ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

STATA Workshop I

Agenda (1)

I. Introduction to STATA

- STATA Interface
- Data Management in STATA

II. Empirical Example 1 : Cross Sectional Data

- Example of a test in Finance for 60 students
- Probability density function
- Basic Distribution parameters (mean, standard deviation, skewness, kurtosis)

Agenda (2)

III. Empirical Example 2 : Time Series Data

- Example of monthly returns for equity indices of G7 countries
- Distribution parameters (covariance, correlation)

IV. Empirical Example 3 : Transformations of data & plots - Time Series Data

- Example of simple and continuous compounding returns for UK Market Index

Agenda (3)

V. Classical Linear Regression Model Estimation

- **Empirical Example 4 :** CAPM model

➢ Model Estimation

➢Hypothesis Testing

➢Wald Test

➢Multiple Hypothesis : the F -test

VI. Multiple Linear Regression Model Estimation

- **Empirical Example 5 :** APT Model

➢ Model Estimation

➢Hypothesis Testing

➢Wald Test

➢Multiple Hypothesis : the F -test

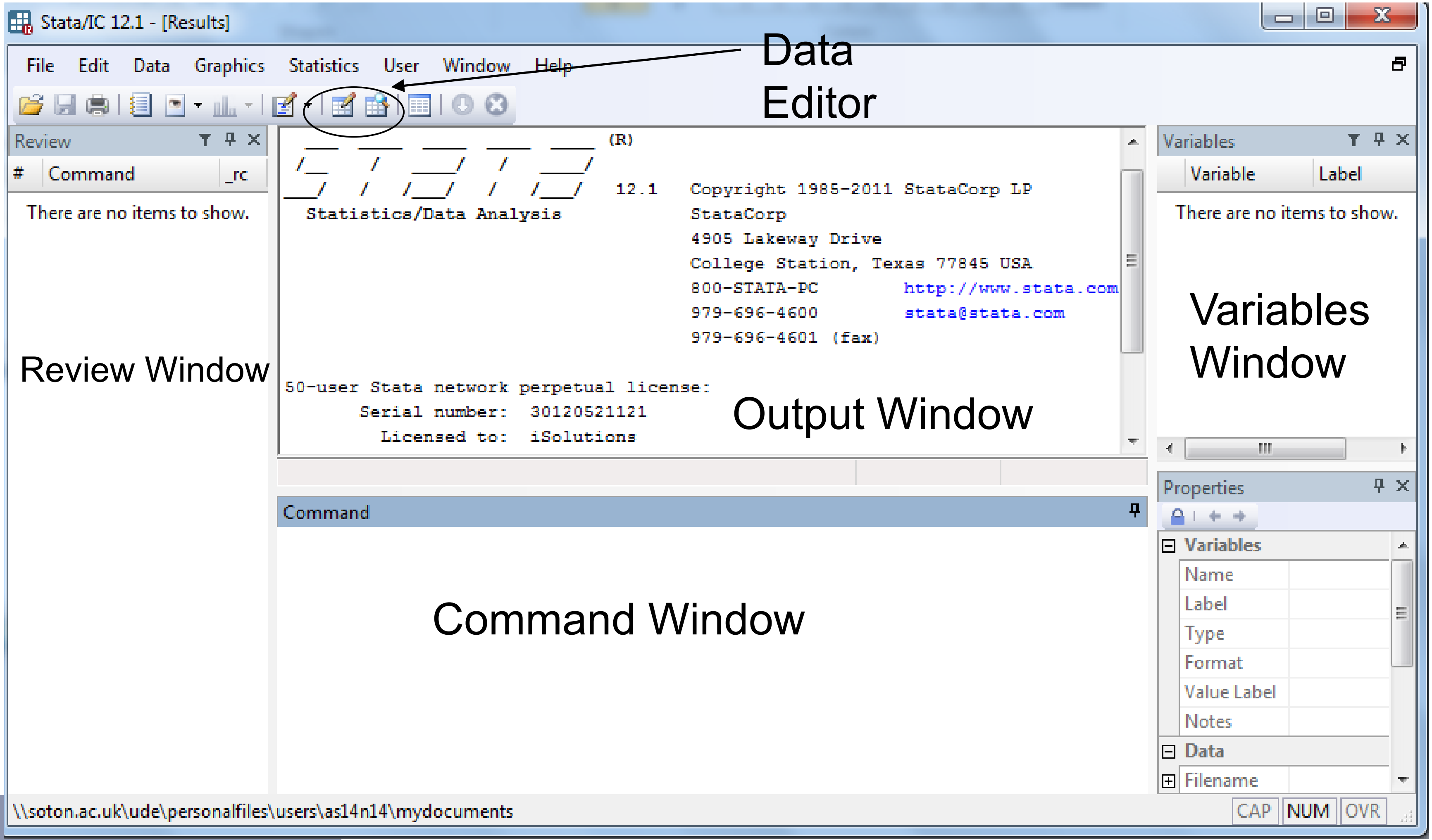
➢Stepwise procedure equation estimation

➢R-squared & F -Statistic

Introduction to STATA

1.Open *STATA* from PC- lab

- Double Click on the STATA on the desktop of your pc



Data Editor

Review Window

Output Window

Variables Window

Command Window

There are no items to show.

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   12.1   Copyright 1985-2011 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
                                      College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

50-user Stata network perpetual license:
       Serial number:  30120521121
         Licensed to:  iSolutions
```

There are no items to show.

\\soton.ac.uk\ude\personalfiles\users\as14n14\mydocuments

How to save your output – Log files

- Stata can record your session into a file called a log file but does not start a log automatically; you must tell Stata to record your session.
- By default, the resulting log file contains what you type and what Stata produces in response, recorded in a format called Stata Markup and Control Language (SMCL).
- To start a log: click on File → Log → Begin
- To temporarily stop logging: click on the Log button, and choose Suspend
- To resume: click on the Log button, and choose Resume
- To stop logging and close the file: click on the Log button, and choose Close
- To print previous or current log: select File > View..., choose file, right-click on the Viewer, and select Print
- www.stata.com/manuals13/u15.pdf

How to save your commands – Do files

- Rather than typing commands at the keyboard, you can create a text file containing commands and instruct Stata to execute the commands stored in that file.

- Such files are called Do-files because the command that causes them to be executed is do

- To create a Do file:
  - Click the "New Do file editor"
  - Type in your commands
  - Save the file\

- To execute a do file:
  - Type: do and then add the path of the do file
    - E.g. "C:\Users\user\Desktop\Untitled.do"
  - Or File → Do
  - Or click the button "Execute" in the Do file editor window

https://www.stata.com/manuals13/u16.pdf

How to save your data – Data files (.dta)

- To save an unnamed dataset (or an old dataset under a new name):
  1. select File > Save as...;
  2. OR type "save filename" in the Command window
- To save a dataset that has been changed (overwriting the original data file)
  1. select File > Save;
  2. OR click on the Save button;
  3. OR type "save, replace" in the Command window.
- To open a Stata dataset:
  1. Double-click on a Stata data file, which is a file whose extension is .dta.
  2. OR Select File > Open... or click on the Open button and navigate to the file.
  3. OR type "use filename" in the Command window

  www.stata.com/manuals/gsw5.pdf

Data

- **Types of Data**

a. Numeric data (i.e. number)

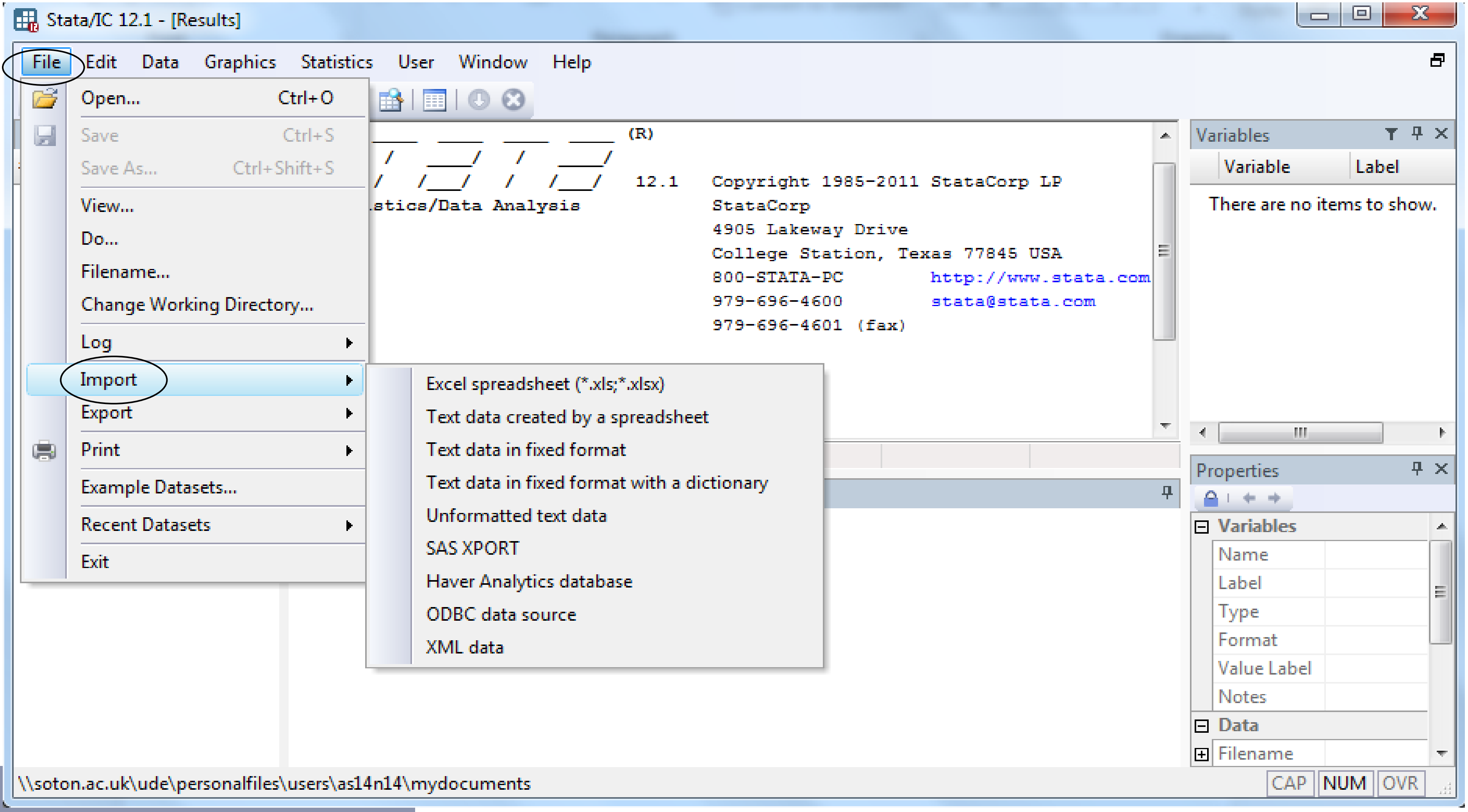
b. String data(i.e. text)

- **Missing Values**
- For numeric data: single dot (.)
- For string data: double quotes (“ ”) or dot double quotes (“. ”)

- **Useful commands for changing string into numeric or other type and vice versa:**
  - encode (www.stata.com/manuals/dencode.pdf)
  - destring (www.stata.com/manuals/u24.pdf#u24.2Categoricalstringvariables)
  - format (www.stata.com/manuals/dformat.pdf)

Workfile

2. Go to File ⟶ Import ⟶ .....



```
 ____  ____  ____  ____  (R)
/ __/ / __/ / __/ / __/
/ /__/ / /__/  12.1   Copyright 1985-2011 StataCorp LP
stics/Data Analysis           StataCorp
                              4905 Lakeway Drive
                              College Station, Texas 77845 USA
                              800-STATA-PC        http://www.stata.com
                              979-696-4600        stata@stata.com
                              979-696-4601 (fax)
```

\\soton.ac.uk\ude\personalfiles\users\as14n14\mydocuments

Empirical Example 1 :

Cross Sectional Data

Empirical Example 1 – *Cross Sectional Data(1)*

1. Go to folder *Empirical Examples* ⟶ *Example_1*

- Shows the results of a test in Finance for 60 students

(Source: “Econometrics for Financial Analysis”, A. G. Merikas, A. A. Merika)

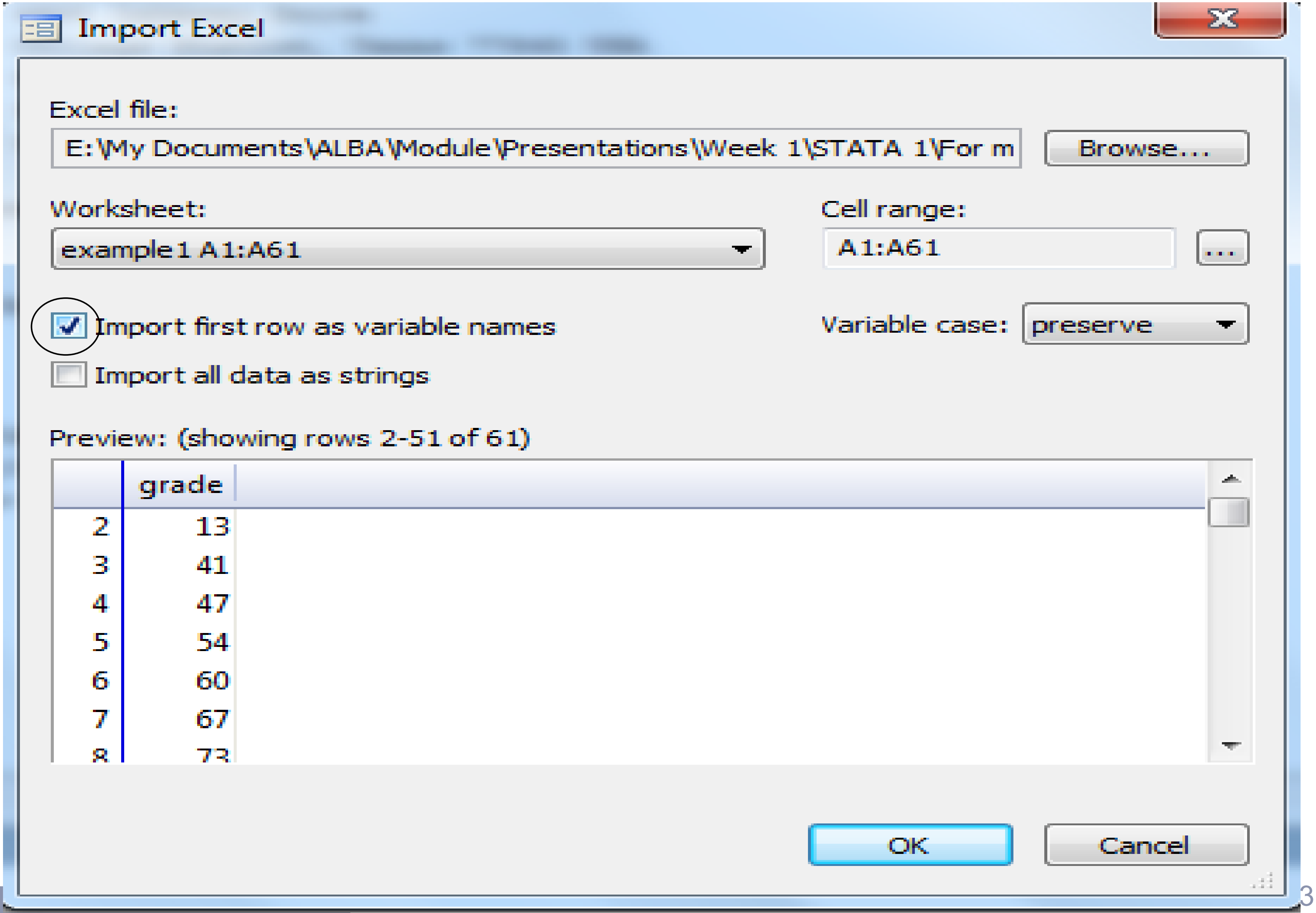
2. Open xlsx file: *example_1.xlsx*

3. Define the type of the data : *Cross Sectional Data*

4. Define the number of observations of the sample: *60*

5. Close xlsx file

Empirical Example 1 – *Cross Sectional Data(2)*



Excel file: E:\My Documents\ALBA\Module\Presentations\Week 1\STATA 1\For m

Worksheet: example1 A1:A61

Cell range: A1:A61

☑ Import first row as variable names

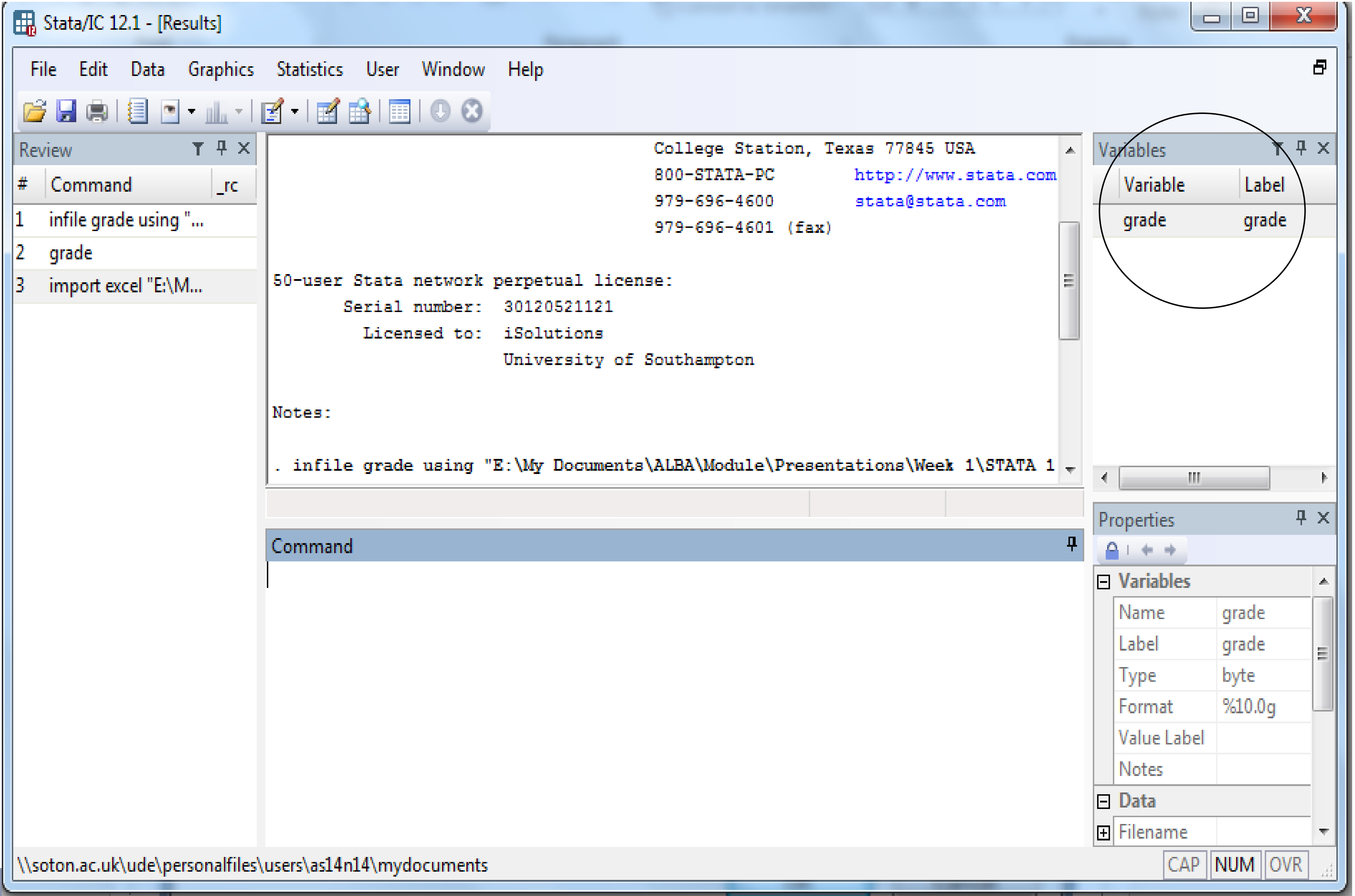
Variable case: preserve

□ Import all data as strings

Preview: (showing rows 2-51 of 61)

	grade
2	13
3	41
4	47
5	54
6	60
7	67
8	73

Empirical Example 1 – *Cross Sectional Data(3)*


Stata/IC 12.1 - [Results]
File Edit Data Graphics Statistics User Window Help
Review
Command
Variables
Properties
CAP NUM OVR

#	Command	_rc
1	infile grade using "...	
2	grade	
3	import excel "E:\M...	

```
                                        College Station, Texas 77845 USA
                                        800-STATA-PC        http://www.stata.com
                                        979-696-4600        stata@stata.com
                                        979-696-4601 (fax)

50-user Stata network perpetual license:
       Serial number:  30120521121
         Licensed to:  iSolutions
                       University of Southampton

Notes:

. infile grade using "E:\My Documents\ALBA\Module\Presentations\Week 1\STATA 1
```

Variable	Label
grade	grade

Variables

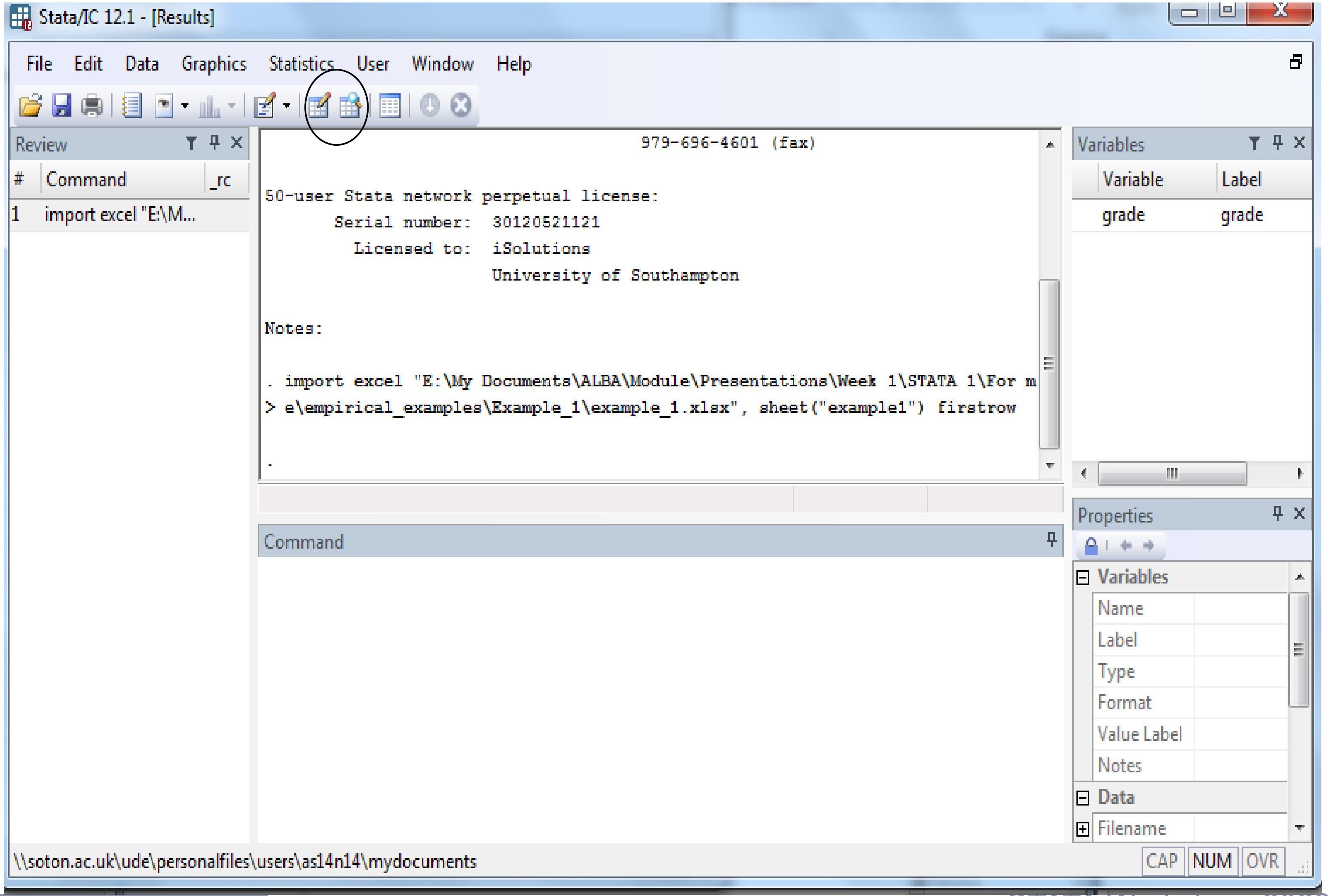
Name	grade
Label	grade
Type	byte
Format	%10.0g
Value Label	
Notes	

Data

Filename

\\soton.ac.uk\ude\personalfiles\users\as14n14\mydocuments

Empirical Example 1 – *Cross Sectional Data(4)*



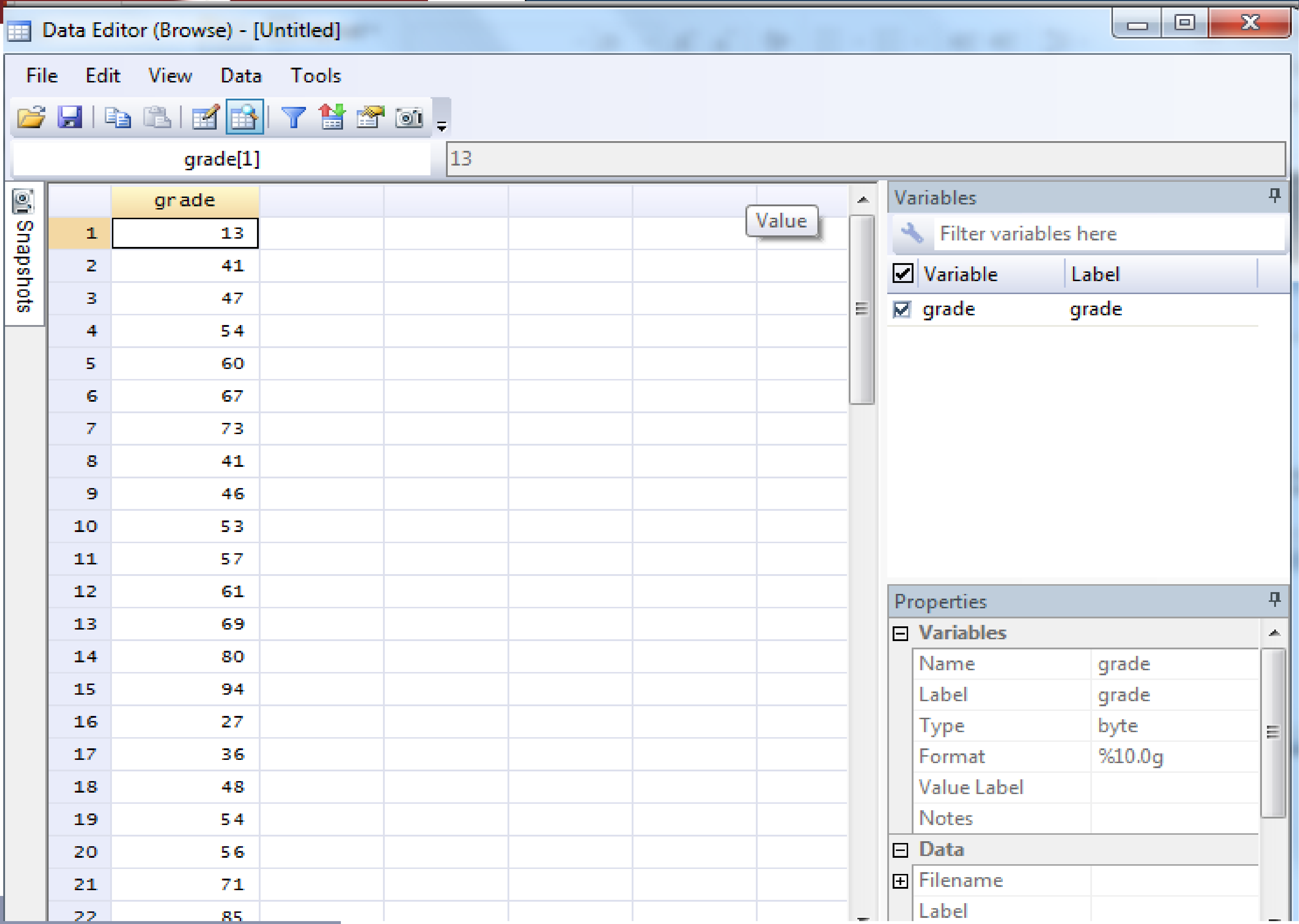
#	Command	_rc
1	import excel "E:\M...	

```
                                             979-696-4601 (fax)

50-user Stata network perpetual license:
       Serial number:  30120521121
         Licensed to:  iSolutions
                       University of Southampton

Notes:

. import excel "E:\My Documents\ALBA\Module\Presentations\Week 1\STATA 1\For m
> e\empirical_examples\Example_1\example_1.xlsx", sheet("example1") firstrow

.
```

Variable	Label
grade	grade

\\soton.ac.uk\ude\personalfiles\users\as14n14\mydocuments

Empirical Example 1 – *Cross Sectional Data(5)*



Data Editor (Browse) - [Untitled]

grade[1]   13

	grade
1	13
2	41
3	47
4	54
5	60
6	67
7	73
8	41
9	46
10	53
11	57
12	61
13	69
14	80
15	94
16	27
17	36
18	48
19	54
20	56
21	71
22	85

Variables

☑ Variable	Label
☑ grade	grade

Properties

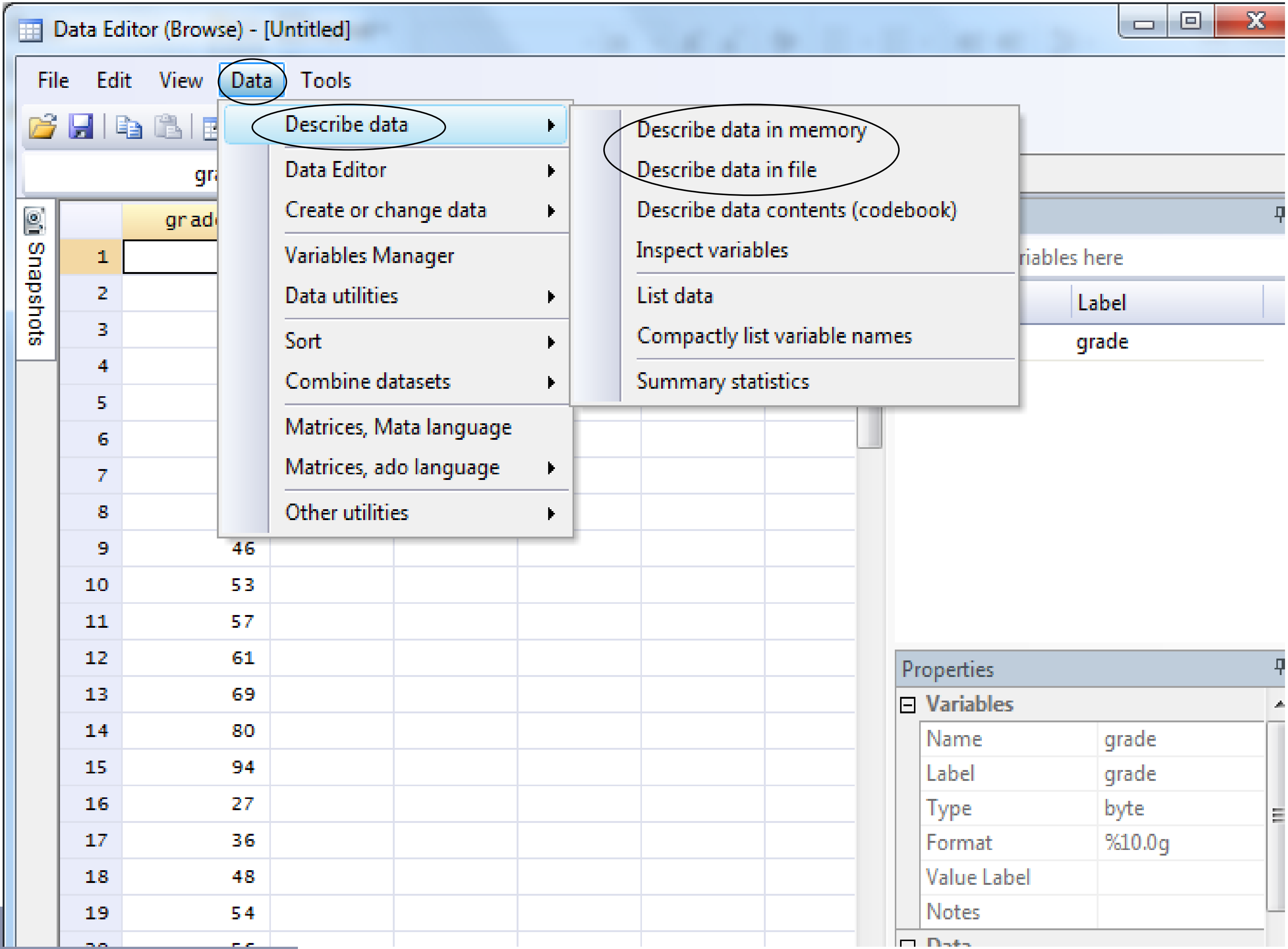
⊟ Variables

Name	grade
Label	grade
Type	byte
Format	%10.0g
Value Label	
Notes	

⊟ Data

⊞ Filename	
Label	

Empirical Example 1 – *Cross Sectional Data(6)*

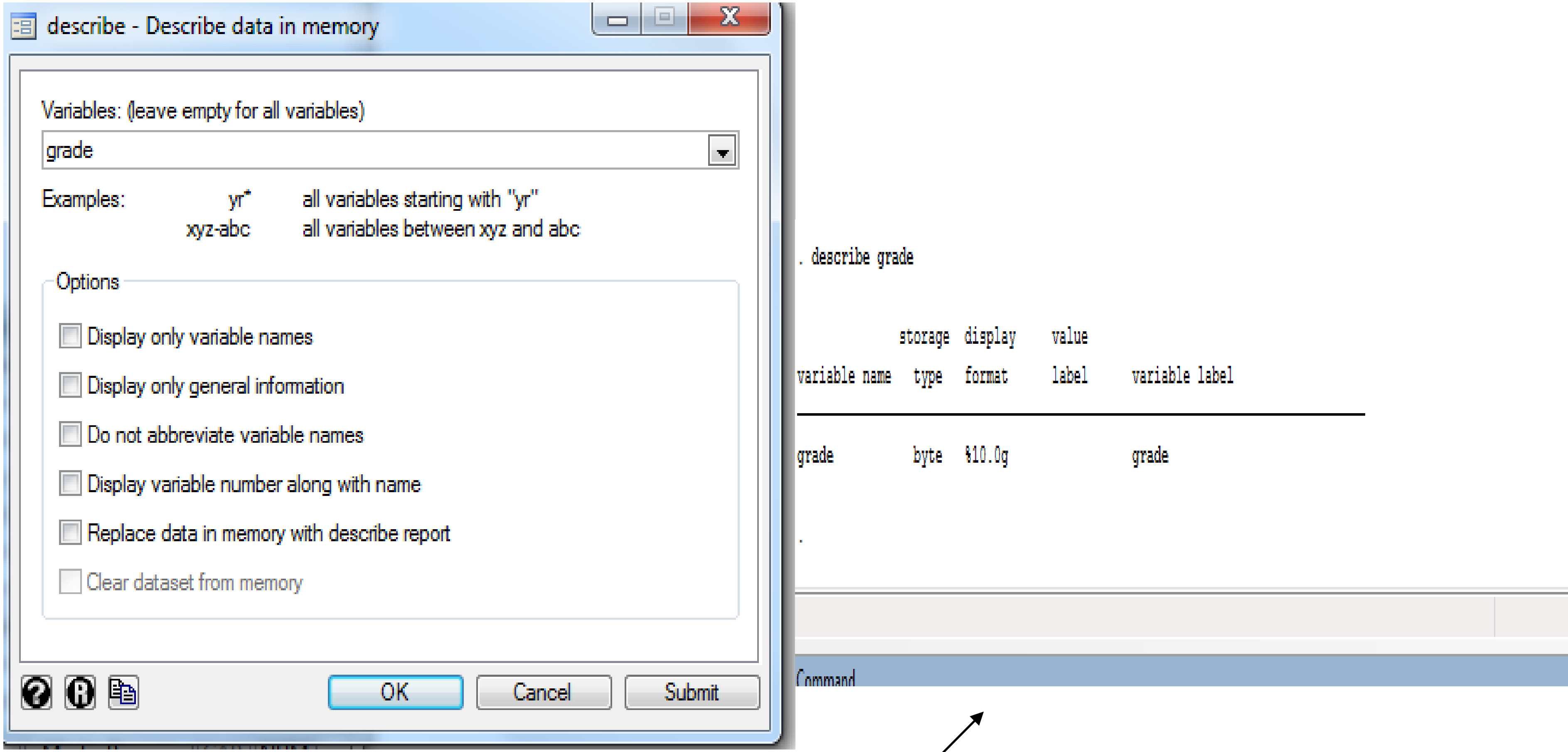


	grade
1	
2	
3	
4	
5	
6	
7	
8	
9	46
10	53
11	57
12	61
13	69
14	80
15	94
16	27
17	36
18	48
19	54

Name	grade
Label	grade
Type	byte
Format	%10.0g
Value Label	
Notes	

Empirical Example 1 – *Cross Sectional Data(7)*

Write the name of the variable (here is grade)



describe - Describe data in memory

Variables: (leave empty for all variables)

grade

Examples: yr* all variables starting with "yr"
xyz-abc all variables between xyz and abc

Options

□ Display only variable names

□ Display only general information

□ Do not abbreviate variable names

□ Display variable number along with name

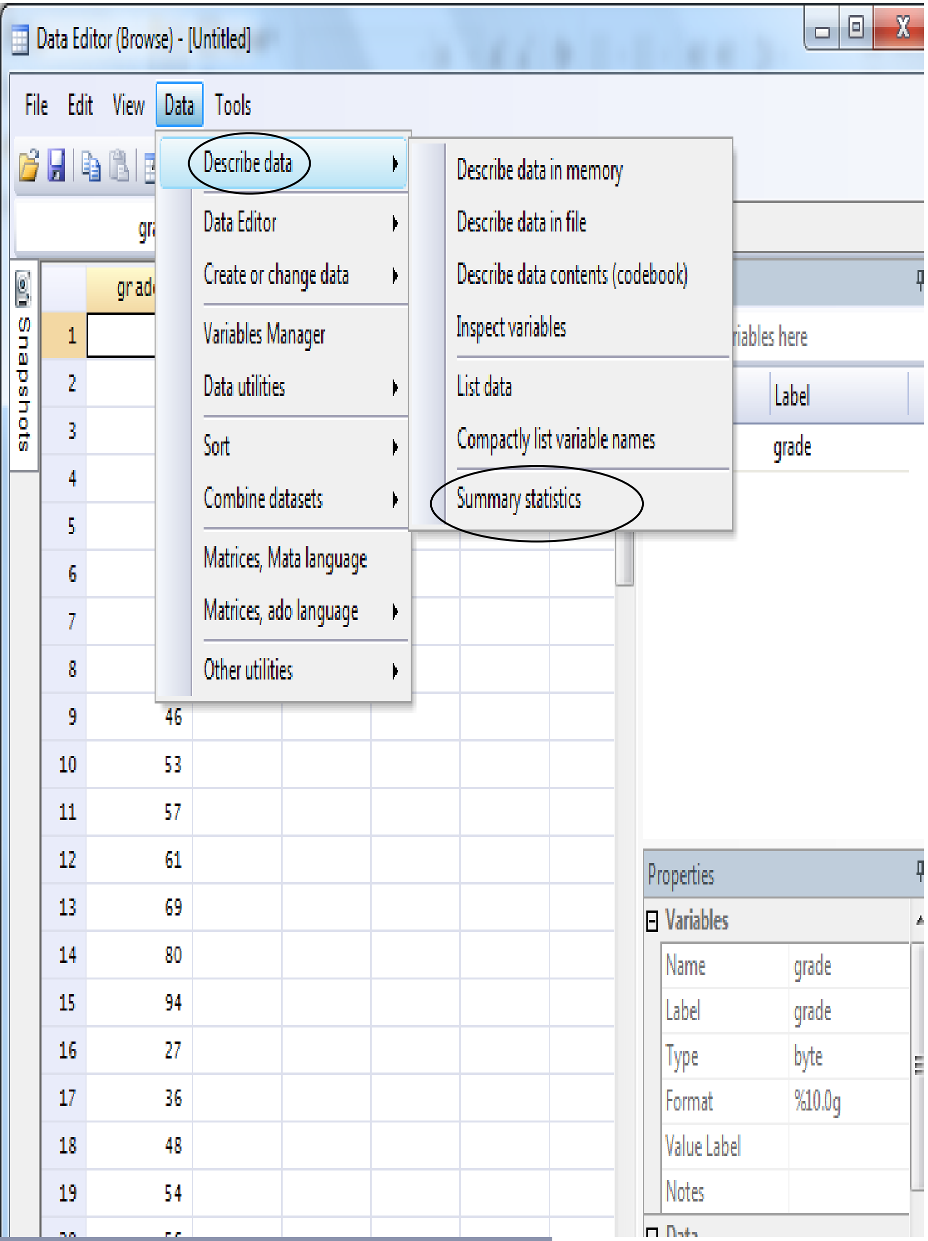
□ Replace data in memory with describe report

□ Clear dataset from memory

```
. describe grade

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------
grade           byte   %10.0g                 grade

.
```

Output Window

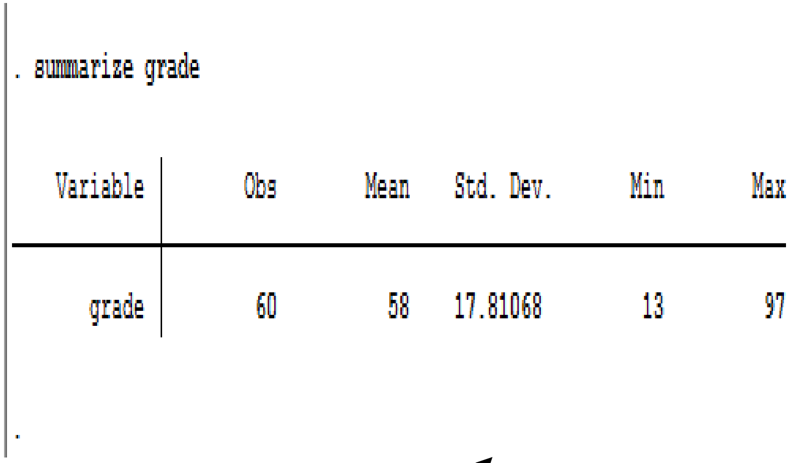
Empirical Example 1 – *Cross Sectional Data(8)*



	grade
1	
2	
3	
4	
5	
6	
7	
8	
9	46
10	53
11	57
12	61
13	69
14	80
15	94
16	27
17	36
18	48
19	54

Name	Label
grade	grade

Name	grade
Label	grade
Type	byte
Format	%10.0g
Value Label	
Notes	

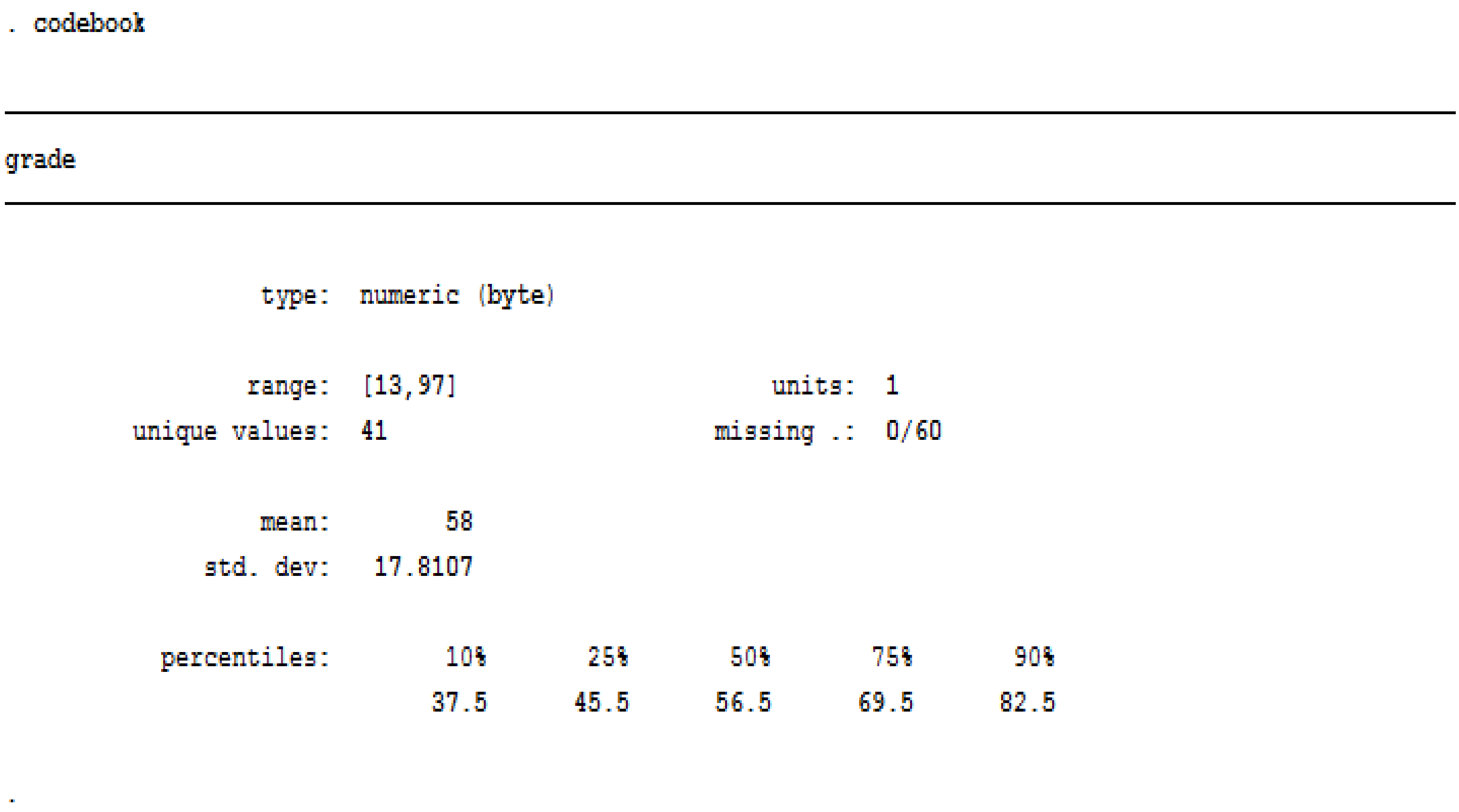


```
. summarize grade

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       grade |        60          58    17.81068         13         97

.
```

Output Window

Empirical Example 1 – *Cross Sectional Data(9)*

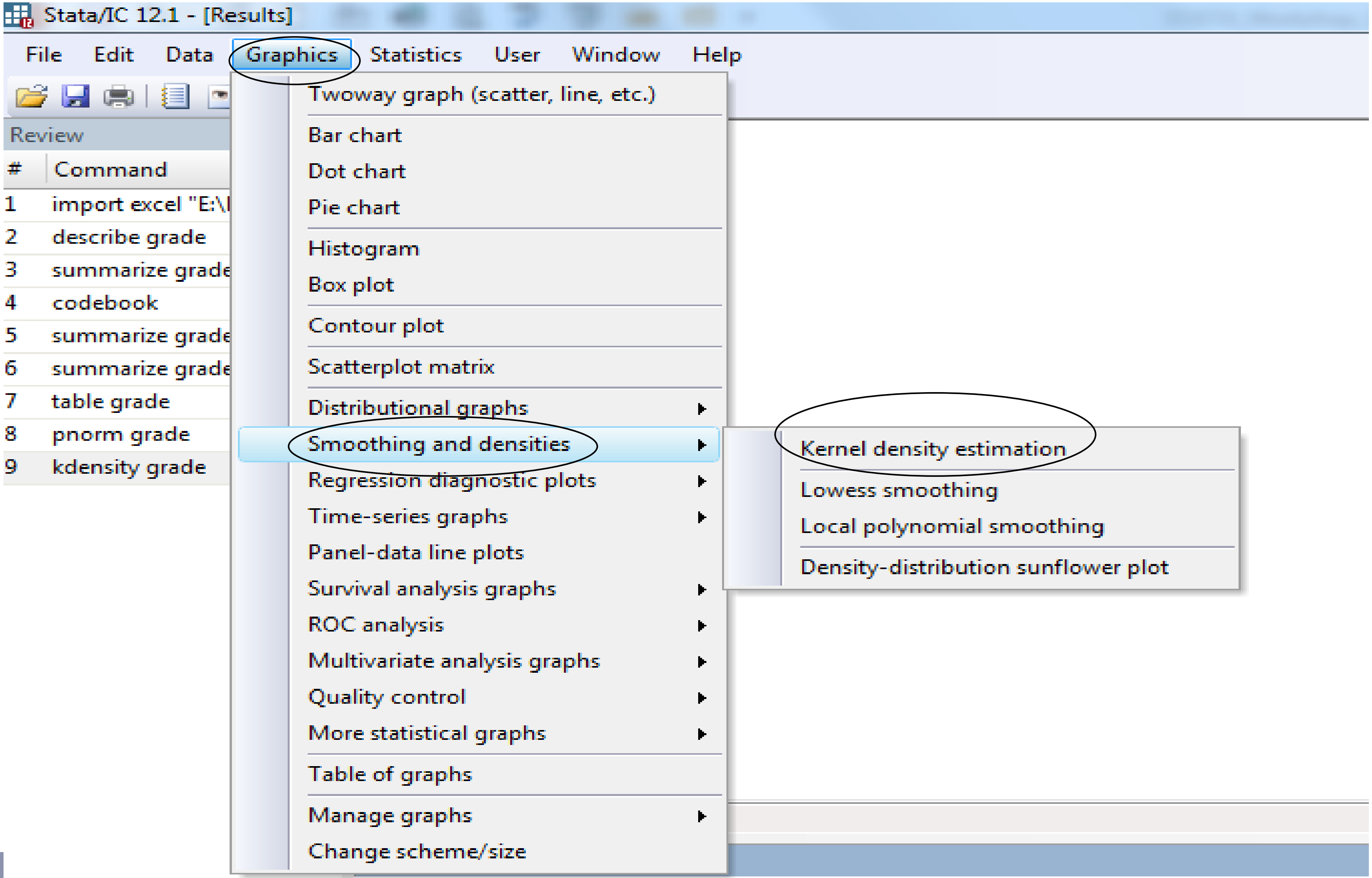
- Write codebook in the command window



```
. codebook

--------------------------------------------------------------------------------
grade
--------------------------------------------------------------------------------

                  type:  numeric (byte)

                 range:  [13,97]                      units:  1
         unique values:  41                       missing .:  0/60

                  mean:        58
             std. dev:    17.8107

           percentiles:        10%       25%       50%       75%       90%
                              37.5      45.5      56.5      69.5      82.5

.
```

Empirical Example 1 – *Cross Sectional Data(10)*

- Graph

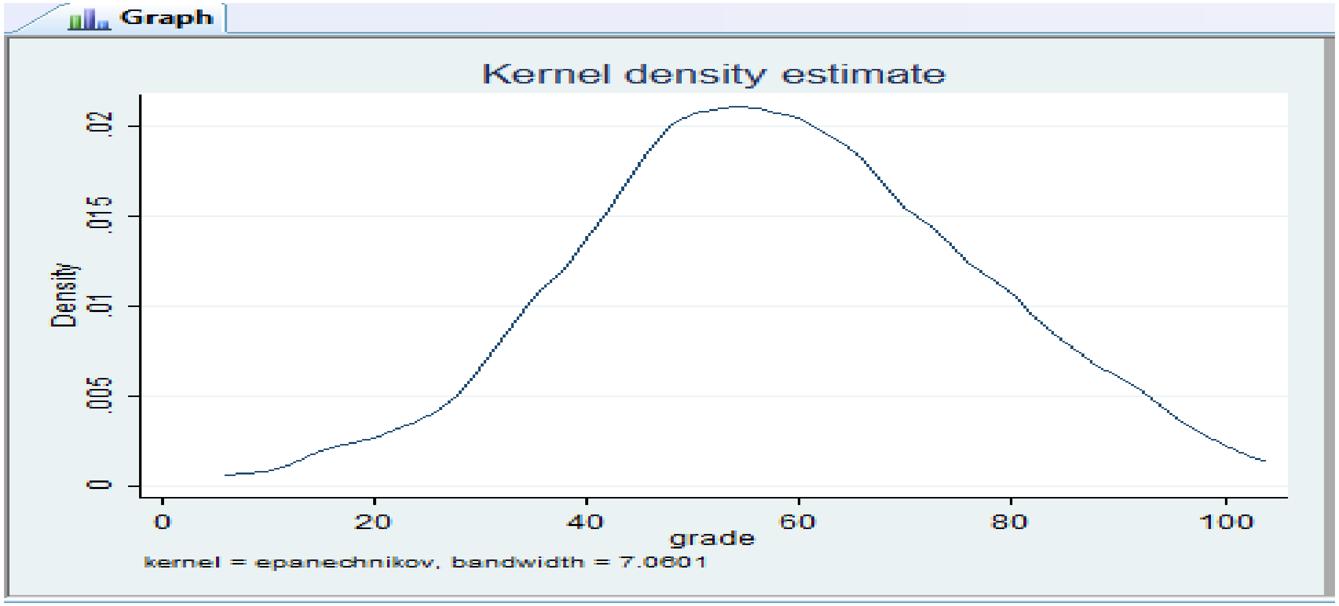


```
#  Command
1  import excel "E:\I
2  describe grade
3  summarize grade
4  codebook
5  summarize grade
6  summarize grade
7  table grade
8  pnorm grade
9  kdensity grade
```

Empirical Example 1 – ***Cross Sectional Data(11)***

We can visualize the shape of distribution

- Distinction between normal and non-normal distributions



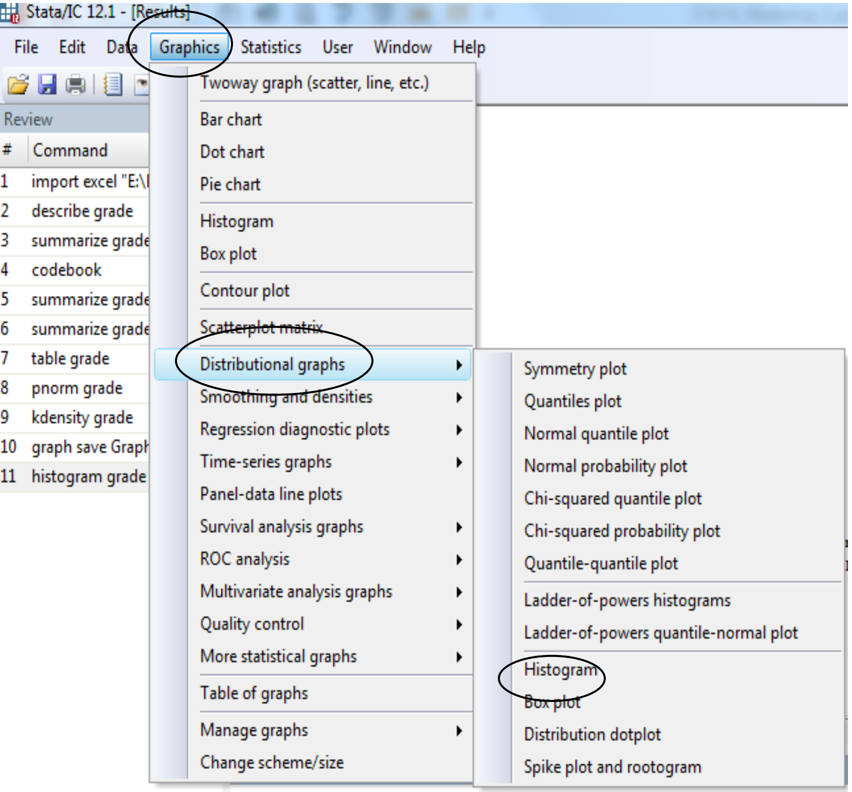
kernel = epanechnikov, bandwidth = 7.0601

The above diagram shows the probability density function (pdf) of the variable “grade”. In a first view resembles the “normal distribution” with pdf function :

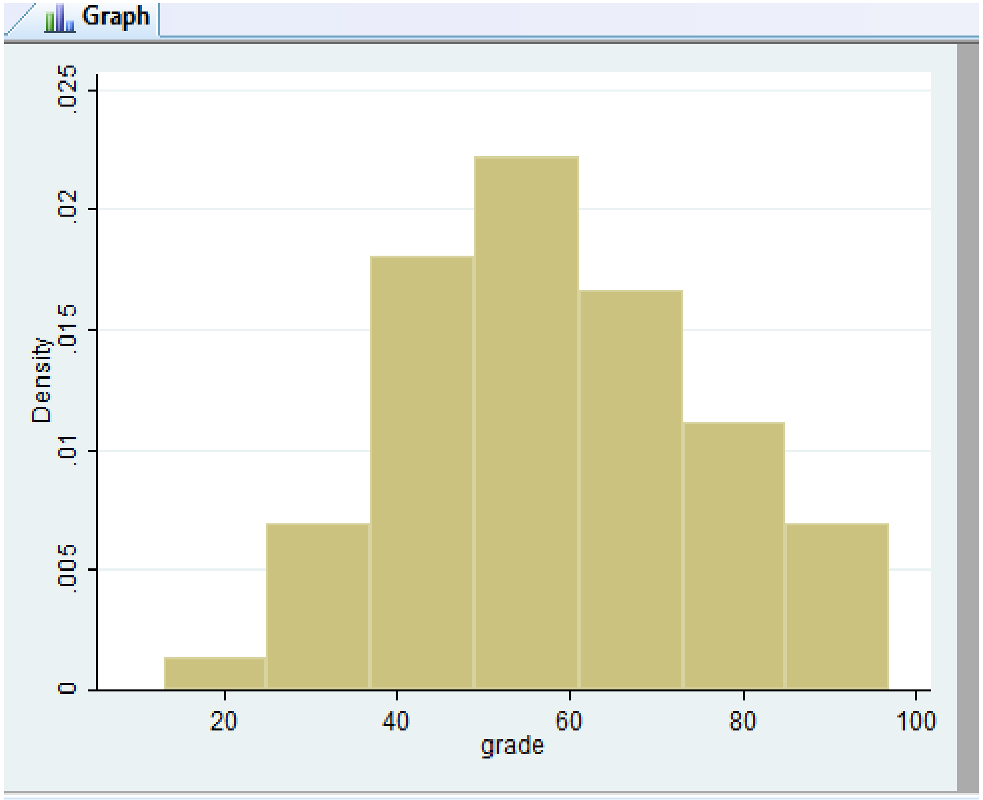
$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \right]$$

Empirical Example 1 – *Cross Sectional Data(12)*

Go to Graphics ⟶ Distributional Graphs ⟶ Histogram



```
#   Command
1   import excel "E:\l
2   describe grade
3   summarize grade
4   codebook
5   summarize grade
6   summarize grade
7   table grade
8   pnorm grade
9   kdensity grade
10  graph save Graph
11  histogram grade
```



Close the above window ⟶ Go to *File* ⟶ *Save as…*

Empirical Example 2 :

Time Series Data

Empirical Example 2- *Time Series Data (1)*

1. Go to folder *Empirical Examples* ⟶ *Example_2*

- Shows the monthly total simple returns(capital + dividends) in $ of the equity indices of G7 countries from 31/01/1980 – 31/10/2012 .

(Source : DataStream)

2. Open .xlsx file: *example_2.txt*

3. Define the type of the data : *Time series data*

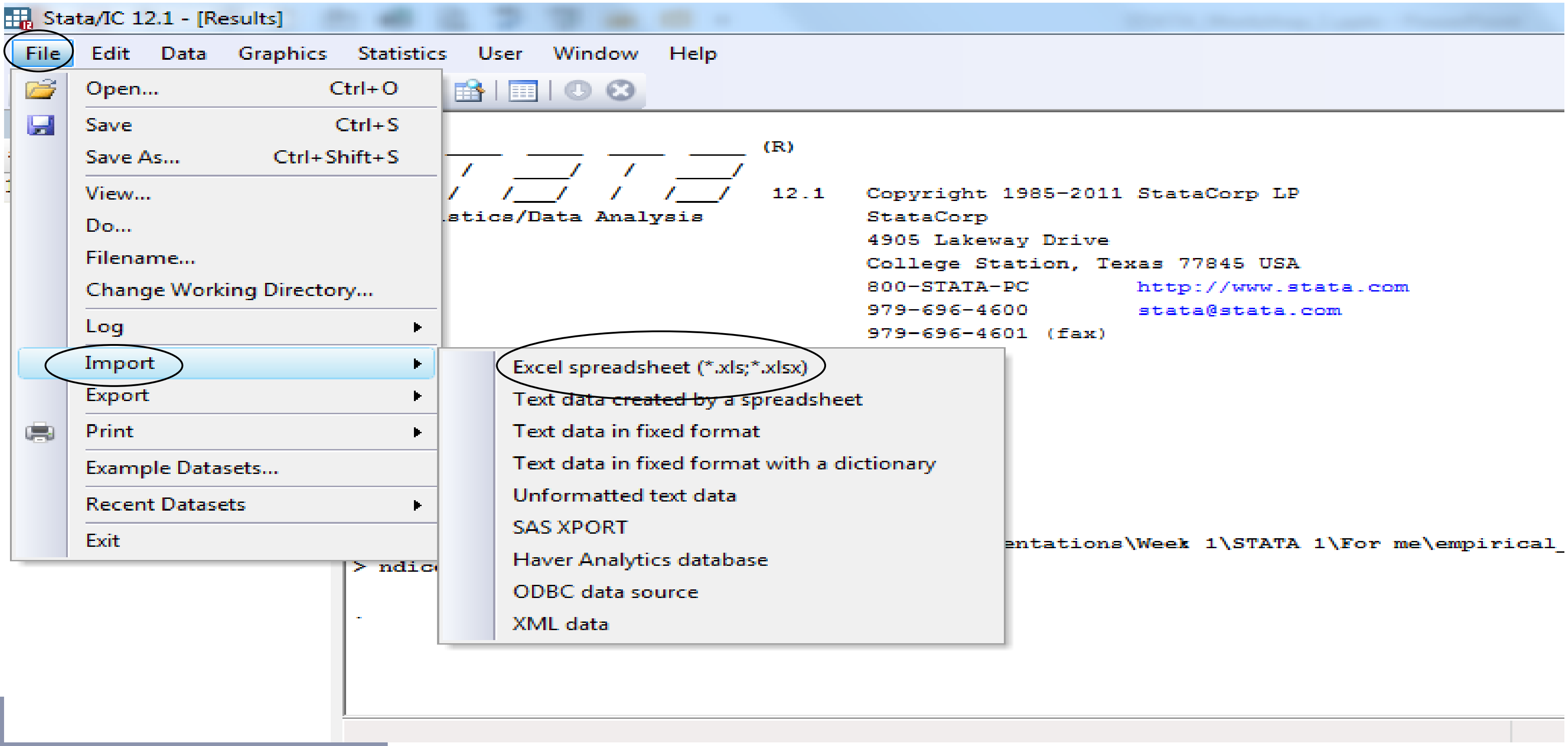
4. Close .xlsx file

Empirical Example 2- *Time Series Data (2)*

5 . Open *STATA* from PC - lab

6. *Go to File* ⟶ *Import* ⟶ Excel Spreadsheet

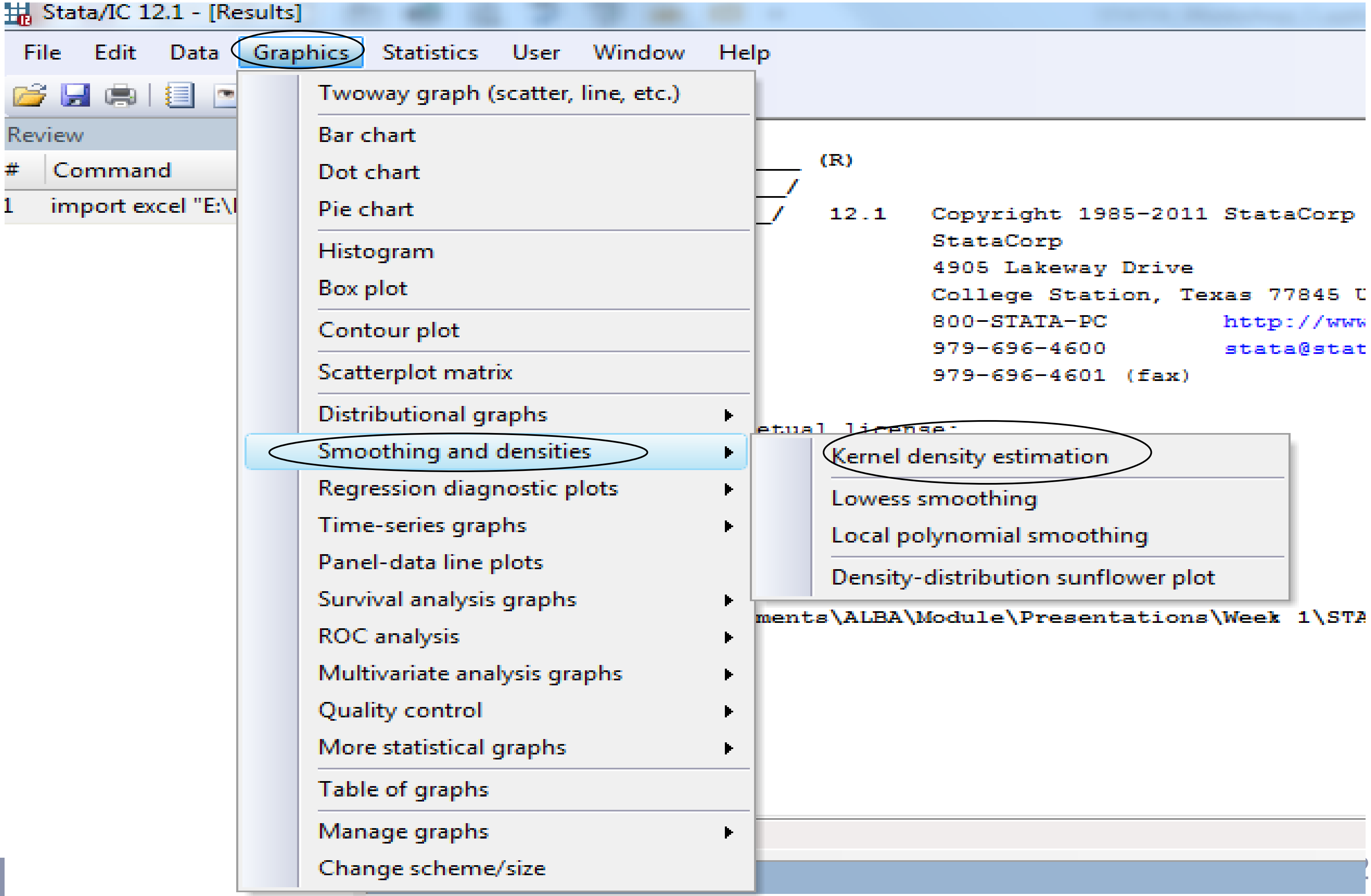
7. Browse example_2.xlsx



```
                          (R)
                            12.1   Copyright 1985-2011 StataCorp LP
stics/Data Analysis                StataCorp
                                   4905 Lakeway Drive
                                   College Station, Texas 77845 USA
                                   800-STATA-PC        http://www.stata.com
                                   979-696-4600        stata@stata.com
                                   979-696-4601 (fax)

                                 entations\Week 1\STATA 1\For me\empirical_
> ndic
.
```

Empirical Example 2- *Time Series Data (3)*

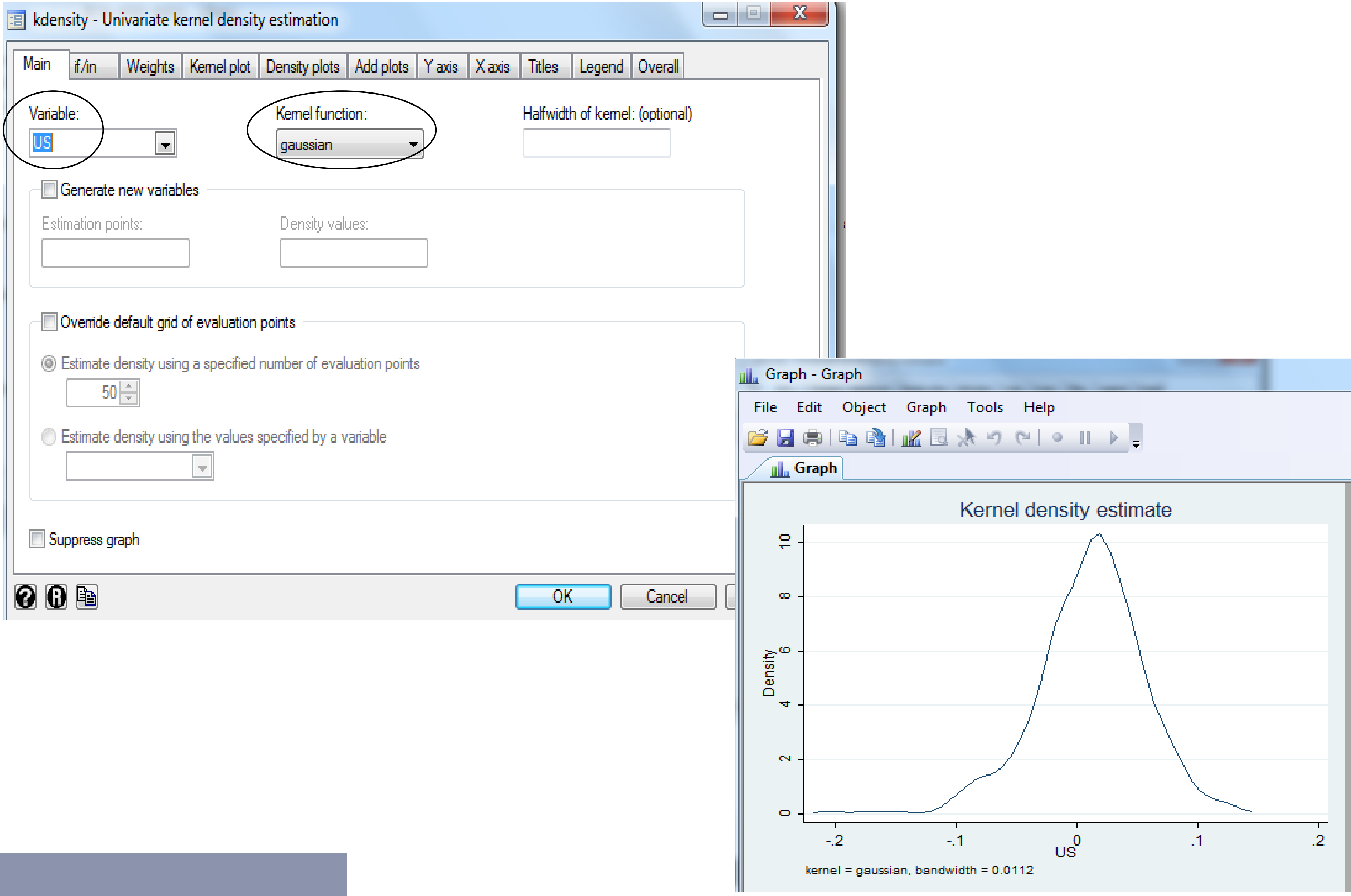
8. Go to Graphics→ smoothing and densities → Kernel Density Function



```
#  Command
1  import excel "E:\
```

```
                  (R)
   /
  /     12.1    Copyright 1985-2011 StataCorp
                StataCorp
                4905 Lakeway Drive
                College Station, Texas 77845 U
                800-STATA-PC          http://www
                979-696-4600          stata@stat
                979-696-4601 (fax)

etual license:

ments\ALBA\Module\Presentations\Week 1\STA
```

Empirical Example 2- *Time Series Data (4)*



kdensity - Univariate kernel density estimation

Main | if/in | Weights | Kernel plot | Density plots | Add plots | Y axis | X axis | Titles | Legend | Overall

Variable: US

Kernel function: gaussian

Halfwidth of kernel: (optional)

□ Generate new variables

Estimation points:

Density values:

□ Override default grid of evaluation points

● Estimate density using a specified number of evaluation points: 50

○ Estimate density using the values specified by a variable

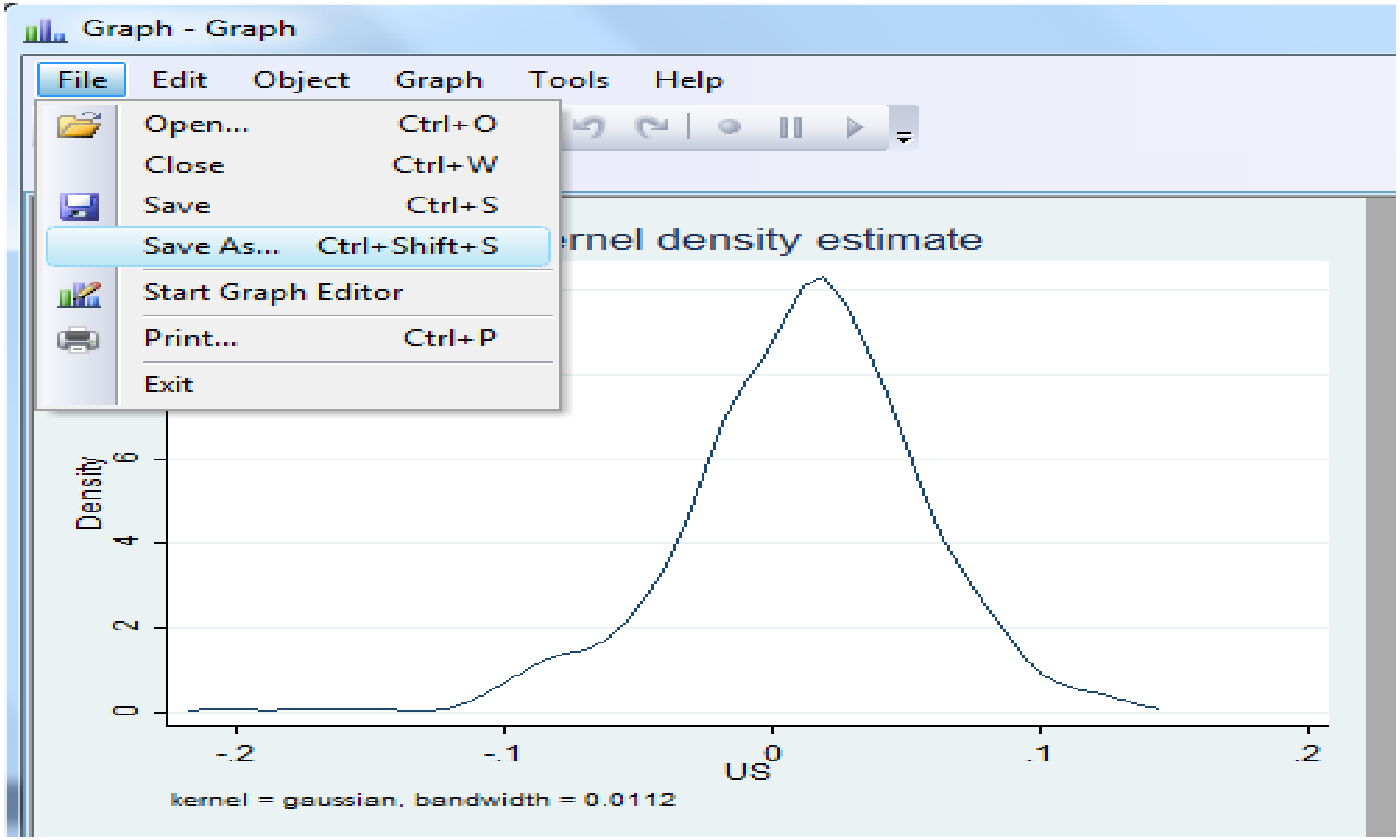
□ Suppress graph

OK | Cancel

Graph - Graph

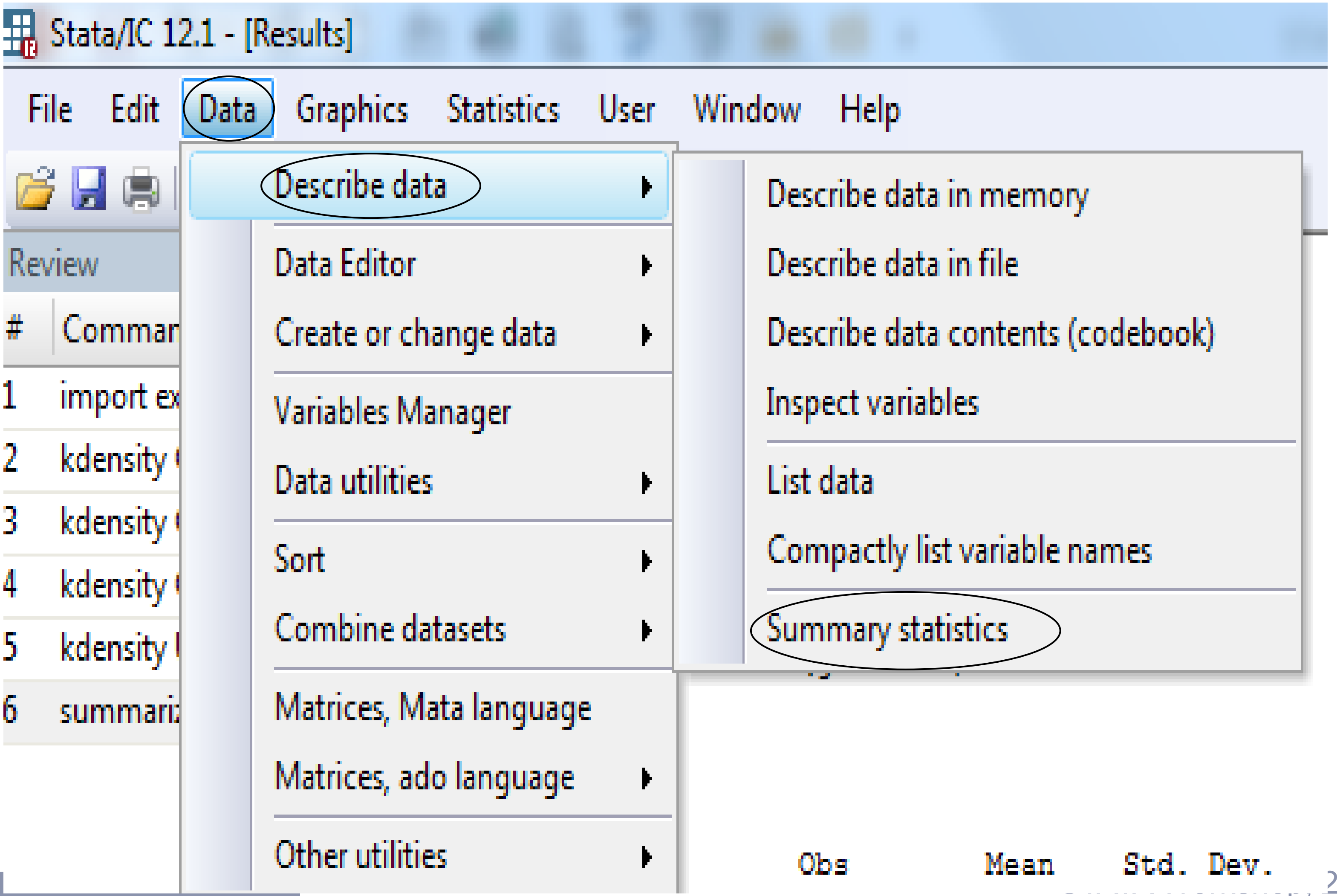
kernel = gaussian, bandwidth = 0.0112

9. File ⟶ Save as ⟶



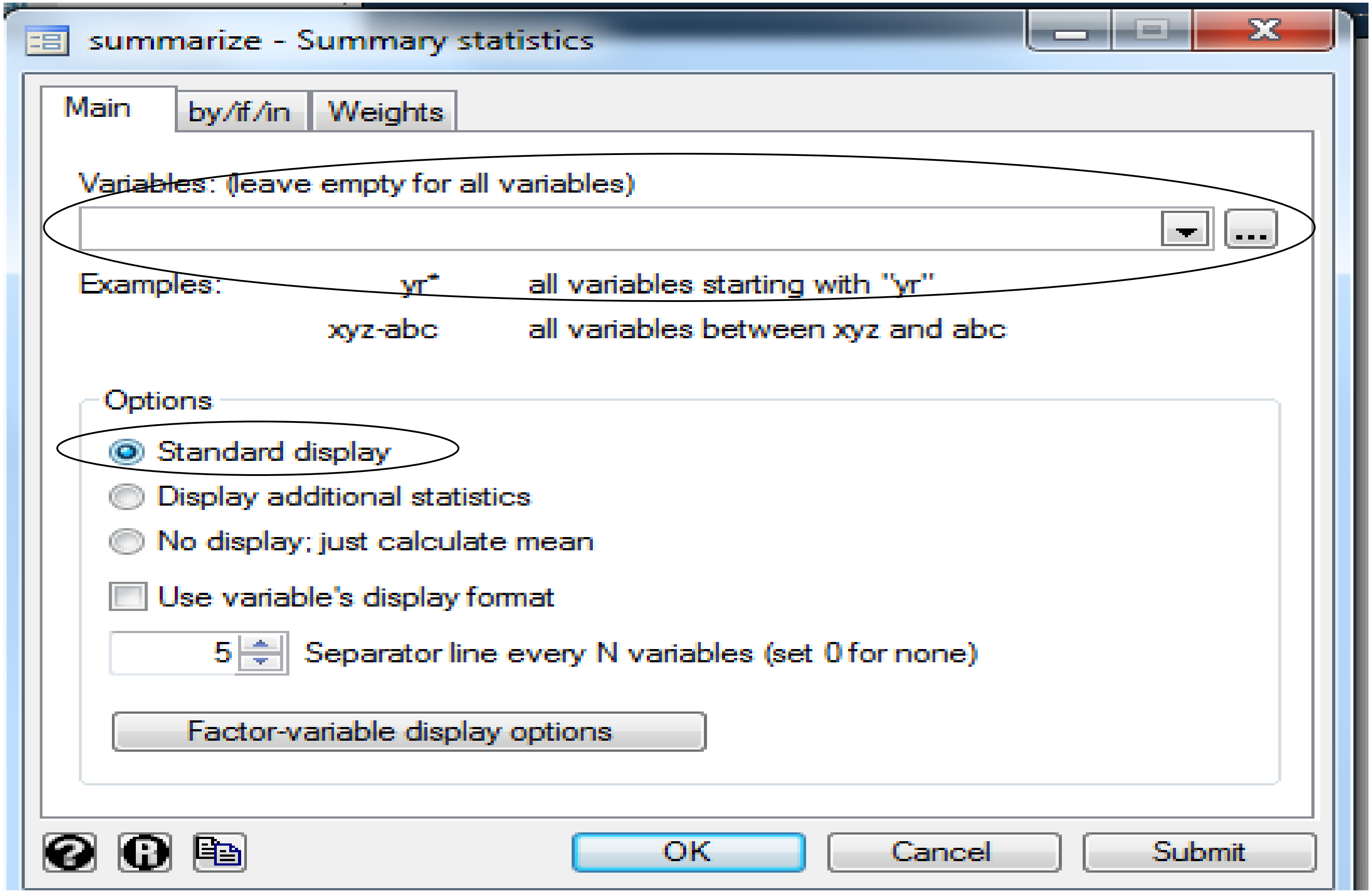
Empirical Example 2- *Time Series Data (6)*

10. Data → Describe data → Summary Statistics



```
         Obs        Mean    Std. Dev.
```

Empirical Example 2- *Time Series Data (7)*



summarize - Summary statistics

Variables: (leave empty for all variables)

Examples: yr* all variables starting with "yr"

xyz-abc all variables between xyz and abc

Options

● Standard display

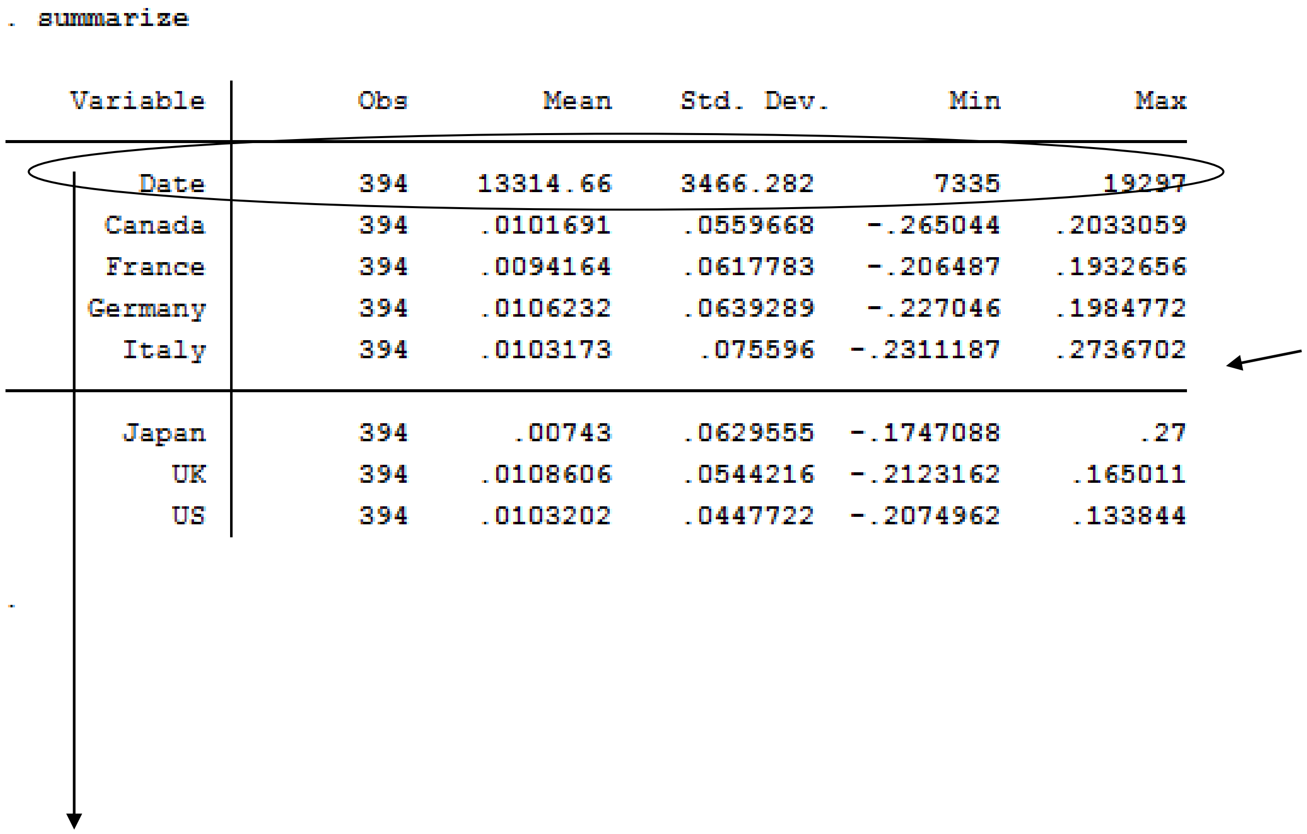
○ Display additional statistics

○ No display; just calculate mean

□ Use variable's display format

5 Separator line every N variables (set 0 for none)

Empirical Example 2- *Time Series Data (8)*

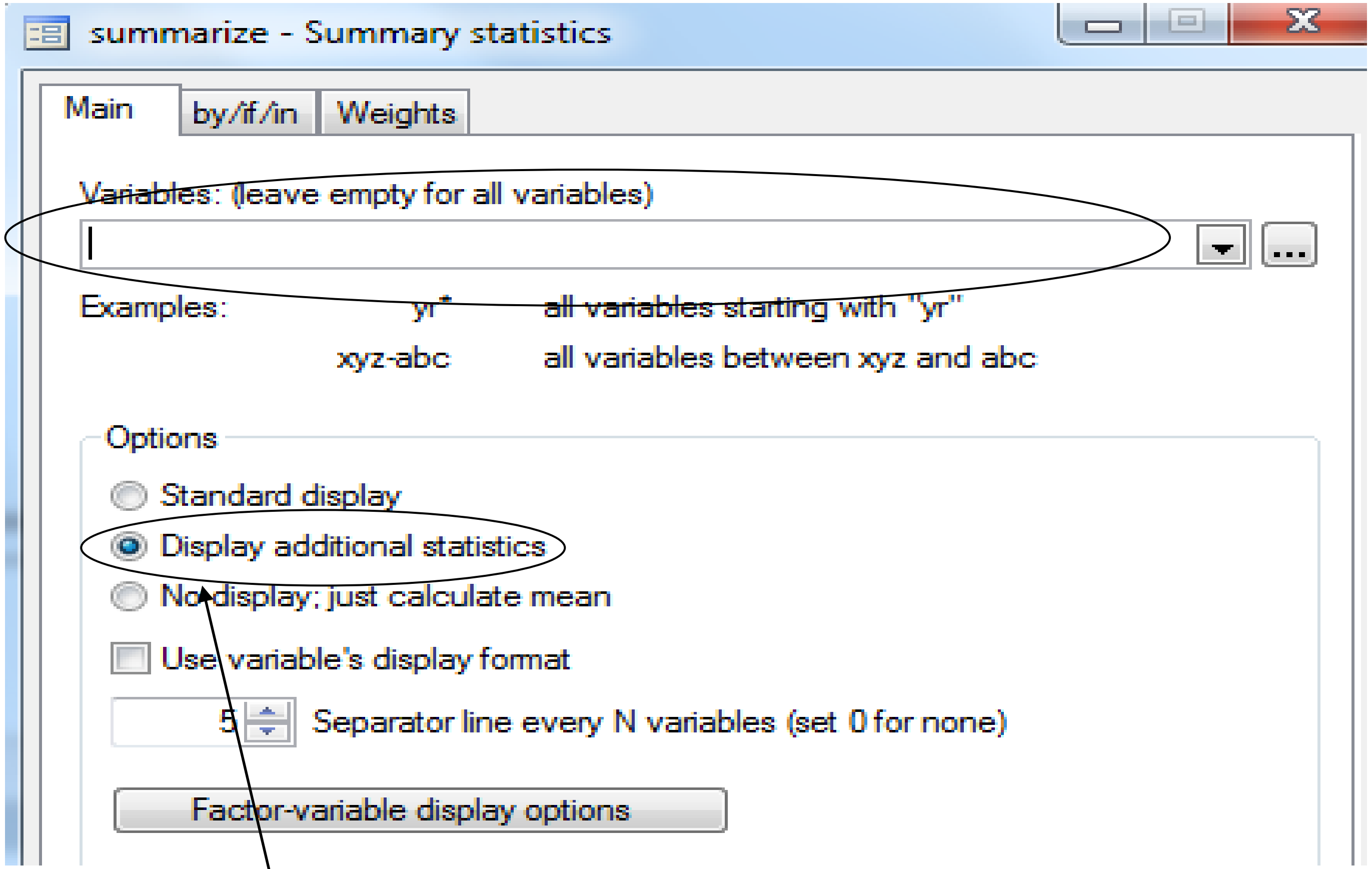


```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        Date |       394    13314.66    3466.282       7335      19297
      Canada |       394    .0101691    .0559668   -.265044   .2033059
      France |       394    .0094164    .0617783   -.206487   .1932656
     Germany |       394    .0106232    .0639289   -.227046   .1984772
       Italy |       394    .0103173     .075596  -.2311187   .2736702
-------------+--------------------------------------------------------
       Japan |       394      .00743    .0629555  -.1747088        .27
          UK |       394    .0108606    .0544216  -.2123162    .165011
          US |       394    .0103202    .0447722  -.2074962    .133844

.
```

Command Window

Go back to the ‘Summarize’ Window

Empirical Example 2- *Time Series Data (9)*



summarize - Summary statistics

Variables: (leave empty for all variables)

Examples: yr* all variables starting with "yr"

xyz-abc all variables between xyz and abc

Options

○ Standard display

● Display additional statistics

○ No display; just calculate mean

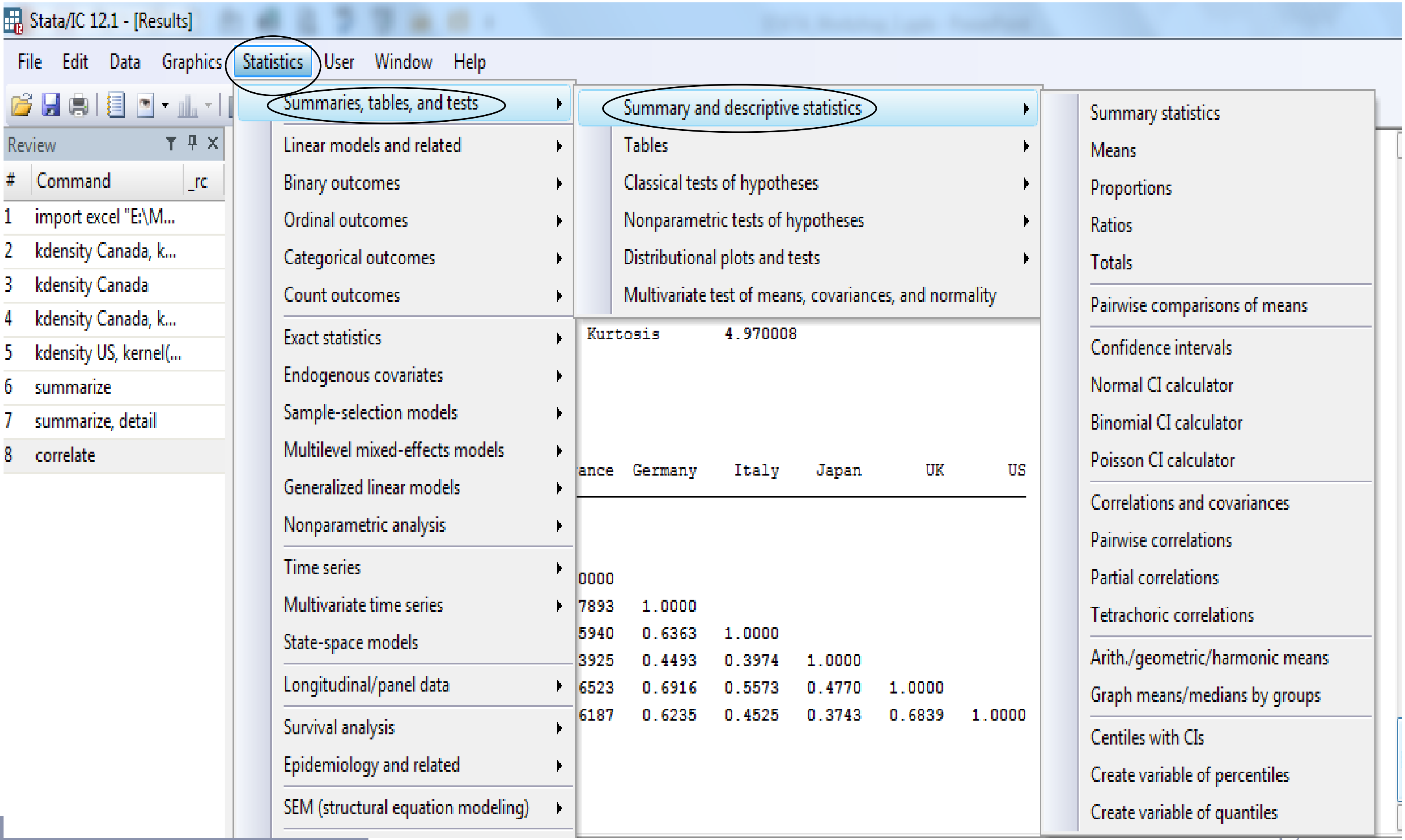
□ Use variable's display format

5 Separator line every N variables (set 0 for none)

For additional statistics measures

Empirical Example 2- *Time Series Data (10)*

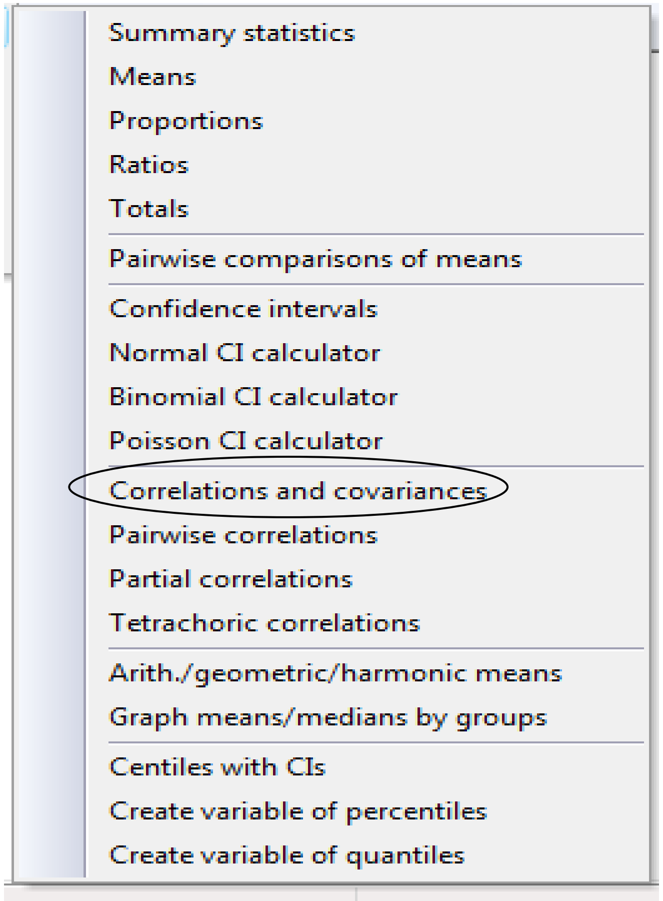
11. Statistics ⟶ Summaries…⟶ Summary and descriptive Statistics

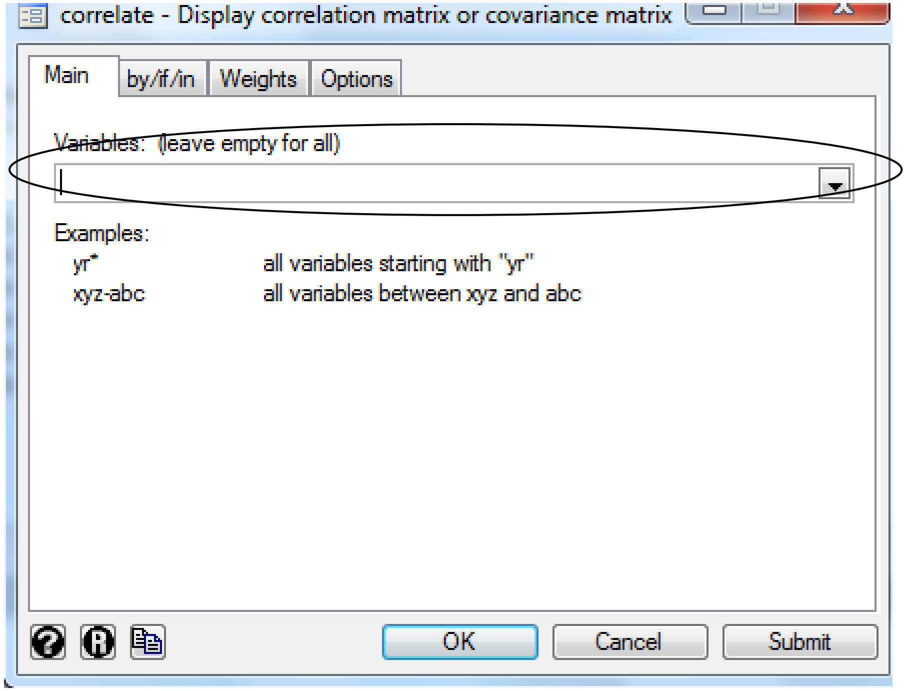


```
#  Command
1  import excel "E:\M...
2  kdensity Canada, k...
3  kdensity Canada
4  kdensity Canada, k...
5  kdensity US, kernel(...
6  summarize
7  summarize, detail
8  correlate
```

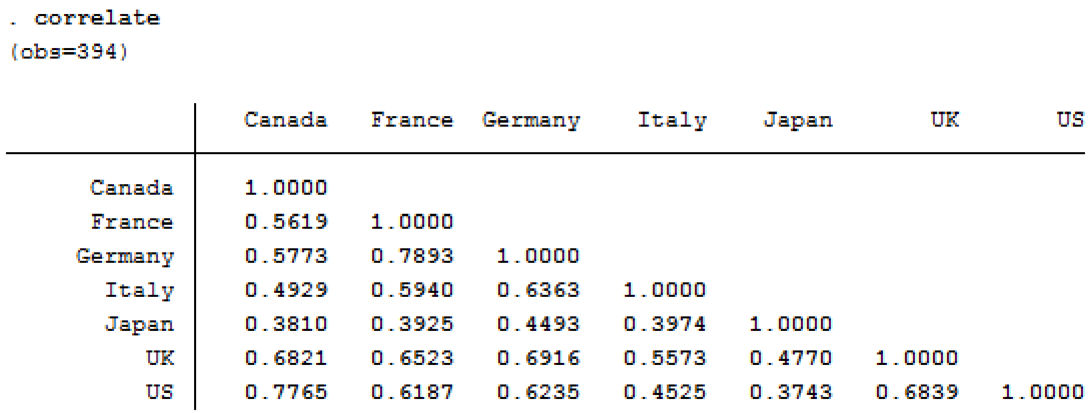
```
Kurtosis       4.970008

ance  Germany    Italy    Japan       UK       US
-------------------------------------------------

0000
7893   1.0000
5940   0.6363   1.0000
3925   0.4493   0.3974   1.0000
6523   0.6916   0.5573   0.4770   1.0000
6187   0.6235   0.4525   0.3743   0.6839   1.0000
```

12. Select Correlations and Covariances



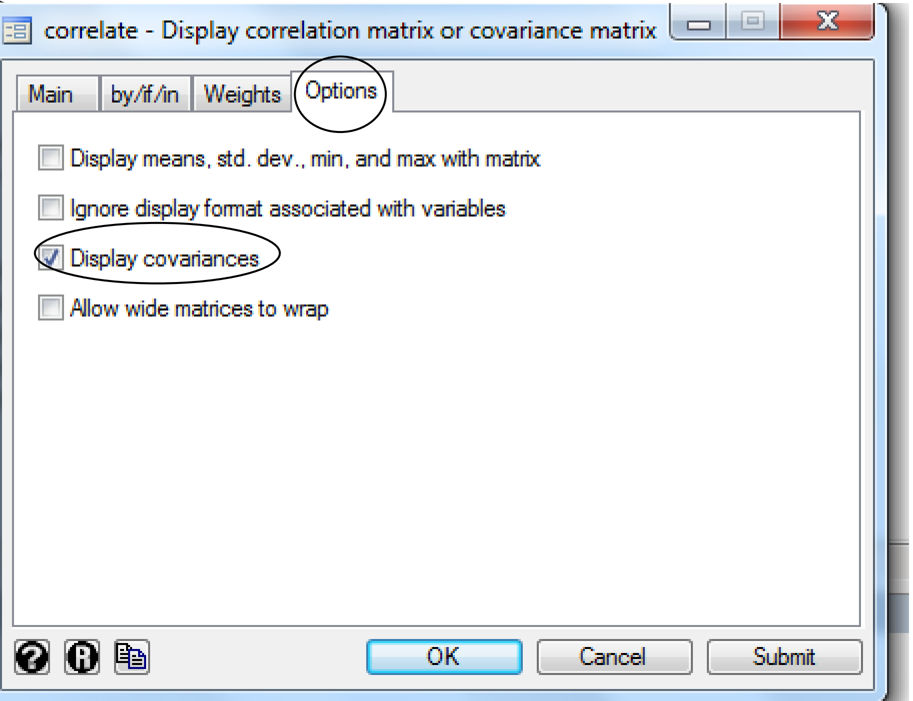

correlate - Display correlation matrix or covariance matrix
Main
by/if/in
Weights
Options
Variables: (leave empty for all)
Examples:
yr*
all variables starting with "yr"
xyz-abc
all variables between xyz and abc
OK
Cancel
Submit

Empirical Example 2- *Time Series Data (12)*



```
. correlate
(obs=394)

             |   Canada   France  Germany    Italy    Japan       UK       US
-------------+---------------------------------------------------------------
      Canada |   1.0000
      France |   0.5619   1.0000
     Germany |   0.5773   0.7893   1.0000
       Italy |   0.4929   0.5940   0.6363   1.0000
       Japan |   0.3810   0.3925   0.4493   0.3974   1.0000
          UK |   0.6821   0.6523   0.6916   0.5573   0.4770   1.0000
          US |   0.7765   0.6187   0.6235   0.4525   0.3743   0.6839   1.0000
```

13. Go to Options and *tick* Display Covariances



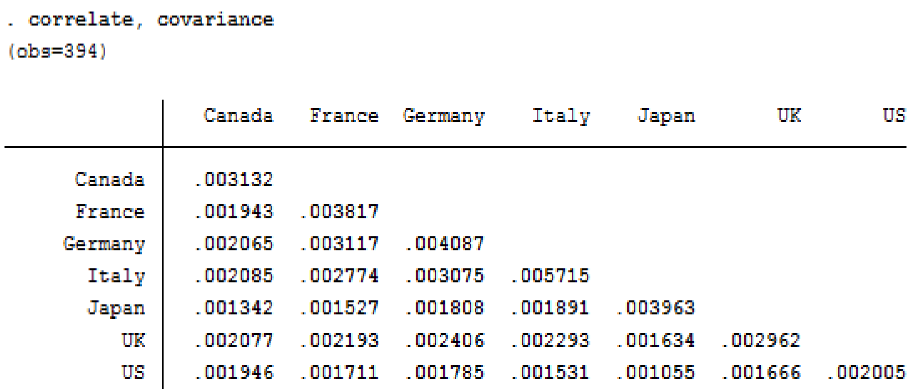
correlate - Display correlation matrix or covariance matrix

□ Display means, std. dev., min, and max with matrix

□ Ignore display format associated with variables

☑ Display covariances

□ Allow wide matrices to wrap



```
. correlate, covariance
(obs=394)

             |   Canada   France  Germany    Italy    Japan       UK       US
-------------+---------------------------------------------------------------
      Canada |  .003132
      France |  .001943  .003817
     Germany |  .002065  .003117  .004087
       Italy |  .002085  .002774  .003075  .005715
       Japan |  .001342  .001527  .001808  .001891  .003963
          UK |  .002077  .002193  .002406  .002293  .001634  .002962
          US |  .001946  .001711  .001785  .001531  .001055  .001666  .002005
```

Empirical Example 2- *Time Series Data (13)*

We can here define

- Covariance between X and Y variables $Cov(X,Y) = E(XY) - E(X)E(Y)$
- Correlation between X and Y variables $\rho = \frac{Cov(X,Y)}{\sqrt{V\ \mathrm{ar}(X)Var(Y)}}$

14 . Close the above window ⟶ Go to *File* ⟶ *Save as…*

Empirical Example 3 :

Transformations of Data & Plots

Time Series Data

Empirical Example 3 : *Transformation of data & Plots(1)*

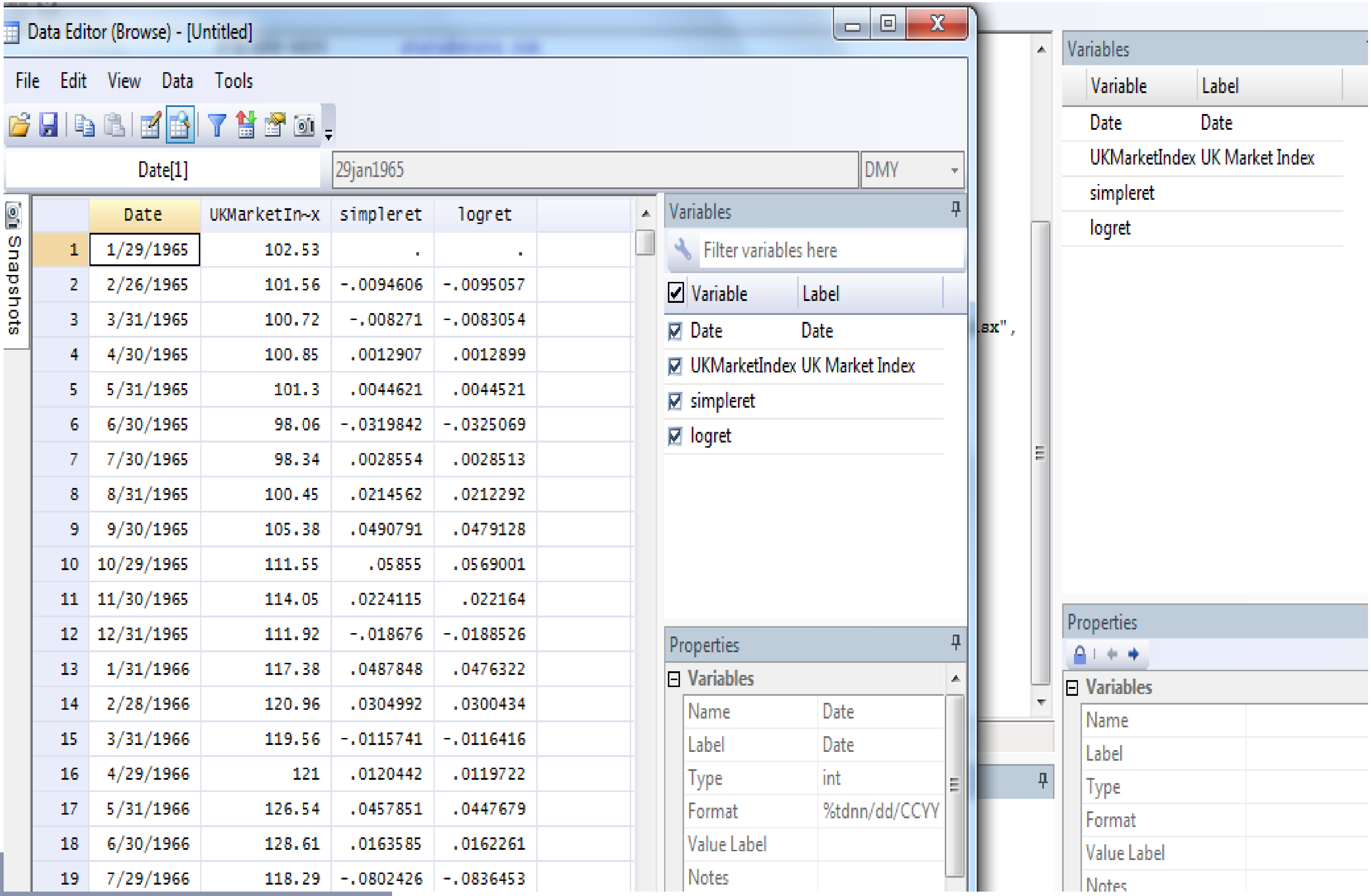
1.Go to folder *Example_3* → *Import the example_3.xlsx*

- We present the price of UK market index from 01/1965 – 06/2015 *(Source : DataStream)*

2.Go to Command window and type the following

```
tsset Date

gen simpleret=(UKMarketIndex/UKMarketIndex[_n-1])-1

gen logret=ln(UKMarketIndex/UKMarketIndex[_n-1])
```

Empirical Example 3 : *Transformation of data & Plots(2)*



Date[1] 29jan1965 DMY

	Date	UKMarketIn~x	simpleret	logret
1	1/29/1965	102.53	.	.
2	2/26/1965	101.56	-.0094606	-.0095057
3	3/31/1965	100.72	-.008271	-.0083054
4	4/30/1965	100.85	.0012907	.0012899
5	5/31/1965	101.3	.0044621	.0044521
6	6/30/1965	98.06	-.0319842	-.0325069
7	7/30/1965	98.34	.0028554	.0028513
8	8/31/1965	100.45	.0214562	.0212292
9	9/30/1965	105.38	.0490791	.0479128
10	10/29/1965	111.55	.05855	.0569001
11	11/30/1965	114.05	.0224115	.022164
12	12/31/1965	111.92	-.018676	-.0188526
13	1/31/1966	117.38	.0487848	.0476322
14	2/28/1966	120.96	.0304992	.0300434
15	3/31/1966	119.56	-.0115741	-.0116416
16	4/29/1966	121	.0120442	.0119722
17	5/31/1966	126.54	.0457851	.0447679
18	6/30/1966	128.61	.016358 5	.0162261
19	7/29/1966	118.29	-.0802426	-.0836453

Variables

Filter variables here

☑ Variable	Label
☑ Date	Date
☑ UKMarketIndex	UK Market Index
☑ simpleret	
☑ logret	

Properties

⊟ Variables

Name	Date
Label	Date
Type	int
Format	%tdnn/dd/CCYY
Value Label	
Notes	

Variables

Variable	Label
Date	Date
UKMarketIndex	UK Market Index
simpleret	
logret	

Properties

⊟ Variables

Name	
Label	
Type	
Format	
Value Label	
Notes	

Continuous compounding or log- returns

Advantages

- Are time additive.
- Assets can be compared since the frequency of compounding return does not play any role.

Disadvantages

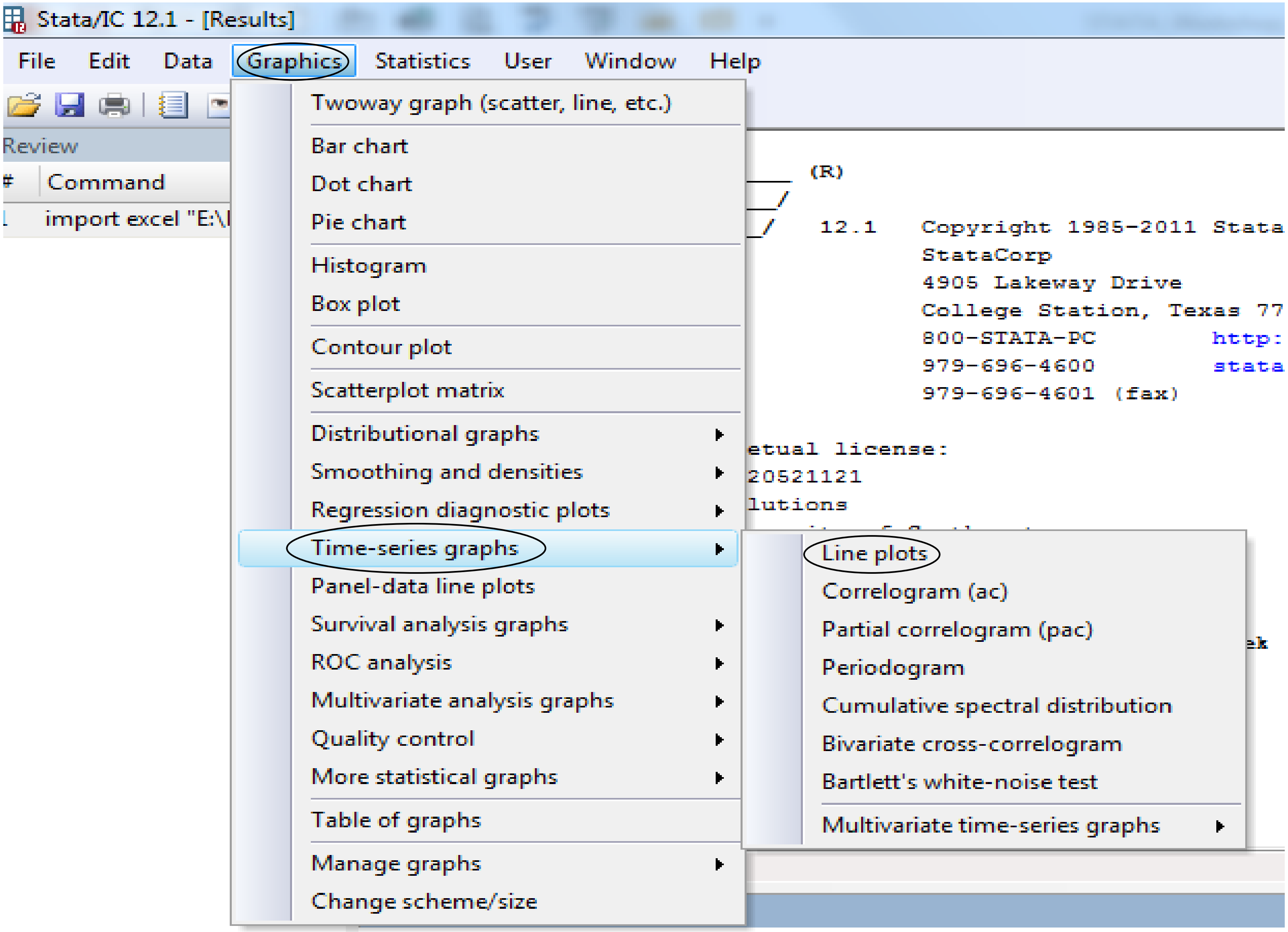
- In Investments , the simple portfolio return is a weighted average of the simple returns on the individual assets. $R_{pt} = \sum_{i=1}^{n} w_i R_{it}$
- **However,** this is not feasible for log returns since the log of a sum is not the same as the sum of a log.

Empirical Example 3 : *Transformation of data & Plots(4)*

Go to Graphics ⟶ Time-series Graphs ⟶ Line Plots

3. Click on Create

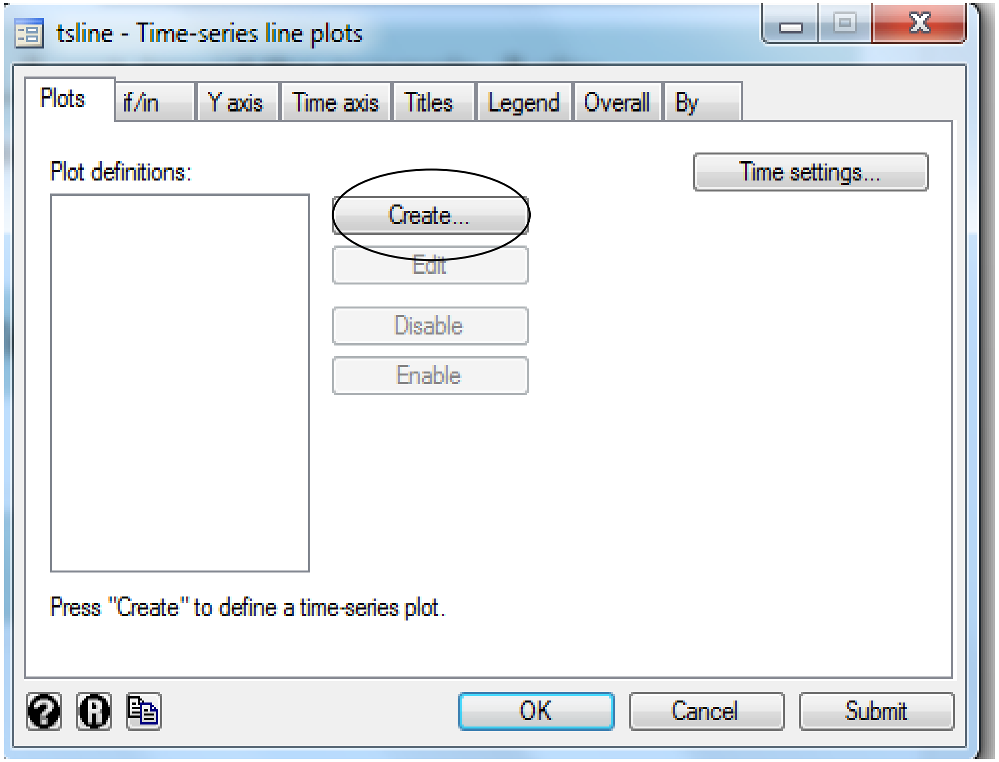
Empirical Example 3 : *Transformation of data & Plots(5)*



```
1   import excel "E:\l
```

```
 (R)
 12.1   Copyright 1985-2011 Stata
        StataCorp
        4905 Lakeway Drive
        College Station, Texas 77
        800-STATA-PC          http:
        979-696-4600          stata
        979-696-4601 (fax)

etual license:
20521121
lutions
ek
```

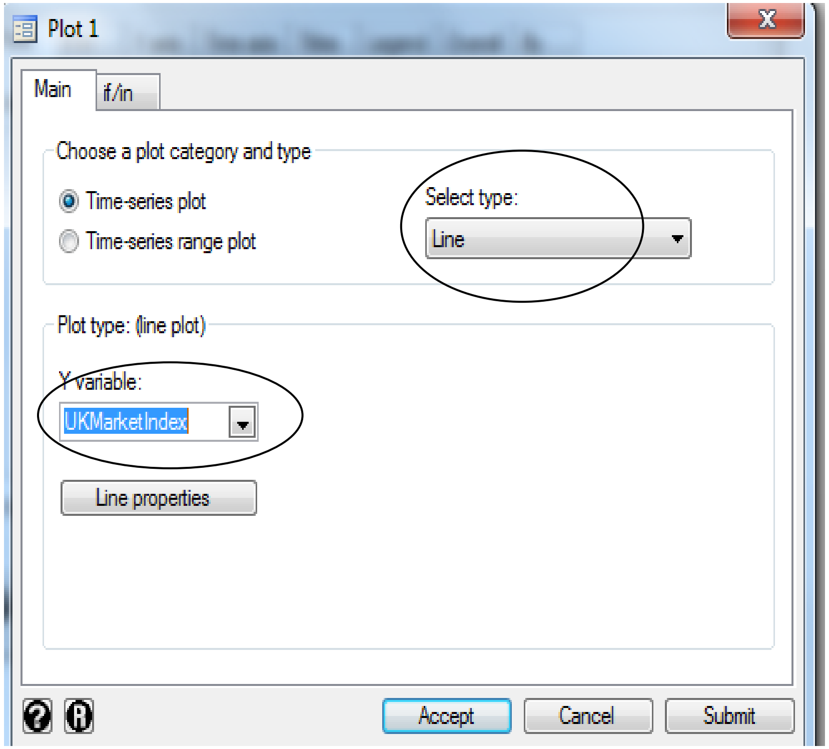
Empirical Example 3 : *Transformation of data & Plots(6)*



tsline - Time-series line plots

Plot definitions:

Press "Create" to define a time-series plot.



Plot 1

Choose a plot category and type

● Time-series plot

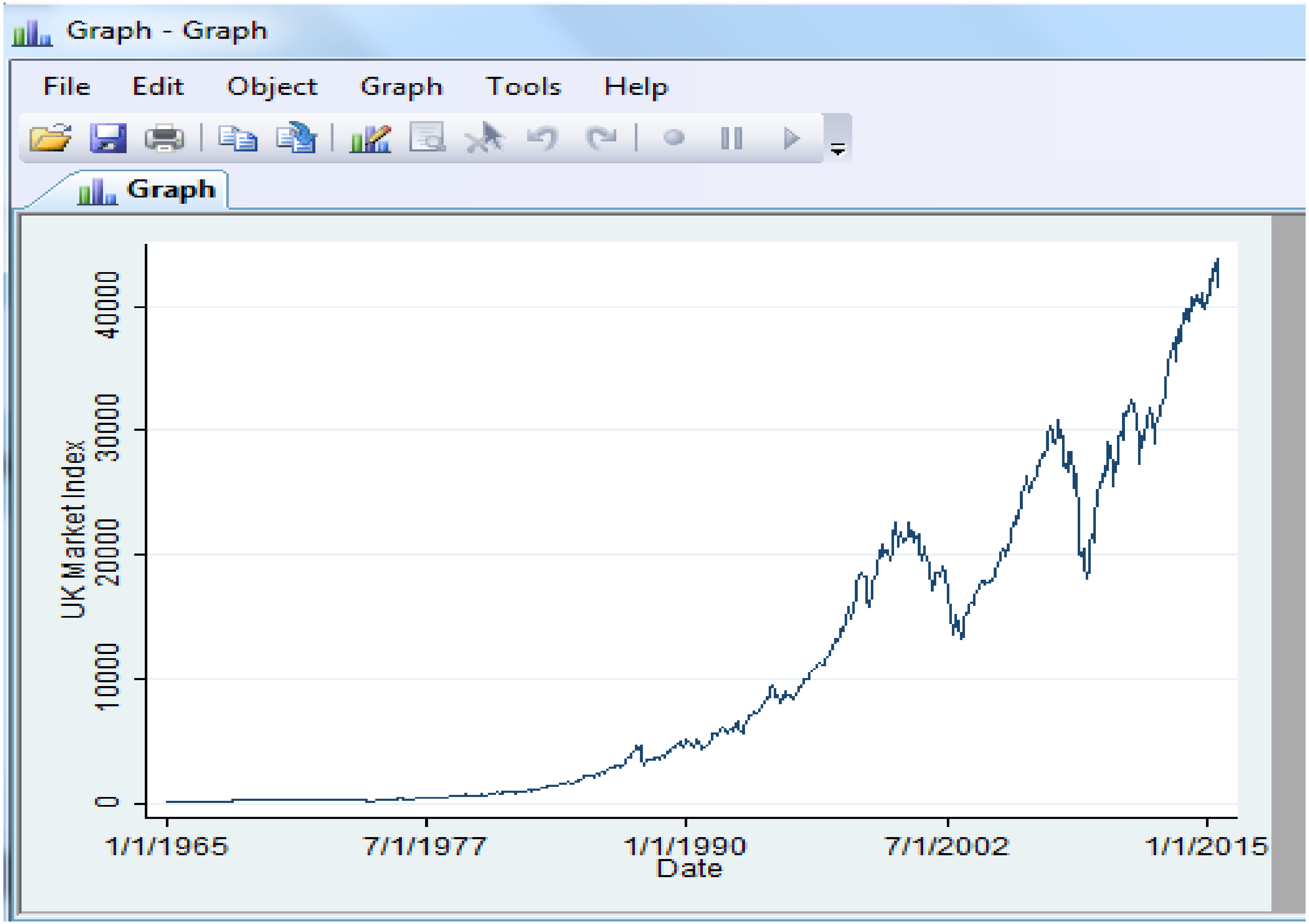
○ Time-series range plot

Select type: Line

Plot type: (line plot)

Y variable: UKMarketIndex

Empirical Example 3 : *Transformation of data & Plots(7)*



Empirical Example 4 :

Data transformation and setup

Panel Data

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Empirical Example 4 : Long and Wide datasets

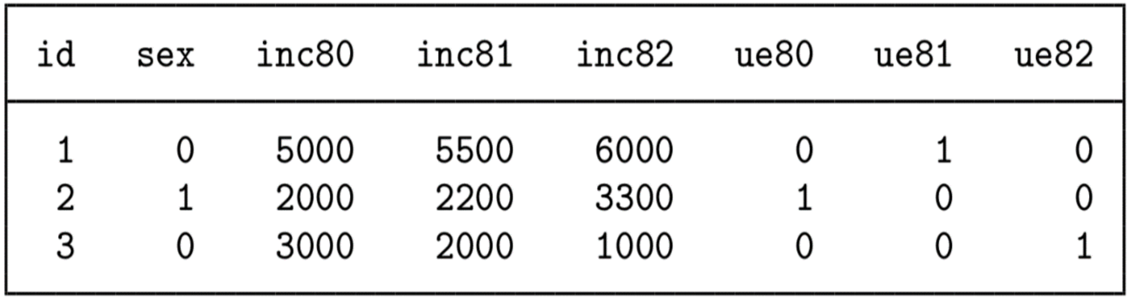
1.Go to folder *Example_4* → *Import the example_4 workbook*

- Prices and Dividend Yields for the share i (i=1,2,3) in year j (j=19,20,21)

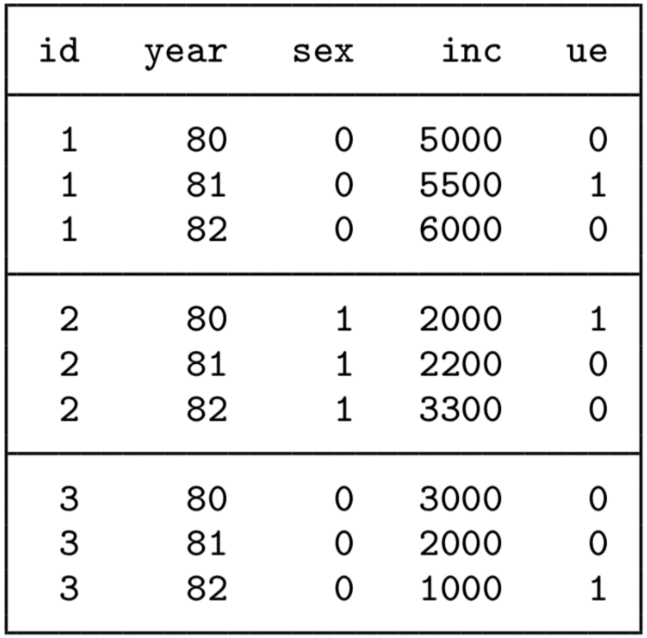
2.Is this Panel dataset long or wide?

Think of the data as a collection of observations Xij, where i is the logical observation, or group identifier, and j is the subobservation, or within-group identifier.

- Wide-form data are organized by logical observation, storing all the data on a particular observation in one row.
- Long-form data are organized by subobservation, storing the data in multiple rows.



id	sex	inc80	inc81	inc82	ue80	ue81	ue82
1	0	5000	5500	6000	0	1	0
2	1	2000	2200	3300	1	0	0
3	0	3000	2000	1000	0	0	1

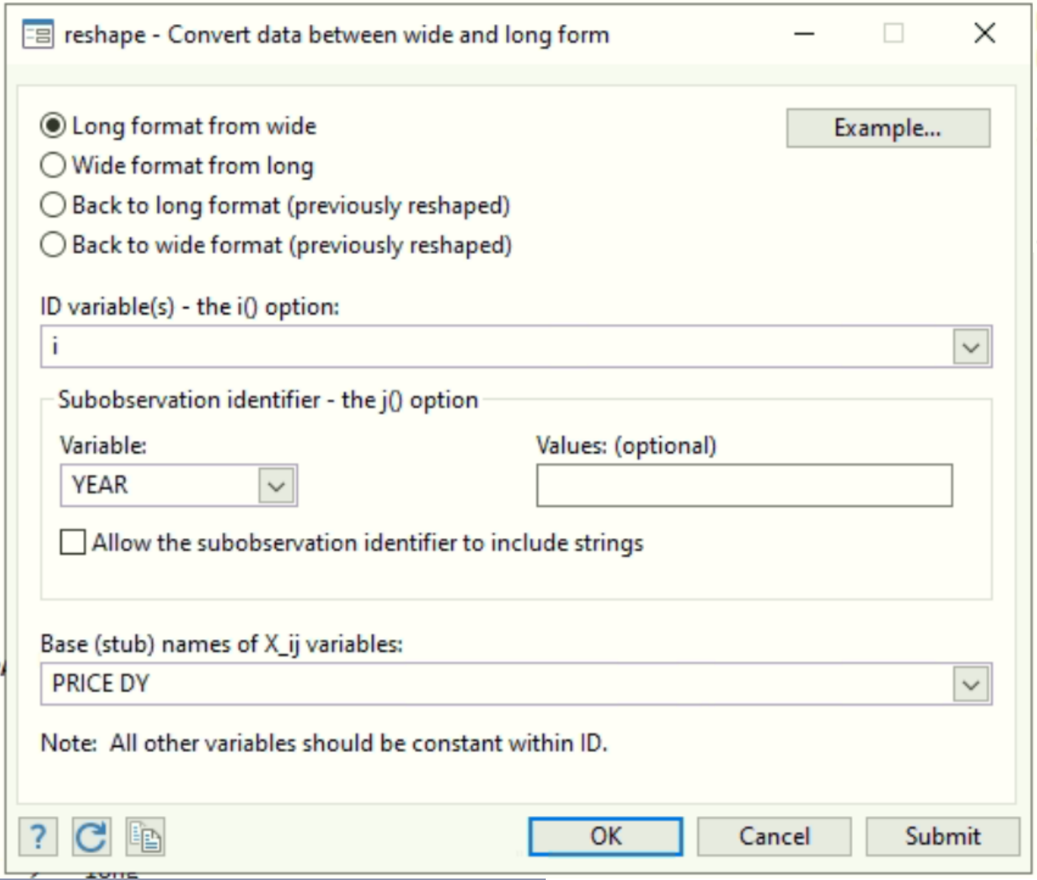


id	year	sex	inc	ue
1	80	0	5000	0
1	81	0	5500	1
1	82	0	6000	0
2	80	1	2000	1
2	81	1	2200	0
2	82	1	3300	0
3	80	0	3000	0
3	81	0	2000	0
3	82	0	1000	1

Convert data between wide and long

Data > Create or change data > Other variable-transformation commands > Convert data between wide and long

```
reshape long PRICE DY, i(i) j(YEAR)
```



reshape - Convert data between wide and long form

- ● Long format from wide
- ○ Wide format from long
- ○ Back to long format (previously reshaped)
- ○ Back to wide format (previously reshaped)

ID variable(s) - the i() option: i

Subobservation identifier - the j() option

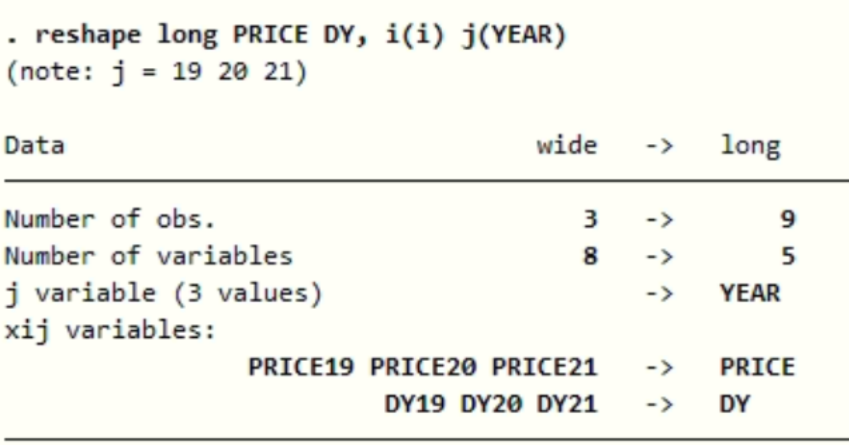
Variable: YEAR

Values: (optional)

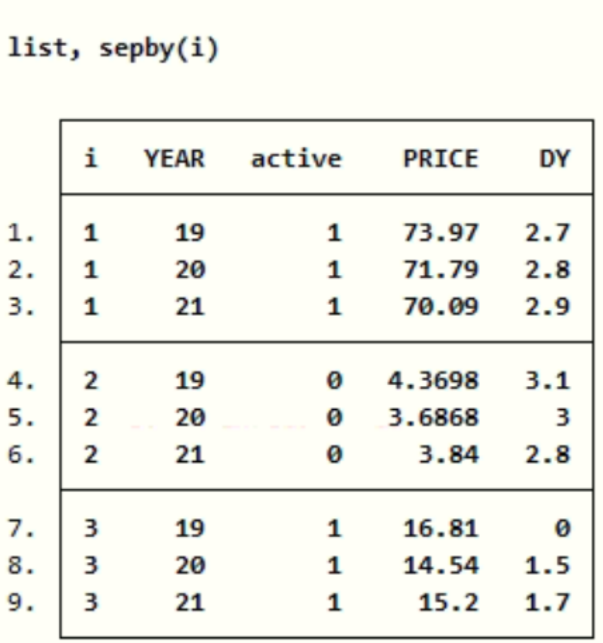
□ Allow the subobservation identifier to include strings

Base (stub) names of X_ij variables: PRICE DY

Note: All other variables should be constant within ID.



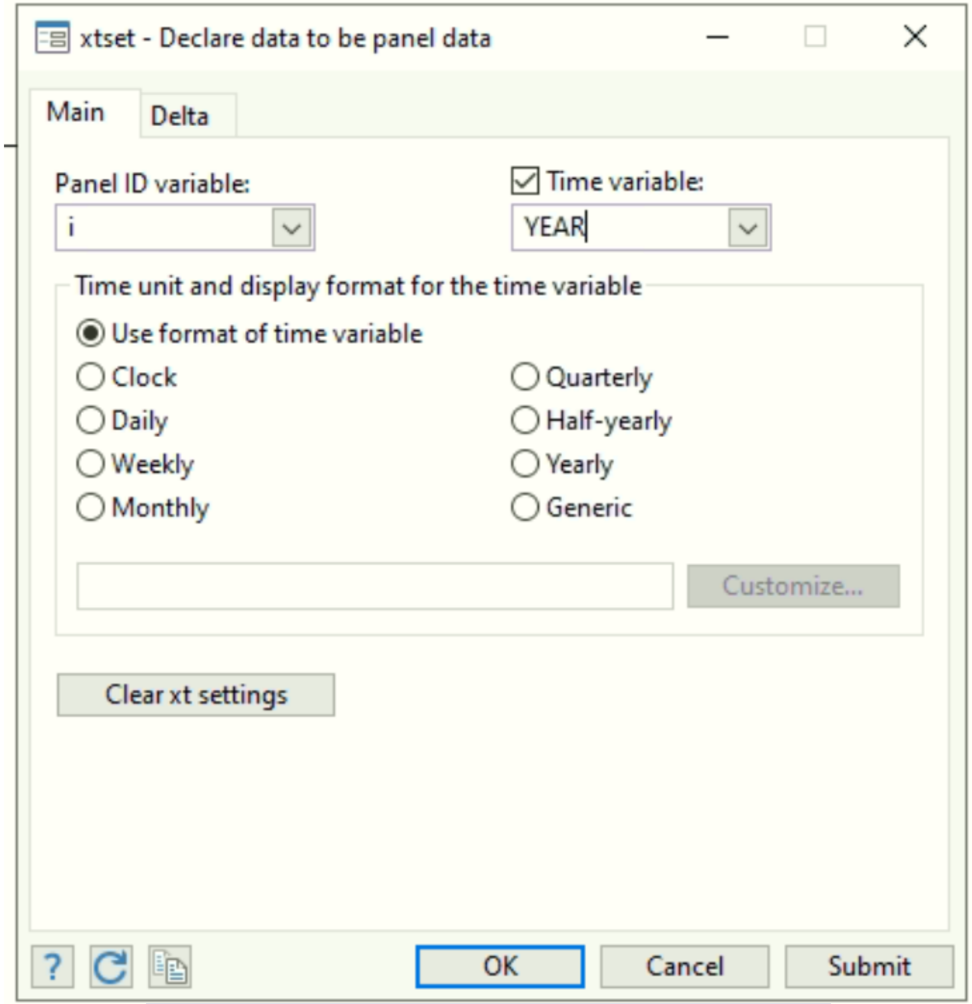
```
. reshape long PRICE DY, i(i) j(YEAR)
(note: j = 19 20 21)

Data                               wide   ->   long
-----------------------------------------------------------
Number of obs.                        3   ->      9
Number of variables                   8   ->      5
j variable (3 values)                     ->   YEAR
xij variables:
                   PRICE19 PRICE20 PRICE21  ->   PRICE
                               DY19 DY20 DY21  ->   DY
-----------------------------------------------------------
```



```
list, sepby(i)

     +--------------------------------------+
     | i   YEAR   active    PRICE     DY |
     |--------------------------------------|
  1. | 1     19        1    73.97    2.7 |
  2. | 1     20        1    71.79    2.8 |
  3. | 1     21        1    70.09    2.9 |
     |--------------------------------------|
  4. | 2     19        0   4.3698    3.1 |
  5. | 2     20        0   3.6868      3 |
  6. | 2     21        0     3.84    2.8 |
     |--------------------------------------|
  7. | 3     19        1    16.81      0 |
  8. | 3     20        1    14.54    1.5 |
  9. | 3     21        1     15.2    1.7 |
     +--------------------------------------+
```

Declare Dataset to be Panel Data

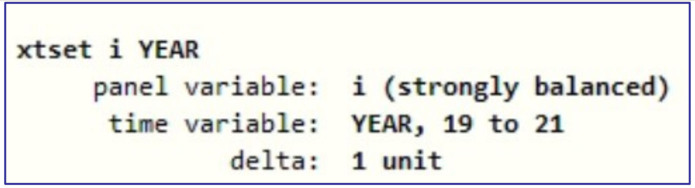
- xtset panelvar declares the data in memory to be a panel in which the order of observations is irrelevant.
- **xtset panelvar timevar** declares the data to be a panel in which the order of observations is relevant.
  - When you specify timevar, you can then use Stata's time-series operators and analyze your data with the ts commands without having to tsset your data.
- Statistics > Longitudinal/panel data > Setup and utilities > Declare dataset to be panel data



Panel ID variable: i

☑ Time variable: YEAR

● Use format of time variable



```
xtset i YEAR
       panel variable:  i (strongly balanced)
        time variable:  YEAR, 19 to 21
                delta:  1 unit
```

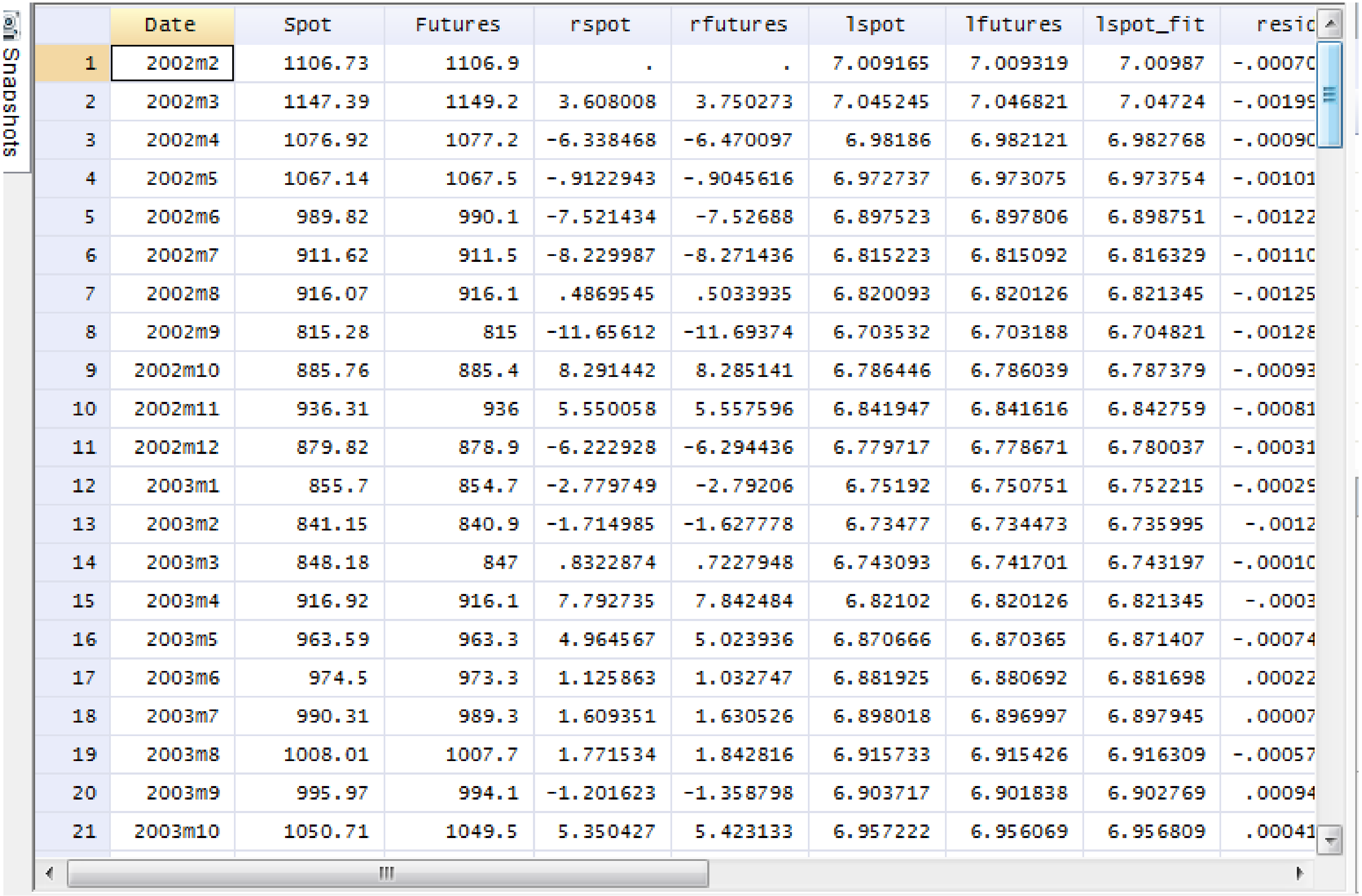
Classical Linear Regression Model Estimation

Empirical Example 5 :

Simple Linear Regression

Simple Linear Regression

Open the file SandPhedge.dta

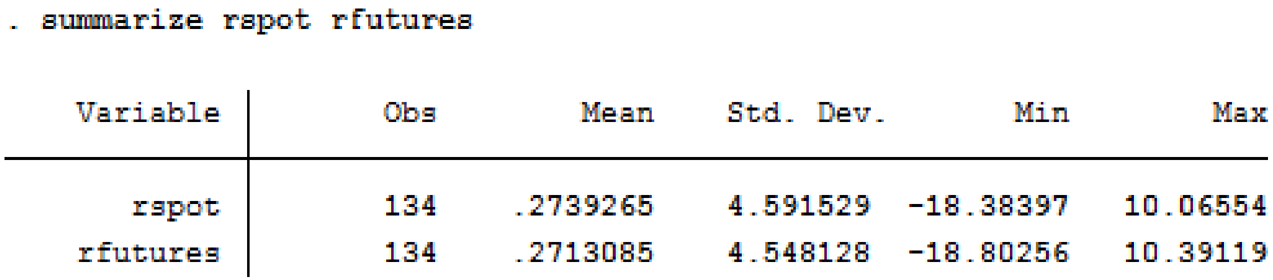


	Date	Spot	Futures	rspot	rfutures	lspot	lfutures	lspot_fit	resid
1	2002m2	1106.73	1106.9	.	.	7.009165	7.009319	7.00987	-.00070
2	2002m3	1147.39	1149.2	3.608008	3.750273	7.045245	7.046821	7.04724	-.00199
3	2002m4	1076.92	1077.2	-6.338468	-6.470097	6.98186	6.982121	6.982768	-.00090
4	2002m5	1067.14	1067.5	-.9122943	-.9045616	6.972737	6.973075	6.973754	-.00101
5	2002m6	989.82	990.1	-7.521434	-7.52688	6.897523	6.897806	6.898751	-.00122
6	2002m7	911.62	911.5	-8.229987	-8.271436	6.815223	6.815092	6.816329	-.00110
7	2002m8	916.07	916.1	.4869545	.5033935	6.820093	6.820126	6.821345	-.00125
8	2002m9	815.28	815	-11.65612	-11.69374	6.703532	6.703188	6.704821	-.00128
9	2002m10	885.76	885.4	8.291442	8.285141	6.786446	6.786039	6.787379	-.00093
10	2002m11	936.31	936	5.550058	5.557596	6.841947	6.841616	6.842759	-.00081
11	2002m12	879.82	878.9	-6.222928	-6.294436	6.779717	6.778671	6.780037	-.00031
12	2003m1	855.7	854.7	-2.779749	-2.79206	6.75192	6.750751	6.752215	-.00029
13	2003m2	841.15	840.9	-1.714985	-1.627778	6.73477	6.734473	6.735995	-.0012
14	2003m3	848.18	847	.8322874	.7227948	6.743093	6.741701	6.743197	-.00010
15	2003m4	916.92	916.1	7.792735	7.842484	6.82102	6.820126	6.821345	-.0003
16	2003m5	963.59	963.3	4.964567	5.023936	6.870666	6.870365	6.871407	-.00074
17	2003m6	974.5	973.3	1.125863	1.032747	6.881925	6.880692	6.881698	.00022
18	2003m7	990.31	989.3	1.609351	1.630526	6.898018	6.896997	6.897945	.00007
19	2003m8	1008.01	1007.7	1.771534	1.842816	6.915733	6.915426	6.916309	-.00057
20	2003m9	995.97	994.1	-1.201623	-1.358798	6.903717	6.901838	6.902769	.00094
21	2003m10	1050.71	1049.5	5.350427	5.423133	6.957222	6.956069	6.956809	.00041

Simple Linear Regression

Summary Statistics

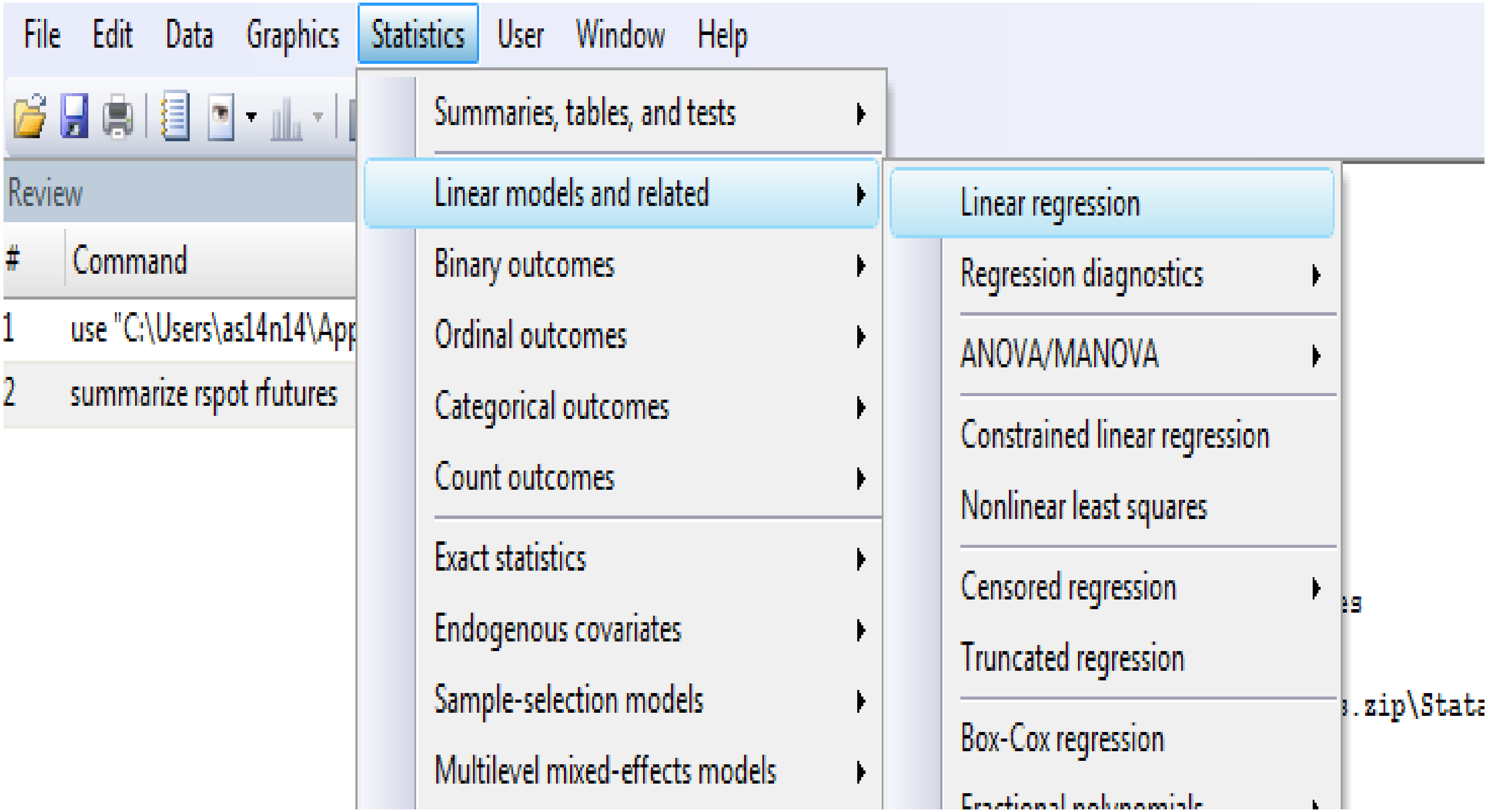
Use summarize in the command window



```
. summarize rspot rfutures

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       rspot |       134    .2739265    4.591529  -18.38397   10.06554
    rfutures |       134    .2713085    4.548128  -18.80256   10.39119
```

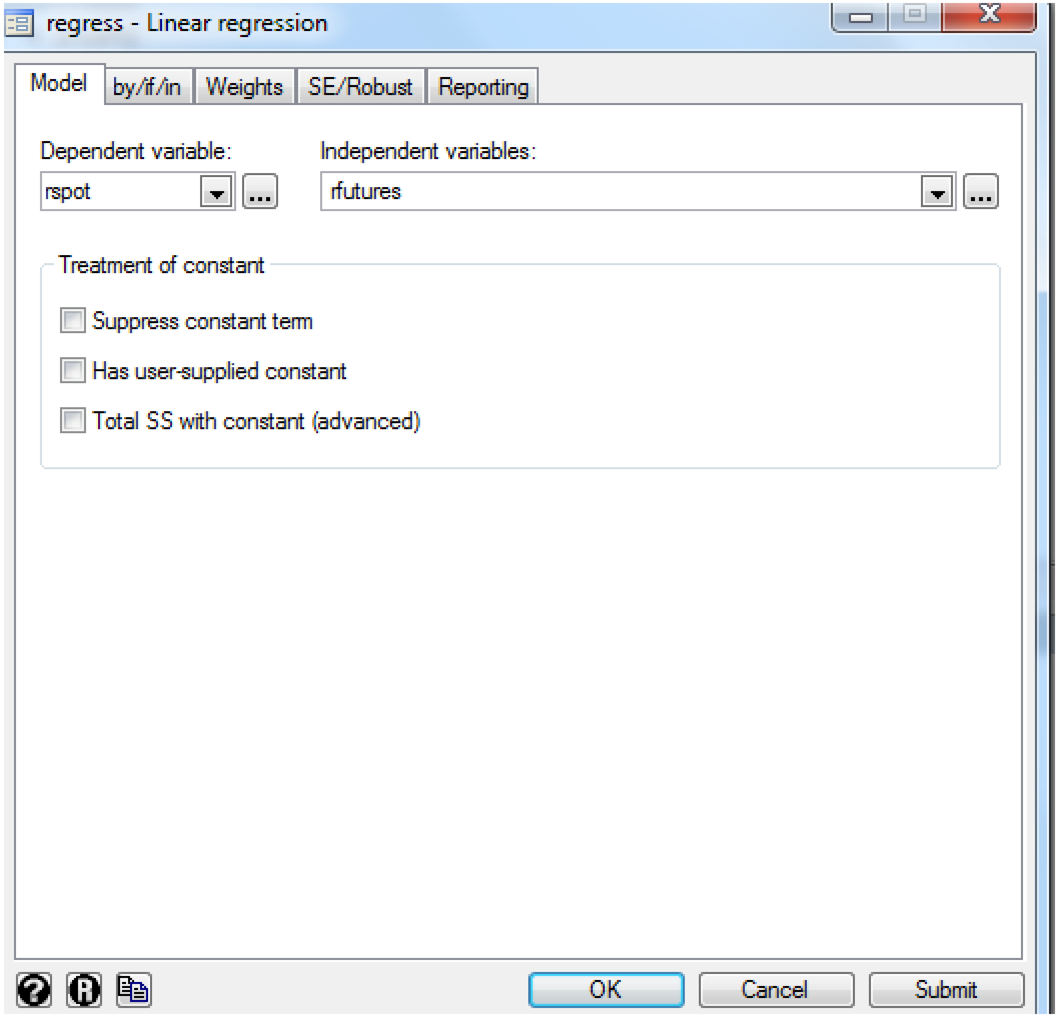
Simple Linear Regression

Go to Statistics→ Linear models and related→ Linear regression



```
1  use "C:\Users\as14n14\App
2  summarize rspot rfutures
```

Simple Linear Regression



Dependent variable: rspot

Independent variables: rfutures

Treatment of constant

□ Suppress constant term

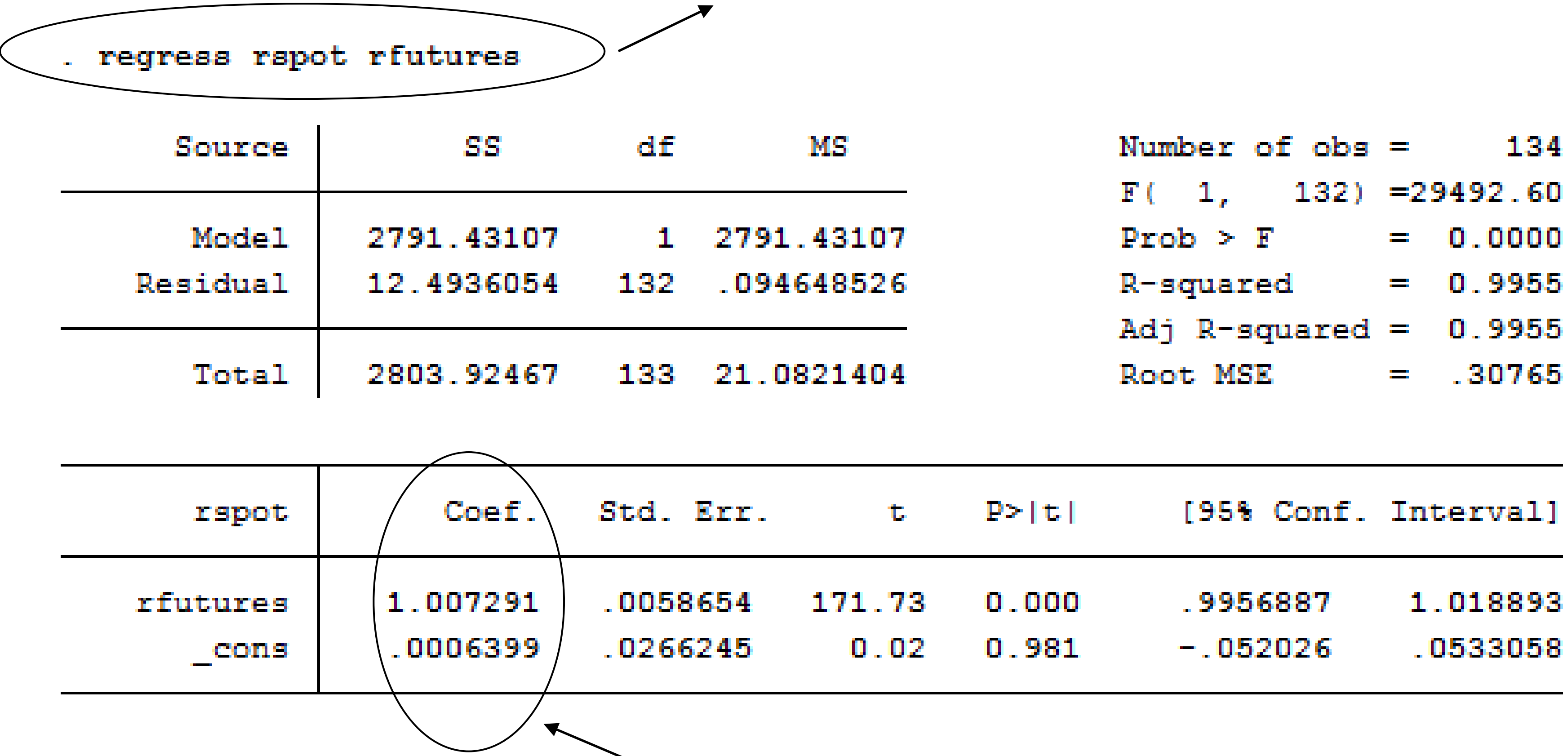
□ Has user-supplied constant

□ Total SS with constant (advanced)

Simple Linear Regression

In the command window you can type

Regress dependent variable independent variable



```
. regress rspot rfutures

      Source |       SS       df       MS              Number of obs =     134
-------------+------------------------------           F(  1,   132) =29492.60
       Model |  2791.43107     1  2791.43107           Prob > F      =  0.0000
    Residual |  12.4936054   132  .094648526           R-squared     =  0.9955
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  2803.92467   133  21.0821404           Root MSE      =  .30765

------------------------------------------------------------------------------
       rspot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rfutures |   1.007291   .0058654   171.73   0.000     .9956887    1.018893
       _cons |   .0006399   .0266245     0.02   0.981     -.052026    .0533058
------------------------------------------------------------------------------
```

Coefficients

1. Hypothesis Testing – Critical value approach

Two –sided Test

$$H_0 : \alpha = 0$$

$$H_A : \alpha \neq 0$$

$$H_0 : \beta = 0$$

$$H_A : \beta \neq 0$$

Critical value approach $\xrightarrow{\hspace{3em}}$

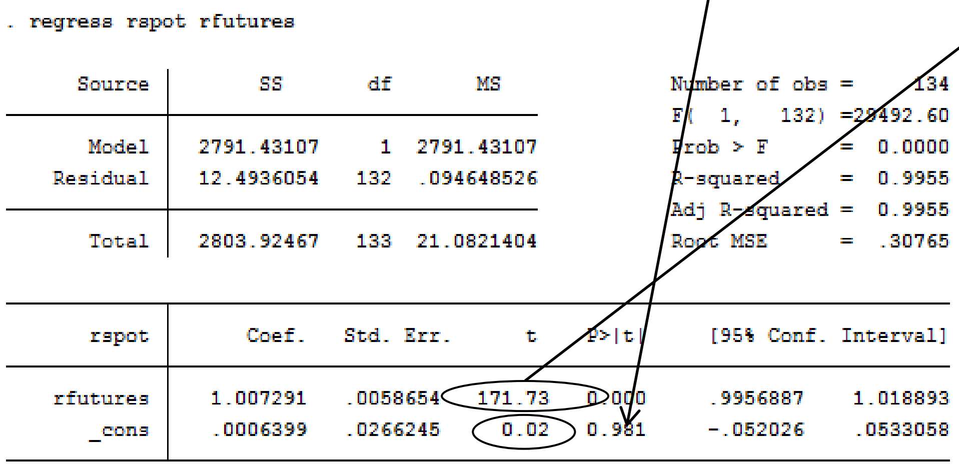
a = 5% significance level

$$\textit{test statistic} = \frac{\hat{a} - a}{SE\left(\hat{a}\right)}$$

We do **not** reject the Null Hypothesis for a ; thus a is **insignificant**

$$\textit{test statistic} = \frac{\hat{\beta} - \beta}{SE\left(\hat{\beta}\right)}$$

We reject the Null Hypothesis for b ; thus b is **significant**



```
. regress rspot rfutures

      Source |       SS       df       MS              Number of obs =     134
-------------+------------------------------           F(  1,   132) =29492.60
       Model |  2791.43107     1  2791.43107           Prob > F      =  0.0000
    Residual |  12.4936054   132  .094648526           R-squared     =  0.9955
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  2803.92467   133  21.0821404           Root MSE      =  .30765

------------------------------------------------------------------------------
       rspot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rfutures |   1.007291   .0058654   171.73   0.000     .9956887    1.018893
       _cons |   .0006399   .0266245     0.02   0.981     -.052026    .0533058
------------------------------------------------------------------------------
```

2. Hypothesis Testing – Confidence interval approach

Two –sided Test

$$H_0: \alpha = 0$$
$$H_A: \alpha \neq 0$$

$$H_0: \beta = 0$$
$$H_A: \beta \neq 0$$

Confidence interval approach

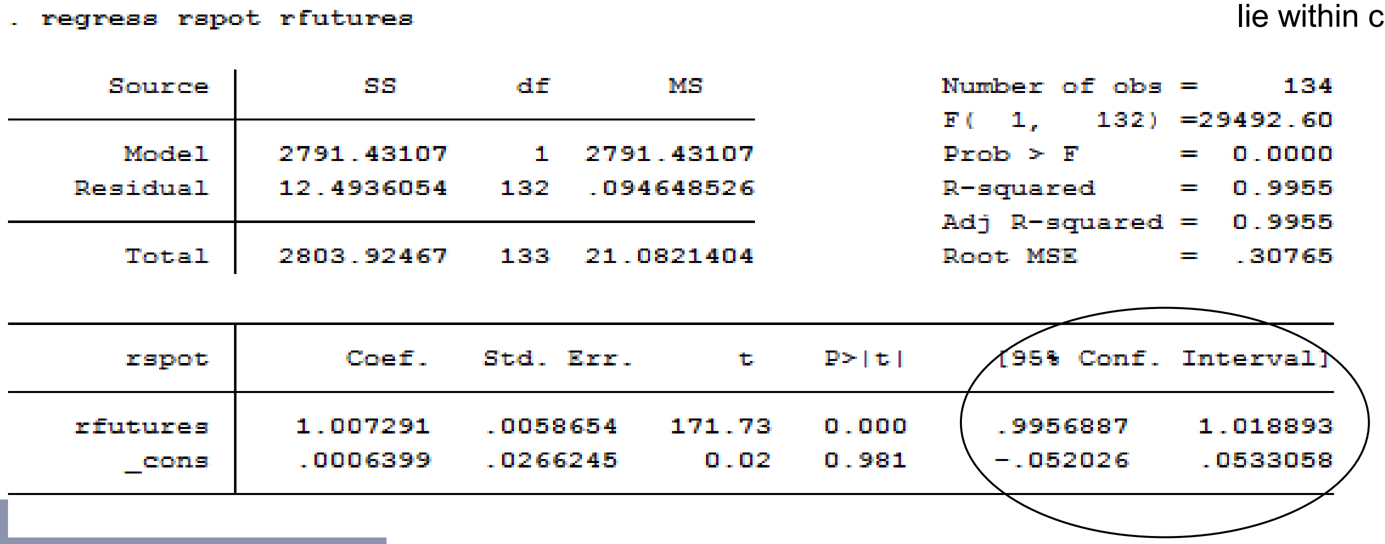
a = 5% significance level

$$\hat{a} \pm t_{crit} SE(\hat{a})$$

$$\hat{\beta} \pm t_{crit} SE(\hat{\beta})$$

(-0.052026,0.0533)
We do **not** reject the Null Hypothesis for a ; thus a is **Insignificant,** since 0 lies within confidence interval

(0.995,1.01889)
We reject the Null Hypothesis for b ; thus b is **significant,** since 0 does **not** lie within confidence interval



```
. regress rspot rfutures

      Source |       SS       df       MS              Number of obs =     134
-------------+------------------------------           F(  1,   132) =29492.60
       Model |  2791.43107     1  2791.43107           Prob > F      =  0.0000
    Residual |  12.4936054   132  .094648526           R-squared     =  0.9955
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  2803.92467   133  21.0821404           Root MSE      =  .30765

------------------------------------------------------------------------------
       rspot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rfutures |   1.007291   .0058654   171.73   0.000     .9956887    1.018893
       _cons |   .0006399   .0266245     0.02   0.981    -.052026    .0533058
------------------------------------------------------------------------------
```

3. Hypothesis Testing – p-value approach

Two –sided Test

$$H_0 : \alpha = 0$$
$$H_A : \alpha \neq 0$$

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

P-value approach

$$\longrightarrow$$

a = 5% significance level

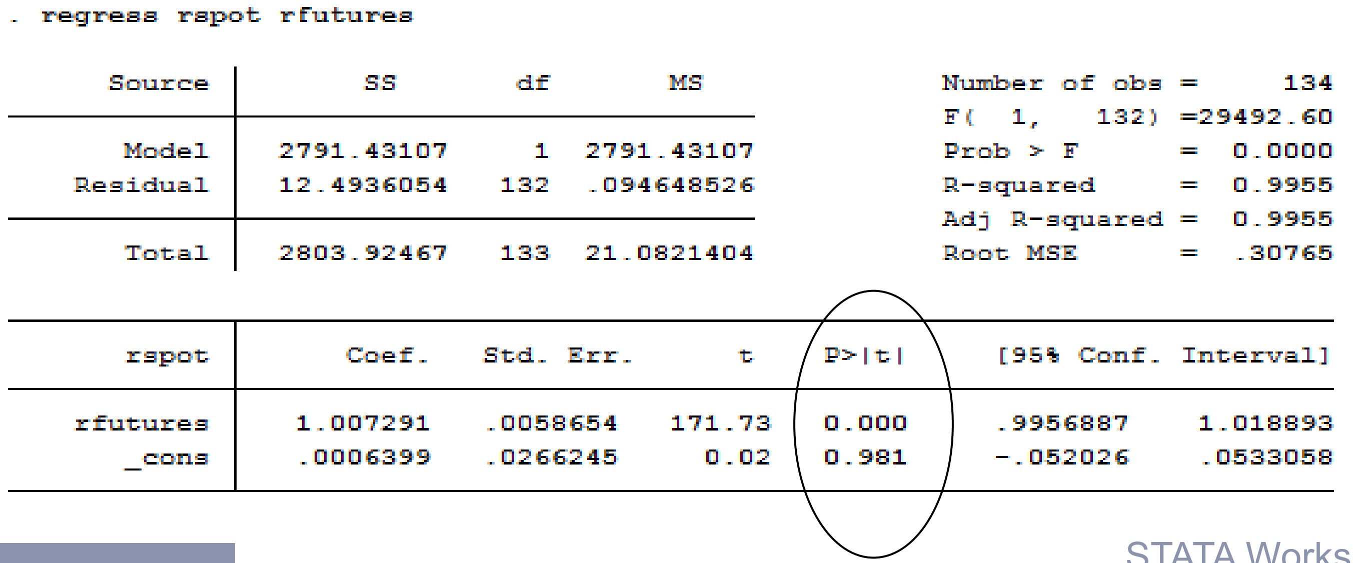
p-value is termed as the

“plausibility” of the Null Hypothesis;

the smaller the p-value, the less plausible is the null hypothesis.

Is the largest significance level at which we fail to reject

the null hypothesis.



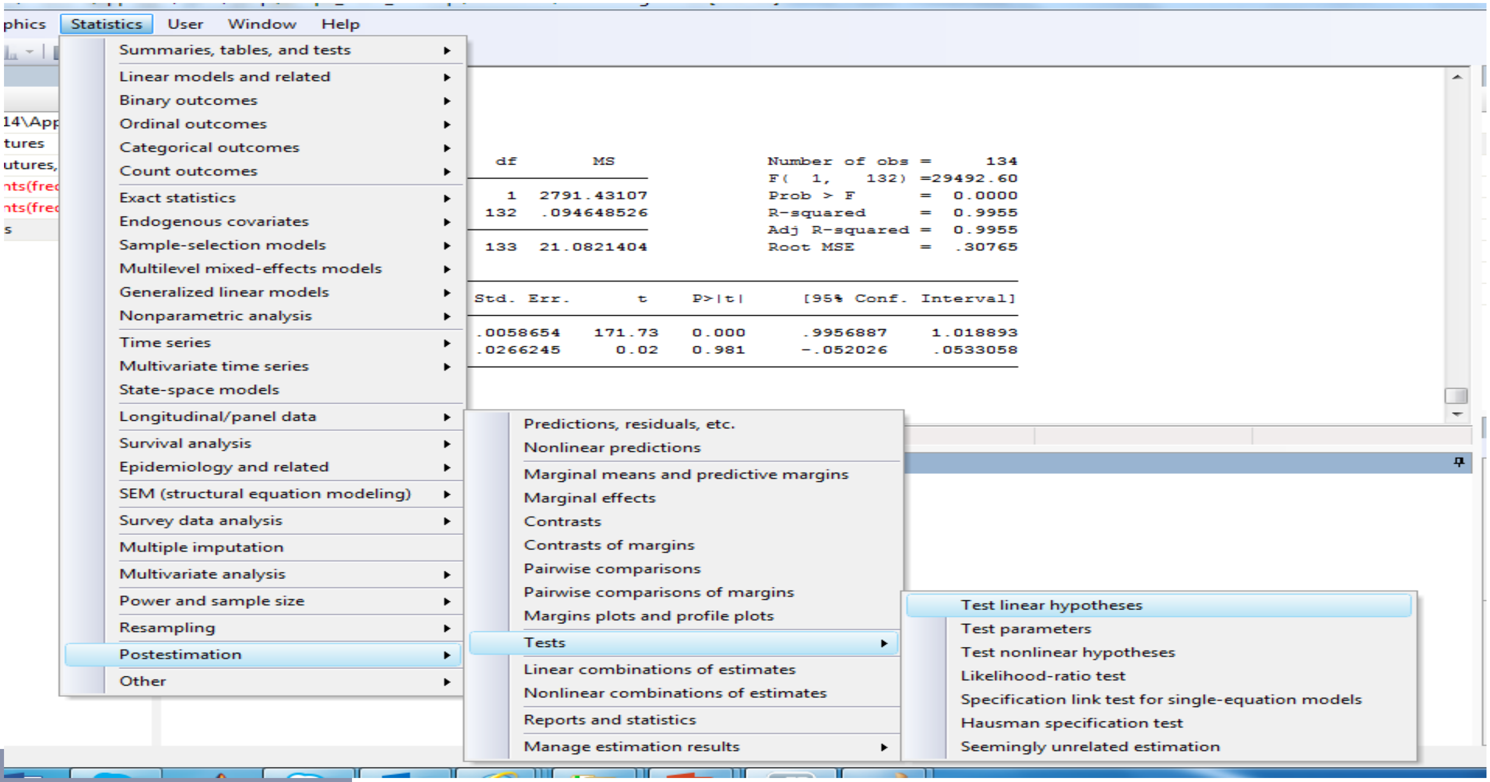
```
. regress rspot rfutures

      Source |       SS       df       MS              Number of obs =     134
-------------+------------------------------           F(  1,   132) =29492.60
       Model |  2791.43107     1  2791.43107           Prob > F      =  0.0000
    Residual |  12.4936054   132  .094648526           R-squared     =  0.9955
-------------+------------------------------           Adj R-squared =  0.9955
       Total |  2803.92467   133  21.0821404           Root MSE      =  .30765

------------------------------------------------------------------------------
       rspot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    rfutures |   1.007291   .0058654   171.73   0.000     .9956887    1.018893
       _cons |   .0006399   .0266245     0.02   0.981     -.052026    .0533058
------------------------------------------------------------------------------
```

Simple Linear Regression

Suppose now we want to test the null hypothesis that

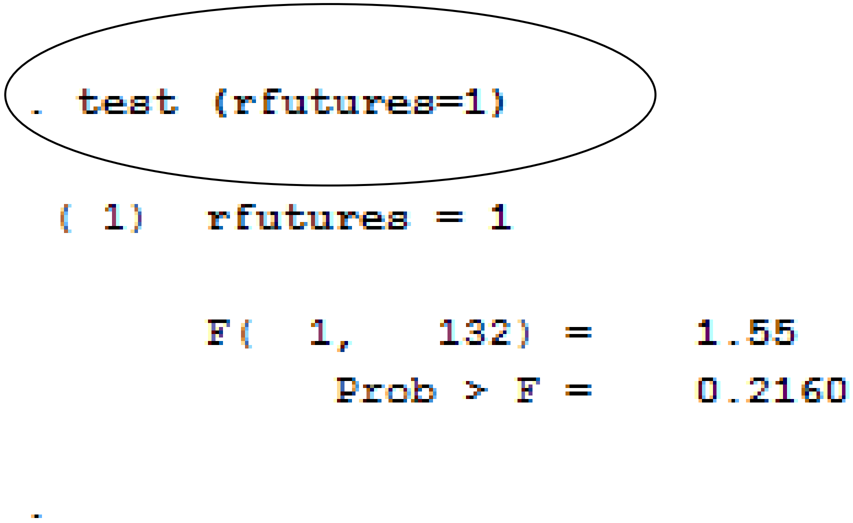
$$H_0 : \beta = 1$$
$$H_A : \beta \neq 1$$

Go to Statistics → Proestimation → Tests → Test linear hypotheses



```
 df        MS                        Number of obs =      134
                                     F(  1,   132) =29492.60
   1  2791.43107                     Prob > F      =  0.0000
 132  .094648526                     R-squared     =  0.9955
                                     Adj R-squared =  0.9955
 133  21.0821404                     Root MSE      =  .30765

Std. Err.      t     P>|t|     [95% Conf. Interval]

.0058654   171.73   0.000     .9956887    1.018893
.0266245     0.02   0.981    -.052026     .0533058
```

Simple Linear Regression



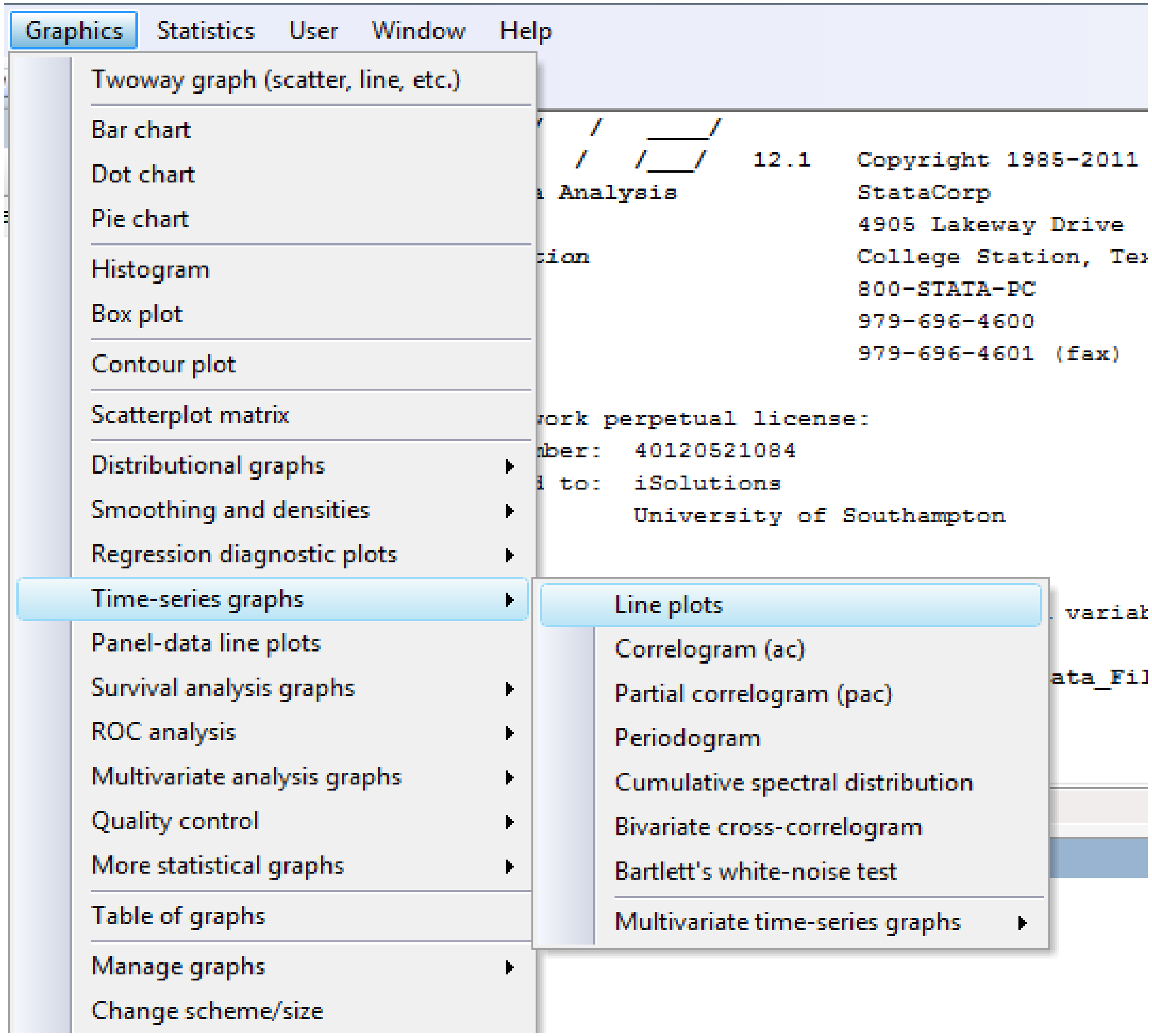
```
. test (rfutures=1)

 ( 1)  rfutures = 1

       F(  1,   132) =    1.55
            Prob > F =    0.2160

.
```

❑ F(1,132) : F-statistic with one restriction and T-k=134-2=132

❑ We cannot reject the Null hypothesis since the

p-value=0.2160>0.05

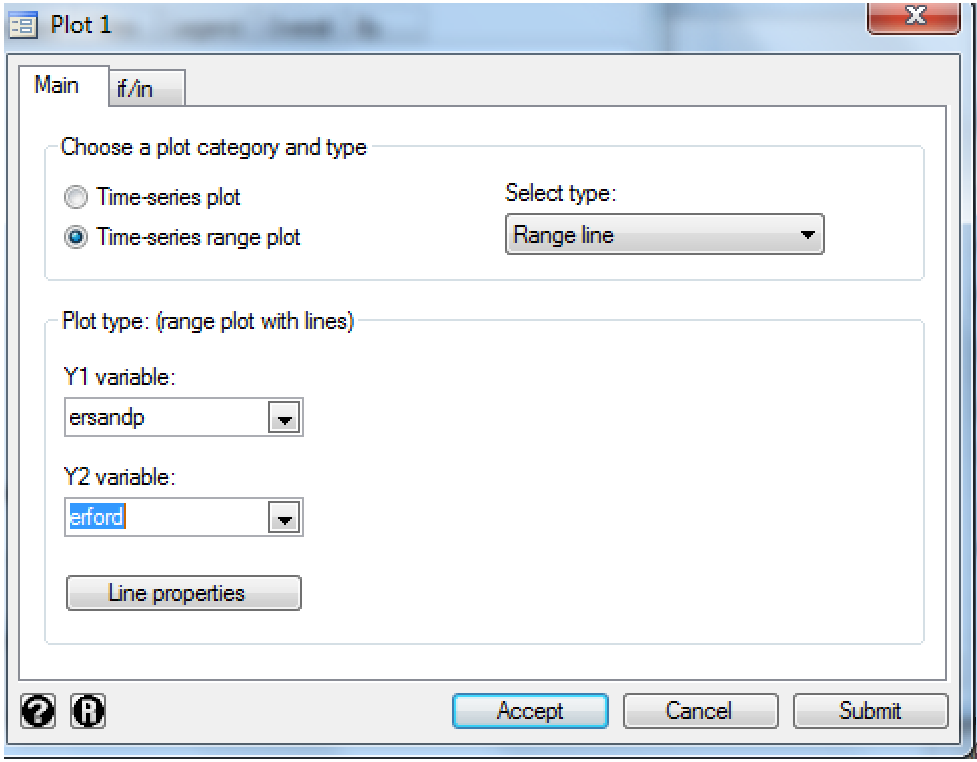
Simple Linear Regression



```
  /   /    ____/
   /   /___/    12.1   Copyright 1985-2011
a Analysis             StataCorp
                       4905 Lakeway Drive
tion                   College Station, Te
                       800-STATA-PC
                       979-696-4600
                       979-696-4601 (fax)

work perpetual license:
nber:   40120521084
d to:   iSolutions
        University of Southampton

                                      variab

                                     ata_Fil
```

Simple Linear Regression

- ❑ Open capm.dta



Choose a plot category and type

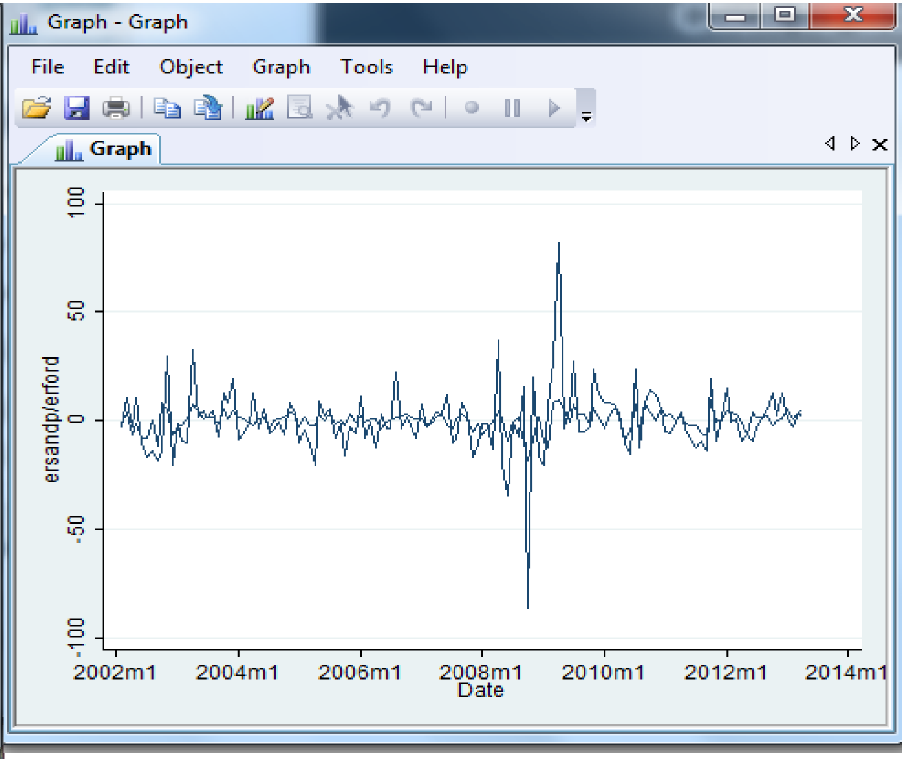
○ Time-series plot
● Time-series range plot

Select type: Range line

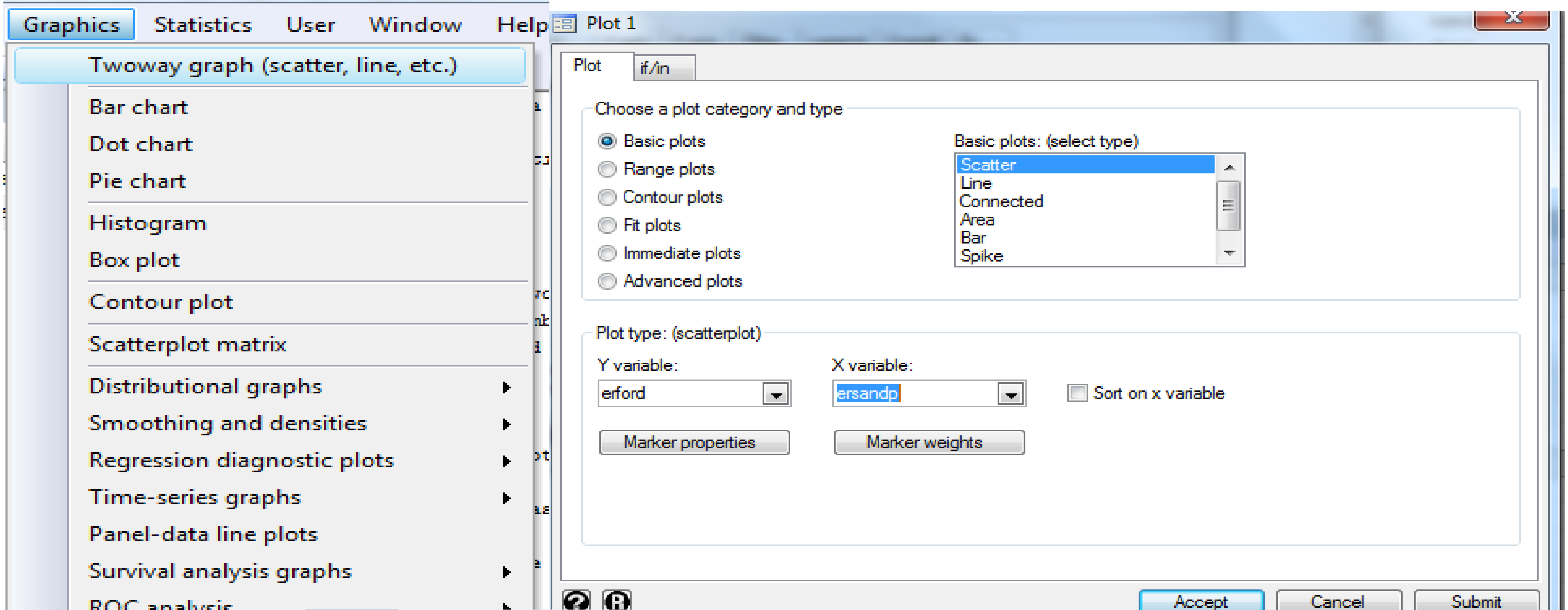
Plot type: (range plot with lines)

Y1 variable: ersandp

Y2 variable: erford



Simple Linear Regression



Graphics | Statistics | User | Window | Help

- Twoway graph (scatter, line, etc.)
- Bar chart
- Dot chart
- Pie chart
- Histogram
- Box plot
- Contour plot
- Scatterplot matrix
- Distributional graphs ▸
- Smoothing and densities ▸
- Regression diagnostic plots ▸
- Time-series graphs ▸
- Panel-data line plots
- Survival analysis graphs ▸
- ROC analysis ▸

Plot 1

Plot | if/in

Choose a plot category and type

- ● Basic plots
- ○ Range plots
- ○ Contour plots
- ○ Fit plots
- ○ Immediate plots
- ○ Advanced plots

Basic plots: (select type)

- Scatter
- Line
- Connected
- Area
- Bar
- Spike

Plot type: (scatterplot)

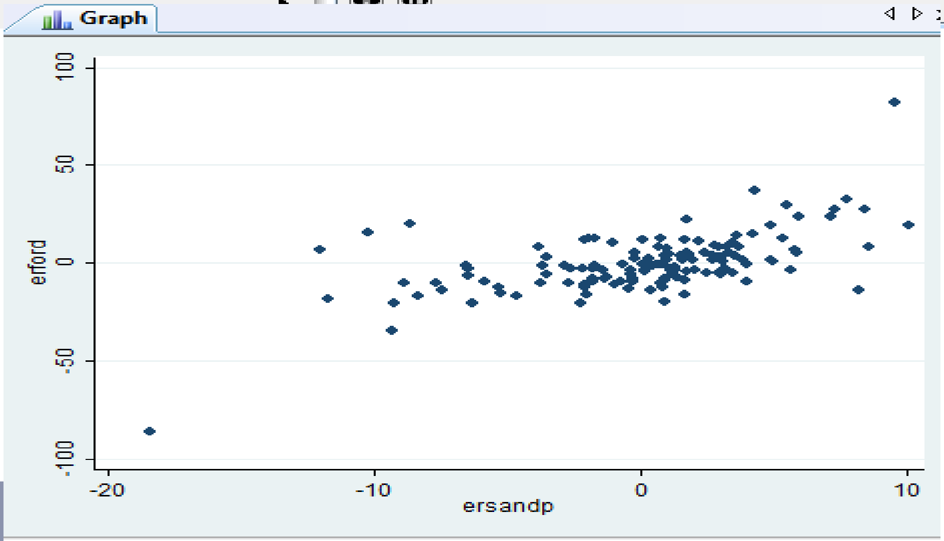
Y variable: erford

X variable: ersandp

□ Sort on x variable

Marker properties | Marker weights

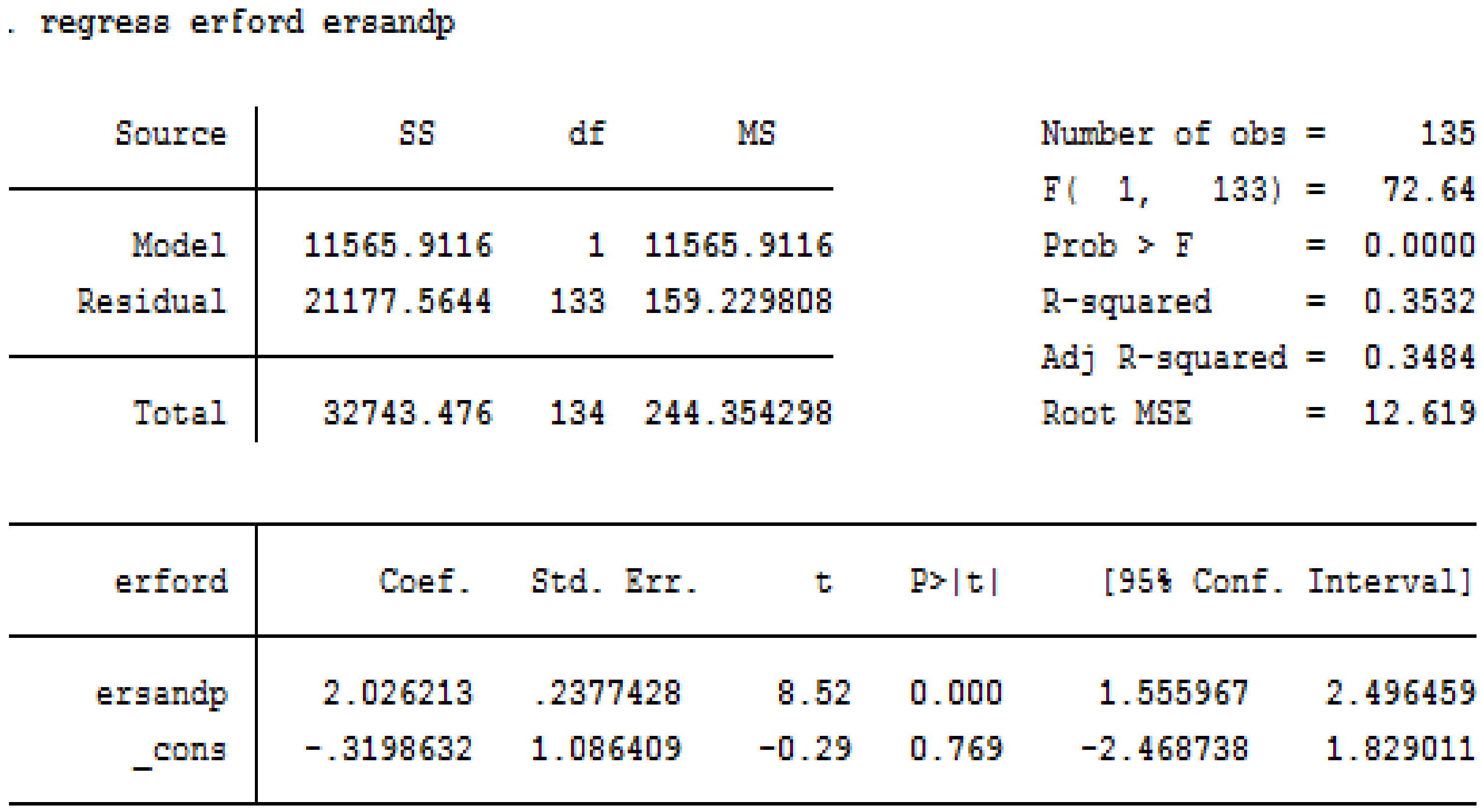
Accept | Cancel | Submit



Scatter plot

Simple Linear Regression

- ❑ Type in the command window **regress** erford ersandp



```
. regress erford ersandp

      Source |       SS       df       MS              Number of obs =     135
-------------+------------------------------           F(  1,   133) =   72.64
       Model |  11565.9116     1  11565.9116           Prob > F      =  0.0000
    Residual |  21177.5644   133  159.229808           R-squared     =  0.3532
-------------+------------------------------           Adj R-squared =  0.3484
       Total |   32743.476   134  244.354298           Root MSE      =  12.619

------------------------------------------------------------------------------
      erford |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   2.026213   .2377428     8.52   0.000     1.555967    2.496459
       _cons |  -.3198632   1.086409    -0.29   0.769    -2.468738    1.829011
------------------------------------------------------------------------------
```

1. Hypothesis Testing – Critical value approach

Two –sided Test

$$H_0 : \alpha = 0$$
$$H_A : \alpha \neq 0$$

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

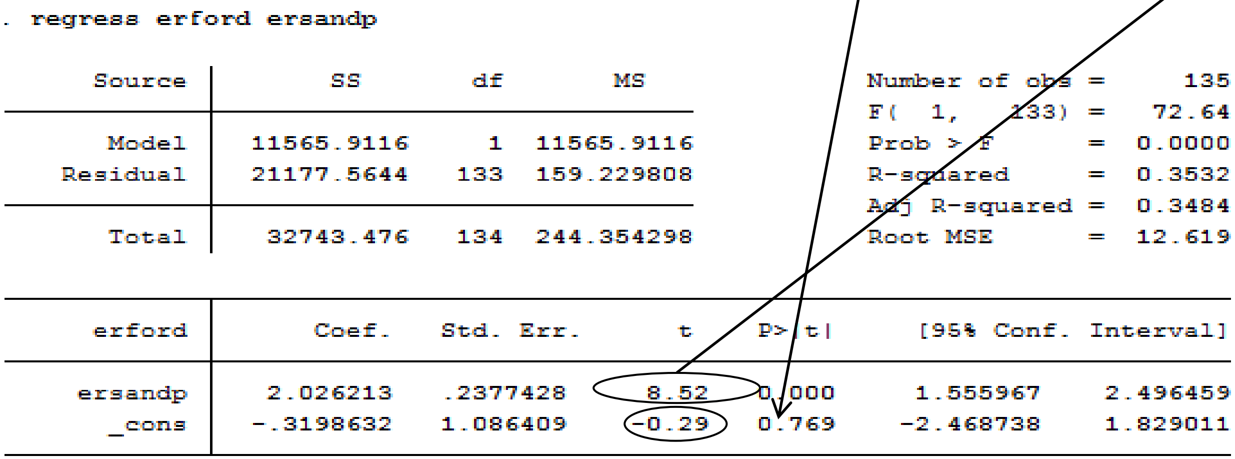
$$\xrightarrow[\text{a = 5\% significance level}]{\text{Critical value approach}}$$

$$test\ statistic = \frac{\hat{a} - a}{SE\left(\hat{a}\right)}$$

We do **not** reject the Null Hypothesis for a ; thus a is **insignificant**

$$test\ statistic = \frac{\hat{\beta} - \beta}{SE\left(\hat{\beta}\right)}$$

We reject the Null Hypothesis for b ; thus b is **significant**



```
. regress erford ersandp

      Source |       SS       df       MS              Number of obs =     135
-------------+------------------------------           F(  1,   133) =   72.64
       Model |  11565.9116     1  11565.9116           Prob > F      =  0.0000
    Residual |  21177.5644   133  159.229808           R-squared     =  0.3532
-------------+------------------------------           Adj R-squared =  0.3484
       Total |   32743.476   134  244.354298           Root MSE      =  12.619

------------------------------------------------------------------------------
      erford |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   2.026213   .2377428     8.52   0.000     1.555967    2.496459
       _cons |  -.3198632   1.086409    -0.29   0.769    -2.468738    1.829011
------------------------------------------------------------------------------
```

2. Hypothesis Testing – Confidence interval approach

Two –sided Test

$$H_0 : \alpha = 0$$
$$H_A : \alpha \neq 0$$

Confidence interval approach →

a = 5% significance level

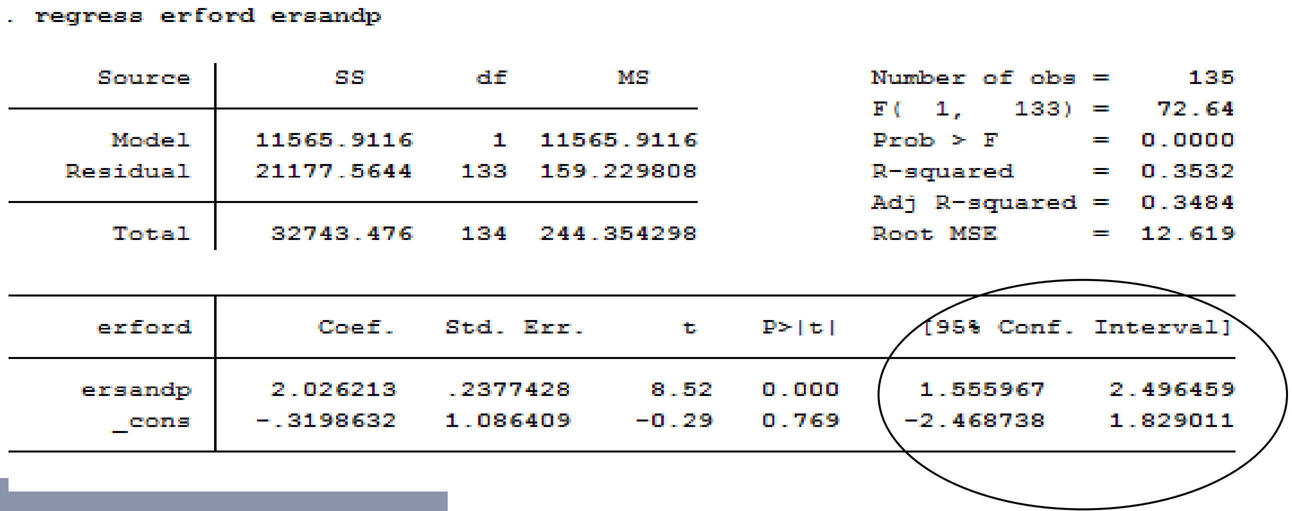
$$\hat{a} \pm t_{crit} SE(\hat{a})$$

(-0.052026,0.0533)
We do **not** reject the Null Hypothesis for a ; thus a is **Insignificant,** since 0 lies within confidence interval

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

$$\hat{\beta} \pm t_{crit} SE(\hat{\beta})$$

(0.995,1.01889)
We reject the Null Hypothesis for b ; thus b is **significant,** since 0 does **not** lie within confidence interval



```
. regress erford ersandp

      Source |       SS       df       MS              Number of obs =     135
-------------+------------------------------           F(  1,   133) =   72.64
       Model |  11565.9116     1  11565.9116           Prob > F      =  0.0000
    Residual |  21177.5644   133  159.229808           R-squared     =  0.3532
-------------+------------------------------           Adj R-squared =  0.3484
       Total |   32743.476   134  244.354298           Root MSE      =  12.619

------------------------------------------------------------------------------
      erford |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   2.026213   .2377428     8.52   0.000     1.555967    2.496459
       _cons |  -.3198632   1.086409    -0.29   0.769    -2.468738    1.829011
------------------------------------------------------------------------------
```

3. Hypothesis Testing – p-value approach

Two –sided Test

$$H_0 : \alpha = 0$$
$$H_A : \alpha \neq 0$$

$$H_0 : \beta = 0$$
$$H_A : \beta \neq 0$$

P-value approach

→

a = 5% significance level

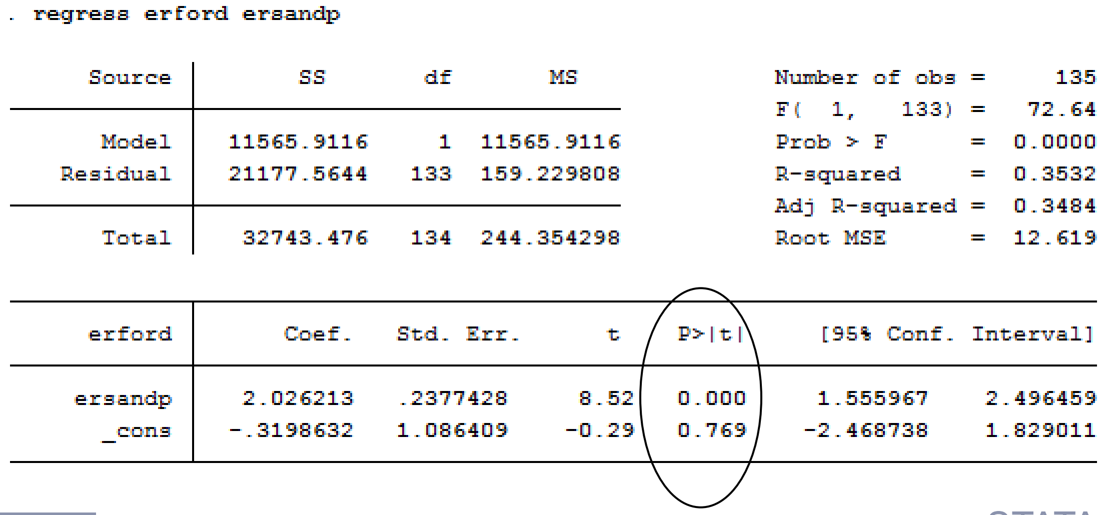
p-value is termed as the

“plausibility” of the Null Hypothesis;

the smaller the p-value, the less plausible is the null hypothesis.

Is the largest significance level at which we fail to reject

the null hypothesis.

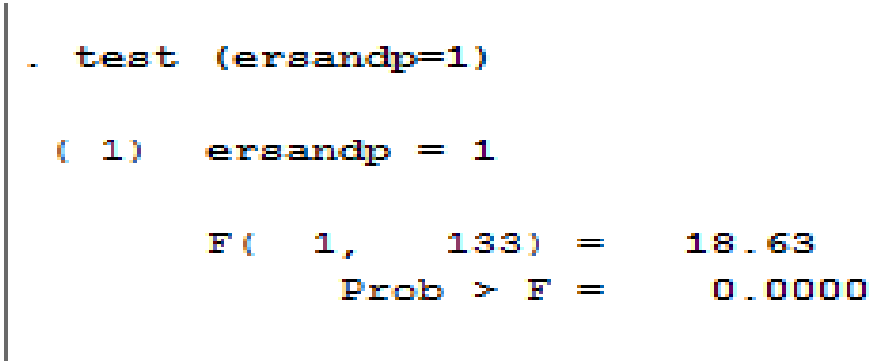


```
. regress erford ersandp

      Source |       SS       df       MS              Number of obs =     135
-------------+------------------------------           F(  1,   133) =   72.64
       Model |  11565.9116     1  11565.9116           Prob > F      =  0.0000
    Residual |  21177.5644   133  159.229808           R-squared     =  0.3532
-------------+------------------------------           Adj R-squared =  0.3484
       Total |   32743.476   134  244.354298           Root MSE      =  12.619

------------------------------------------------------------------------------
      erford |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   2.026213   .2377428     8.52   0.000     1.555967    2.496459
       _cons |  -.3198632   1.086409    -0.29   0.769    -2.468738    1.829011
------------------------------------------------------------------------------
```

Simple Linear Regression

Suppose now we want to test the null hypothesis that

$$H_0 : \beta = 1$$
$$H_A : \beta \neq 1$$



```
. test (ersandp=1)

 ( 1)  ersandp = 1

       F(  1,   133) =   18.63
            Prob > F =    0.0000
```

- ❑ F(1,133) : F-statistic with one restriction and T-k=135-2=133
- ❑ We reject the Null hypothesis since the p-value=0.000

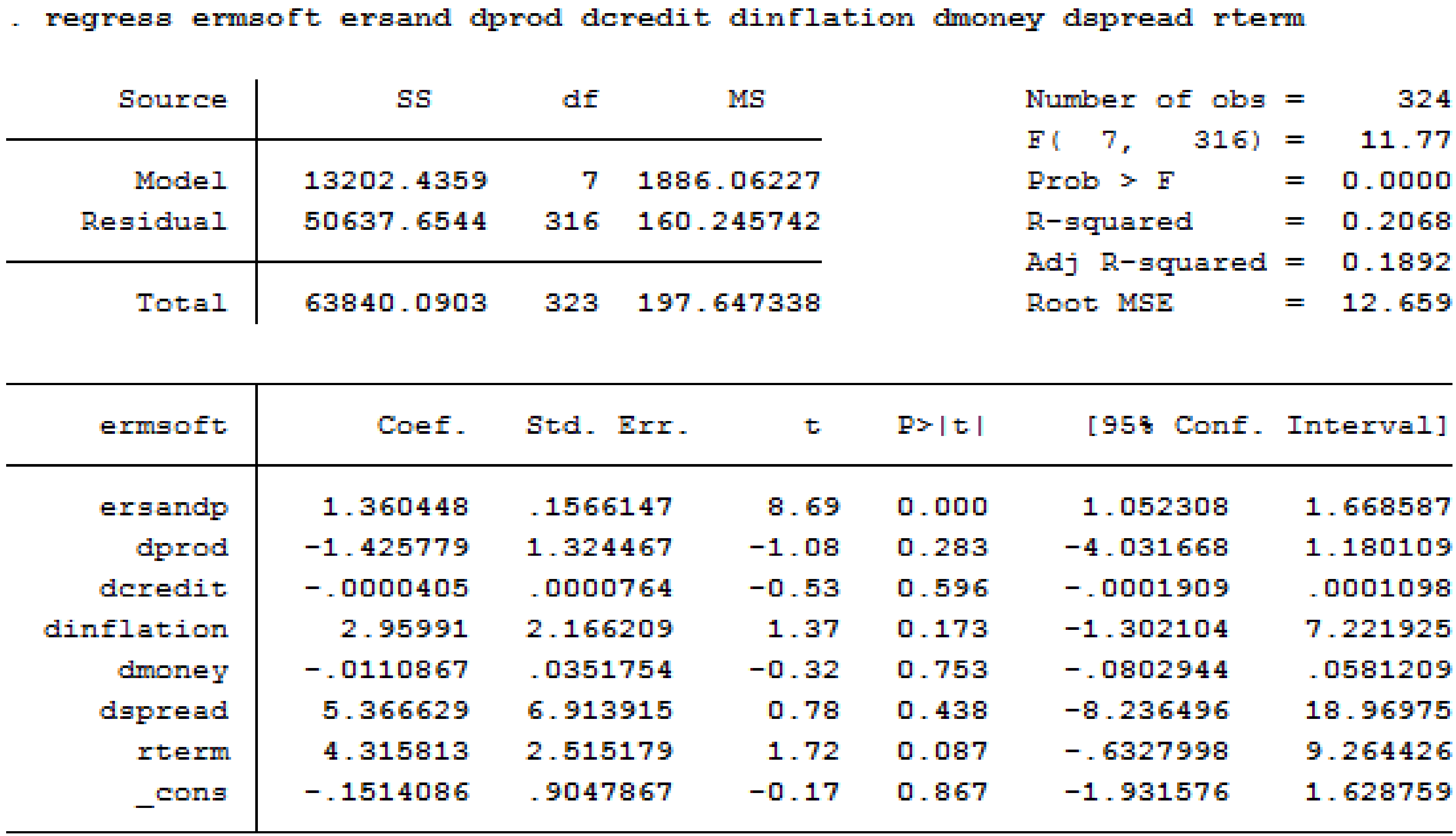
Sata manual on testing linear hypotheses after estimation:
www.stata.com/manuals/rtest.pdf

Empirical Example 6 :

Multivariate Linear Regression

Multivariate Linear Regression

- ❑ Open macro.dta
- ❑ Run the regression



```
. regress ermsoft ersand dprod dcredit dinflation dmoney dspread rterm

      Source |       SS       df       MS              Number of obs =     324
-------------+------------------------------           F(  7,   316) =   11.77
       Model |  13202.4359     7  1886.06227           Prob > F      =  0.0000
    Residual |  50637.6544   316  160.245742           R-squared     =  0.2068
-------------+------------------------------           Adj R-squared =  0.1892
       Total |  63840.0903   323  197.647338           Root MSE      =  12.659

------------------------------------------------------------------------------
     ermsoft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   1.360448   .1566147     8.69   0.000     1.052308    1.668587
       dprod |  -1.425779   1.324467    -1.08   0.283    -4.031668    1.180109
     dcredit |  -.0000405   .0000764    -0.53   0.596    -.0001909    .0001098
  dinflation |    2.95991   2.166209     1.37   0.173    -1.302104    7.221925
      dmoney |  -.0110867   .0351754    -0.32   0.753    -.0802944    .0581209
     dspread |   5.366629   6.913915     0.78   0.438    -8.236496    18.96975
       rterm |   4.315813   2.515179     1.72   0.087    -.6327998    9.264426
       _cons |  -.1514086   .9047867    -0.17   0.867    -1.931576    1.628759
------------------------------------------------------------------------------
```

6. **Testing Multiple Hypothesis : The F- test**

The t-test was used to test single- hypothesis (one coefficient hypothesis)

For more than one parameter hypothesis we use F - statistic

$$t - staticitc = \frac{RRSS - URSS}{URSS} \times \frac{T-k}{m} \sim F(m, T-k)$$

$$Z \sim t_{T-k}$$

$$Z^2 \sim t^2_{T-k} \sim F(1, T-k)$$

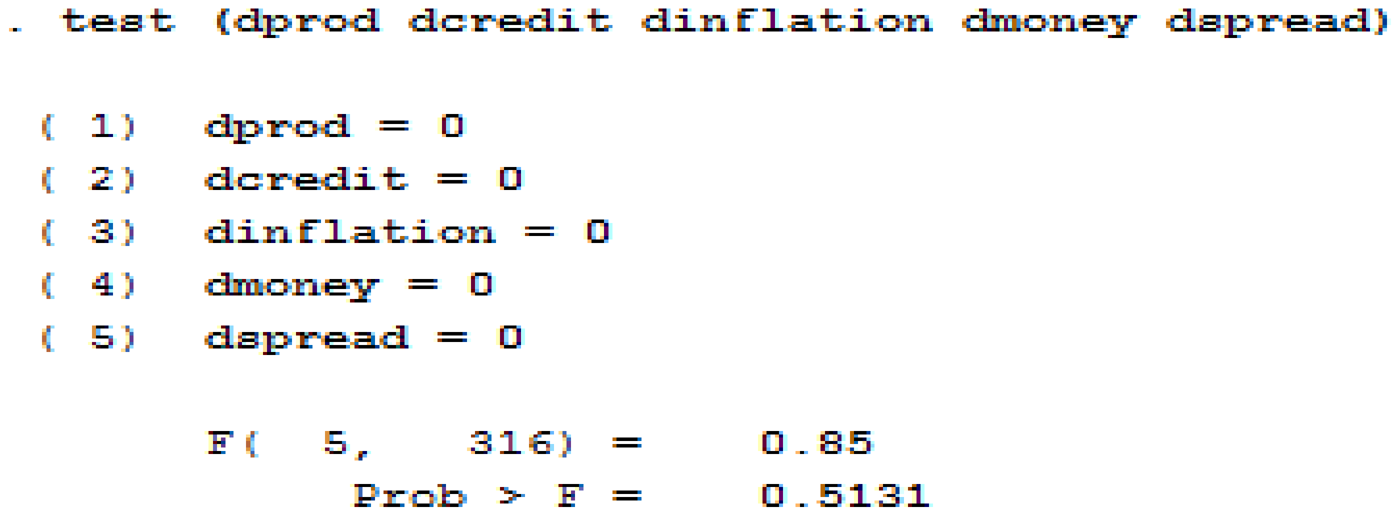
- •URSS: Residual sum of squares from unrestricted regression
- •RRSS : Residual sum of squares from restricted regression
- •m : number of restrictions
- •T : number of observations
- •k : number of regressors in unrestricted regression

Reject the Null when $F \succ t_{crit}$

Multivariate Linear Regression

Test whether *dprod dcredit dinflation dmoney dspread* are jointly zero using F-test

$$H_0 : \beta_2 = 0 and \beta_3 = 0 and \beta_4 = 0 and \beta_5 = 0 and \beta_6 = 0$$
$$H_A : \beta_2 \neq 0 or \beta_3 \neq 0 or \beta_4 \neq 0 or \beta_5 \neq 0 or \beta_6 \neq 0$$



```
. test (dprod dcredit dinflation dmoney dspread)

 ( 1)  dprod = 0
 ( 2)  dcredit = 0
 ( 3)  dinflation = 0
 ( 4)  dmoney = 0
 ( 5)  dspread = 0

       F(  5,   316) =    0.85
            Prob > F =    0.5131
```

The Null Hypothesis cannot be rejected

STATA Workshop II

Agenda (1)

I. Testing for heteroskedasticity

- ➢Wald Test
- ➢Breusch-Pagan- Godfrey Test

II. Testing for serial correlation

- ➢ Durbin- Watson Test
- ➢Breusch-Godfrey Test

III. Testing for non normality

- ➢Jarque – Bera Test
- ➢Dummies

IV. Testing for multicollinearity

- ➢Correlation Matrix
- ➢Add/Remove of Explanatory variable

Agenda (2)

V. Testing for linear relationship between Y and X

- ➢Ramsey RESET Test

VI. Univariate Time Series Modelling of US Home Prices

- ➢Autoregressive Process (AR)
- ➢Moving Average Process (MA)
- ➢ARMA model

CLRM

Assumptions underlying the CLR model

$E(u_t)=0$ The errors have zero mean (Mean Independence)

$\mathrm{var}(u_t)=\sigma^2$ The variance of the errors is constant (Homoskedasticity)

$\mathrm{cov}(u_i,u_j)=0$ The errors are linearly independent of one other

$\mathrm{cov}(u_t,x_t)=0$ There is no relationship between the error and the corresponding variate x

u_t □ $N(0,\sigma^2)$ The errors are normally distributed (Normality)

Violation of one of the above assumptions may lead to

1. Biased coefficient estimates
2. Biased standard errors
3. Inappropriate distributions

Thus, we need to test and solve for these violations

Statistical Distributions for diagnostic tests

The tests that detect any violation are based on the calculation of test statistic

LM test

- Chi-squared distribution
- df equal to the number of restrictions

Wald Test

- F-distribution
- df equal to $(m, T-k)$

$$\frac{\chi^2(m)}{m} \overset{A}{\sim} F(m, T-k)$$

1st Assumption: Mean Independence

$E(u_t)=0$ The errors have zero mean (Mean Independence)

•If we include a constant term in the regression equation, this assumption **will never be** violated.

•If financial theory suggest a model **without** intercept then

a. R-squared may be negative (the sample average of y explains more of the variation in y than the explanatory variables x).

b. Severe biases in slope coefficients.

Testing for Heteroskedasticity

2nd Assumption: Homoskedasticity

$\text{var}(u_t)=\sigma^2$ The variance of the errors is constant (Homoskedasticity)

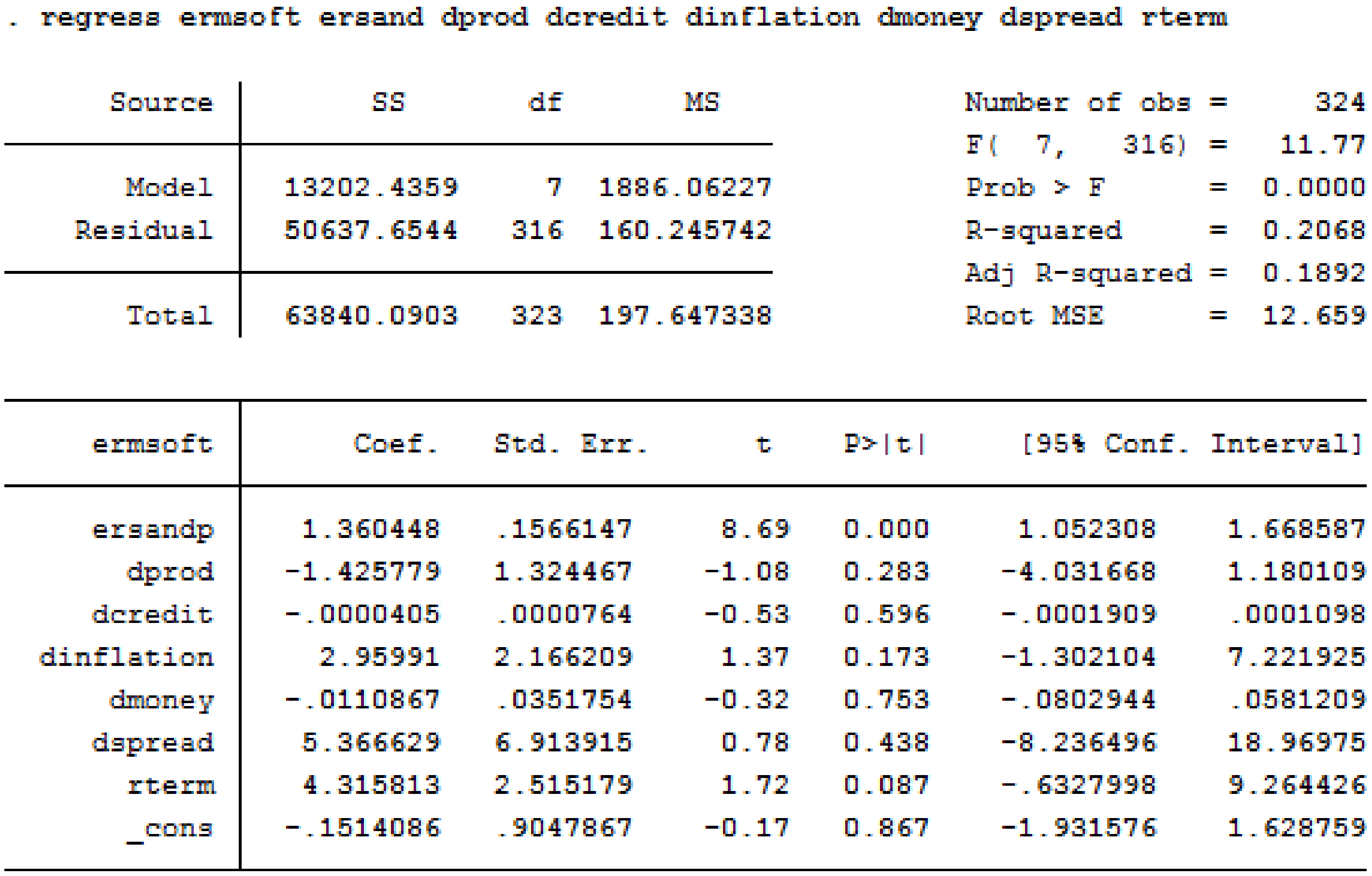
•You can plot the residuals with an explanatory variable; however, it is difficult to detect the presence or not of heteroskedasticity, since we do not know the form of the latter.

Thus, we use a number of tests that detect heteroskedasticity

here in STATA: ***White Test***

Testing for heteroskedasticity

- ❑ Load macro.dta



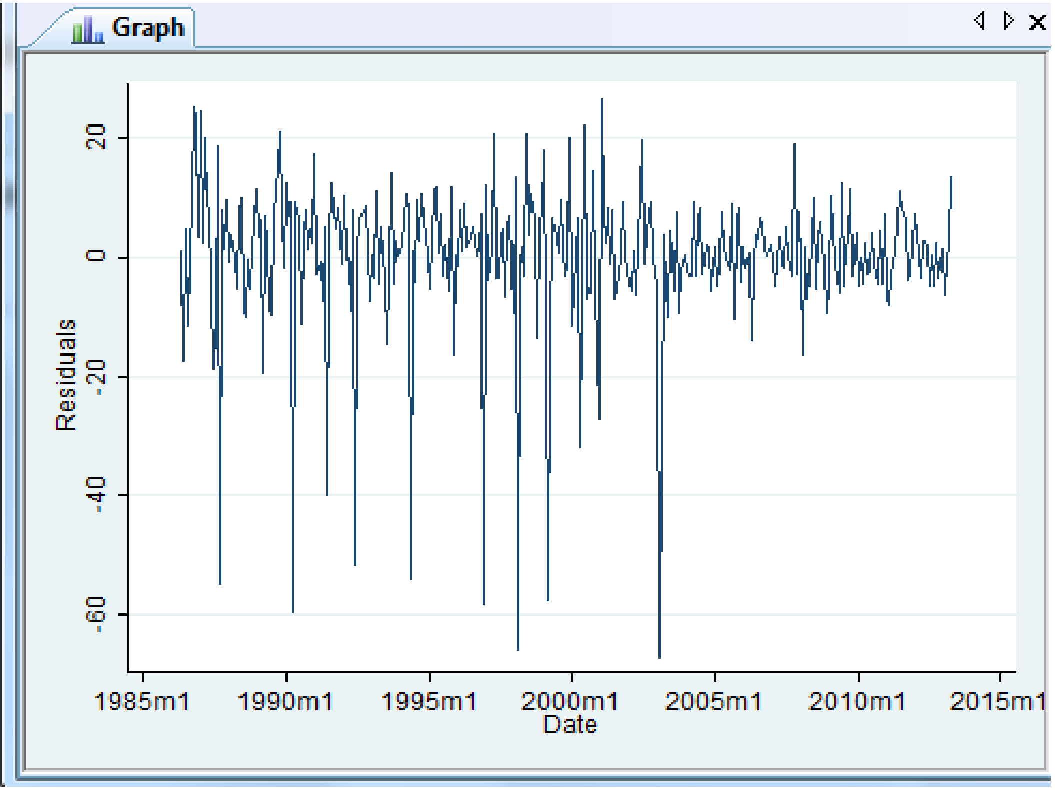
```
. regress ermsoft ersand dprod dcredit dinflation dmoney dspread rterm

      Source |       SS       df       MS              Number of obs =     324
-------------+------------------------------           F(  7,   316) =   11.77
       Model |  13202.4359     7  1886.06227           Prob > F      =  0.0000
    Residual |  50637.6544   316  160.245742           R-squared     =  0.2068
-------------+------------------------------           Adj R-squared =  0.1892
       Total |  63840.0903   323  197.647338           Root MSE      =  12.659

------------------------------------------------------------------------------
     ermsoft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   1.360448   .1566147     8.69   0.000     1.052308    1.668587
       dprod |  -1.425779   1.324467    -1.08   0.283    -4.031668    1.180109
     dcredit |  -.0000405   .0000764    -0.53   0.596    -.0001909    .0001098
  dinflation |    2.95991   2.166209     1.37   0.173    -1.302104    7.221925
      dmoney |  -.0110867   .0351754    -0.32   0.753    -.0802944    .0581209
     dspread |   5.366629   6.913915     0.78   0.438    -8.236496    18.96975
       rterm |   4.315813   2.515179     1.72   0.087    -.6327998    9.264426
       _cons |  -.1514086   .9047867    -0.17   0.867    -1.931576    1.628759
------------------------------------------------------------------------------
```

Testing for heteroskedasticity

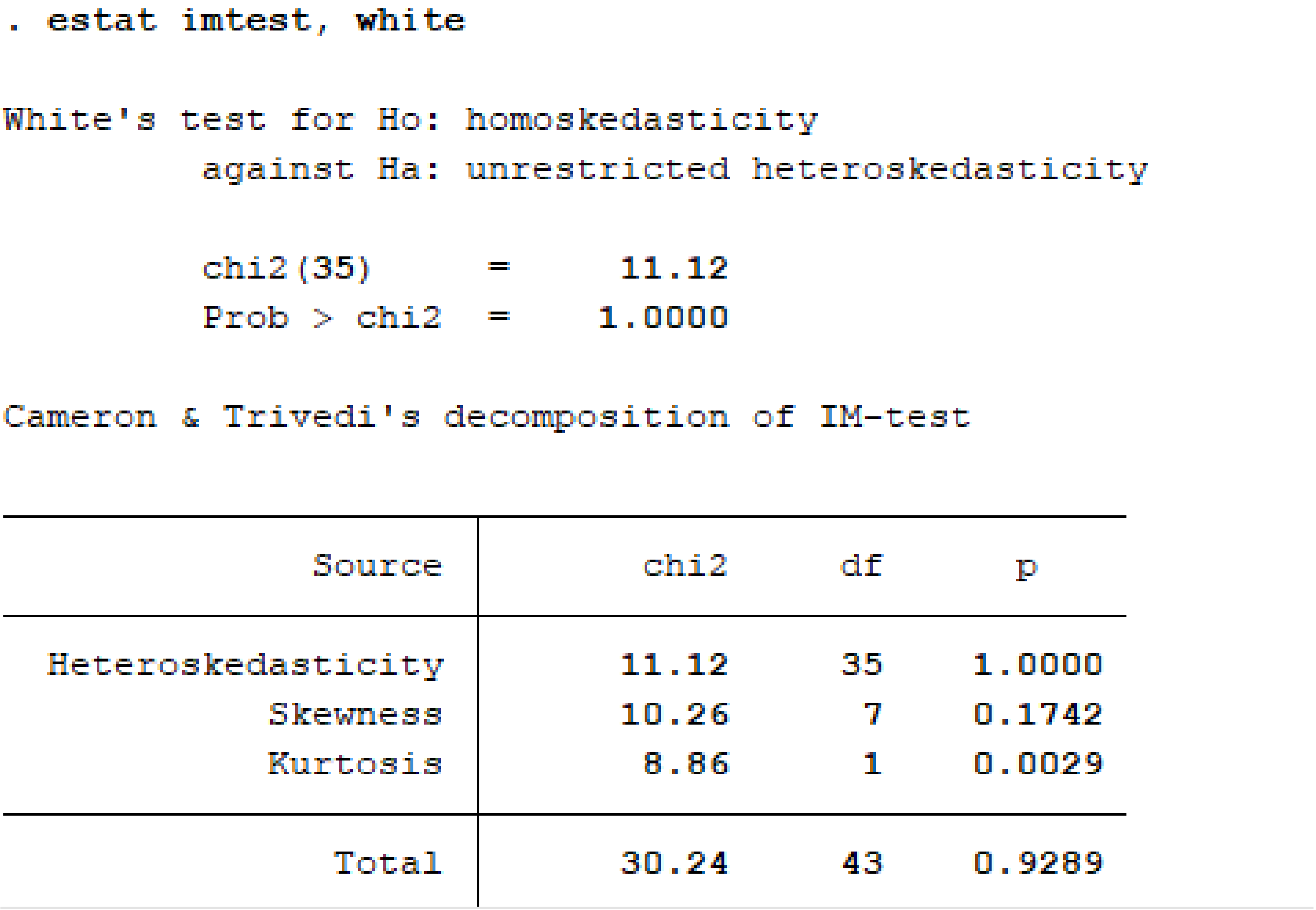
Graphical Illustration of possible heteroskedasticity

In the command window write

twoway (tsline resid)



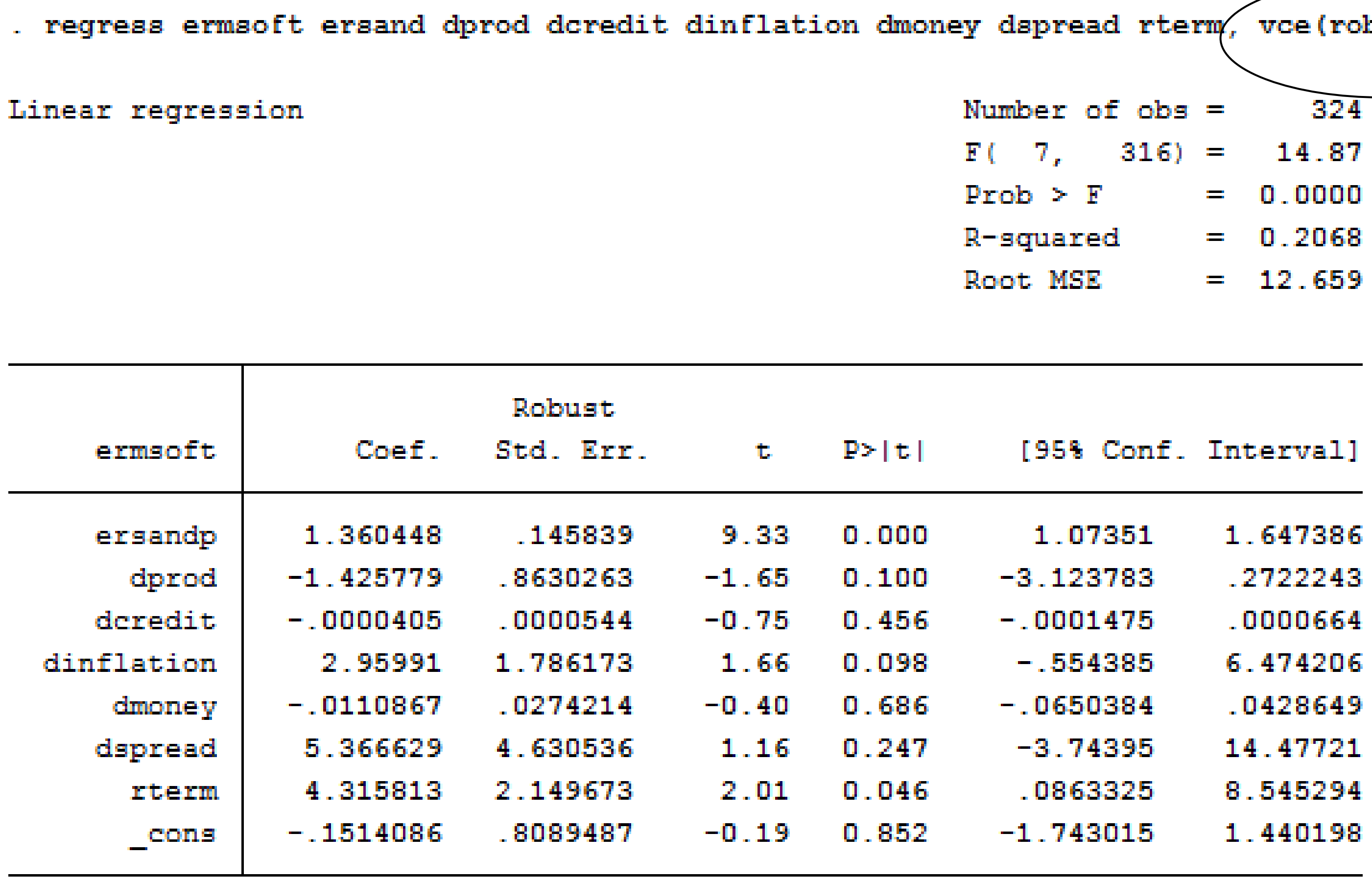
If the residuals of the regression have systematically changing variability over the sample, that is a sign of heteroskedasticity



```
. estat imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(35)     =     11.12
         Prob > chi2  =    1.0000

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      11.12     35    1.0000
            Skewness |      10.26      7    0.1742
            Kurtosis |       8.86      1    0.0029
---------------------+-----------------------------
               Total |      30.24     43    0.9289
---------------------------------------------------
```

Correcting for heteroskedasticity

White standard errors



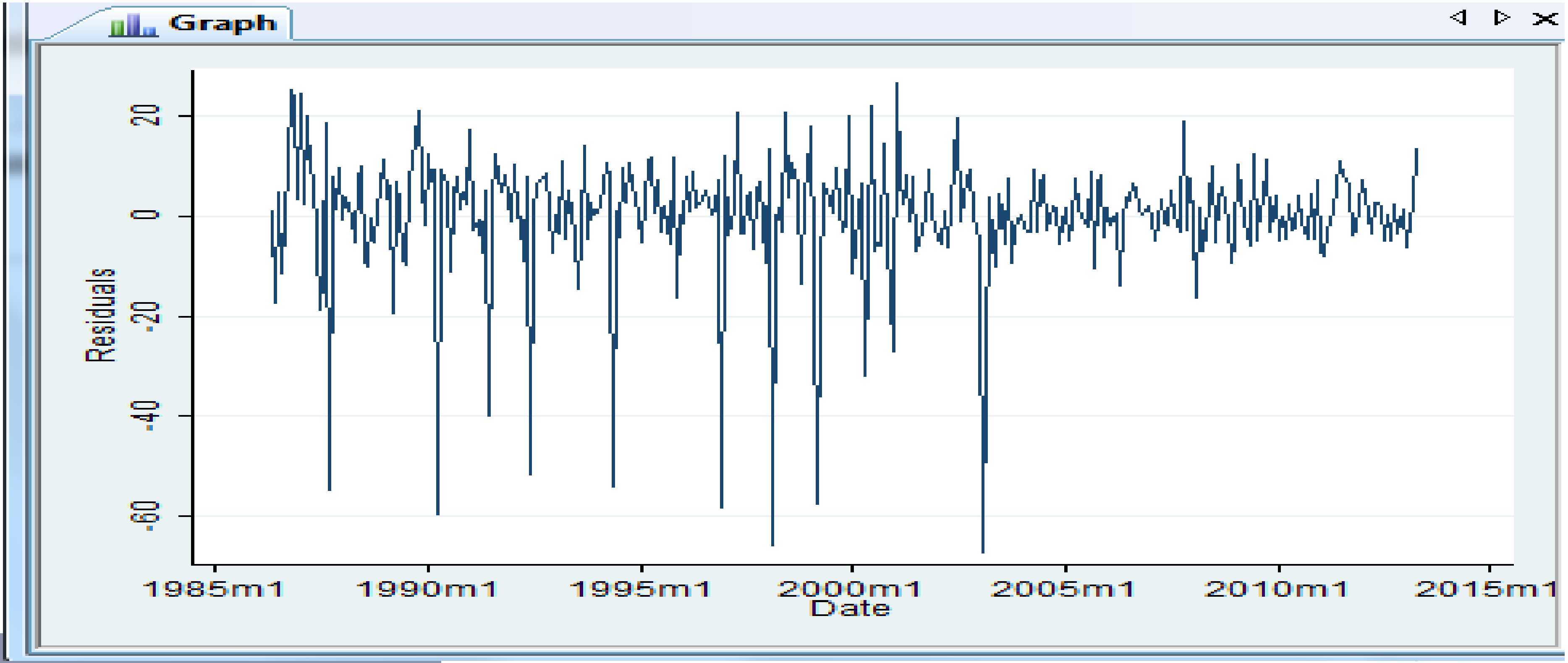
```
. regress ermsoft ersand dprod dcredit dinflation dmoney dspread rterm, vce(robust)

Linear regression                                      Number of obs =     324
                                                       F(  7,   316) =   14.87
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2068
                                                       Root MSE      =  12.659

------------------------------------------------------------------------------
             |               Robust
     ermsoft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   1.360448    .145839     9.33   0.000      1.07351    1.647386
       dprod |  -1.425779   .8630263    -1.65   0.100    -3.123783    .2722243
     dcredit |  -.0000405   .0000544    -0.75   0.456    -.0001475    .0000664
  dinflation |    2.95991   1.786173     1.66   0.098     -.554385    6.474206
      dmoney |  -.0110867   .0274214    -0.40   0.686    -.0650384    .0428649
     dspread |   5.366629   4.630536     1.16   0.247     -3.74395    14.47721
       rterm |   4.315813   2.149673     2.01   0.046     .0863325    8.545294
       _cons |  -.1514086   .8089487    -0.19   0.852    -1.743015    1.440198
------------------------------------------------------------------------------
```

Testing for Serial Correlation/Autocorrelation

3rd Assumption: No serial autocorrelation

$\text{cov}(u_i, u_j) = 0$ The errors are linearly independent of one other

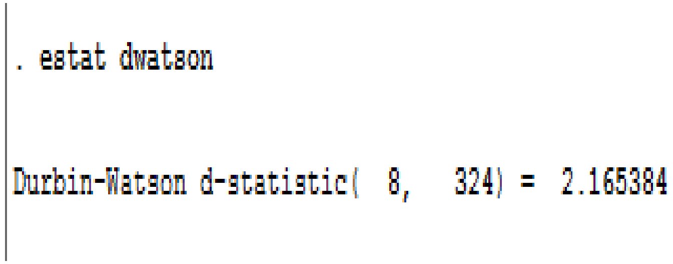
- Errors are uncorrelated with one another
- If errors are not uncorrelated with one another, it would be stated that they are ***autocorrelated or serially correlated.***



Testing for serial correlation

How detect autocorrelation??

From the estimation output a simple test is Durbin –Watson Test



```
. estat dwatson

Durbin-Watson d-statistic(  8,   324) =  2.165384
```

The Durbin- Watson test statistic is 2.19, close to 2

Durbin – Watson(DW) is a test for **first order autocorrelation.**(tests the relationship between an error and its immediately previous value).

$$u_t = \rho u_{t-1} + v_t$$

$$H_0 : \rho = 0 (\text{No Autocorrelation})$$

$$H_A : \rho \neq 0 (\text{Autocorrelation})$$

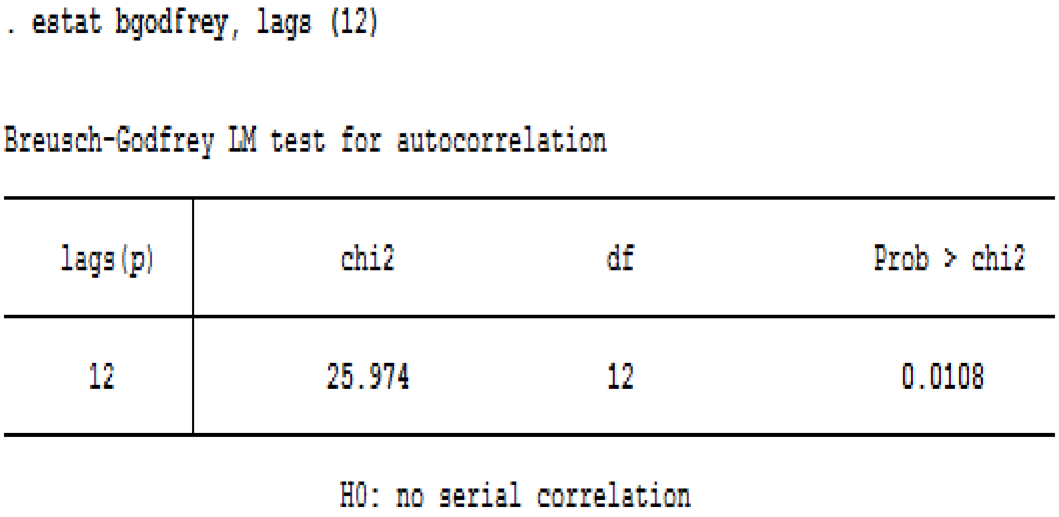
$$DW \approx 2(1-\rho)$$

Conditions for DW to be a valid Test

1. Existence of a constant term.
2. Non –stochastic regressors.
3. **No** lags of dependent variable.

Testing for serial correlation

Another more robust test than DW is **Breush – Godfrey Test**



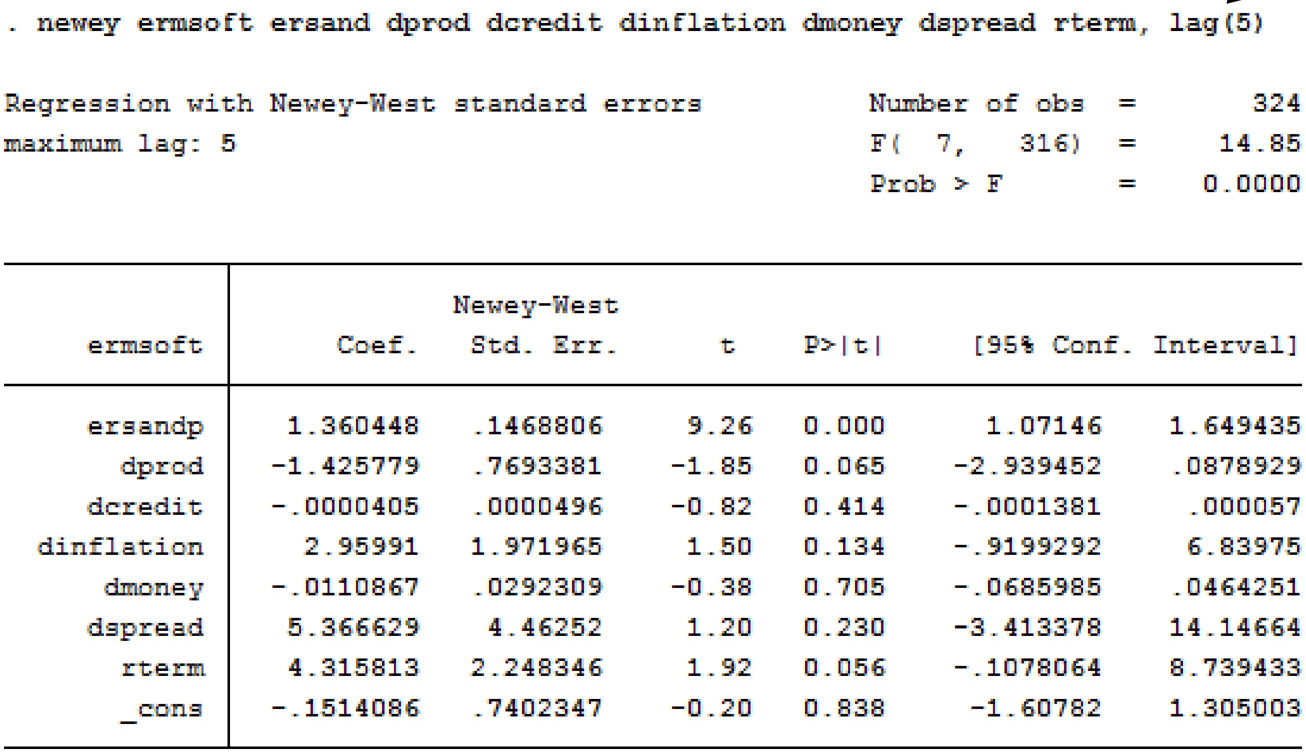
```
. estat bgodfrey, lags (12)

Breusch-Godfrey LM test for autocorrelation
---------------------------------------------------------------------
    lags(p)  |          chi2               df                 Prob > chi2
-------------+-------------------------------------------------------
      12     |         25.974              12                   0.0108
---------------------------------------------------------------------
                        H0: no serial correlation
```

Specify the number of lags equal to12. There is no an obvious answer to this, you can experiment on a range of number. You can relate the number of lags with the frequency of your data. (for monthly data use 12, for quarterly data 4, etc)

Correction

Newey & West for both ***heteroskedasticity and autocorrelation***

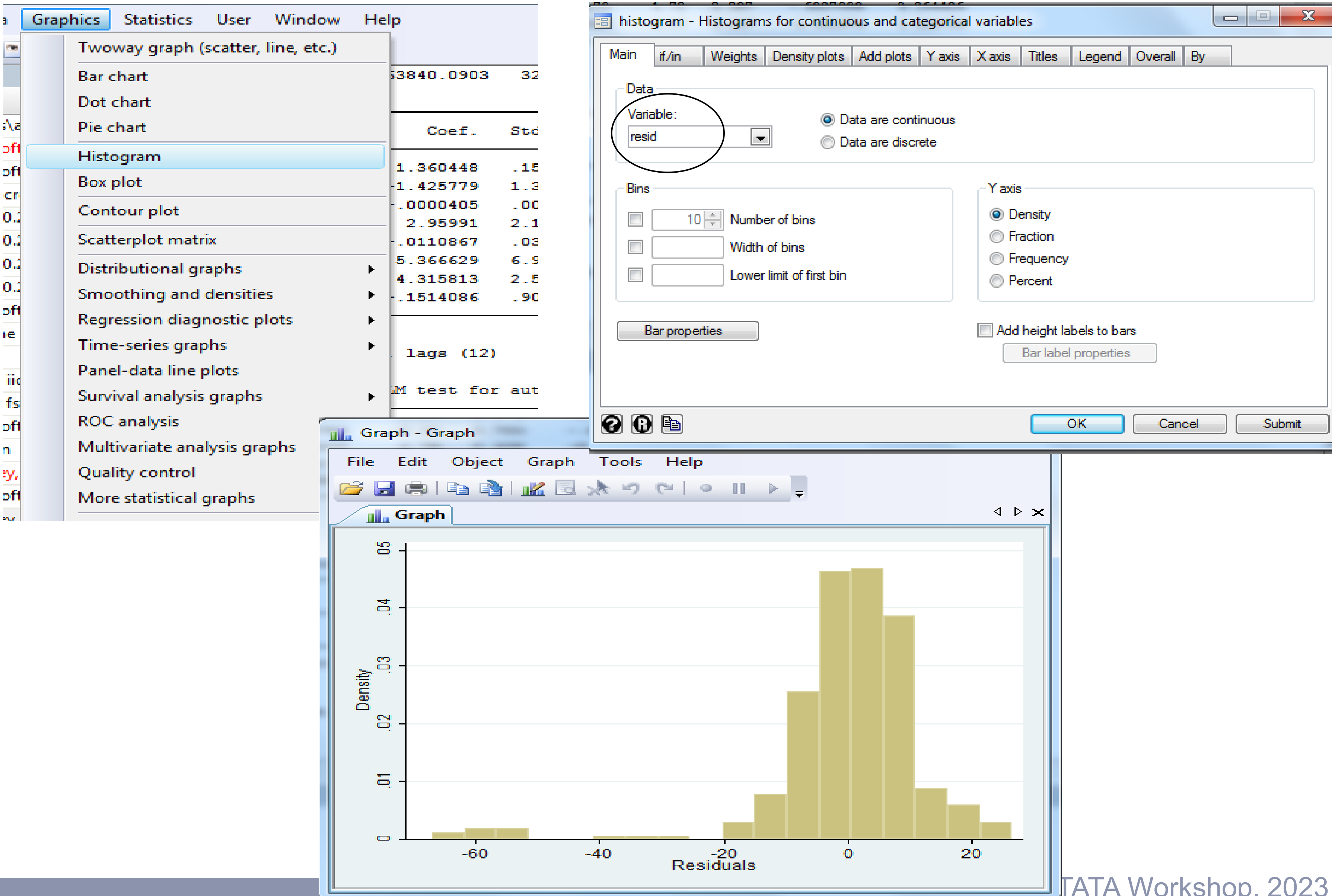
$$m(T) = floor[4(T/100)^{2/9}].$$



```
. newey ermsoft ersand dprod dcredit dinflation dmoney dspread rterm, lag(5)

Regression with Newey-West standard errors          Number of obs  =       324
maximum lag: 5                                      F(  7,    316) =     14.85
                                                    Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
     ermsoft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   1.360448   .1468806     9.26   0.000      1.07146    1.649435
       dprod |  -1.425779   .7693381    -1.85   0.065    -2.939452    .0878929
     dcredit |  -.0000405   .0000496    -0.82   0.414    -.0001381     .000057
  dinflation |    2.95991   1.971965     1.50   0.134    -.9199292     6.83975
      dmoney |  -.0110867   .0292309    -0.38   0.705    -.0685985    .0464251
     dspread |   5.366629    4.46252     1.20   0.230    -3.413378    14.14664
       rterm |   4.315813   2.248346     1.92   0.056    -.1078064    8.739433
       _cons |  -.1514086   .7402347    -0.20   0.838     -1.60782    1.305003
------------------------------------------------------------------------------
```

Testing for Non- Normality

Testing for Non- Normality



Variable: resid

● Data are continuous
○ Data are discrete

□ 10 Number of bins
□ Width of bins
□ Lower limit of first bin

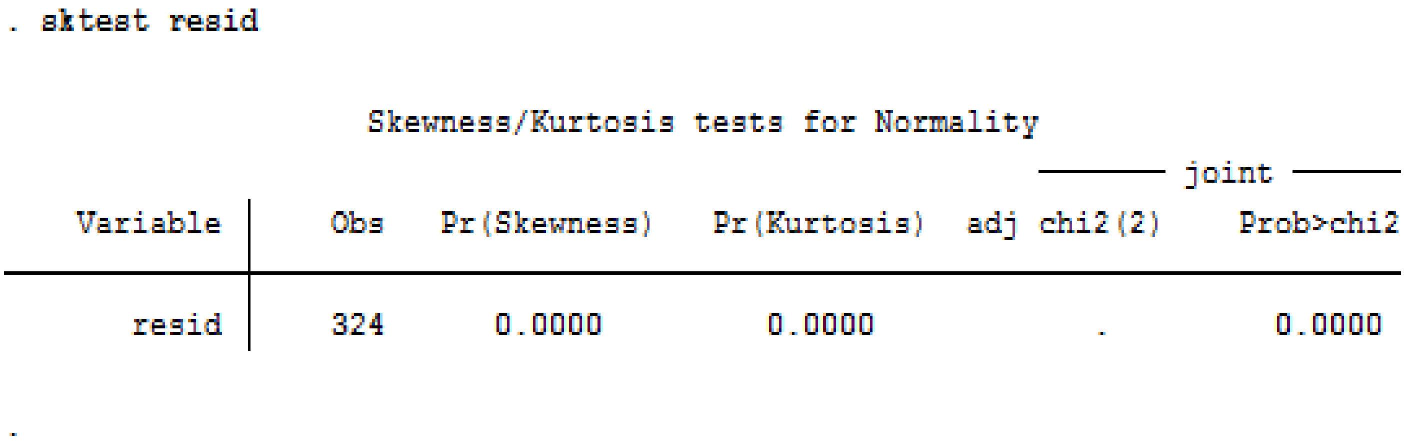
● Density
○ Fraction
○ Frequency
○ Percent

□ Add height labels to bars

Testing for Non- Normality

Null Hypothesis : Normality (Both Kurtosis and Skewness are those of the normal distribution, Skewness =0 and Kurtosis =3

Skewness and Kurtosis Test : A variation of Jarque Bera test



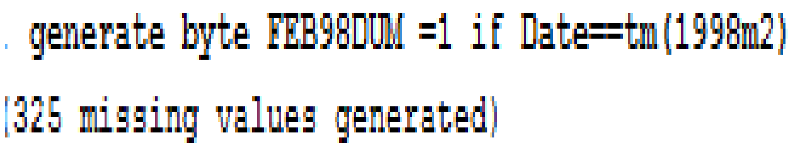
```
. sktest resid

                   Skewness/Kurtosis tests for Normality
                                                        ------ joint ------
    Variable |    Obs   Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+---------------------------------------------------------------
       resid |    324      0.0000         0.0000            .         0.0000

.
```

Testing for Non- Normality

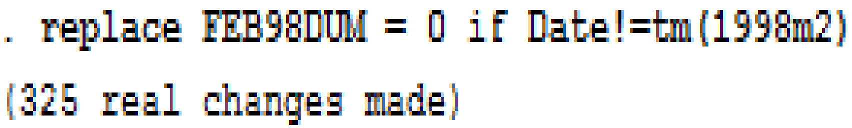
What to do if evidence of non-normality is found?

- Central Limit Theory: The test statistics will asymptotically follow the appropriate distribution even in the absence of error normality ; the sample mean converges to a normal distribution.
- Financial/ Economic theory : One or two very extreme residuals cause a rejection of normality assumption (outliers)

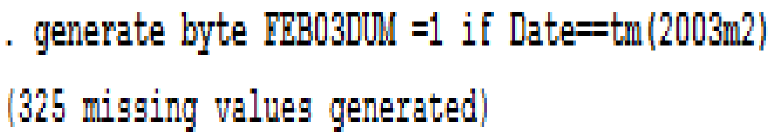
A plausible solution : Use of dummy variables



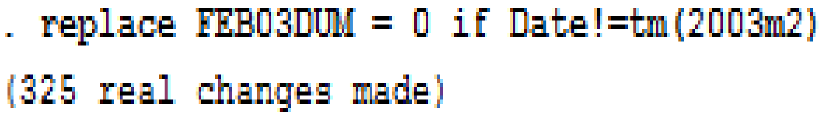
```
. generate byte FEB98DUM =1 if Date==tm(1998m2)
(325 missing values generated)
```



```
. replace FEB98DUM = 0 if Date!=tm(1998m2)
(325 real changes made)
```

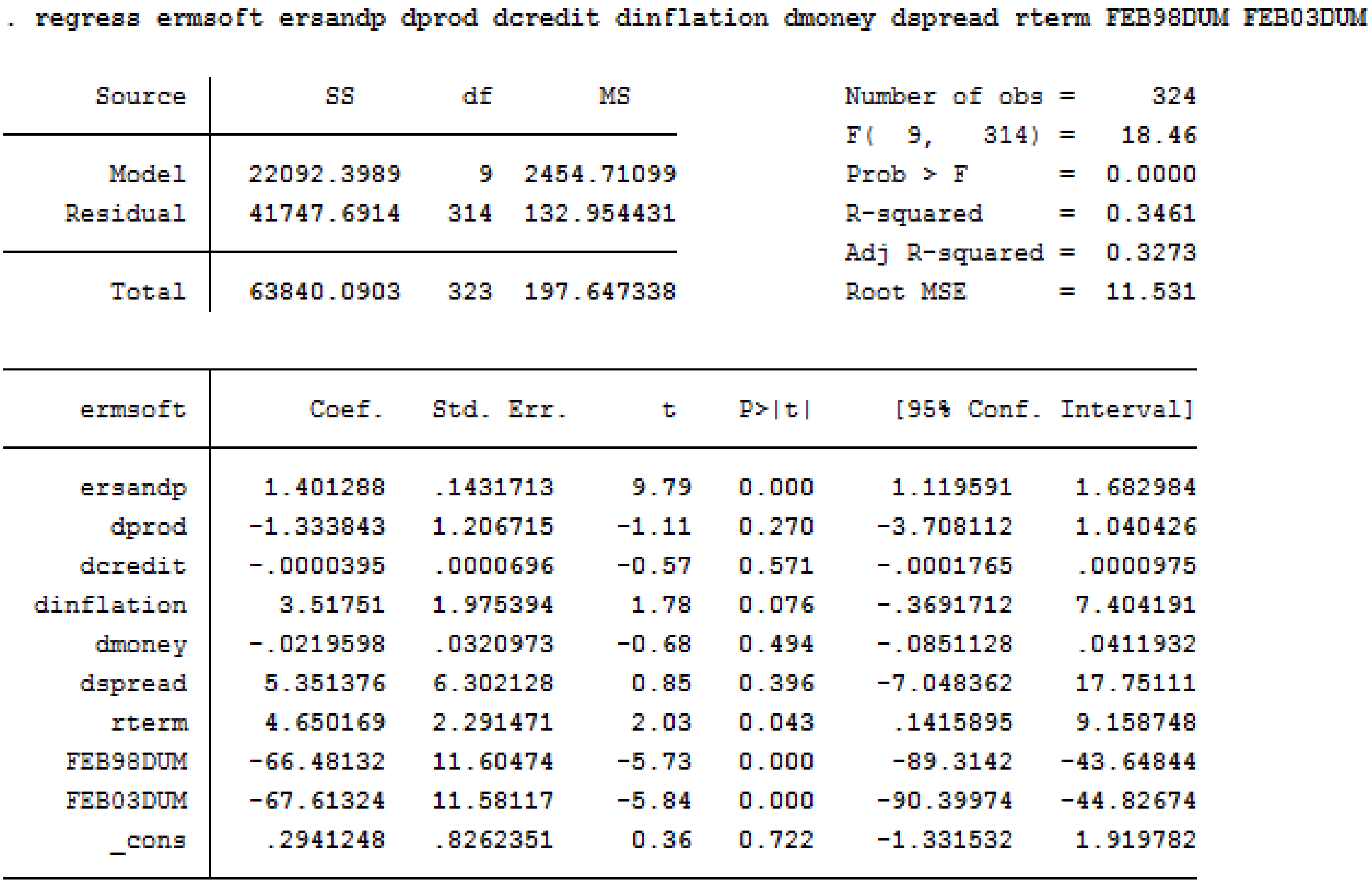


```
. generate byte FEB03DUM =1 if Date==tm(2003m2)
(325 missing values generated)
```



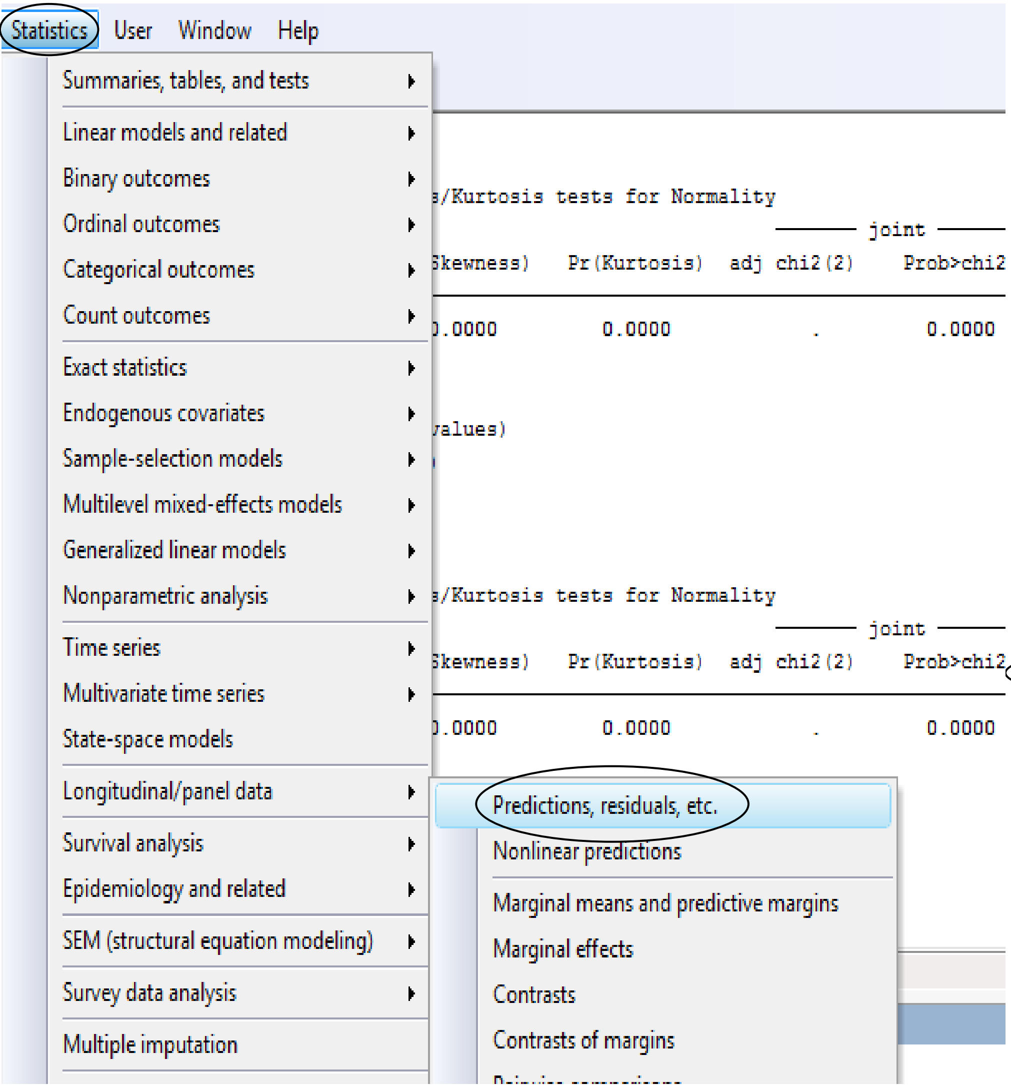
```
. replace FEB03DUM = 0 if Date!=tm(2003m2)
(325 real changes made)
```

Testing for Non- Normality

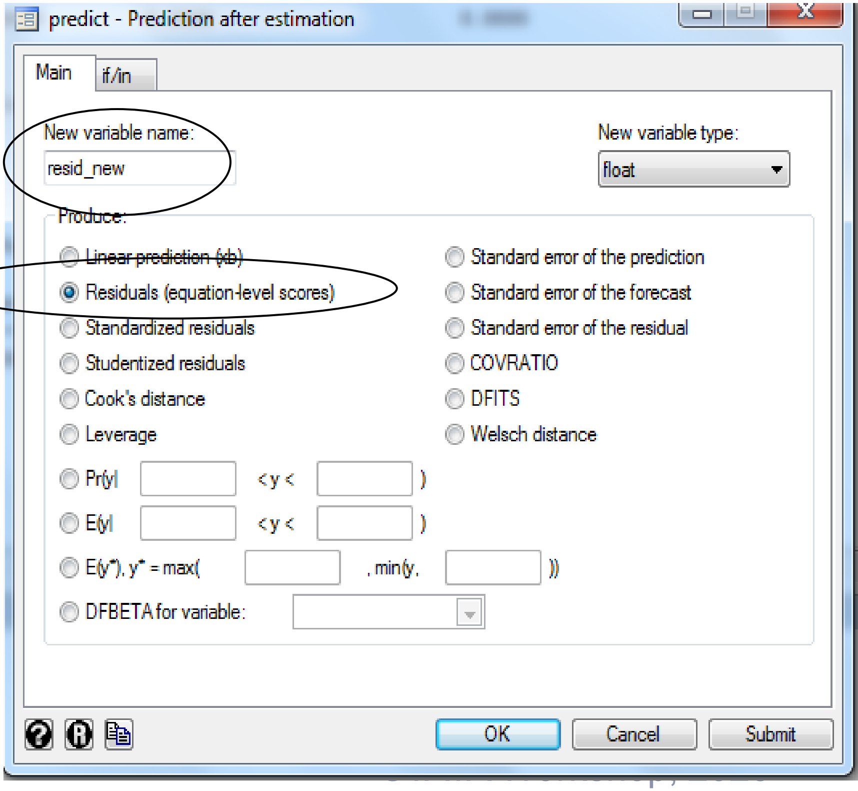


```
. regress ermsoft ersandp dprod dcredit dinflation dmoney dspread rterm FEB98DUM FEB03DUM

      Source |       SS       df       MS              Number of obs =     324
-------------+------------------------------           F(  9,   314) =   18.46
       Model |  22092.3989     9  2454.71099           Prob > F      =  0.0000
    Residual |  41747.6914   314  132.954431           R-squared     =  0.3461
-------------+------------------------------           Adj R-squared =  0.3273
       Total |  63840.0903   323  197.647338           Root MSE      =  11.531

------------------------------------------------------------------------------
     ermsoft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ersandp |   1.401288   .1431713     9.79   0.000      1.119591    1.682984
       dprod |  -1.333843   1.206715    -1.11   0.270     -3.708112    1.040426
     dcredit |  -.0000395   .0000696    -0.57   0.571     -.0001765    .0000975
  dinflation |    3.51751   1.975394     1.78   0.076     -.3691712    7.404191
      dmoney |  -.0219598   .0320973    -0.68   0.494     -.0851128    .0411932
     dspread |   5.351376   6.302128     0.85   0.396     -7.048362    17.75111
       rterm |   4.650169   2.291471     2.03   0.043      .1415895    9.158748
    FEB98DUM |  -66.48132   11.60474    -5.73   0.000      -89.3142   -43.64844
    FEB03DUM |  -67.61324   11.58117    -5.84   0.000     -90.39974   -44.82674
       _cons |   .2941248   .8262351     0.36   0.722     -1.331532    1.919782
------------------------------------------------------------------------------
```

Testing for Non- Normality



```
s/Kurtosis tests for Normality
                                        ------ joint ------
Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2

).0000         0.0000            .          0.0000

values)

s/Kurtosis tests for Normality
                                        ------ joint ------
Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2

).0000         0.0000            .          0.0000
```



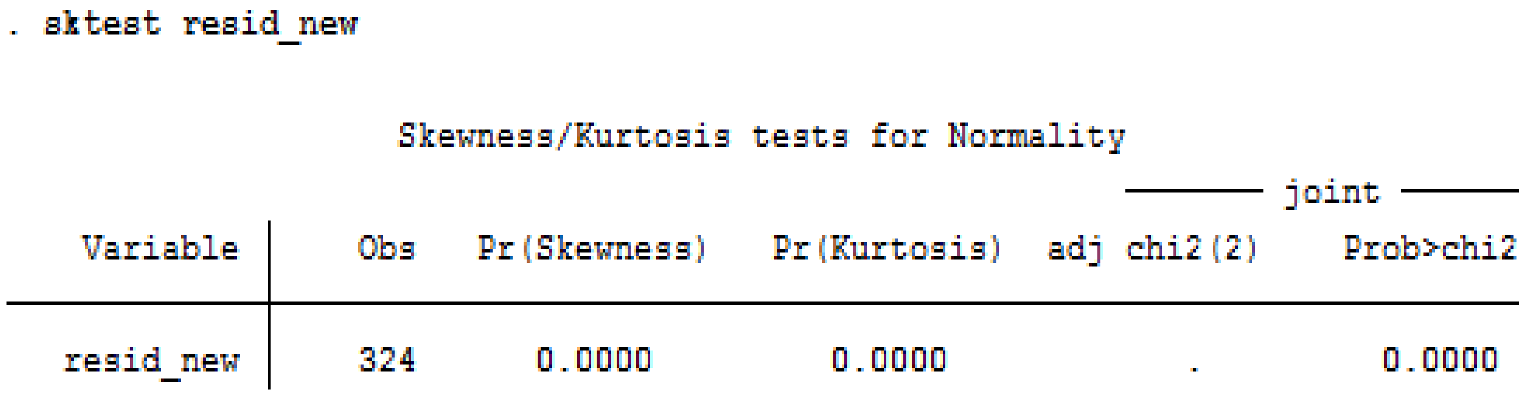
New variable name: resid_new

New variable type: float

Produce:

○ Linear prediction (xb)
● Residuals (equation-level scores)
○ Standardized residuals
○ Studentized residuals
○ Cook's distance
○ Leverage
○ Pr(y| <y<)
○ E(y| <y<)
○ E(y*), y* = max(, min(y,))
○ DFBETA for variable:
○ Standard error of the prediction
○ Standard error of the forecast
○ Standard error of the residual
○ COVRATIO
○ DFITS
○ Welsch distance

Testing for Non- Normality



```
. sktest resid_new

                    Skewness/Kurtosis tests for Normality
                                                          ------ joint ------
    Variable |        Obs  Pr(Skewness)  Pr(Kurtosis) adj chi2(2)   Prob>chi2
-------------+---------------------------------------------------------------
   resid_new |        324     0.0000        0.0000            .        0.0000
```

A long way for residuals to follow a normal distribution…

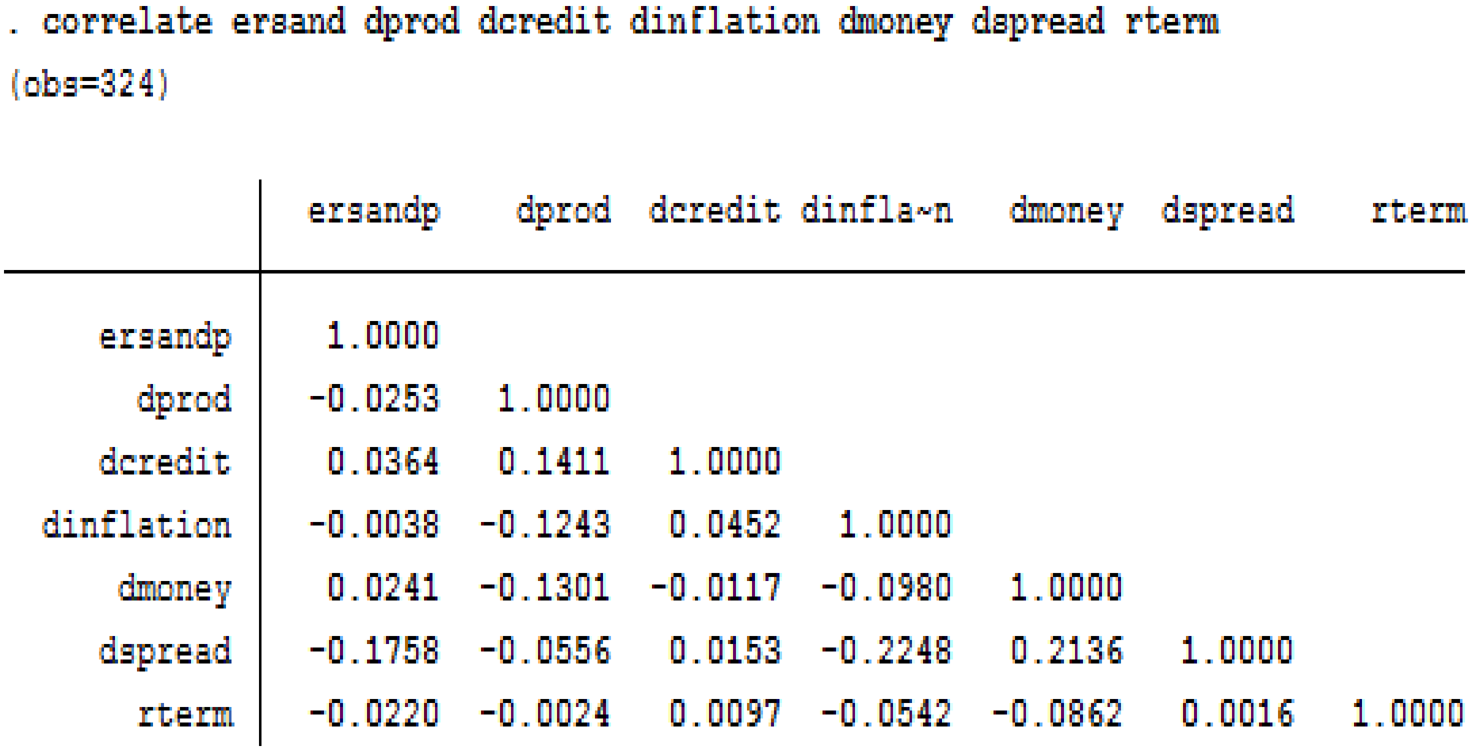
Testing for Multicollinearity

Testing for multicollinearity

Implicit assumption: explanatory variables not correlated/orthogonal with one another.

How detect multicollinearity?? Two easy ways:

1. Use the correlation matrix of the explanatory variables



```
. correlate ersand dprod dcredit dinflation dmoney dspread rterm
(obs=324)

             |  ersandp    dprod  dcredit dinfla~n   dmoney  dspread    rterm
-------------+---------------------------------------------------------------
     ersandp |   1.0000
       dprod |  -0.0253   1.0000
     dcredit |   0.0364   0.1411   1.0000
  dinflation |  -0.0038  -0.1243   0.0452   1.0000
      dmoney |   0.0241  -0.1301  -0.0117  -0.0980   1.0000
     dspread |  -0.1758  -0.0556   0.0153  -0.2248   0.2136   1.0000
       rterm |  -0.0220  -0.0024   0.0097  -0.0542  -0.0862   0.0016   1.0000
```

Testing for multicollinearity

- ***Problems if near Multicollinearity is present but ignored***

1. R-squared will be high, but the individual coeff. will have high standard errors, so that regression "looks good" as a whole, but the individual variables are not significant.

Remark: Multicollinearity does ***not*** *affect the value of R-squared in the regression.*

2. Regression becomes very sensitive to small changes in the specification; add/remove an independent variable leads to large changes in the coeff. values or significances of other variables.

3. Wide confidence intervals for the parameters; inappropriate results for significance tests.

Testing for multicollinearity

- ***Solutions to the problem of multicollinearity***

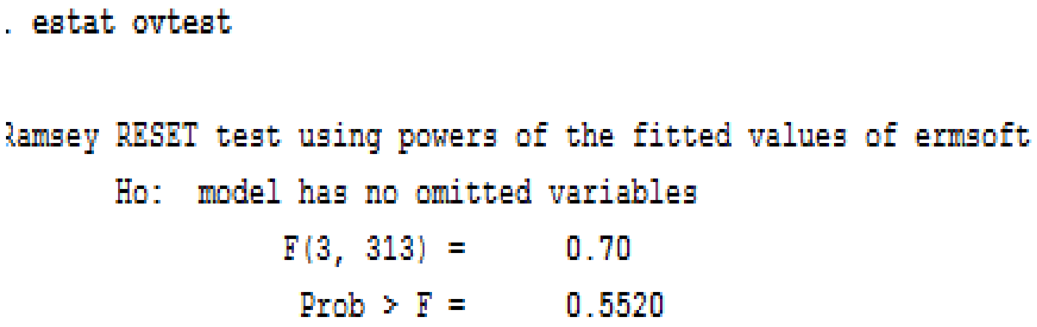
1. Use of ridge Regressions
2. Use of Principal Component Analysis.
3. **Ignorance** of multicollinearity **if the model is statistically appropriate.**
4. **Drop** one of the collinear variables
5. **Transform** the highly correlated variables into a **ratio** and **include** the **ratio** and **not** the individual explanatory variables.
6. A sufficient history of data : longer run of data/ higher frequent data/pooled data.

Testing for linear relationship between Y and X

Testing for linear relationship between Y and X

Linearity or not???

Ramsey RESET test : View → Stability Diagnostics → Ramsey RESET Test



```
. estat ovtest

Ramsey RESET test using powers of the fitted values of ermsoft
       Ho:  model has no omitted variables
                 F(3, 313) =      0.70
                  Prob > F =      0.5520
```

$$H_0 : Linearity$$

$$H_A : Non-Linearity$$

Thus, we cannot reject the null hypothesis that the model has no omitted variables. In other words, we do not find strong evidence that the chosen linear functional form of the model is incorrect.

The end