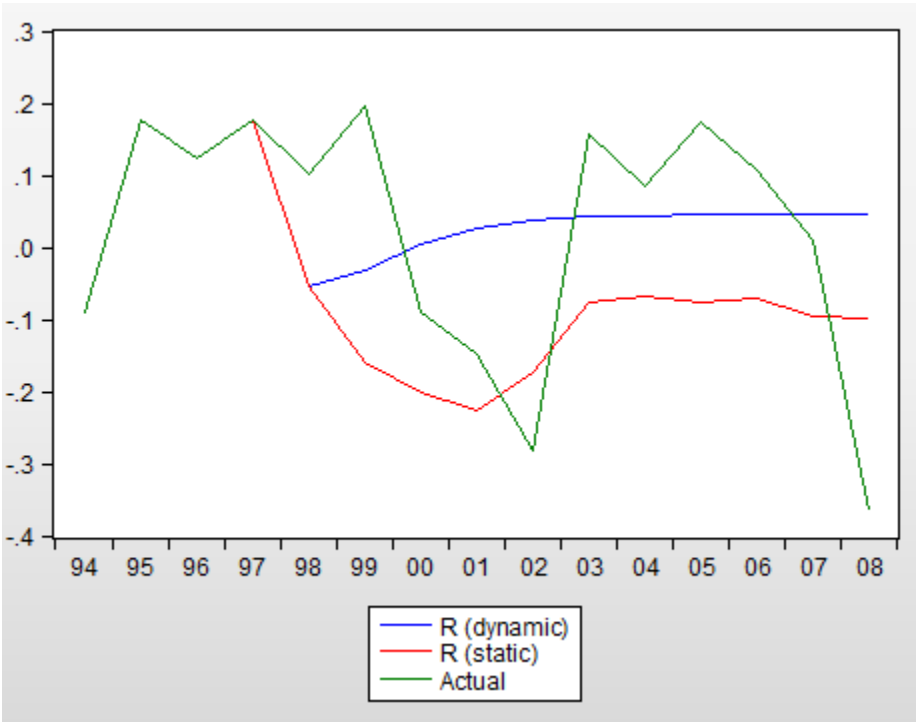


Solution sample should be set to the time period we want to forecast. For this example we use "1998 2008". By clicking OK, we will get 3 new variables named after the original variable, but now we an added "_0" after. It is possible to rename the new variables by double clicking on them, and choosing *Name* in the menu, by doing that we secure that these forecasts are not being overwritten by new forecasts.

It is therefore possible to make two new variables for every variable that we want to forecast. One variable contains the static forecasts and one containing the dynamic forecasts. Graphing these forecasts is done by clicking *Quick/Graph*



Entering the variables we want to graph in the above window, where the dynamic forecast for *r* has been renamed to *r_0_dyn*. Clicking OK two times yields



The static and dynamic forecasts get very different results as shown above. The dynamic forecast converges to the conditional mean in the long run.

10.7 Lag Length

The model specified should have the "right" number of lags included. Too many included lags will cause that we lose a lot of degrees of freedom. Too few included lags will cause our model to be imprecise.

The determination of lag length is a trade-off between the curse of dimensionality[10] and reduced models, which are not appropriate to indicate the dynamic adjustment.

If the lag length is too short, autocorrelation of the error terms could lead to apparently significant and inefficient estimators. Therefore, one would receive wrong results.

One way to examine if the lag length is correctly specified: *View/Lag structure/Lag length Criteria*

The appropriate amount of lags to be included could be 5 or 10, including more lags causes the loss of more data, and will change the information criteria. We would like to minimize the information criteria. The information criteria function are functions of the log-likelihood function (the more negative the better) and some positive addition that tries to control for the inclusion of more variables. Thereby the information criteria seeks to handle the tradeoff between a parsimonious model and a comprehensive model.

Var: VAR01 Workfile: UK::Uk\

View | Proc | Object | Print | Name | Freeze | Estimate | Stats | Impulse | Resids

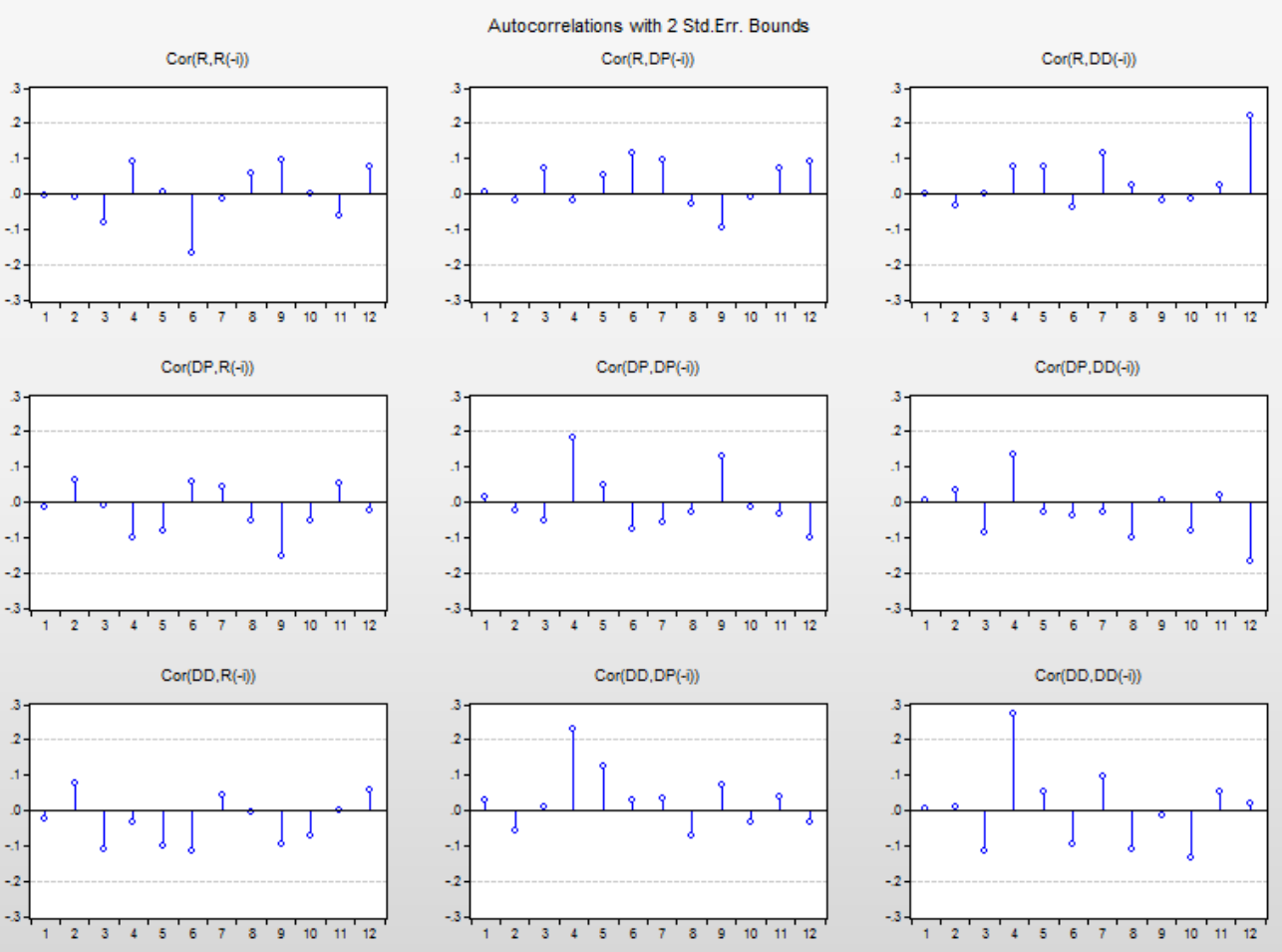
VAR Lag Order Selection Criteria
Endogenous variables: R DP DD
Exogenous variables: C
Date: 02/04/13 Time: 15:32
Sample: 1901 2000
Included observations: 95

Lag	LogL	LR	FPE	AIC	SC	HQ
0	67.36115	NA	5.18e-05	-1.354972	-1.274323	-1.322383
1	499.7480	828.3622	6.97e-09	-10.26838	-9.945784*	-10.13803*
2	508.8579	16.87722	6.96e-09	-10.27069	-9.706151	-10.04258
3	522.7577	24.87332*	6.28e-09*	-10.37385*	-9.567358	-10.04796
4	531.6505	15.35182	6.32e-09	-10.37159	-9.323156	-9.947943
5	534.4838	4.712204	7.23e-09	-10.24176	-8.951384	-9.720353
6	539.8208	8.539304	7.86e-09	-10.16465	-8.632323	-9.545475

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

The SC and HQ information criterion chooses 1 lag, and LR, FPE and AIC chooses to include 3 lags. If we only include one lag, we can check the correlogram for autocorrelation.

[10] Even with inclusion of a small lag length interval we will have to estimate many parameters. Increasing number of parameters causes the degrees of freedom to decrease.



We test autocorrelation with 5 % confidence interval. Our goal is to eliminate all autocorrelation. We only have three autocorrelation values outside of our confidence intervals and therefore we have sufficiently dealt with the problem of autocorrelation.

10.8 Johanson Cointegration test

If we have a VAR(p) model which can be rewritten as

$$Y_t = \delta + \Theta Y_{t-1} + \epsilon_t$$

Then we can further rewrite it as

$$\Delta Y_t = \Pi(Y_{t-1} - \mu) + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + \epsilon_t$$

Where $\Pi = \sum_{i=1}^{p} \Theta_i - I$ and $\Gamma_i = -\sum_{j=i+1}^{p} \Theta_j$

In this VAR framework we are able to test for cointegration in a way that does not have the shortcomings of the Engle-Granger approach. In the Johansen cointegration test the result does not depend on which variable we normalize with regards to, and it is possible for us include to more cointegration relationships.

In the Johansen cointegration test we exploit that the number of non-zero eigenvalues is at most the rank of the matrix Π, meaning that we can interpret the number of significant eigenvalues as the number of cointegration relations.

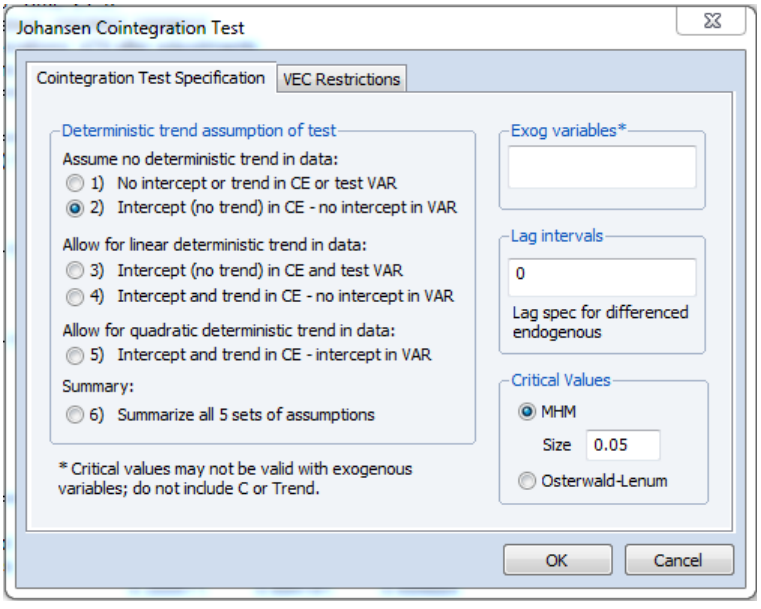
For the likelihood ratio test we can exploit that the maximum log-likelihood can be expressed as a function of the eigenvalues of Π.

For this example we use a new data set with 3 non-stationary I(1) variables.

Variable	**Description**
r3m	Annualized interest rate on 3-month treasury bills
r1y	Annualized interest rate on 1-year treasury bills
r10y	Annualized interest rate on 10-year treasury bills

The dataset is called "tbills.wf1". We would like to find a linear relationship between our variables that is I(0) and thus stationary. Finding a linear relationship that is I(0) can solve the problem of spurious regression.

We should carry out the Johansen test with the correct number of lags to eliminate serial correlation. As showed earlier we could check the correlogram of the functions, to see if we have eliminated the serial-correlation. To carry out the test for cointegration we for instance first estimate our VAR(p) (in this example p=1) model in the usual way, by marking the variables, right clicking and selecting open/as VAR. In this example we choose the lag interval to be "1 1", by clicking **OK** we get our VAR(1) model, and the vector autoregressive estimates. Before making the Johansen test we should always make a test to see if our variables are I(1) (see chapter about unit root testing) The Johansen cointegration test is carried out by clicking *View/cointegration test.* In the vector auto-regression estimates window, and the following window will appear:



In this test window we have the option of many different assumptions about the deterministic trend in the model.

The used series may have nonzero means and deterministic trends as well as stochastic trends. Similarly, the cointegrating equations may have intercepts and deterministic trends. The asymptotic distribution of the LR test statistic for cointegration does not have the usual χ^2 distribution and depends on the assumptions made with respect to deterministic trends. Therefore, in order to carry out the test, it is needed to make an assumption regarding the trend underlying the data.[11]

Summary of the 5 options in the test:

1. The level data y_t have no deterministic trends and the cointegrating equations do not have intercepts
2. The level data y_t have no deterministic trends and the cointegrating equations have intercepts
3. The level data y_t have linear trends but the cointegrating equations have only intercepts
4. The level data y_t and the cointegrating equations have linear trends
5. The level data y_t have quadratic trends and the cointegrating equations have linear trends:

Note: including extra exogenous variables in the exog variables will lead to wrong critical values.

The lag interval specified (for example 1 1) will include one differenced lag Δy_{t-1}, 1 2 will include Δy_{t-1} and Δy_{t-2}.

By selecting 0 in lag and option 2, one should get the following output by clicking *OK*:

Date: 02/05/13 Time: 22:24
Sample (adjusted): 1960M02 1999M12
Included observations: 479 after adjustments
Trend assumption: No deterministic trend (restricted constant)
Series: R10Y R1Y R3M
Lags interval (in first differences): No lags

Unrestricted Cointegration Rank Test (Trace)

Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.**
None *	0.106013	80.81795	35.19275	0.0000
At most 1 *	0.049109	27.13933	20.26184	0.0048
At most 2	0.006282	3.018789	9.164546	0.5772

Trace test indicates 2 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

Unrestricted Cointegration Rank Test (Maximum Eigenvalue)

Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.**
None *	0.106013	53.67861	22.29962	0.0000
At most 1 *	0.049109	24.12054	15.89210	0.0020
At most 2	0.006282	3.018789	9.164546	0.5772

Max-eigenvalue test indicates 2 cointegrating eqn(s) at the 0.05 level
* denotes rejection of the hypothesis at the 0.05 level
**MacKinnon-Haug-Michelis (1999) p-values

And some more output. So we have the two Johansen test, Trace and maximum Eigenvalue test, and these yield the same results in this example. One can see the sequential test procedure used in the Johansen test. First we are testing the null hypothesis of zero cointegration relationships against the alternative hypothesis of 1 or more cointegration relationships. In the first step we can see that the null hypothesis of none cointegration relationsships is rejected -> next we test H_0: at most 1, against H_1: at most 2 and so on, until we cannot reject H_0 anymore. In the example given above the test finds 2 cointegration relationships in both tests.

[11] Eviews manual – Johansen test

We could also make a plot of the series to get an idea of whether they share a stochastic or deterministic trend, Click on *open as group/view/graph.*

10.9 Vector Error Correction Model (VECM)

In this section we use the same dataset as the previous section.

Variable	Description
r3m	Annualized interest rate on 3-month treasury bills
r1y	Annualized interest rate on 1-year treasury bills
r10y	Annualized interest rate on 10-year treasury bills

The general vector error correction model with deterministic trend is

$$\Delta Y_t = \phi + \Pi Y_{t-1} + \alpha t + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + \epsilon_t$$

This can be rewritten into our test equation:

$$\Delta Y_t = \phi_1 + \alpha_1 t + \gamma(\beta' Y_{t-1} - \phi_2 - \alpha_2 t) + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i} + \epsilon_t$$

Where $\phi = \phi_1 - \gamma\phi_2$ and $\alpha = \alpha_1 - \gamma\alpha_2$

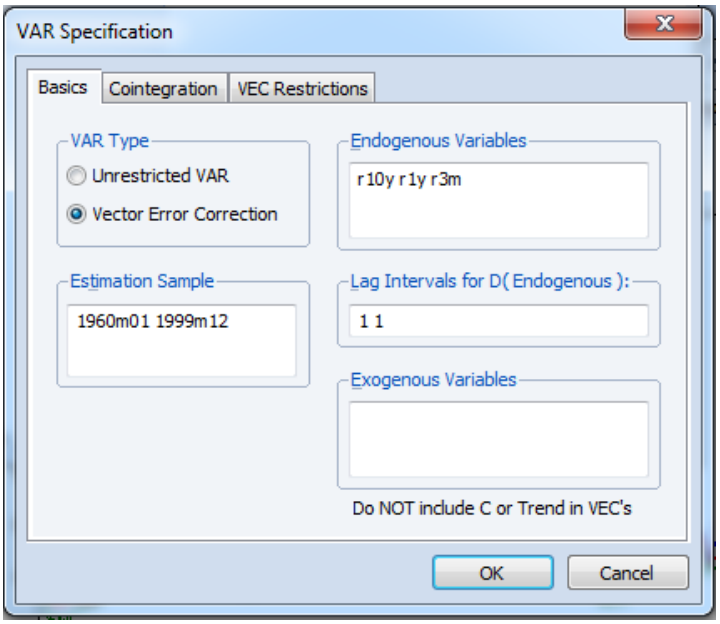
The intuition of this expression is that a change in Y_t can come from the time trend, or the error correction part of the expression (the error correction part is the only in parenthesis). The last part of the expression with a summation from $i = 1$ up to $p - 1$ of lagged values of the differenced dependent variable is used to eliminate serial correlation.

Since our dependent variable is differenced in the test equation, we note that:

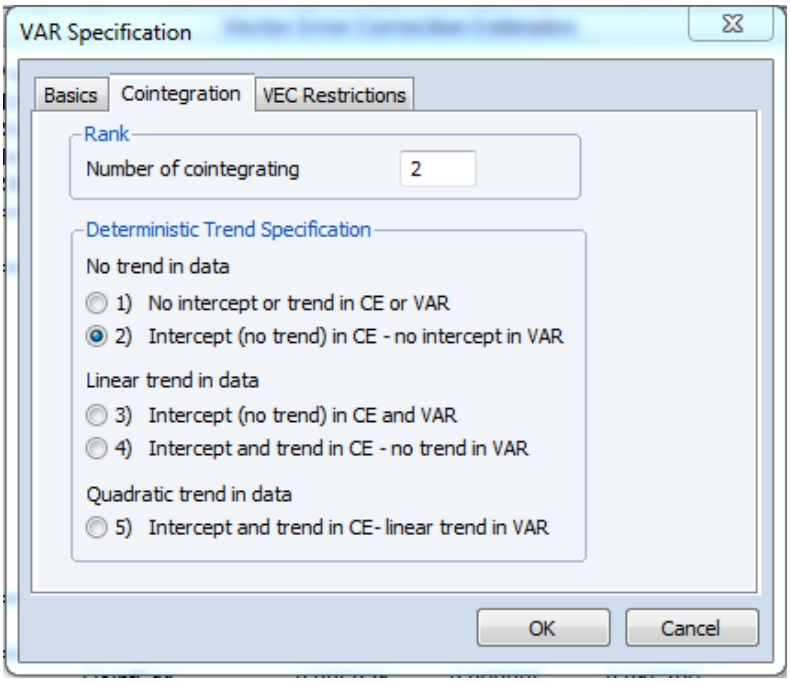
$\phi_1 \neq 0$: Deterministic trends in Y_t
$\alpha_1 \neq 0$: Quadratic trends in Y_t
$\phi_2 \neq 0$: The linear combinations $\beta' Y_{t-1}$ have a non-zero equilibrium value
$\alpha_2 \neq 0$: The linear combinations $\beta' Y_{t-1}$ are trending deterministically

10.10 Estimate the VECM (vector error correction model)

To estimate the VECM, one marks all the variables that should be included in the VECM, right click on one of the marked variables and chooses *Open /as VAR*..... The following window should appear



The lag interval in the window above (1 1) corresponds to $p = 2$, this can be seen by looking at the sum operator in our test equation expression: $\sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-i}$. So if we wanted to test this for p=1, we should set lag interval equal to 0. In the tab *cointegration* we should specify the correct number of cointegration relationships. We found the number of cointegratio relationships to be two in the Johansen cointegration test in the previous section. By clicking the tab *Cointegration* we get to include this information in the test and determine which trend should be included.



An explanation for all 5 options can be found under the section *Johansen cointegration test*. If we want to estimate the model using p=1 (lag interval = 0) and r=2 (meaning two cointegration intercepts) and option 2) in the cointegration tab, we should get the following output:

Vector Error Correction Estimates
Date: 02/08/13 Time: 15:45
Sample (adjusted): 1960M02 1999M12
Included observations: 479 after adjustments
Standard errors in ( ) & t-statistics in [ ]

Cointegrating Eq:	CointEq1	CointEq2
R10Y(-1)	1.000000	0.000000
R1Y(-1)	0.000000	1.000000
R3M(-1)	-1.108270	-1.102046
	(0.09529)	(0.02434)
	[-11.6304]	[-45.2863]
C	-0.811882	-0.040813
	(0.62414)	(0.15939)
	[-1.30080]	[-0.25606]

1.

Error Correction:	D(R10Y)	D(R1Y)	D(R3M)
CointEq1	0.006031	0.028885	-0.015792
	(0.01565)	(0.02573)	(0.02602)
	[0.38547]	[1.12272]	[-0.60693]
CointEq2	-0.116023	-0.057328	0.167502
	(0.04667)	(0.07675)	(0.07762)
	[-2.48579]	[-0.74697]	[2.15796]

2.

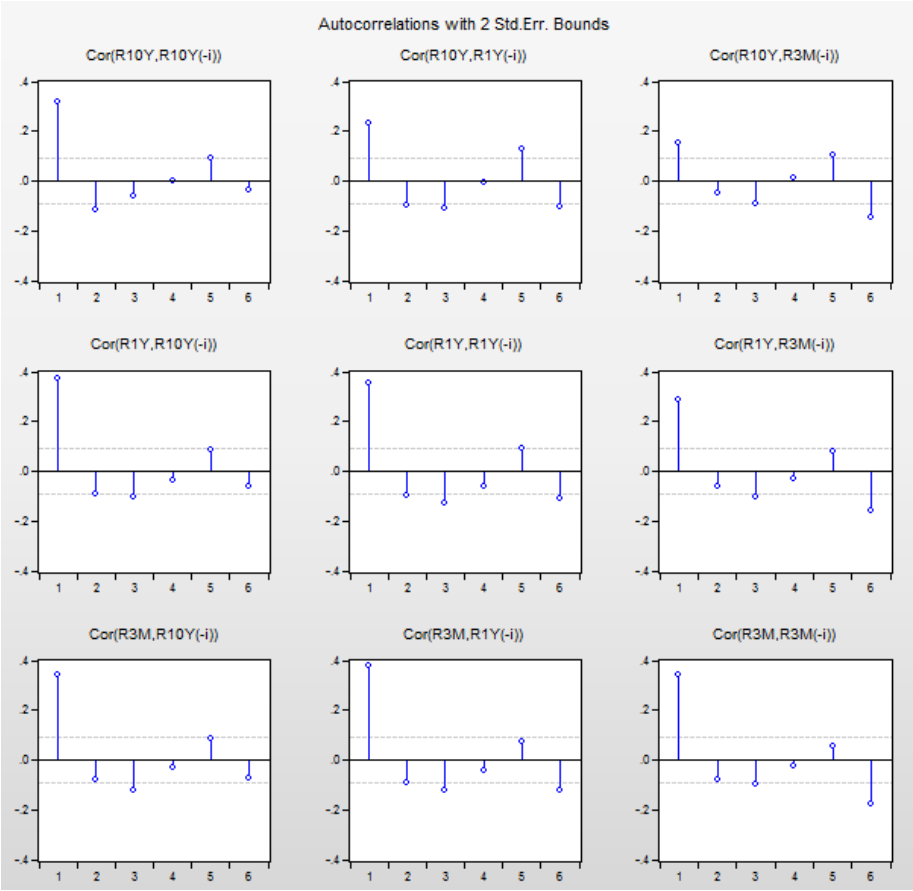
R-squared	0.023238	0.002675	0.014686
Adj. R-squared	0.021190	0.000584	0.012620
Sum sq. resids	43.15365	116.6790	119.3481
S.E. equation	0.300780	0.494581	0.500206
F-statistic	11.34831	1.279396	7.109626
Log likelihood	-103.2110	-341.4319	-346.8490
Akaike AIC	0.439294	1.433954	1.456572
Schwarz SC	0.456713	1.451372	1.473990
Mean dependent	0.003257	0.001691	0.001775
S.D. dependent	0.304019	0.494725	0.503392

Determinant resid covariance (dof adj.)	0.000209	
Determinant resid covariance	0.000206	
Log likelihood	-6.204266	
Akaike information criterion	0.084360	3.
Schwarz criterion	0.206289	

1. Represents the long run equilibrium relations. The first cointegration equation, CointEq1, is estimated as $R10Y_{t-1} - 0{,}81188 - 1{,}1083 \cdot R3M_{t-1} = 0$ which can be rewritten as: $R10Y_{t-1} = 0{,}81188 + 1{,}1083 \cdot R3M_{t-1}$ The second cointegration equation, CointEq2, is estimated as $R1Y_{t-1} = 0{,}04081 + 1{,}1021 \cdot R3M_{t-1}$.

2. The error correction part represents the short run relations. CointEq1 relates R10Y with R3M. What we can see is that if the value of R10Y lies above its long run equilibrium, then R3M will increase next period. CointEq2 relates R1Y with R3M. Here the short run effect is that if 1Y interest rate lies above its long run equilibrium, the R3M interest rate will fall next period.
3. We notice that the information criteria which we want to minimize is reported as well.

Again we could test for serial-correlation, to see if we have included the right amount of lags (or if we should increase p in our test equation). This is done as before by choosing *View/Residual test/Correlogram*, if we do that it will be clear to us that we should use at least one more lag in the test ($p \geq 2$).



Clearly we still have a lot of serial correlation, especially in the first period.

We will not show it here, but one should repeat the process of estimating the VECM model, but now with $p \geq 2$.

11 ARCH and GARCH Models

11.1 The basics

In financial time series one often observes what is referred to as volatility clustering. In this case big shocks (residuals) tend to be followed by big shocks in the opposite direction, and small shocks tend to follow small shocks. For example, stock markets are typically characterized by periods of high volatility and more "relaxed" periods of low volatility. This is particularly true at high frequencies. One way to model such patterns is to allow the variance of ε_t to depend upon its history.

The basic idea of ARCH and GARCH model is to test whether the conditional variance depend on time (variance of the error). In other words; is the variance at a given time point different from the past variances?

$$\sigma_t{}^2 \equiv E[\varepsilon_t^2 | I_{t-1}] = \varpi + \alpha \varepsilon_{t-1}^2$$

Where I_{t-1} denotes the information set, typically including ε_{t-1} and its entire history. This sepcification is called an ARCH(1) process. If the conditional variance depends on time, then we say that there are ARCH/GARCH effects.

This can be tested with a Breusch-Pagan test for heteroskedasticity (chapter 4 Verbeek 4'th edition). Here an auxiliary regression is run on the squared OLS residuals ε_t^2 upon lagged squares $\varepsilon_{t-1}^2 \ldots \ldots \varepsilon_{t-p}^2$ and a constant and compute T times the R^2. Under the null hypothesis of homoskedasticity $(\alpha_1 = \cdots = \alpha_p = 0)$ the test statistic is asymptotically Chi-squared distributed with p degrees of freedom.

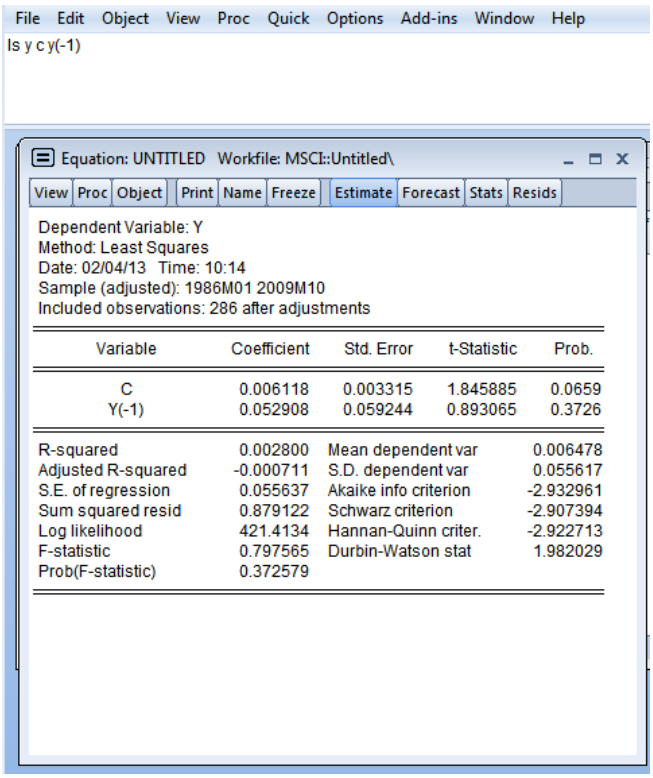
11.2 Testing for ARCH/GARCH effects

In this example we consider monthly observation on the Danish stock market index. The Eviews workfile msci.wf1 contains data from 1985:12 to 2009:10. Y_t denote the change in the log price index.

$Y_t = \delta + \beta Y_{t-1} + \epsilon_t$

Now we would like to test whether this process suffers from ARCH/GARCH effects, and if this is the case, how can we best model the variance?

1. First we estimate the above model by OLS. This is done by typing *ls Y c Y(-1)* in the command field.



2. Second we must save the OLS residuals from the above estimation. You can use the short cut and type *genr e=resid* in the command field or click *proc/ make a residual series*.
3. Third we have to compute the squared residuals, which is done by typing *genr e2=e^2* and press enter in the command field.
4. The last thing is to run an auxliary regression by the following command *ls e2 c e2(-1)*.

File Edit Object View Proc Quick Options Add-ins Window Help

```
genr e=resid
genr e2=e^2
ls e2 c e2(-1)
```

Equation: UNTITLED Workfile: MSCI::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: E2
Method: Least Squares
Date: 02/04/13 Time: 10:28
Sample (adjusted): 1986M02 2009M10
Included observations: 285 after adjustments

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.002553	0.000351	7.276166	0.0000
E2(-1)	0.163117	0.058532	2.786788	0.0057

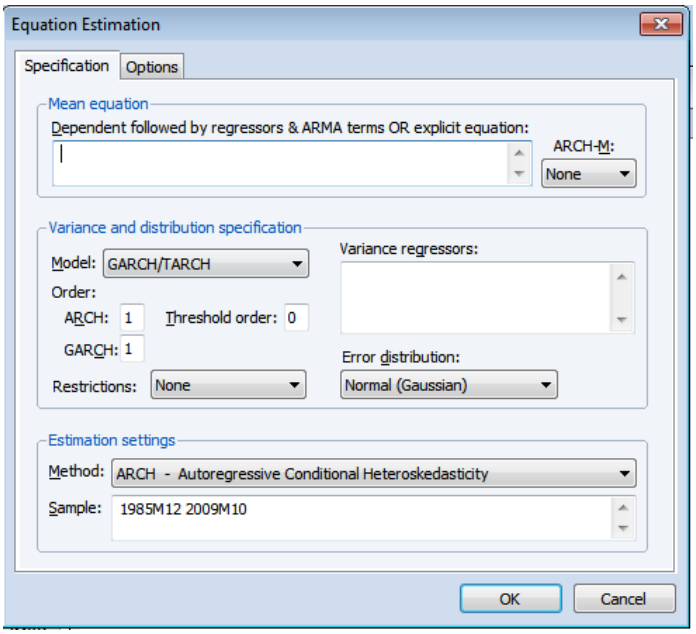
R-squared	0.026709	Mean dependent var	0.003055
Adjusted R-squared	0.023270	S.D. dependent var	0.005142
S.E. of regression	0.005082	Akaike info criterion	-7.719157
Sum squared resid	0.007310	Schwarz criterion	-7.693526
Log likelihood	1101.980	Hannan-Quinn criter.	-7.708882
F-statistic	7.766185	Durbin-Watson stat	1.995262
Prob(F-statistic)	0.005683		

The R^2 from the auxiliary regression is 0,026709 this should be multiplied with the number of observations.

$$0{,}026709 * 285 = 7{,}6$$

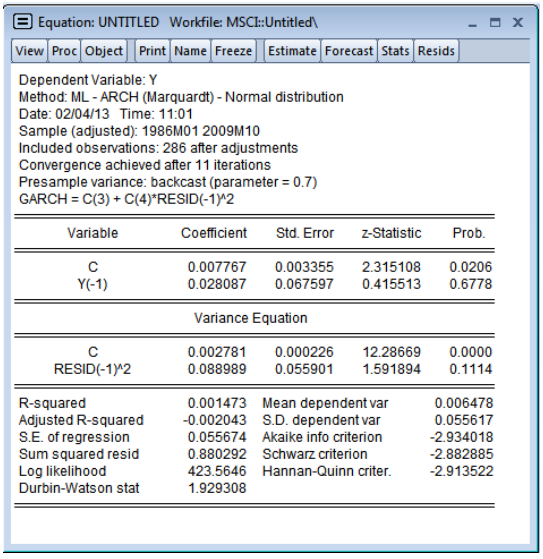
This should be compared with a critical value of $\chi^2(1) = 3{,}48$. This implies that we clearly reject the null of no arch effects.

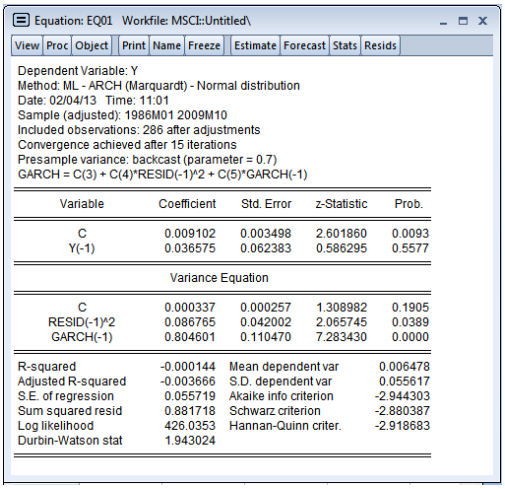
Now we have established that the conditional variance varies over time. We can now try to estimate the conditional variance by the build in function in EViews. This function will appear by typing *ARCH* in the command field.



1. Fist we have to inform EViews which regression we have previously worked with. Type in specification of the estimation *Y c Y(-1)*.
2. Then we have to specify the ARCH or GARCH model. You need to make a number of different ARCH and GARCH estimations and compare the information criterion and statistical significance of the ARCH and GARCH terms in order to evaluate the best fit.

In this example we will only estimate an ARCH(1) and a GARCH(1,1) model and evaluate which of the two has the best fit. You should compare a larger number of ARCH/GARCH specifications when you evaluate the best fit.





Equation: EQ01 Workfile: MSCI::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

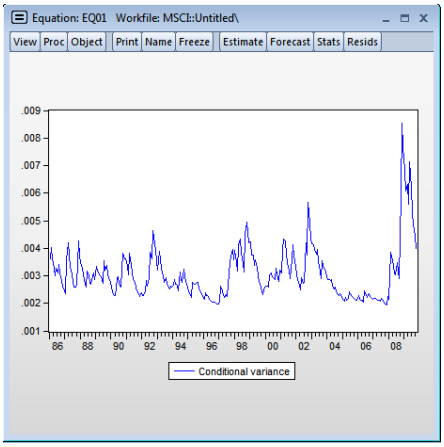
Dependent Variable: Y
Method: ML - ARCH (Marquardt) - Normal distribution
Date: 02/04/13 Time: 11:01
Sample (adjusted): 1986M01 2009M10
Included observations: 286 after adjustments
Convergence achieved after 15 iterations
Presample variance: backcast (parameter = 0.7)
GARCH = C(3) + C(4)*RESID(-1)^2 + C(5)*GARCH(-1)

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	0.009102	0.003498	2.601860	0.0093
Y(-1)	0.036575	0.062383	0.586295	0.5577
Variance Equation				
C	0.000337	0.000257	1.308982	0.1905
RESID(-1)^2	0.086765	0.042002	2.065745	0.0389
GARCH(-1)	0.804601	0.110470	7.283430	0.0000

R-squared	-0.000144	Mean dependent var	0.006478
Adjusted R-squared	-0.003666	S.D. dependent var	0.055617
S.E. of regression	0.055719	Akaike info criterion	-2.944303
Sum squared resid	0.881718	Schwarz criterion	-2.880387
Log likelihood	426.0353	Hannan-Quinn criter.	-2.918683
Durbin-Watson stat	1.943024		

Based on the information criterion and the fact that the ARCH effect is insignificant in the ARCH(1) model we will choose a GARCH(1,1) to model the conditional variance. The last thing we are going to do is make a plot of the conditional variance.

1. Click on "view" and "GARCH graph" then you can choose between variance and standard deviation.



12 Panel data

Panel data models combine the time dimension and the cross sectional dimension. In panel data we sample the *same* cross sectional units over time, and thus get a two-dimensional data set.

Important advantages of panel data compared with time series or cross-sectional data set is that they allow identification of certain parameters or questions, without the need to make restrictive assumption. For example, panel data make it possible to analyze changes on an individual level. Consider a situation in which the average consumption level rises by 2 % from one year to another. Panel data can identify the result of, an increase of 2 % for all individuals or an increase of 4 % for approximately one-half of the individuals and no change for the other half. In more general terms we try to capture all unobservable time invariant differences across individuals.

In this section we will go through four different panel data models, pooled cross section, fixed effect estimation, random effect estimation and first difference estimation.

The basic panel data model for cross sectional unit i is:

$$y_{it} = x'_{it}\beta + c_i + u_{it}\,, \quad t = 1,2,\ldots,T$$

Where x_{it} is a $Kx1$ vector of observable explanatory variables, u_{it} is the idiosyncratic error with mean 0 (the error varies both over time and individual). c_i is the unobserved time constant characteristics of an individual, and that is the effect we specifically want to control for in our panel data models. A classic example of unobserved characteristics could be an individual's ability (since it's non-measurable).

Normally we will use random effects models if $Corr(x_{it}, c_i) = 0$ – the unobserved characteristics are uncorrelated with the explanatory variables. If $Corr(x_{it}, c_i) \neq 0$ we would use either fixed effects estimation or first difference estimation. It is possible for us to use instruments and robust standard errors in panel data models.

12.1 The data set & setting panel data in EViews

In this panel data section we'll be using the data set jtrain.wf1. The dataset can be used to investigate if job training grants can reduce firm scrap rates. Our data originates from Michigan and consist of $\boldsymbol{N = 54}$ firms over $\boldsymbol{T = 3}$ years: 1987-89. The dataset includes the following variables:

Variable	Description
$lscrap_{it}$	Log of scrap rate for firm i at time t
$grant_{it}$	Dummy variable, equal to 1 if firm i received a grant at time t
$grant_{i,t-1}$	Dummy variable, equal to 1 if firm i received a grant at time t-1
$union_{it}$	Dummy variable, equal to 1 if firm i was a member of a union at time t
$d88_{it}$	Dummy variable, equal to 1 if year 1988
$d89_{it}$	Dummy variable, equal to 1 if year 1989
$fcode_{it}$	Identifier of firm i
$year_{it}$	Year

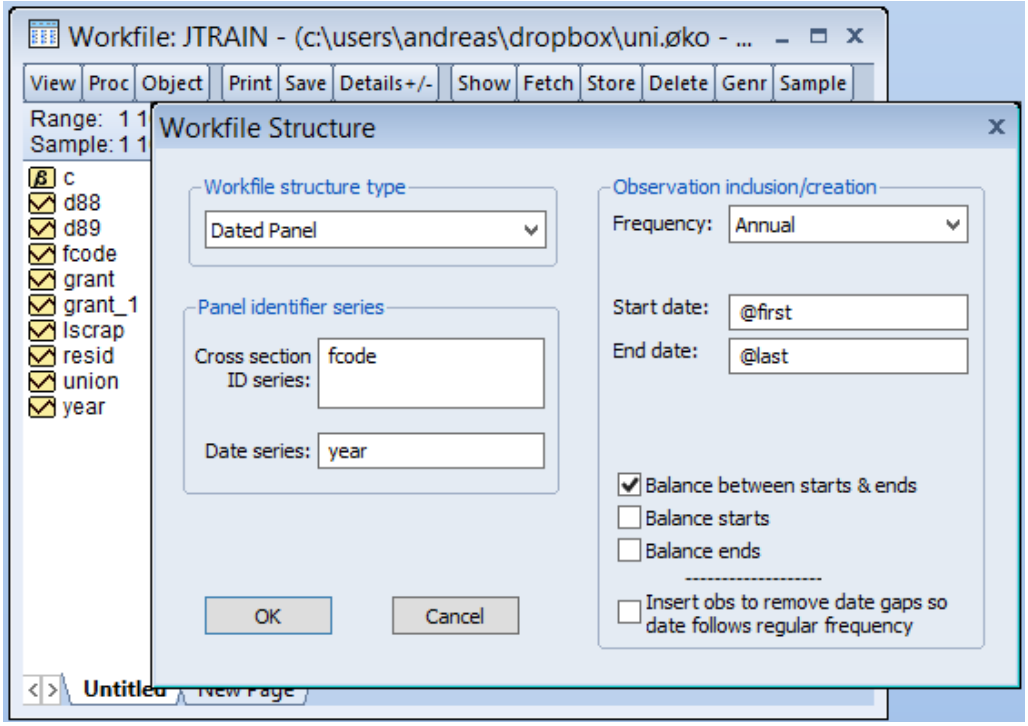
Our equation of interest is

$$lscrap_{it} = \beta_0 + \beta_1 d88_{it} + \beta_2 d89_{it} + \beta_3 union_{it} + \beta_4 grant_{it} + \beta_5 grant_{i,t-1} + c_i + u_{it}$$

The unobserved effects c_i in this model could be worker and managerial ability (nearly impossible to measure). In this job training program we are sampling the job training grants, which were approved on a first-come, first-serve basis, meaning that the unobserved effect might be correlated with $grant$.

12.2 Setting EViews up for panel data

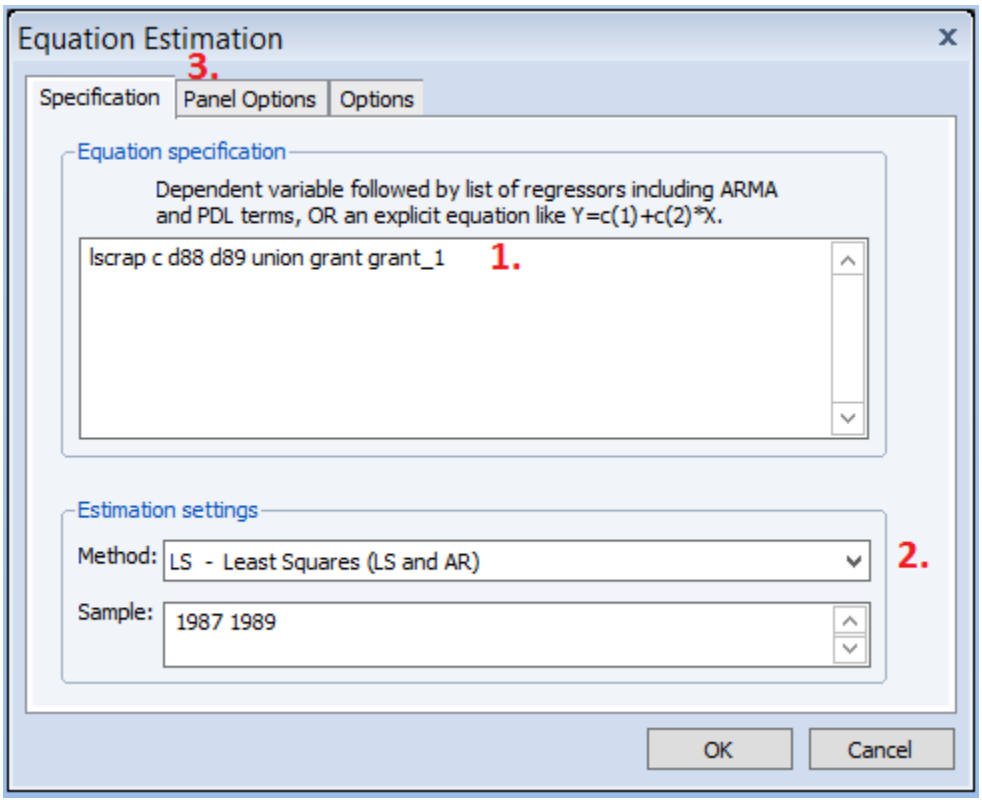
We have to inform EViews that we are working with panel data. EViews need to identify each individual, and a time variable. After opening the dataset we go to **Proc→ Structure/Resize Current Page**, the following window should appear:



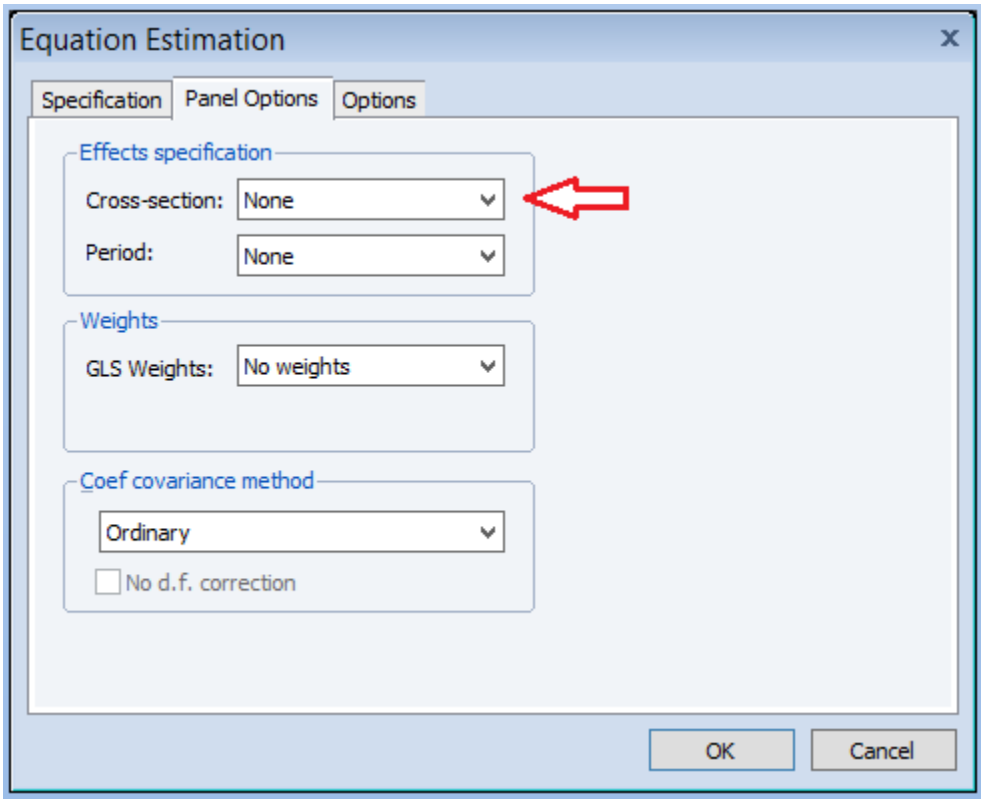
In this dataset our panel identifier series is the *fcode* variable (ID value for each cross sectional unit), and our Date series is the *year* variable. The frequency is set to *Annual*, because we have yearly data. With the start date set to *@first* and end date set to *@last* EViews automatically finds out the start and end date. Alternatively we could manually have written 1987 in start date and 1989 in end date. Clicking OK – EViews will create a new variable called *dateid*, and EViews will now be able to produce panel data model estimates.

12.3 Fixed effect estimation

To estimate our model we go to *Quick/Estimate Equation*.

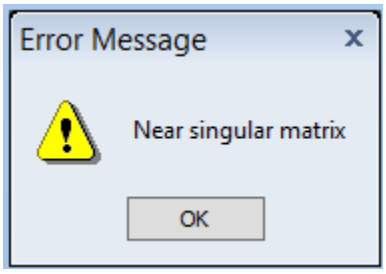


1. We write up our model of interest
2. Normally we just choose to estimate the model by Least Squares, but if we have problems with endogeneity we can choose TSLS, and then a new tab called Instruments will appear where it will be possible to specify instruments.
3. We click on the Panel options tab, and the following window appears:



In the effects specification we can choose between *fixed* and *random* effects estimation by choosing Fixed or Random in the *Cross-section* option. If we choose *none* we will just get *pooled OLS*.

We start by choosing *fixed effect* estimation. And click OK - we get an error: near singular matrix.



This is because fixed effect estimation does not accept an explanatory variable that is constant over time, and union is constant over time (mathematically it's impossible to invert a singular matrix, and that is what causes the error). We fix this by removing *union* from the regression, and run it again. Output produced:

Dependent Variable: LSCRAP
Method: Panel Least Squares
Date: 02/10/13 Time: 21:07
Sample: 1987 1989
Periods included: 3
Cross-sections included: 54
Total panel (balanced) observations: 162

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.597434	0.067734	8.820245	0.0000
D88	-0.080216	0.109475	-0.732730	0.4654
D89	-0.247203	0.133218	-1.855622	0.0663
GRANT	-0.252315	0.150629	-1.675075	0.0969
GRANT_1	-0.421590	0.210200	-2.005659	0.0475

Effects Specification

Cross-section fixed (dummy variables)

R-squared	0.927572	Mean dependent var	0.393681
Adjusted R-squared	0.887876	S.D. dependent var	1.486471
S.E. of regression	0.497744	Akaike info criterion	1.715383
Sum squared resid	25.76593	Schwarz criterion	2.820819
Log likelihood	-80.94603	Hannan-Quinn criter.	2.164207
F-statistic	23.36680	Durbin-Watson stat	1.996983
Prob(F-statistic)	0.000000		

The constant in the regression assures that the unobserved effect c_i has a mean of zero. It is actually possible for us to obtain the c_i's by clicking *View/Cross-section Effects.*

The data in the column Effect is our unobserved effect given for each individual in the sample:

Cross-section Fixed Effects

	FCODE	Effect
	FCODE	Effect
1	410523	-3.423253
2	410538	0.482006
3	410563	1.294075
4	410565	1.020414
5	410566	1.198181
6	410567	-1.143666
7	410577	-0.000106
8	410592	2.703367

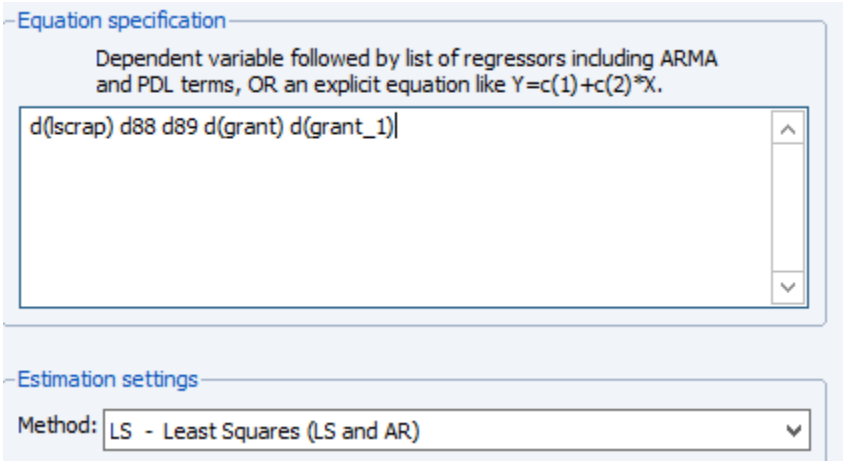
12.4 First difference estimation

In first difference estimation we simply estimate our model in first difference:

$$\Delta lscrap_{it} = \beta_1 \Delta d88_{it} + \beta_2 \Delta d89_{it} + \beta_3 \Delta union_{it} + \beta_4 \Delta grant_{it} + \beta_5 \Delta grant_{i,t-1} + u_{it}$$

The intercept and the unobserved effect are differenced away, since we have lagged the model one period and subtracted it from the original model to obtain the first difference model. As before we have a problem with the explanatory variable, *union*, therefore we have to exclude it because it is constant over time. We have the choice of including a con-

stant and deleting a dummy, or keeping both dummies but leave out the constant in the estimation. An easy way to take first difference of a variable is writing $d(var\ name)$ in the equation specification:



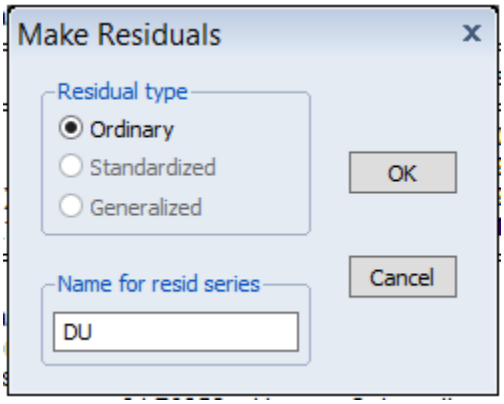
The effects specification in the tab *Panel Options/Cross-section* should be set to "*None*". Estimation output with and without constant:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.090607	0.090970	-0.996017	0.3216
D89	-0.096208	0.125447	-0.766923	0.4449
D(GRANT)	-0.222781	0.130742	-1.703970	0.0914
D(GRANT_1)	-0.351246	0.235085	-1.494124	0.1382

Variable	Coefficient	Std. Error	t-Statistic	Prob.
D88	-0.090607	0.090970	-0.996017	0.3216
D89	-0.186815	0.104395	-1.789508	0.0764
D(GRANT)	-0.222781	0.130742	-1.703970	0.0914
D(GRANT_1)	-0.351246	0.235085	-1.494124	0.1382

12.5 Choosing between fixed effect and first difference estimation

If we have a strong serial correlation in the residuals the first difference estimator will be most efficient, but if we only have weak or none serial correlation in the residuals we should use the fixed effect estimator. We can test this by making an AR(1) model (without constant) of the residuals from the first difference estimator. We get the residuals from the first difference estimation by pressing *Proc/Make Residual Series*.



Then we estimate the AR(1) model either by using the *Quick/Estimate Equation*, or just typing *ls du du(-1)* in the command window. EViews produces the following output:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
DU(-1)	0.236906	0.133357	1.776481	0.0814

The coefficient estimate on DU(-1) should be 0 for the first difference estimator to be efficient, and -0,5 for the fixed effect estimator to be efficient.

One should carry out a wald test of DU(-1) being equal to either 0 or -½.

Overview table:

$Corr(x_{it}c_i) \neq 0$	
Fixed effect est.	Consistent; efficient if coef. on DU(-1)=0
First difference est.	Consistent; efficient if coef. on DU(-1)=-½

Normally we choose the estimator that is closest to have fulfilled the efficiency condition.

12.6 Random effects estimation

We can include all variables in random effects estimation. The only thing we should choose differently is the effects specification in the tab *Panel Option/Cross sectional*, which is set to *Random* instead of *None or Fixed*. Below it the estimation output:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.414833	0.242965	1.707379	0.0897
D88	-0.093452	0.108946	-0.857779	0.3923
D89	-0.269834	0.131397	-2.053577	0.0417
UNION	0.547802	0.409837	1.336635	0.1833
GRANT	-0.214696	0.147500	-1.455565	0.1475
GRANT_1	-0.377070	0.204957	-1.839747	0.0677

12.7 Random effects or fixed effects/first difference

It will now be possible for us to make a test that indirectly tests whether or not the $corr(x_{it}c_i) = 0$ condition is satisfied, and thereby if the appropriate estimation model is random effects/pooled OLS, OR fixed effect/first difference.

We have that both the random and fixed effect estimator is consistent if $corr(x_i c_{it}) = 0$ (though only the random effect would be efficient), and they should therefore give roughly the same estimates if $corr(x_i c_{it}) = 0$.

The Hausman test can be carried out by EViews *after* the random effects estimation, by clicking *View/Fixed/Random Effects Testing /Correlated Random Effects Hausman Test*. The output produced:

Correlated Random Effects - Hausman Test
Equation: Untitled
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	0.000000	4	1.0000

* Cross-section test variance is invalid. Hausman statistic set to zero.

The test yields an error because of non-invertible differences in covariances between the two estimators. To be able to carry out the test we need to re-estimate the random effects model without all dummies and time-constant variables like union in our example.

The new Random effects specification is: *lscrap c grant grant_1*. Now it's possible to make a valid Hausman test, and it yields the following output:

Correlated Random Effects - Hausman Test
Equation: Untitled
Test cross-section random effects

Test Summary	Chi-Sq. Statistic	Chi-Sq. d.f.	Prob.
Cross-section random	2.848361	2	0.2407

Cross-section random effects test comparisons:

Variable	Fixed	Random	Var(Diff.)	Prob.
GRANT	-0.384852	-0.355935	0.000320	0.1057
GRANT_1	-0.694953	-0.662883	0.000574	0.1807

The null hypothesis of the Hausman test is that both estimators are consistent and thus $corr(x_{it}c_i) = 0$. We cannot reject the null on a 5% significance level, and therefore we have that the unobserved characteristics and the explanatory variables *grant* and *grant_1* is uncorrelated. Therefore we conclude that the random effects estimator is the most efficient to use in this case.

13 The Generalized Method of Moments (GMM)

This approach estimates the model parameters directly from the moment conditions that are imposed by the model. These conditions can be linear in the parameters, but quite often are these nonlinear. To enable identification, the number of moment conditions should be at least as large as the number of unknown parameters. This can be done by EViews. Consider the following moments:

$$M_{t+1} = CG_{t+1}^{-\gamma}$$

Where M_{t+1} is the stochastic discount factor, and R_{t+1}^{e} is a vector of excess returns. Assume that the stochastic discount factor is given by:

$$E[M_{t+1}R_{t+1}^{e}] = 0$$

Where CG is the US annual fourth quarter consumption growth rate (per capita and in real units), and γ is the relative risk aversion.

The vector of annual excess returns is based on 10 decile portfolios sorted by the book-to-market ratio. The first decile contains the stocks with the smallest book-to- market ratios (growth stocks), while the tenth decile contains the stocks with the largest book-to-market ratios (value stocks). The portfolios are constructed using stocks listed on NYSE, AMEX and NASDAQ. More information is available from Kenneth French's website from which the portfolios have been downloaded.

First we have to specify our moment conditions. The population moment conditions are

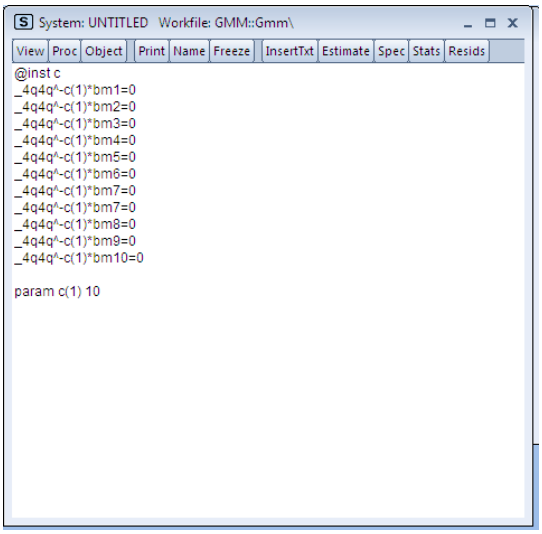
$$E\left[CG_{t+1}^{-\gamma}R_{j,t+1}^{e}\right] = 0$$

Where $j = 1,2,\dots,10$

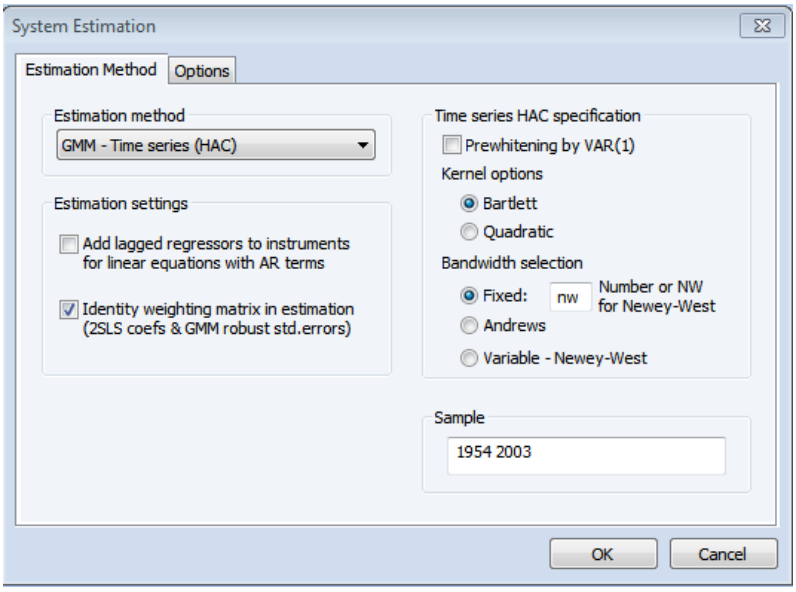
We have 10 moment condition and only 1 parameter to estimate, implying our system is overidentified. In the data set $CG_{t+1}^{-\gamma} = \_4q4q$ *and* $R_{j,t+1}^{e} = BM_i$.

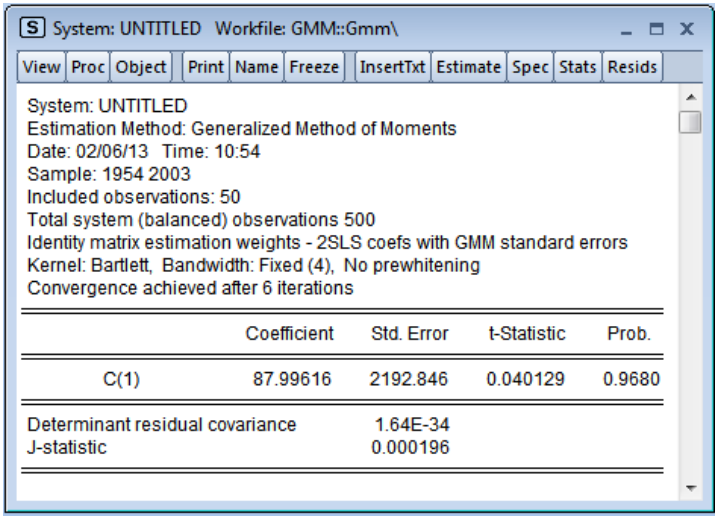
To type the moment conditions into EViews you have to click on *object /New object/system* and then type in the moments.

Then we need to specify which instrument we are using in all our moments. In the system you *write @inst c* (c is equal to γ), then the moments follows. You can set a starting value for c, by typing param c(1) 10 after the moments. In this example the starting value for c(1) is set to 10, which is shown below:



Now our system and specifications are done and we can estimate c(1), Click *proc /estimate and choose the estimation method GMM HAC* .We also need to tell EViews which weighting matrix we desire (identity matrix or the optimal weighted matrix). In this example we will use identity matrix because we want equal weight of the 10 portfolios.





As you can see the EViews estimates of standard errors are rather high. In the 7th edition of E-Views there are problems when estimating GMM with a system. It seem like EViews can't handle the estimation of standard errors with an identity matrix. A solution to this will come as soon as possible. The right estimates can be seen below. This estimation is done in the 5th edition of EViews.

System: UNTITLED Workfile: PROBLEM1::Problem1\

View Proc Object Print Name Freeze MergeText Estimate Spec Stats Resids

System: UNTITLED
Estimation Method: Generalized Method of Moments
Date: 12/03/12 Time: 13:53
Sample: 1954 2003
Included observations: 50
Total system (balanced) observations 500
Identity matrix estimation weights - 2SLS coefs with GMM standard errors
Kernel: Bartlett, Bandwidth: Fixed (3), No prewhitening
Convergence achieved after 6 iterations

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	87.99616	43.85692	2.006437	0.0453
Determinant residual covariance		1.64E-34		
J-statistic		7.84E-08		

The estimated coefficient of c is 87,996 with a s.e. of 43,86. Given asymptotic normality, we are able to reject (marginally), meaning that the coefficient of relative risk aversion is significant.

14 Programming in EViews

EViews is capable of doing various programming tasks. Here we will give a brief introduction to the basic commands, syntax and methodology.

EViews help function is very useful and should be used if you are in doubt about any command, and/or function. The help-function can be found in EViews: *Help/EViews Help Topics*. The most efficient way to search for a given command or item, is under "Index", where it is possible to type the search query.

14.1 Open program in EViews

In the coding area/prompt we type

new program program_title

This will open up the program editor, where we will type our code. Various options are available in the program window: Run, Save, Save As and so on. It is recommended to save the program by pressing *SaveAs*.

Run will open a dialogue box, where you can select additional options. Typically we want to choose *Quiet (fast)* under the *Reporting* tab, to enhance the speed of the compiling process. To run the program we simply press the OK button in the bottom of the dialogue box.

14.2 Create Workfile

We need a workfile to declare and store our variables, matrices and vectors in.

To declare and create a new workfile, you may use the `wfcreate` command. You may enter the keyword `wfcreate` followed by a name for the workfile, an option for the frequency of the workfile, and the start and end dates.

a	annual.
s	semi-annual.
q	quarterly.
m	monthly.
w	weekly.
d	daily (5 day week).
7	daily (7 day week).
u	undated/unstructured.

Example

Yearly data

wfcreate my_workfile1 a 1991 1999

Monthly data
wfcreate my_workfile2 m 1991m1 1999m4

Or simply a workfile without any dated structure, here with 1000 observations.

wfcreate my_workfile_unstructured u 1 1000

14.3 Comments

Another useful thing is the possibility of making comments in the code, by using an apostrophe.

wfcreate my_workfile_unstructured u 1 1000 'this is a comment, and will be ignored by the compiler

Using comments to comment on your code is best practice. It will ease the reading of the code.

14.4 Scalar, vector and matrix declarations and manipulations

To declare a scalar, we use the command scalar followed by the name.

scalar count

We could assign a value to the scalar after the declaration

count = 1

To declare a vector of size n, we type the command vector(n) followed by the name of the vector we want to declare .

vector(n) my_vector

After running this code, we will have an n-sized vector in our workfile. Other similar declarations:

matrix (5,5) my_matrix 'creating a 5 by 5 matrix

rowvector(5) my_rowvector 'creating a rowvector instead of a columnvector

To fill a value into the vector, we write the name of the vectorobject followed by a dot and *fill*

vector(5) myvector

my_vector.fill 1,2,3,4,5 'filling the values 1,2,3,4,5 into the columnvector my_vector

14.5 Generating series and new variables

Scalar, vector and matrix declarations and manipulations

To generate a new series/variable we use the command *genr.*

genr price = 0

The generated series will have the size of the chosen workfile size. If we have chosen our workfile as unstructured with 1000 observations, price will be a series of size 1000.

If we use a variable in the program that is not intended to be saved in the workfile, we declare it by using an exclamation mark in front of the name of the variable. It could for example be a count variable in a loop (see under loop, next section).

!count = 1 'setting the variable count equal to one
!count = !count + 1 'setting the variable count equal to the previous value plus 1

14.6 Setting sample size

If we want to manipulate only a part of the data set, we can set the sample size to a certain period.

smpl 1992m1 1993m2
smpl 1 1000
smpl @all 'reversing the sample size to all observations

14.7 Equation objects

To save an equation, we use an equation object. To save a standard OLS (least square = *ls*) regression (where we regress y on a constant and x) in an equation object called reg1 we type:

equation reg1.ls y c x

equation reg2.ls(h) y c x 'now with white robust standard errors

These equation objects will now be saved in our workfile.

It is possible to extract certain information from our equation object. We can for example extract the r-squared or the coefficient estimates from our regression.

scalar r2 'declaring a scalar to save the r squared in.
r2 = reg1.@r2

Selected Keywords that Return Scalar Values

@aic	Akaike information criterion
@coefcov(i,j)	Covariance of coefficient estimates
@coefs(i)	*i*-th coefficient value
@dw	Durbin-Watson statistic
@f	*F*-statistic
@fprob	*F*-statistic probability.
@hq	Hannan-Quinn information criterion
@jstat	*J*-statistic - value of the GMM objective function (for GMM)

@logl	value of the log likelihood function
@meandep	mean of the dependent variable
@ncoef	number of estimated coefficients
@r2	R-squared statistic
@rbar2	adjusted R-squared statistic
@rlogl	retricted (constant only) log-likelihood.
@regobs	number of observations in regression
@schwarz	Schwarz information criterion
@sddep	standard deviation of the dependent variable
@se	standard error of the regression
@ssr	sum of squared residuals
@stderrs(i)	standard error for coefficient
@tstats(i)	*t*-statistic value for coefficient
c(i)	*i*-th element of default coefficient vector for equation (if applicable)

Selected Keywords that Return Vector or Matrix Objects

@coefcov	matrix containing the coefficient covariance matrix
@coefs	vector of coefficient values
@stderrs	vector of standard errors for the coefficients
@tstats	vector of *t*-statistic values for coefficients

It is of course also possible to use other estimators than OLS (*ls*)

14.8 Equation Methods

arch autoregressive conditional heteroskedasticity (ARCH and GARCH).

binary binary dependent variable models (includes probit, logit, gompit) models.

censored censored and truncated regression (includes tobit) models.

cointreg estimate cointegrating equation using FMOLS, CCR, or DOLS.

count count data modeling (includes poisson, negative binomial and quasi-maximum likelihood count models).

glm estimate a Generalized Linear Model (GLM).

gmm estimate an equation using generalized method of moments (GMM).

liml estimate an equation using Limited Information Maximum Likelihood and K-class Estimation.

logit logit (binary) estimation.

ls equation using least squares or nonlinear least squares.

ordered ordinal dependent variable models (includes ordered probit, ordered logit, and ordered extreme value models).

probit probit (binary) estimation.

qreg estimate an equation using quantile regression.

stepls estimate an equation using stepwise regression.

tsls estimate an equation using two-stage least squares regression.

Help for the different estimation methods can be found in Eviews. In TSLS estimation we will have to specify our instruments for the estimation to work.

equation twostep.tsls y c x h @ c j h

The instruments for the regression is given after the @-symbol. We have that c and h and instruments for themselves and x have j as instrument.

14.9 Loops

The idea of making either a *for* or *while*-loop is to iterate through the same code, doing a certain task many times. It could for example be that we wanted Eviews to estimate 1000 OLS-regressions in our data set. A *for* loop that places the values from 51 to 1050 in a vector.

```
wfcreate my_workfile_unstructured u 1 1000 'creating workfile
vector(1000) values 'declaring vector of size 1000

for !i=1 to 1000 'for-loop beginning
                    values(!i) = 50 + !i
next
```

The same output can be produced by a while-loop

```
wfcreate my_workfile_unstructured u 1 1000 'creating workfile
vector(1000) values 'declaring vector of size 1000

!i = 1
while !i <= 1000 'while-loop beginning
                    values(!i) = 50 + !i
                    !i=!i+1
wend
```

After the while-statement we have a logic expression, and as long as the logic expression is true, we will loop. In the while loop we need to manipulate our controlvariable *i* ourselves.

14.10 Simulation study – Monte Carlo Simulation

Monte Carlo simulations can be used to assess the properties of a given estimator. We will here use the simulation study to say something about the properties (consistency, bias, distribution) of OLS estimation compared to TSLS estimation when our estimation suffers from endogenity.

To make a simulation study, we will need all the command from the previous sections.

In this simulation study we would like to create a workfile with N=1000 observations and use a loop to pick out the sample sizes (we have different sample sizes to assess the properties of the estimator). The sample sizes we will be working with is: $SS = \{10, 50, 100, 500, 1000\}$

First of all we need a given data generating process.

$$y_i = 1 + x_i + \epsilon_i$$

Where

```
'remember to run program in quiet mode (fast)

wfcreate Monte_Carlo u 1 1000 'creating workfile
vector(1000) values 'declaring vector of size 1000

!m = 1000 'number of simulations for each sample size
vector(5) samplesize
samplesize.fill 10, 50, 100, 500, 1000
rndseed 1 'seeding the random number generator, making it possible to reproduce results
matrix(!m , 5) results 'creating a matrix with dimension 10000 by 5 to store slope coefficients in

for !i=1 to 5
    'loop for picking out sample size
    !ss = samplesize(!i)

    for !j=1 to !m
        'data generation
        genr x = nrnd 'the values of x is drawn from a normaldistribution
        genr z = nrnd
        genr e = -x(-1)+z 'generation of our error term according to the specification
        genr y = 1+x+e

        smpl 2 1000 'restricting the sample size, because of the lagged value

        'estimation
        equation ols.ls y c x
        results(!j , !i) = ols.@coefs(2) 'storing the slope coefficient in the j'th row, and the i'th column.
    next
next
```

$$\epsilon_i = -x_{i-1} + z_i$$
$$x_i \sim N(0,1)$$
$$z_i \sim N(0,1)$$

Now we can assess the properties of the estimator by double clicking the *malpha* and *mbeta* matrix respectively and choose *View/Descriptive stats by column.*

15 Appendix A - Variables in the dataset rus98.wf1

Here is a list describing the variables in the rus9

8.wf1 dataset which is using for the first part of the manual.

Hold: *Education*;

1 = HA1-6, 2 = HA7-10,dat; 3=HA7-10,dat; 4=HA jur; 5=BSc B

Sp01: *Sex*;

1=Female; 2=Male

Sp02: *Expect income > 300.000*;

1=Yes; 2=No

Sp03: *Party*

0= Undecided; 1= Soc.Dem.; 2=Rad.V.; 3=Kons.; 4=SF; 5=CD; 6=Da.Fo; 7=Venstre; 8=Fremskr.; 9=Enhedsl.; 10=Andre

Sp04: *Your weight*

Sp05: *Your height*

Sp06: *Your mother's height*

Sp07: *Your father's height*

Sp08: *Drinks (Genstande), number of in week 34*

Sp09: *Average marks (Karakter) at qualifying exam*

Sp10: *Social order*

1=Rather important; 2=Important; 3=Very important; 4=Extremely important

Sp11: *Social justice*

1=Rather important; 2=Important; 3=Very important; 4=Extremely important

Sp12: *Belong to somebody*

1=Rather important; 2=Important; 3=Very important; 4=Extremely important

Sp13: *Selt-realization*

1=Rather important; 2=Important; 3=Very important; 4=Extremely important

Sp14: *Fun and joy in life*

1=Rather important; 2=Important; 3=Very important; 4=Extremely important

16 Appendix B – The dataset FEMALEPRIVATEWAGE.wf1

The following list describes each variable in the FEMALEPRIVATEWAGE.wf1 which is used for the SLR and MLR part of this manual.

Age -the age measured in total years

Childd – dummy variable – 0 = no children, 1=one or more children

Educatio– number years of education completed, including elementary school

Exper – amount of experience in current job measured in years,

Hourwage – personal pretax income per hour measured in krones

ln_hwage – LN(hourwage) that is the natural logarithm on the above described variable

marriedd – dummy variable – 0 = if not married, 1= if married

provinced – dummy variable – 0 = if the respondent does not work the in province, 1= if respondent work in the province.

Note that we used Excel to modify the existing SPSS file FEMALEPRIVATEWAGE.sav to Eviews format (.wf1). This was done to create the above mentioned dummy variables.

17 Appendix C – Installing Windows on a Mac

First off all we want to stress that the ICT Department does not support or in any way assist students in installing Windows on their Macs. But we will try to point you in the right direction. To install Windows on you Mac you will need the following:

- A working CD or DVD containing Windows (XP or later is recommended)
- The application Bootcamp (included in Leopard and later versions of OS X)

Apple has already made a complete guide illustrating the use of Bootcamp – download it at: http://manuals.info.apple.com/en_US/Boot_Camp_Install-Setup.pdf

THIS GUIDE HAS BEEN PRODUCED BY

ANALYTICS GROUP

Analytics Group, a division comprised of student instructors under AU IT, primarily offers support to researchers and employees.

Our field of competence is varied and covers questionnaire surveys, analyses and processing of collected data etc. AG also offers teaching assistance in a number of analytical resources such as SAS, SPSS and Excel by hosting courses organised by our student assistants. These courses are often an integrated part of the students' learning process regarding their specific academic area which ensures the coherence between these courses and the students' actual educational requirements.

In this respect, AG represents the main support division in matters of analytical software.

ADVANCED MULTIMEDIA GROUP

Advanced Multimedia Group is a division under AU IT supported by student instructors. Our primary objective is to convey knowledge to relevant user groups through manuals, courses and workshops.

Our course activities are mainly focused on MS Office, Adobe CS and CMS. Furthermore we engage in e-learning activities and auditive and visual communication of lectures and classes. AMG handles video assignments based on the recording, editing and distribution of lectures and we carry out a varied range of ad hoc assignments requested by employees.

In addition, AMG offers solutions regarding web development and we support students' and employees' daily use of typo3.

PLEASE ADDRESS QUESTIONS OR COMMENTS REGARDING THE CONTENTS OF THIS GUIDE TO

ANALYTICS@ASB.DK