

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

🔜 Stata/IC 12.1 - [Results]		Dete	
File Edit Data Graphics	Statistics User Window Help	Data	Ð
🞽 🗐 🌒 🗐 🔹 🛄 🗐		Editor	
Review T # × # Command _rc There are no items to show.	Control (Control (Contro) (Contro) (Contro) (Contro) (Contro) (Contro) (Contro) (Con	Copyright 1985-2011 StataCorp LP 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) hse: Output Window	Variables T ¥ × Variable Label There are no items to show. Variables Wariables Window
	Licensed to: iSolutions	••••••••••••••••••••••••••••••••••••••	Comparison
	Command	д	
	Command V	Vindow	Variables ▲ Name ▲ Label ■ Type ■ Format ↓ Value Label ↓ Notes ■ Data ▼
\\soton.ac.uk\ude\personalfiles\	users\as14n14\mydocuments		CAP NUM OVR



- 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 \rightarrow Log \rightarrow 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



- 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 \rightarrow Do
 - Or click the button "Execute" in the Do file editor window

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



- 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



- 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 (<u>www.stata.com/manuals/dencode.pdf</u>)
 - destring (<u>www.stata.com/manuals/u24.pdf#u24.2Categoricalstringvariables</u>)
 - format (<u>www.stata.com/manuals/dformat.pdf</u>)





2. Go to File \longrightarrow Import \longrightarrow

Stata/IC 12.1 - [Results]

		Second Se	
File	Edit Data Graphics Sta	istics User Window Help	8
	Open Ctrl+0		
	Save Ctrl+	(R)	Variables T 4 ×
-	Save As Ctrl+Shift+	// //	Variable Label
	View	stics/Data Analysis StateCorp	There are no items to show.
	P.	4905 Lakeway Drive	
	Do	College Station, Texas 77845 USA	
	Filename	800-STATA-PC http://www.stata.com	
	Change Working Directory	979-696-4600 stata@stata.com	
		979-696-4601 (fax)	
	Log		
	Import	Excel spreadsheet (*.xls;*.xlsx)	
	Export	Text data created by a spreadsheet	
	Print	Text data in fixed format	۰
	Evample Datasets	Text data in fixed format with a dictionary	Properties P ×
	Example Datasets		≙ : + +
	Recent Datasets	Unformatted text data	□ Variables ▲
	Exit	SAS XPORT	Name
		Haver Analytics database	Label 😑
		ODBC data source	Туре
		XML data	Format
			Value Label
			Notes
			🗆 Data
			🛨 Filename 🔻
\\soto	on.ac.uk\ude\personalfiles\users	as14n14\mydocuments	CAP NUM OVR



Empirical Example 1 :

Cross Sectional Data



- 1. Go to folder *Empirical Examples* \longrightarrow *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



Import Excel										
	Even	fler								
	E:\M	lije: Iy Docum	nents\ALBA\Module\Presentations\Week 1\STATA 1\For m Browse							
vvorksneet: Ceil range:										
example1 A1:A61 A1:A61										
Import first row as variable names Variable case: preserve										
Import all data as strings										
F	Previe	ew: (sho	wing rows 2-51 of 61)							
		grade								
	2	13								
	3	41								
	4	47								
	5	54								
	6	60								
	7	67								
	8	73	<u>र</u>							
			OK Cancel	đ						



Empirical Example 1 – <u>Cross Sectional Data(3)</u>

🔜 Stata/IC 12.1 - [Results]				
File Edit Data Graphics	Statistics User Window Help			8
📁 🖬 🗐 🗐 💽 • 📠 • 🗌	🛃 🕶 🔜 🔜 🕕 😣			
Review T # × # Command _rc 1 infile grade using " 2 2 grade 3 3 import excel "E:\M	50-user Stata network perpetual licens Serial number: 30120521121 Licensed to: iSolutions University of S Notes:	College Station, Texas 77845 USA 800-STATA-PC http://www.stata.co 979-696-4600 stata@stata.com 979-696-4601 (fax) se:	The second secon	Label grade
	. infile grade using "E:\My Documents\	ALBA\Module\Presentations\Week 1\STATA	1 -	
			Properties	Ŧх
	Command		₽ 🔒 i + →	
			□ Variables	<u>^</u>
			Name	grade
			Label	grade E
			Туре	byte
			Format	%10.0g
			Value Label	
			Notes	
\\soton ac uk\ude\nersonalfiles\	users\as14n14\mydocuments			
(soconaciak (ade (personalities)	users (userner4 (mydocuments		CAr	



🏭 Stata/IC 12.1 - [Results]			
File Edit Data Graphics	Statistics User Window Help		8
😂 🗟 🖨 🗐 🖸 • 📠 • 🛯			
Review T # ×	979-696-4601 (fax)	Variables	▼ ₽ ×
# Command _rc		Variable	Label
1 import excel "E:\M	Serial number: 30120521121	grade	grade
	Licensed to: iSolutions		
	University of Southampton	1	
	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		
	· · · · · · · · · · · · · · · · · · ·	<	•
	Command	Properties	4 X
		Variables	
		Name	
		Label	-
		Type	=
		Format	
		Value Label	
		Notes	
		🗆 Data	
			Ŧ
\\soton.ac.uk\ude\personalfiles\	users\as14n14\mydocuments	CAP	NUM OVR
	SIA	ATA Works	shop, 2023

OIKONOMIKO ATHENS UNIVERSITY Empirical Example 1 – <u>Cross</u> Sectional Data(5) ΠΑΝΕΠΙΣΤΗΜΙΟ OF ECONOMICS AND BUSINESS AOHNON х Data Editor (Browse) - [Untitled] File Edit View Data Tools 🍸 🚼 😤 🔟 💂 🚰 🗔 I 🗈 🛍 I 🛃 🔂 grade[1] 13 Variables **무** Snapshots grade ۰ Value 🔧 Filter variables here 13 1 2 41 ✓ Variable Label 47 з 🗹 grade grade Ξ 4 54 5 60 6 67 7 73 41 8 9 46 53 10 57 11 12 61 **中** Properties 69 13 Variables æ. 14 80 Name grade 15 94 Label grade 16 27 Type byte Ξ 17 36 %10.0g Format Value Label 18 48 Notes 54 19 Data 20 56 Filename 71 21

22

85

Label



Empirical Example 1 – <u>Cross Sectional Data(6)</u>

х

		Data Editor	(Browse) -	[Untitled]
--	--	-------------	------------	------------

File	e Edi	it View	Data	a Tools									
2		d (2) <u>e</u>		Describe da	ta	►		Describe dat	a in memory				
		gri		Data Editor		•	(Describe dat	a in file	\mathcal{I}			
		grad		Create or ch	hange data	•		Describe dat	a contents (co	odeb	oook)		Ą
Sn	1			Variables M	anager			Inspect varia	bles		r	iables here	
apsh	2			Data utilitie	s	•		List data				Label	
ots	3			Sort		•		Compactly li	st variable na	mes		grade	
	4			Combine d	atasets	•		Summary sta	atistics				
	5			Matrices, M	lata language		F			Г			
	7			Matrices, ad	do language	•				9			
	8			Other utiliti	es	•							
	9		46				1						
	10		53										
	11		57										
	12		61							P	roperties		Ą
	13		69							E	Variables		
	14		80								Name	grade	
	15		94								Label	grade	
	16		27								Туре	byte	Ξ
	17		36								Format	%10.0g	
	18		48								Value Labe	1	
	19		54								Notes		
_													



Write the name of the variable (here is grade)

describe - Describe data in memory Variables: (leave empty for all variables) grade Examples: vr* all variables starting with "vr"	×						
xyz-abc all variables between xyz and abc		. describe gra	de				
 Display only variable names Display only general information 		variable name	storage type	display format	value label	variable label	_
 Do not abbreviate variable names Display variable number along with name 		grade	byte	€10.0g		grade	-
Replace data in memory with describe report Clear dataset from memory							
OK Cancel Su	omit	Command	/	*			
		Outp	out	Wi	ndo	W	



ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Empirical Example 1 – <u>Cross Sectional Data(8)</u>



. summarize g	rade				
Variable	Obs	Mean	Std. Dev.	Min	Max
grade	60	58	17.81068	13	97

Output Window



• 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



• Graph

œ	Stata/IC 12.1 - [Re	esults]			
F	ile Edit Data	Grap	ohics Statistics User Window	He	elp
Ċ	i 🛃 🖷 I 🗐 🖻		Twoway graph (scatter, line, etc.)		
Rev	view		Bar chart		
#	Command		Dot chart		
1	import excel "E:\l		Pie chart		
2	describe grade		Histogram		
3	summarize grade		Box plot		
4	codebook		Contour plat		-
5	summarize grade		Contour plot		-
6	summarize grade		Scatterplot matrix		
7	table grade		Distributional graphs	•	
8	pnorm grade		Smoothing and densities	×	Kernel density estimation
9	kdensity grade		Regression diagnostic plots	►	Lowess smoothing
			Time-series graphs	•	Local polynomial smoothing
			Panel-data line plots		Density distribution sunflower plot
			Survival analysis graphs	٦	Density-distribution sunnower prot
			ROC analysis	►	
			Multivariate analysis graphs	►	
			Quality control	►	
			More statistical graphs	►	
			Table of graphs		
			Manage graphs	►	
			Change scheme/size		



We can visualize the shape of distribution

•Distinction between normal and non-normal distributions



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

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



Go to Graphics \rightarrow Distributional Graphs \rightarrow Histogram



Close the above window \longrightarrow Go to File \longrightarrow Save as...



Empirical Example 2 :

Time Series Data

1. Go to folder *Empirical Examples* \longrightarrow *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



5. Open STATA from PC - lab

6. Go to File \longrightarrow Import \longrightarrow Excel Spreadsheet

7. Browse example_2.xlsx

	Sta	ta/IC 12.1 - [Re	sults]	-								The second second second
Fil	e)	Edit Data	Graphics	s Statistic	s User	Window	Help					
C	7	Open		Ctrl+O								
		Save		Ctrl+S								
-		Save As	Ctrl+	Shift+S	<u> </u>			(R)				
-		View			1 1	i i	1	12.1	Copyright	1985-20	011 StataCorp	LP
		Do			.stics/	Data Ana	lysis		StataCorp			
		Filename							4905 Lakew	ay Driv	/e	
		ritename							College St	ation,	Texas 77845 U	SA
		Change Worki	ing Direct	ory					800-STATA-	PC	http://www	stata.com
		Log		•					979-696-46	00 01 (fav	stata@stat	a.com
									575 656 46	l (Lan	.,	
	-	Import		•	E	cel spread	sheet (*.xls;	*.xlsx)				
		Export		•	Т	ext data cre	ated by a s	preadshe	et			
		Print		•	Т	ext data in f	fixed forma	ıt				
		Example Datas	sets		Т	ext data in f	fixed forma	it with a d	lictionary			
		Recent Datase	ts	•	U	nformatteo	l text data					
					S	AS XPORT						
		Exit			н	aver Analy	tics databa	ce.		entatio	ons\Week 1\STA	TA 1\For me\empirical_
				> ndic		DBC data						
						DBC data s	ource					
				1.	X	ML data						



8. Go to Graphics \rightarrow smoothing and densities \rightarrow Kerne

Kernel Density Function

抗 Stata/IC 12.1 - [Re	sults]					and the second second second
File Edit Data	Grap	ohics Statistics User Window	Help			
💕 🛃 🖷 I 🗐 🖻		Twoway graph (scatter, line, etc.)				
Review		Bar chart				
# Command		Dot chart	- F	- '	(R)	
1 import excel "E:\l		Pie chart		7	12.1	Copyright 1985-2011 StataCorp
		Histogram				StataCorp
		Box plot				College Station, Texas 77845
		Contour plot				800-STATA-PC http://www
		Scatterplot matrix				979-696-4600 stata@sta 979-696-4601 (fax)
		Distributional graphs	•	- 7	11-	5 ·
	<	Smoothing and densities	►		Kernel	density estimation
		Regression diagnostic plots	•		Lowess	s smoothing
		Time-series graphs			Local p	polynomial smoothing
		Panel-data line plots			Density	v-distribution sunflower plot
		Survival analysis graphs) mei	enta	ALBA	Module\Presentations\Week 1\S
		ROC analysis	•			
		Multivariate analysis graphs	•			
		Quality control	•			
		More statistical graphs	•			
		Table of graphs				
		Manage graphs	•			
		Change scheme/size				



📧 kdensity - Univariate kernel density estimation	
Main if /in Weights Kemel plot Density plots Add plots Y axis X axis Titles Legend Overall	
Variable: US 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	III Graph - Graph
	File Edit Object Graph Tools Help
Estimate density using the values specified by a variable	
	Kernel density estimate
Suppress graph	2
OK Cancel	
OK Cancel	
OK Cancel	



9. File \longrightarrow Save as \longrightarrow





10. Data→ Describe data → Summary Statistics

Results						
F	ile Edit	Data	Graphics Statistics U	ser	Window Help	
C	i 📙 🖷 i		Describe data	•	Describe data in memory	
Rev	/iew		Data Editor	۲	Describe data in file	
#	Comman		Create or change data	F	Describe data contents (codebook)	
1	import ex		Variables Manager		Inspect variables	
2	kdensity (Data utilities	F	List data	
3	kdensity (Sort	•	Compactly list variable names	
4	kdensity (-		
5	kdensity		Combine datasets	•	Summary statistics	
6	summari:		Matrices, Mata language			
			Matrices, ado language	×		
			Other utilities	F	Obs Mean Std. Dev. 20)23



🗉 summarize - Summary statistics
Main by/if/in Weights
Variables: (leave empty for all variables)
Examples: vr* all variables starting with "vr"
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) Factor-variable display options
OK Cancel Submit



. 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



🗐 summarize - Summary statistics					
Main by/if/in Weights					
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)					
Factor-variable display options					
For additional statistics mea	sures				



11. Statistics — Summaries... Summary and descriptive Statistics

🔒 Stata/IC 12.1 - [Results]	1778.0.					
File Edit Data Graphics Statist	ics User Window Help					
	Summaries, tables, and tests	Þ	<	Summary and descriptive statistics	1	Summary statistics
Review T P ×	Linear models and related	•		Tables >		Means
# Command _rc	Binary outcomes	•		Classical tests of hypotheses		Proportions
1 import excel "E:\M	Ordinal outcomes	•		Nonparametric tests of hypotheses		Ratios
2 kdensity Canada, k	Categorical outcomes	+		Distributional plots and tests		Totals
3 kdensity Canada	Count outcomes	•		Multivariate test of means, covariances, and normality		Pairwise comparisons of means
5 kdensity US kernel(Exact statistics	•	Kur	tosis 4.970008		Confidence intervals
6 summarize	Endogenous covariates	→				Normal CI calculator
7 summarize, detail	Sample-selection models	→				Binomial CI calculator
8 correlate	Multilevel mixed-effects models					Poisson CI calculator
	Generalized linear models	•	ance	e Germany Italy Japan OK OS		Correlations and covariances
	Nonparametric analysis	→				Painwise correlations
	Time series	•	0000	1		Partial correlations
	Multivariate time series	•	7893	3 1.0000		Tetrachoric correlations
	State-space models	5	5940 3925	0 0.6363 1.0000 5 0.4493 0.3974 1.0000		Arith./geometric/harmonic means
	Longitudinal/panel data	•	6523	3 0.6916 0.5573 0.4770 1.0000		Graph means/medians by groups
	Survival analysis	•	6187	0.6235 0.4525 0.3743 0.6839 1.0000		Centiles with CIs
	Epidemiology and related	•				Create variable of percentiles
	SEM (structural equation modeling)	•				Create variable of quantiles



12. Select Correlations and Covariances



Main by/if/in	isplay correlation matrix or covariance matrix UP P P A
Variables: (lea	ve empty for all)
уг* xyz-abc	all variables starting with ''yr'' all variables between xyz and abc
00	OK Cancel Submit


. 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 🗔 🔲 🗮 💌								
Main by/if/in Weights Options	. correlate, (obs=394)	covariance						
Display means, std. dev., min, and max with matrix		Canada	France	Germany	Italy	Japan	UK	US
Ignore display format associated with variables	Canada	.003132						
Display covariances	France	.001943	.003817					
	Germany	.002065	.003117	.004087				
Allow wide matrices to wrap	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
OK Cancel Submit					STA	TA Wo	rkshop	, 2023

We can here define

OIKONOMIKO

AOHNON

ΠΑΝΕΠΙΣΤΗΜΙΟ

Covariance between X and Y variables

ATHENS UNIVERSITY

OF ECONOMICS

AND BUSINESS

• Correlation between X and Y variables

Cov(X,Y) = E(XY) - E(X)E(Y) $\rho = \frac{Cov(X,Y)}{\sqrt{V \operatorname{ar}(X)Var(Y)}}$

14. Close the above window \longrightarrow Go to File \longrightarrow Save as...



Empirical Example 3 :

Transformations of Data & Plots

Time Series Data



1.Go to folder *Example_3* \longrightarrow *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 : <u>Transformation of data & Plots(2)</u>

D	ata Edit	tor (Browse) - [U	ntitled]							*	Variables	
ile	Edit	View Data	Tools								Variable	Label
3		B 🖬 👔	7 🚼 😤 🔟 🛛								Date	Date
		Date[1]		29ian1965				DMY	-		UKMarketInd	ex UK Market Index
Т		Dutc[1]		201000			Veriebles				simpleret	
		Date	UKMarketIn~x	simpleret	logret	^	variables		Ŧ		logret	
	1	1/29/1965	102.53	•	•		🔧 Filter variabl	es here				
	2	2/26/1965	101.56	0094606	0095057		✓ Variable	Label				
	3	3/31/1965	100.72	008271	0083054		Date	Date	.sx",			
	4	4/30/1965	100.85	.0012907	.0012899		✓ UKMarketInde	x UK Market Index				
	5	5/31/1965	101.3	.0044621	.0044521		✓ simpleret					
	6	6/30/1965	98.06	0319842	0325069		☑ logret					
	7	7/30/1965	98.34	.0028554	.0028513		E logici			Ξ		
	8	8/31/1965	100.45	.0214562	.0212292							
	9	9/30/1965	105.38	.0490791	.0479128							
	10	10/29/1965	111.55	.05855	.0569001							
ľ	11	11/30/1965	114.05	.0224115	.022164							
ľ	12	12/31/1965	111.92	018676	0188526		Droperties		д		Properties	
ľ	13	1/31/1966	117.38	.0487848	.0476322		Variables		·			
	14	2/28/1966	120.96	.0304992	.0300434		Name	Date		Ŧ	Variables	
	15	3/31/1966	119.56	0115741	0116416		Label	Date			Name	
	16	4/29/1966	121	.0120442	.0119722		Type	int	-	д	Label	
	17	5/31/1966	126.54	.0457851	.0447679		Format	%tdnn/dd/CCYY		'	Format	
	18	6/30/1966	128.61	.0163585	.0162261		Value Label	, ,			Value Label	
	19	7/29/1966	118.29	0802426	0836453		Notes				Notos	



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.



Go to Graphics \longrightarrow Time-series Graphs \longrightarrow Line Plots

3. Click on Create



🗜 Stata/IC 12.1 - [R	esults]								
File Edit Data	Gra	ohics	Statistics	User	Window	He	lp			
📂 🛃 🖷 🗐 🖻	•	Two	way graph	(scatter,	line, etc.)					
Review		Bario	chart							
# Command		Dot	chart				-	(R)		
import excel "E:\	J	Pie o	:hart				5	12.1	Copyright 1985-2011	Stata
		Histo	ogram						StataCorp	
		Box	plot						4905 Lakeway Drive College Station Tex	ag 77
		Cont	tour plot						800-STATA-PC	http:
				•			- I		979-696-4600	stata
		Scat	terplot mat	nx					979-696-4601 (fax)	
		Distr	ibutional g	raphs		►	etua	l lice	nse:	
		Smo	othing and	densitie	es	►	2052	1121		
		Regr	ression diag	nostic p	olots	►	luti	ons		
		Time-series graphs			Line plots					
		Pane	el-data line	plots				Correlo	ogram (ac)	
		Surv	ival analysi	s graphs	;	►		Partial	correlogram (pac)	
		ROC	analysis			►		Period	ogram	Ē^
		Mult	tivariate ana	alysis gra	aphs	►		Cumul	lative spectral distribution	
		Qua	lity control			►		Bivaria	te cross-correlogram	
		Mor	e statistical	graphs		►		Bartlet	t's white-noise test	
		Tabl	e of graphs					Multiv	ariate time-series graphs	•
		Man	age graphs			•				_
		Char	nge scheme	e/size						



🗉 tsline - Time-series line plots	Plot 1
Plots if /in Y axis Time axis Titles Legend Overall By Plot definitions:	Main if /in Choose a plot category and type © Time-series plot Select type:
Disable	Plot type: (line plot) Vvariable: UKMarketIndex Ine properties
Press "Create" to define a time-series plot.	Accept Cancel Submit







Empirical Example 4 :

Data transformation and setup

Panel Data



1.Go to folder $Example_4 \longrightarrow Import the example_4 workbook$

• 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

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

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

(note: j = 19 20 21) reshape long PRICE DY, i(i) j(YEAR) Data wide -> long Number of obs. з -> 9 \times 😑 reshape - Convert data between wide and long form Number of variables 8 5 -> i variable (3 values) -> YEAR Long format from wide Example... xij variables: O Wide format from long PRICE19 PRICE20 PRICE21 PRICE -> Back to long format (previously reshaped) DY19 DY20 DY21 DY -> Back to wide format (previously reshaped) ID variable(s) - the i() option: list, sepby(i) \sim Subobservation identifier - the j() option i YEAR active PRICE DY Variable: Values: (optional) YEAR \sim 73.97 19 1 2.7 1. 1 2. 20 71.79 2.8 Allow the subobservation identifier to include strings 1 1 з. 1 21 1 70.09 2.9 Base (stub) names of X_ij variables: 4,3698 3.1 4. 2 19 Ø PRICE DY \sim 5. 2 20 0 3.6868 3 6. 2 21 Ø 3.84 2.8 Note: All other variables should be constant within ID. 7. 3 19 1 16.81 0 8 3 20 1 14.54 OK Cancel Submit 1.5 9. 3 21 1 15.2 1.7



- 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

😑 xtset - Declare data to be panel data	- 🗆 X
Main Delta	
Panel ID variable:	✓ Time variable: YEAR ✓
Time unit and display format for the	time variable
 Use format of time variable Clock Daily Weekly Monthly 	 Quarterly Half-yearly Yearly Generic
Clear xt settings	Customize OK Cancel Submit
<pre>xtset i YEAR panel variable: time variable: delta:</pre>	i (strongly balanced) YEAR, 19 to 21 1 unit



Classical Linear Regression Model Estimation



Empirical Example 5 :

Simple Linear Regression



Open the file SandPhedge.dta

		Date	Spot	Futures	rspot	rfutures	lspot	lfutures	lspot_fit	resic 🔺
ŝ	1	2002m2	1106.73	1106.9			7.009165	7.009319	7.00987	0007C
n l	2	2002m3	1147.39	1149.2	3.608008	3.750273	7.045245	7.046821	7.04724	00195
of s	3	2002m4	1076.92	1077.2	-6.338468	-6.470097	6.98186	6.982121	6.982768	0009C
-	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 🚽
	4		1	11						•



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



Go to Statistics→ Linear models and related→ Linear regression

File Ed	lit Data Graphics	Statistics	User Window Help				
🗳 🔒 🤅) 🛛 🖻 • 🔟 • II	Sur	nmaries, tables, and tests	Þ			
Review		Line	ear models and related	Þ	Linear regression		
# Con	nmand	Bin	ary outcomes	•	Regression diagnostics	F	
1 use	"C:\Users\as14n14\App	Ord	linal outcomes	•	ANOVA/MANOVA	•	
2 sum	marize rspot rfutures	Cat	egorical outcomes	•	Constrained linear regression		
		Соц	unt outcomes	•	Nonlinear least squares		
		Exa	ct statistics	•			
		End	logenous covariates	•	Censored regression	ł	33
		San	- nple-selection models	•	Truncated regression		zin\Stata
		Mu	Itilevel mixed-effects mode	le i	Box-Cox regression		
		1410	nares es mixed, en éces mode		Eractional nolynomials		



regress - Linear regression	
Model by/if/in Weights SE/Robust Reporting	
Dependent variable: Independent variables: rspot rfutures	
Treatment of constant	
Has user-supplied constant	
Total SS with constant (advanced)	
🕑 😈 🗎 🛛 🛛 🖸 Cancel	Submit

In the command window you can type

Regress dependent variable independent variable

regress rspot rfutures

ATHENS UNIVERSITY

OF ECONOMICS AND BUSINESS

OIKONOMIKO

AOHNON

ΠΑΝΕΠΙΣΤΗΜΙΟ

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	₽> t	[95% Conf.	Interval]
rfutures _cons	1.007291 .0006399	.0058654 .0266245	171.73 0.02	0.000 0.981	.9956887 052026	1.018893 .0533058
			Coeff	icients		



1. Hypothesis Testing – Critical value approach





2. Hypothesis Testing – Confidence interval approach

 $H_0: \alpha = 0$ $H_{A}: \alpha \neq 0$ $H_0:\beta=0$ $H_{A}: \beta \neq 0$

Two -sided Test

- Confidence interval approach a = 5% significance level
- regress rspot rfutures

Source	ss	df	MS
Model Residual	2791.43107 12.4936054	1 132	2791.43107 .094648526
Total	2803.92467	133	21.0821404

					signifi lie with
Num	ber	of	obs	=	134
F (1,	1	L32)	=2	9492.60

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

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

Λ

(-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 nificant, since 0 does not vithin confidence interval

Model	2791.43107	1 2791	L.43107		Prob ≻ F	=	0.0000
Residual	12.4936054	132 .094	4648526		R-squared	=	0.9955
					Adj R-squared	=	0.9955
Total	2803.92467	133 21.0	0821404		Root MSE	=	.30765
							_
rspot	Coef.	Std. Err.	t	₽> t	[95% Conf.	Int	terval]
rfutures	1.007291	.0058654	171.73	0.000	. 9956887	1	.018893
cons	.0006399	.0266245	0.02	0.981	052026	. (0533058

Workshop, 2023



3. Hypothesis Testing – p-value approach

Two _sided Test		
		p-value is termed as the
$H_0: \alpha = 0$		"plausibility" of the Null Hypothesis;
$H_A: \alpha \neq 0$	P-value approach	the smaller the p-value, the less plausible is the null hypothesis.
	>	Is the largest significance level at which we fail to reject
$H_0: \beta = 0$	a = 5% significance level	the null hypothesis.
$H_A: \beta \neq 0$		

. 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
				\frown		
rspot	Coef.	Std. I	Err. t	P> t	[95% Conf.	Interval]
rfutures	1.007291	.0058	654 171.73	0.000	. 9956887	1.018893
_cons	.0006399	.02662	245 0.02	0.981	052026	.0533058
					/	



Suppose now we want to test the null hypothesis that

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

hics Statistics User	Window Help						
Summaries, ta	bles, and tests	•					
Linear models	and related	•					*
Binary outcor	nes	•					
4\App Ordinal outco	mes 🕨	•					
ures Categorical o	utcomes •	•		March and a first a			
Count outcor	nes 🕨	·	II MS	F(1, 132)	= 134 =29492.60		
ts(free Exact statistics	; •	•	1 2791.43107	Prob > F	= 0.0000		
Endogenous	ovariates	13	32 .094648526	R-squared · · · · · · · · · · · · · · · · · · ·	= 0.9955 = 0.9955		
Sample-select	tion models	13	33 21.0821404	Root MSE	= .30765		
Multilevel mi	ed-effects models	•					
Generalized li	near models	Std	i.Err. t P> t	[95% Conf.	Interval]		
Nonparametr	ic analysis	•					
Time series	•	.00	058654 171.73 0.00 266245 0.02 0.98	0 .9956887	1.018893		
Multivariate t	me series						
State-space m	odels						
Longitudinal/	panel data 🔹 🕨	•	Predictions residuals etc.				-
Survival analy	sis 🕨	•	Nonlinear predictions				
Epidemiology	and related	-	Marginal means and predi	ictive marging			 д
SEM (structur	al equation modeling)		Marginal effects	cuve margins			
Survey data a	nalvsis		Contrasts				
Multiple imp	itation	-	Contrasts of margins				
Multivariate	nahois	_	Pairwise comparisons				
	nalysis •		Pairwise comparisons of n	nargins			
Power and sa	mple size		Margins plots and profile	plots	Test linear hypot	heses	
Resampling	•		Tests	•	Test parameters		
Postestimatio	n 🕨		Linear combinations of co	timentes	Test nonlinear hy	/potheses	
Other	,	•	Nonlinear combinations of es	f estimates	Likelihood-ratio	test	
			Noninear combinations o	i estimates	Specification link	test for single-equation mode	ls





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



ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Simple Linear Regression

Grap	hics Statistics User Window	Help						
	Twoway graph (scatter, line, etc.)							
	Bar chart	ſ	/	/				
	Dot chart			//	12.1	Copyright	198	5-2011
	Pie chart	Ì	Ana	17818		4905 Lakew	ay	Drive
	Histogram	;;	ion			College St 800-STATA-	ati PC	on, Te>
	Box plot					979-696-46	00	
	Contour plot					979-696-46	01	(fax)
	Scatterplot matrix	70	ork perpetual license:					
	Distributional graphs	▶ nk	er:	4012052	1084			
	Smoothing and densities	→ [.00	Univers	ity of :	Southampton	L	
	Regression diagnostic plots	+						
	Time-series graphs			Line plots				variak
	Panel-data line plots			Correlogra	m (ac)			
	Survival analysis graphs	+		Partial corr	elogram (pac)		ata_Fil
	ROC analysis	+		Periodogra	m			
	Multivariate analysis graphs	+		Cumulative	e spectral	distribution		
	Quality control	•		Bivariate cr	oss-corre	logram		
	More statistical graphs	•		Bartlett's w	hite-noise	e test		
	Table of graphs			Multivariat	e time-sei	ries graphs	•	
	Manage graphs	•						
	Change scheme/size							



Open capm.dta

Plot 1	Graph - Graph
Main if/in	File Edit Object Graph Tools Help
Choose a plot category and type	Graph ↓ ×
 Time-series plot Select type: ■ Time-series range plot Range line 	Ę -
Plot type: (range plot with lines) Y1 variable: ersandp Y2 variable: erford Line properties	B 2002m1 2004m1 2006m1 2008m1 2010m1 2012m1 2014m1
Accept Cancel Submit	Date



ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Simple Linear Regression

Grap	hics Statistics User Window	Help 🗄 Plot 1	x
	Twoway graph (scatter, line, etc.)	Plot if/in	
	Bar chart	Choose a plot category and type	_
	Dot chart	Basic plots Basic plots: (select type)	
	Pie chart	Contour plots	
	Histogram	○ Fit plots	
	Box plot	© Immediate plots Spike ▼	
	Contour plot	rc Advanced plots	
	Scatterplot matrix	Plot type: (scatterplot)	$\neg \parallel$
	Distributional graphs	Y variable: X variable:	
	Smoothing and densities	erford ersandp Sort on x variable	
	Regression diagnostic plots	Marker properties Marker weights	
	Time-series graphs		
	Panel-data line plots	1.E	
	Survival analysis graphs		-
	ROC analysis	Accept Cancel Subm	it
	Graph		
	ē -		
		•	
	- 20		
		Scatter plot	
	_		
	- 20	•	
	8		
	-20	-10 ersandp 0 10 STATA Workshop, 202	23



□ 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





2. Hypothesis Testing – Confidence interval approach

Two -sided Test

$$H_0: \alpha = 0$$
$$H_1: \alpha \neq 0$$

A

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

regress erford ersandp

Source	ss	df	MS	Number of obs =	13
Model	11565.9116	1	11565.9116	F(1, 133) = 72 Prob > F = 0.0	2.6
Residual	21177.5644	133	159.229808	R-squared = 0.3	353
Total	32743.476	134	244.354298	Root MSE = 12.	. 61

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

(-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** he within confidence interval

JTATA Workshop, 2023

Confidence interval approach

a = 5% significance level

Λ $\beta \pm t_{crit} SE($

5

0 2

9

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



3. Hypothesis Testing – p-value approach

Two —sided Test		
		p-value is termed as the
$H_0: \alpha = 0$		"plausibility" of the Null Hypothesis;
$H_A: \alpha \neq 0$	P-value approach	the smaller the p-value, the less plausible is the null hypothesis.
	>	Is the largest significance level at which we fail to reject
$H_0:\beta=0$	a = 5% significance level	the null hypothesis.
$H_A: \beta \neq 0$		

. 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

				\frown		
erford	Coef.	Std. Err.	t	P> t	[95% Conf.	Interval
ersandp _cons	2.026213 3198632	.2377428 1.086409	8.52 -0.29	0.000 0.769	1.555967 -2.468738	2.496459 1.829013



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) = 18.63Prob > 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: <u>www.stata.com/manuals/rtest.pdf</u>



Empirical Example 6 :

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. E	rr. t	P≻ t	[95% Conf.	In	terval]
ersandp	1.360448	.15661	47 8.69	0.000	1.052308	1	. 668587
dprod	-1.425779	1.3244	67 -1.08	0.283	-4.031668	1	.180109
dcredit	0000405	.00007	64 -0.53	0.596	0001909	_1	0001098
dinflation	2.95991	2.1662	09 1.37	0.173	-1.302104	7	.221925
dmoney	0110867	.03517	54 -0.32	0.753	0802944	_	0581209
dspread	5.366629	6.9139	15 0.78	0.438	-8.236496	1	8.96975
rterm	4.315813	2.5151	79 1.72	0.087	6327998	9	.264426
_cons	1514086	. 90478	67 -0.17	0.867	-1.931576	1	. 628759


Multivariate Linear Regression

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

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

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

•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 $F \succ t_{crit}$



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

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

. test (dprod dcredit dinflation dmoney dspread)

```
(1) dprod = 0
```

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

F(5, 316) = 0.85Prob > 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





Assumptions underlying the CLR model

 $E(u_t)=0$ The errors have zero mean (Mean Independence) var(u_t)= σ^2 The variance of the errors is constant (Homoskedasticity) $cov(u_i, u_j)=0$ The errors are linearly independent of one other $cov(u_t, x_t)=0$ There is no relationship between the error and the corresponding variate x $u_t \square N(0, \sigma^2)$ 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

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

LM test

OF ECONOMIC

OIKONOMIKO

ΑΘΗΝΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ

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

Wald Test

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

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



 $E(u_t)=0$ 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
- 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



 $var(u_t) = \sigma^2$ 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*



Load macro.dta

. regress ermsoft ersand dprod dcredit dinflation dmoney dspread rterm

Source	SS	df	MS	Number of obs =	
				F(7, 316) =	1:
Model	13202.4359	7	1886.06227	Prob > F =	0.0
Residual	50637.6544	316	160.245742	R-squared =	0.3
				Adj R-squared =	0.:
Total	63840.0903	323	197.647338	Root MSE =	12

ermsoft	Coef.	Std. Err.	t	₽> 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



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 Skewness	11.12 10.26	35 7	1.0000
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	₽> 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



 $cov(u_i, u_j)=0$ 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

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

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).

 $u_{t} = \rho u_{t-1} + v_{t}$ $H_{0}: \rho = 0$ (No Autocorrelation) $H_{A}: \rho \neq 0$ (Autocorrelation)

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)





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	₽> 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

ОІКОНОМІКО Панепізтнміо А́они́он ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Testing for Non- Normality

. [Gran	hics Statistics User Windo	w Help	histogram - Histograms for continuous and estensical variables	x
	orap	Twoway graph (seatter line at			
-		rwoway graph (scatter, line, etc		Main if /in Weights Density plots Add plots Yaxis Xaxis Titles Legend Overall By	
		Bar chart	3840.0903 3	32 Data	
		Dot chart		Variable:	
i\ā		Pie chart	Coef. Sto	t c resid Data are discrete	
off		Histogram	1 360448 1		
ы		Box plot	1.425779 1.3	-3 Bins Yaxis	
		Contour plot	0000405 .00	OC O Density	
		Scatterplot matrix	2.95991 2.1	□ I I I I I I I I I I I I I I I I I I I	
0.3		Distributional analysis	5.366629 6.9	S Width of bins	
0.3			4.315813 2.5	. 5 Lower limit of first bin O Percent	
oft		Smoothing and densities	• .1514086 .90	90	
e		Regression diagnostic plots	•	Bar properties Add height labels to bars	
		Time-series graphs	lags (12)	Bar label properties	
iic		Panel-data line plots	N		
fs		Survival analysis graphs	M test for au	<u> </u>	
oft		ROC analysis	Graph - Graph	OK Cancel S	ubmit
n		Multivariate analysis graphs	File Edit Object (Graph Tools Help	_
у,		Quality control			
oft		More statistical graphs			
nr l					
			8		
			8		
			l ≩ ^ë		
			Deni		
			8		
			5-		
			-60	-40 -20 0 20	
				Residuals	2002
			<u></u>		2023



sktest resid

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





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) . replace FEB98DUM = 0 if Date!=tm(1998m2)
(325 missing values generated) (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)
```

. 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	₽> 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

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Testing for Non- Normality

Statistics User Window Help	
Summaries, tables, and tests	▶
Linear models and related	×
Binary outcomes Ordinal outcomes Categorical outcomes Count outcomes	<pre>> #/Kurtosis tests for Normality > 3kewness) Pr(Kurtosis) adj chi2(2) Prob>chi2 > 0.0000 0.0000 . 0.0000</pre>
Exact statistics Endogenous covariates Sample-selection models Multilevel mixed-effects models Generalized linear models	ralues) Main Main
Nonparametric analysis Time series Multivariate time series	<pre>> 3/Kurtosis tests for Normality > 3/Kurtosis tests for Normality > 3/Kurtosis tests for Normality > joint > 3kewness) Pr(Kurtosis) adj chi2(2) Prob>chi2</pre>
State-space models	0.0000 0.0000 . 0.0000 Image: Standardized residuals Image: Standard error of the residual Image: Standardized residuals Image: Standard error of the residual Image: Standard error of the residual
Longitudinal/panel data	Predictions, residuals, etc.
Survival analysis Epidemiology and related SEM (structural equation modeling) Survey data analysis	 Nonlinear predictions Marginal means and predictive margins Marginal effects Contrasts DFBETA for variable:
Multiple imputation	Contrasts of margins



. sktest resid_new

Skewness/Kurtosis tests for Normality											
Variable	Obs	Pr(Skewness)	Pr(Kurtosis)	adj chi2(2)	joint —— 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:

ATHENS UNIVERSITY

OF ECONOMICS

- 1. Use the correlation matrix of the explanatory variables
 - . correlate ersand dprod dcredit dinflation dmoney dspread rterm

(obs=324)

OIKONOMIKO

ΠΑΝΕΠΙ

	ersandp	dprod	dcredit	dinfla~n	dmoney	dspread	rterm
ersandp	1.0000						
dprod	- <mark>0.025</mark> 3	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



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

- 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.
- 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.



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

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