Econometrics & Financial Markets

INTRODUCTION

Toulouse Business School
MSc BIF

Anna CALAMIA
a.calamia@tbs-education.fr

Course contents

- Overview of some classical statistical methods to *model strategic financial data*.

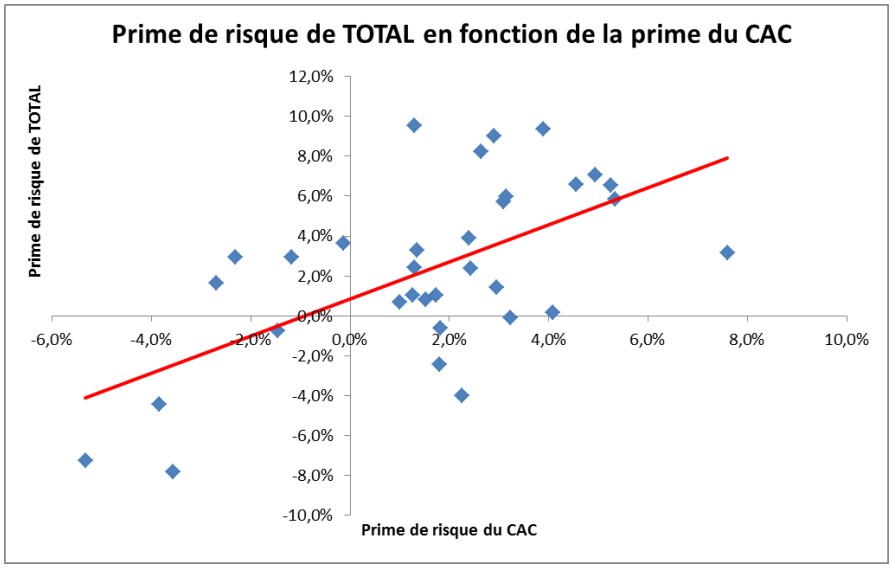
- Theoretical concepts applied on practical empirical cases in finance, using the software XLSTAT.

- The main models we will work on are:
  - Linear regression models (Simple and Multiple)
  - Time series models

- Course outline:
  - ➔ General presentation
  - ➔ Descriptive statistics and Introduction to XLSTAT
  - ➔ Linear regression model and tutorial on XLSTAT
  - ➔ Time series analysis and tutorial on XLSTAT
  - ➔ Other tools and methods: qualitative variables; panel data, event study

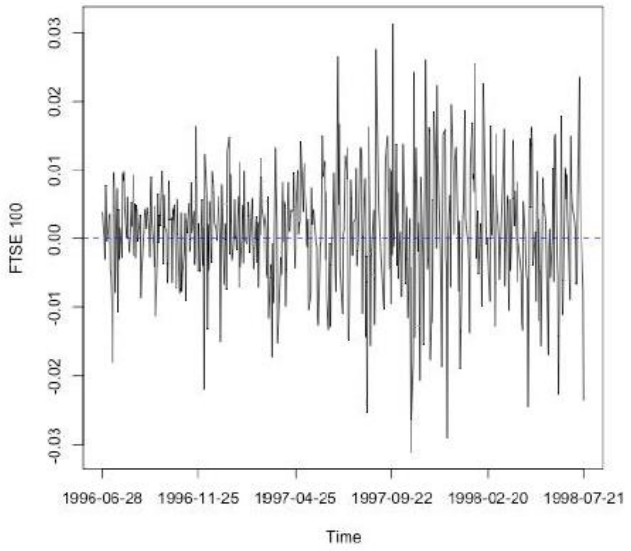
Objectives

The objective of this course is to learn how use statistical tools and quantitative methods to model and forecast finance data and test empirical relations in finance.



CAPM validity? Test asset pricing models and check the ability of a portfolio manager to outperform the market

Return forecasts?
Model and predict strategic variables



Teaching method and evaluation

- Theoretical classes + Quizz
- Tutorials with XLSTAT
- Further reading, Personal work
- Group project

Evaluation:

Empirical project, by groups of 3 or 4 (maximum) students

- ➔ Data to be downloaded using Bloomberg, specialized financial websites (yahoo finance, google finance, …), or companies websites.
- ➔ Submission date: **22 April 2022**

- ➔ Refer to *''Econometrics and financial markets project''*

Group project

Empirical project, by groups (3 students)

- Download data from web
- Data and variable description (presentation, descriptive statistics, plots)
- Test for stationarity (ACF)
- CAPM regression: model estimation, table of results, results interpretation and discussion (t-tests, R2)
- Residuals check: plot residuals and check assumptions
- Interpret and discuss results

Group project submission

- Each group will have to submit on Campus (or by email):

  - ➔ An Excel file containing the data and analysis (the name of the file should be data_namesof students.xls)

  - ➔ A report file (Word, Pdf, PowerPoint) reporting tables and plots and summarizing and interpreting the results (report_namesof students.docx)

    - ❖ Submit both files (excel+report). One file alone will not be considered.
    - ❖ One project per group, with the names of all participants.

Documents on Campus

- All the documents are available on **C@mpus:**
  - Lecture Notes:
    I. Introduction
    II. Financial data and descriptive statistics
    III. Linear regression model
    IV. Time series analysis
    V. Other tools and methods
  - Data file for tutorials: Data2022
  - Econometrics and financial markets GROUP PROJECT

- The software **XLSTAT** can be downloaded on **C@mpus**
  - https://campus.tbs-education.org/documents/informatique/Logiciels/Logiciels.pdf
  - follow the installation procedure and use the code provided to activate the licence

Readings

- Textbook for this course:

***Introductory Econometric for Finance*, C.BROOKS, Cambridge, 2nd Edition**

Other recommended books:

- *Applied Econometrics*, Dimitri Asteriou, Stephen G. Hall, Palgrave, 2nd edition
- *Introductory Econometrics*: A Modern Approach, Wooldridge Jeffrey M. (2009),4th ed., South-Western
- *The Econometrics of Financial Markets*, John Y. Campbell, Andrew W. Lo, & A. Craig MacKinlay, Princeton University Press, 1997

Useful Data links

- TBS Library Services provides a broad portfolio of specialist business resources, including **Bloomberg** (Alaric library in Toulouse or Paris Campus), **Datastream** (Refinitiv online access – ask Helpdesk), **Infinancials**, **Orbis**, etc
- https://bibliotheque.tbs-education.fr/nos-bases-de-donnees.aspx
- Also, available on the web:
  - Yahoo Finance: https://finance.yahoo.com/
  - Google Finance: https://www.google.com/finance/
  - Fama French Factors: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html
  - Federal Reserve Bank: https://fred.stlouisfed.org/
  - Fanque de France: https://www.banque-france.fr/page-sommaire/taux-et-cours

Econometrics & Financial Markets

Financial data and descriptive statistics

Toulouse Business School
MSc BIF

Anna CALAMIA
a.calamia@tbs-education.fr

Outline

- Financial data
- Descriptive statistics: mean, variance, skewness kurtosis
- Brief overview of Normal distribution, tools to assess normality and some other standard distributions (Khi2, Student, Fisher).

Financial Data

There are 3 types of data which econometricians might use for analysis:

1. **Times series data**
   - How the value of a country's stock index has varied with that country's macroeconomic fundamentals.
   - How the value of a company's stock price has varied with that country's market index.

Date	S&P500	Microsoft
01/01/2020	3225.52	168.45
01/02/2020	2954.22	160.31
01/03/2020	2584.59	156.48
01/04/2020	2912.43	177.82
01/05/2020	3044.31	181.82
01/06/2020	3100.29	202.49
01/07/2020	3271.12	203.98
01/08/2020	3500.31	224.40
01/09/2020	3363.00	209.78
01/10/2020	3269.96	201.94
01/11/2020	3638.35	214.67
01/12/2020	3662.45	216.21

Financial Data

2. Cross-Sectional data

- Data collected at a single point in time
- Relationship between board size (board independence) and size (age) for quoted firms
- Relationship between gross investment and firm's value and capital

Firm	Investment	Value	Capital
1	1486.7	5593.6	2226.3
2	459.3	2115.5	669.7
3	189.6	2759.9	888.9
4	172.49	703.2	414.9
5	81.43	365.7	804.9
6	135.72	927.3	238.7
7	89.51	192.7	511.3
8	68.6	1188.9	213.5
9	49.34	474.5	468
10	5.12	58.12	14.33

Financial and Business Data

3. Panel Data

- Relationship between returns and earnings for several stocks over time (dimensions of both time series and cross-sections)

Quarter	Stock	EBIT	Return
2017-Q1	AIRBUS	533	-0.00433
2017-Q2	AIRBUS	529	-0.00305
2017-Q3	AIRBUS	414	0.00513
2017-Q4	AIRBUS	878	-0.01073
2018-Q1	AIRBUS	168	-0.00192
2018-Q2	AIRBUS	871	0.02181
2018-Q3	AIRBUS	1524	-0.00661
2018-Q4	AIRBUS	2155	0.00239
2017-Q1	CARREFOUR	1385	0.005
2017-Q2	CARREFOUR	546	0.00317
2017-Q3	CARREFOUR	546	0.00826
2017-Q4	CARREFOUR	427	0.00194
2018-Q1	CARREFOUR	427	0.00328
2018-Q2	CARREFOUR	-169	-0.00431
2018-Q3	CARREFOUR	-169	-0.00151
2018-Q4	CARREFOUR	931	0.00573

Financial data

- Data may be **quantitative** (e.g. exchange rates, stock prices, number of shares outstanding).
  - ➔ Continuous data can take on any value and are not confined to take specific numbers.
  - ➔ Discrete data can only take on certain values, which may be integers (e.g. the number of shares traded during a day).
- or **qualitative** (e.g. day of the week)
  - ➔ Ordinal : a figure of 12 may be viewed as `better' than a figure of 6, but could not be considered twice as good
  - ➔ Nominal there is no natural ordering of the values at all.

Data Cleaning

- Data cleaning is an essential part of statistical analysis.
- Often time-consuming.

- Raw data: the data as it comes in (may lack headers, contain wrong data types, wrong category labels, unknown or unexpected character…)
- To get technically correct data, you have to organise data (columns) and assign variable names and types (text variables, number variable...).
- Technically correct data may still have missing values, outliers or (obvious) errors (e.g. negative age or bid-ask spread). These inconsistencies should either be removed, corrected or imputed.
- You may also choose to apply other possible filters before you implement specific methodologies (e.g. the first observations of a new stock, the 1% highest/lowest data…)

Econometric model - Steps

- General statement of the problem and formulation of an estimable model (from a theoretical model or intuition that 2 or more variables should be related)
- Collection and cleaning of data relevant to the model
- Choice of relevant estimation method and model estimation
- Statistical evaluation of the model
  - yes
    - interpret (and evaluate according to theory)
    - use for analysis
  - no => reformulate

Descriptive Statistics

Returns

P_t is the price of a stock or a portfolio evaluation at time t.

The return of the stock or portfolio between time t and time t-1 is :

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}$$ or $$R_t = \frac{P_t - P_{t-1} + D_t}{P_{t-1}}$$ if dividend

Log returns:

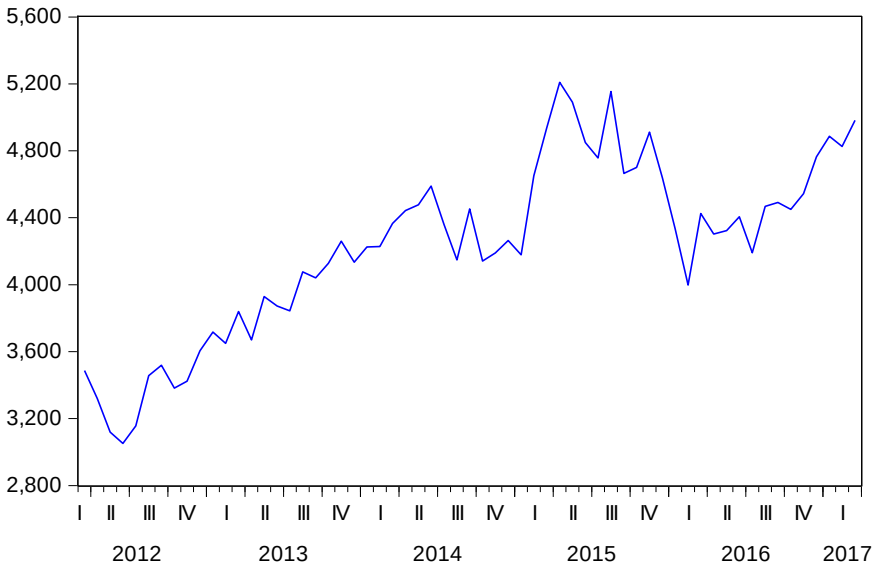
We can write $$R_t = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 \Leftrightarrow R_t + 1 = \frac{P_t}{P_{t-1}}$$

Log's property : when R_t is « small », $\ln(R_t + 1) = R_t$ then $\ln(\frac{P_t}{P_{t-1}}) = \ln(R_t + 1) = R_t$

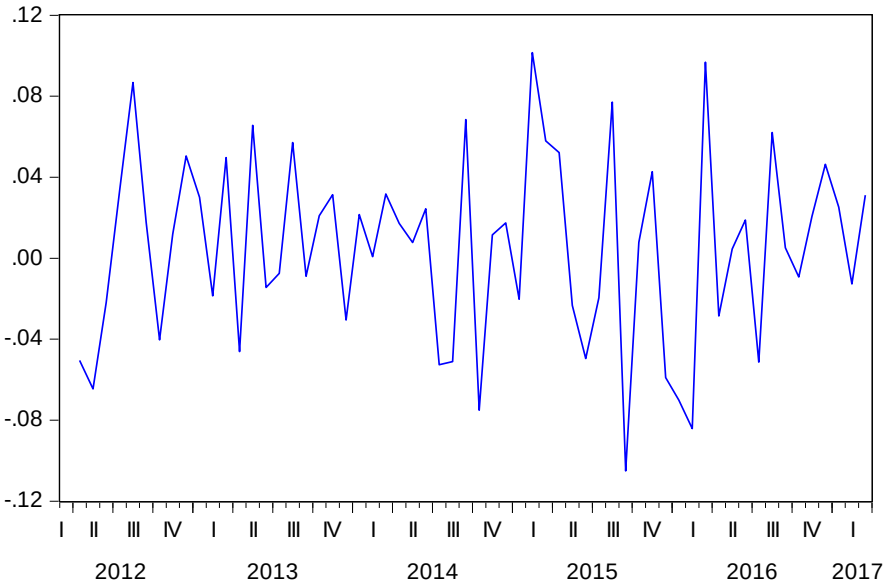
$$R_t = \ln(\frac{P_t}{P_{t-1}}) = \ln(P_t) - \ln(P_{t-1})$$

Returns

CAC40 : Monthly Index Prices from March 2012 to March 2017



CAC40 : Monthly Returns from March 2012 to March 2017



Descriptive Statistics: Mean and variance

- Mean (expected value): $\bar{R} = \frac{1}{T}\sum_{t=1}^{T} R_t$

- Volatility measures:

  - ➔ Variance : $Var(R) = \frac{1}{T-1}\sum_{t=1}^{T}(R - \bar{R})^2$

  - ➔ Standard deviation: $\sigma(R) = \sqrt{V(R)}$

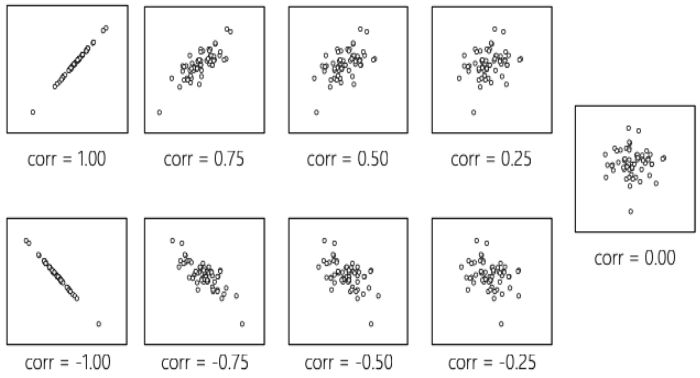
    - *T* is the number of observations
    - *Rt* is return between the dates *t-1* and *t*
    - $\bar{R}$ is the mean of the returns

- Covariance and correlation (btw 2 variables, R_1 and R_2)

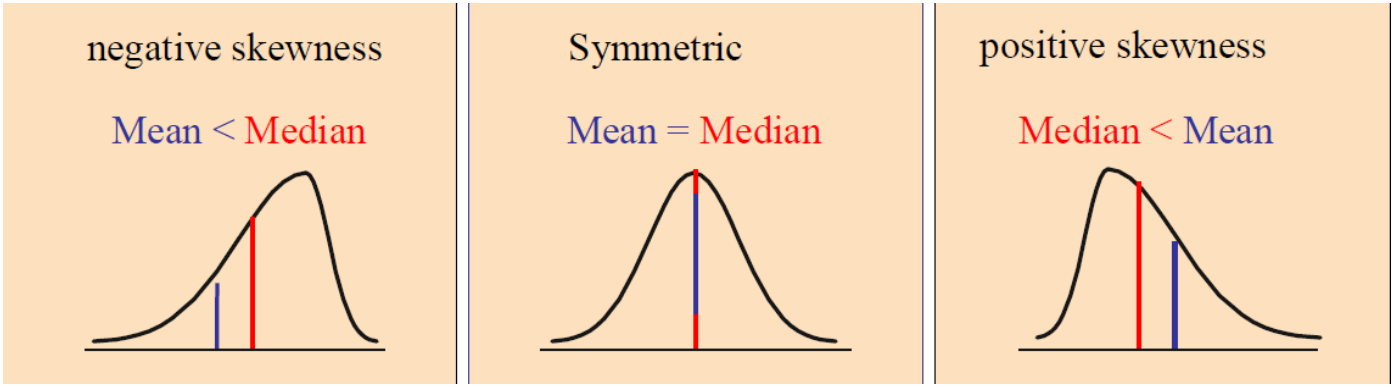
  - ➔ $Cov(R_1,R_2) = \frac{1}{T-1}\sum_{t=1}^{T} (R_1 - \bar{R}_1)(R_2 - \bar{R}_2)$

  - ➔ $Corr(R_1,R_2) = Cov(R_1,R_2) / \sigma_1\sigma_2$

Both the covariance and the correlation measure how the two variables change together ("co-vary", "co-relate"). The correlation is easier to interpret, since it is restricted to lie between -1 and 1.



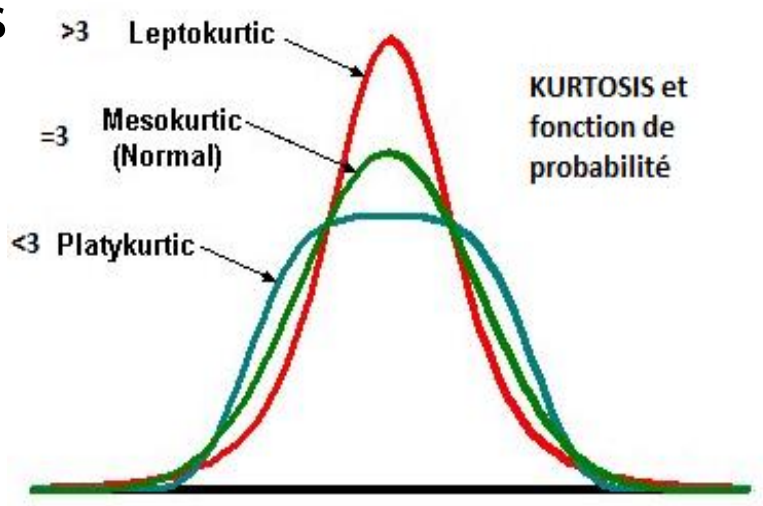
Descriptive Statistics: Skewness and Kurtosis



=> If the distribution is symmetric around the mean, S=0

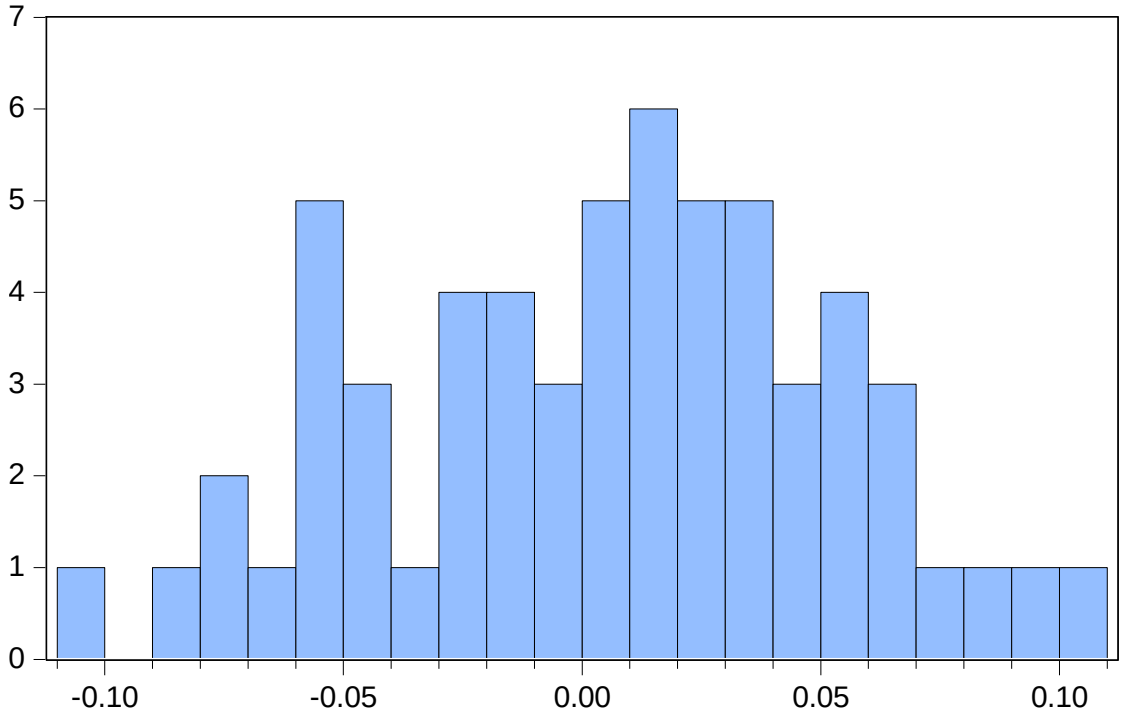
Kurtosis allows to detect extreme values

- ➔ **K-3>0**: *leptokurtic* distribution
  (Heavier tails and a higher peak)
  => presence of extreme values
- ➔ **K-3<0**: *platikurtic* distribution
  (Lighter tails and a lower peak)
  => very few extreme values
- ➔ **K-3= 0**: Normal distribution
- ➔ Kurtosis vs excess kurtosis (K-3)



Descriptive Statistics

Monthly returns of the CAC40 from March 2012 to March 2017



Series: R_CAC40
Sample 2012M03 2017M03
Observations 60

Mean	0.004840
Median	0.009694
Maximum	0.101484
Minimum	-0.105081
Std. Dev.	0.046888
Skewness	-0.119506
Kurtosis	2.475716
Excess kurtosis	-0.524284

Descriptive Statistics

	R_CAC40	R_BNP_PA...	R_DANONE	R_LVMH
Mean	0.004840	0.004674	0.001894	0.006396
Median	0.009694	0.006302	-0.002979	0.000343
Maximum	0.101484	0.170860	0.090256	0.152058
Minimum	-0.105081	-0.266491	-0.146024	-0.158264
Std. Dev.	0.046888	0.086160	0.046594	0.066035
Skewness	-0.119506	-0.522340	-0.253658	-0.165522
Kurtosis	2.475716	3.321080	3.293069	2.760234
Excess kurtosis	-0.524284	0.321080	0.293069	-0.239766

Question 1 -The riskiest equity is :

A-R_CAC40

B-R_BNP_PA

C-R_DANONE

D-R_LVMH

Question 2

A-50% of the CAC_40 returns are lower than 0,48%

B-50% of the CAC_40 returns are lower than 0,96%

C-R_CAC40 is right skewed

D-R_CAC-40 is leptokurtic

Normality

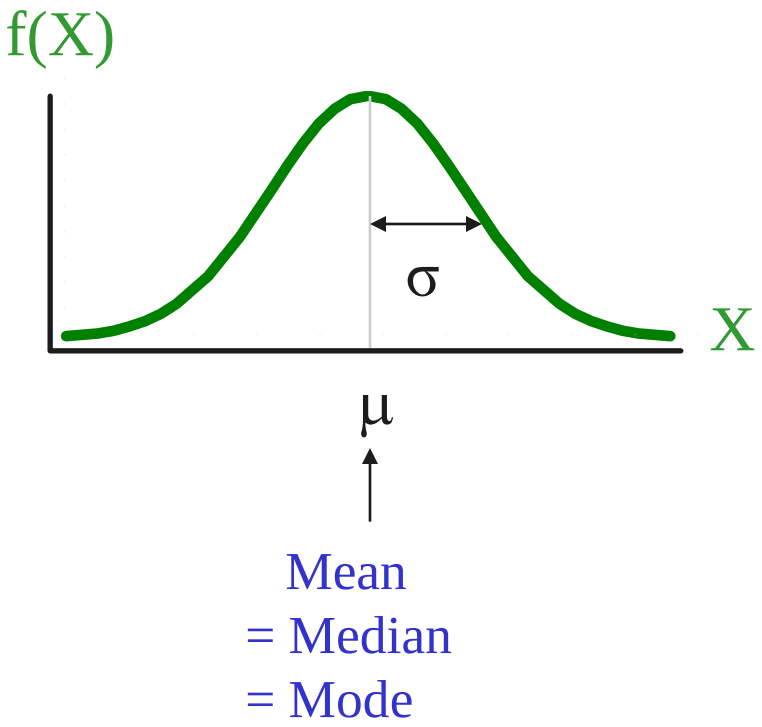
Returns and the normal distribution

Returns and the normal distribution (= loi normale): **Apple Inc** – daily returns 1990-2019

The Normal Distribution

Bell Shaped

- Symmetrical (Skewness=0)
- Kurtosis =3
- Mean, Median and Mode are Equal



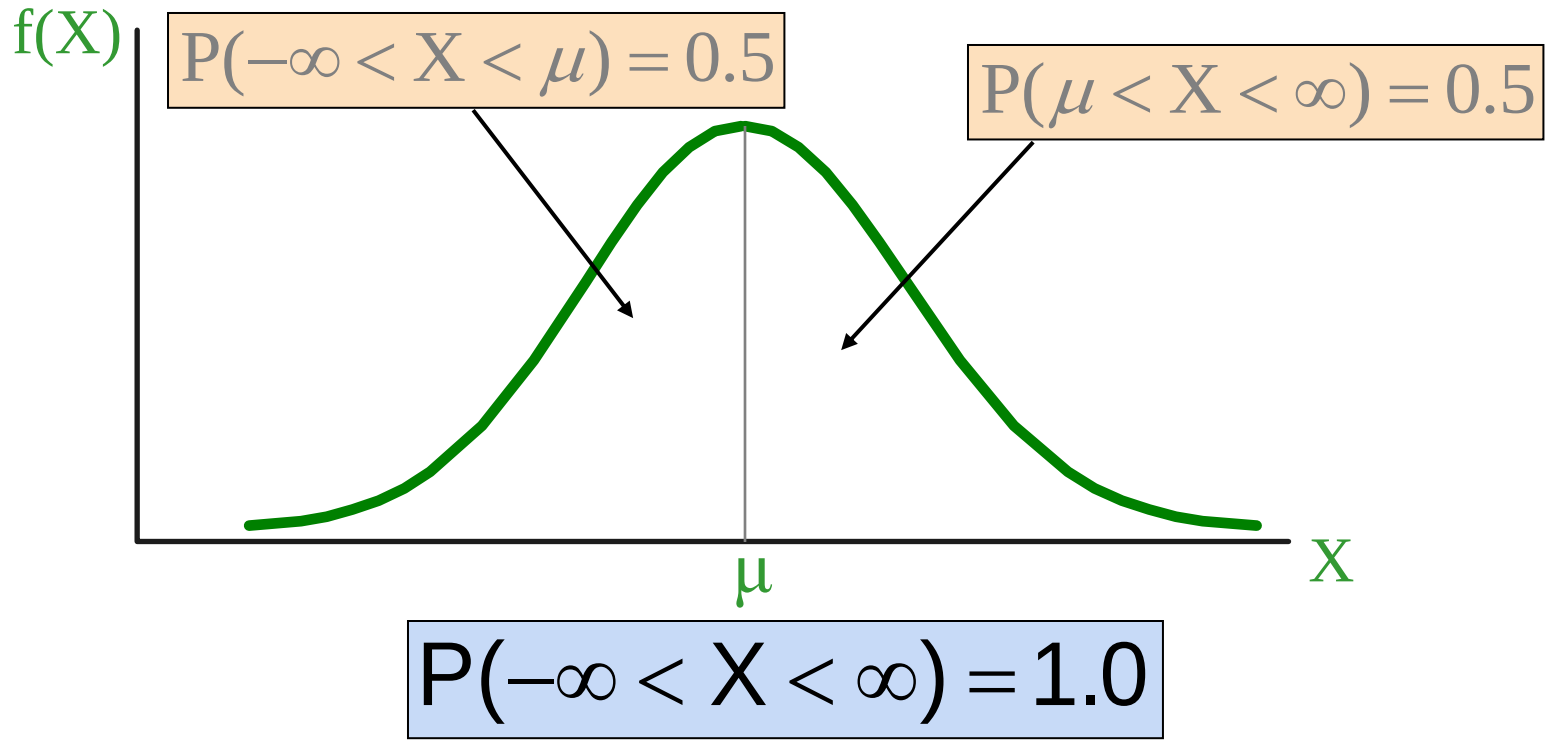
- Location is determined by the mean, μ (changing μ shifts the distributions left or right)
- Spread is determined by the standard deviation, σ
- The random variable has an infinite theoretical range: $-\infty$ to $+\infty$

Definition: random variable = a variable that can take on any value from a given set. Most commonly used distribution to characterize a random variable is a normal distribution.

Normal probabilities

Probability is measured by the area under the curve.

The total area under the curve is 1.0, and the curve is symmetric, so half is above the mean, half is below



Why normality is important?

- The probability law is known ➔ calculus of probability

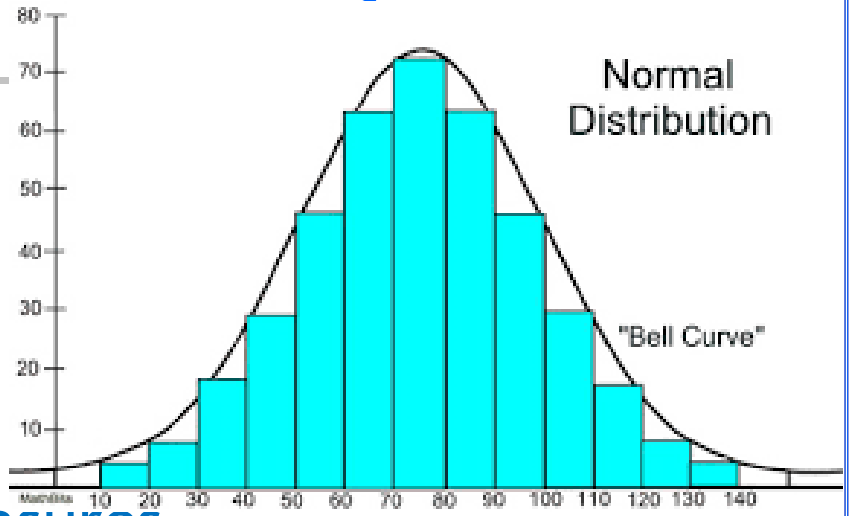
What is the probability that the returns be positive?

- No outliers ➔ Easier to model
- Mean-variance analysis and CAPM
- Specific calculus (VaR)

BUT… Not all continuous random variables are normally distributed

➔ It is important to evaluate how well the data set is approximated by a normal distribution

Assessing Normality



Construct charts or graphs

- ➔ Does the histogram appear bell-shaped?
- ➔ Is the normal probability plot approximately linear with positive slope?

Compute descriptive summary measures

- ➔ Do the mean, median and mode have similar values?
- ➔ Is the Skewness close to 0? Is the Kurtosis close to 3?

Jarque Bera Normality Test:

Based on values of skewness and excess kurtosis

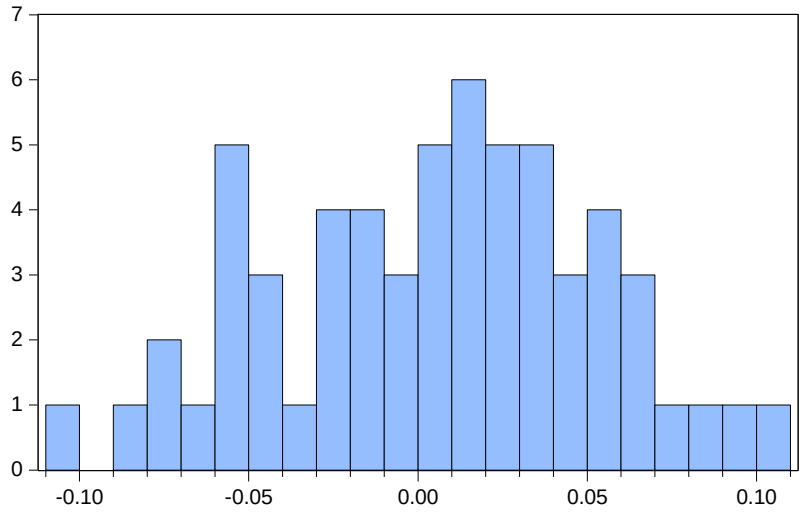
H0 : the series is normally distributed (S and EK jointly not different from 0)

H1 : the series is not normally distributed

JB ~ χ^2(2 dof)

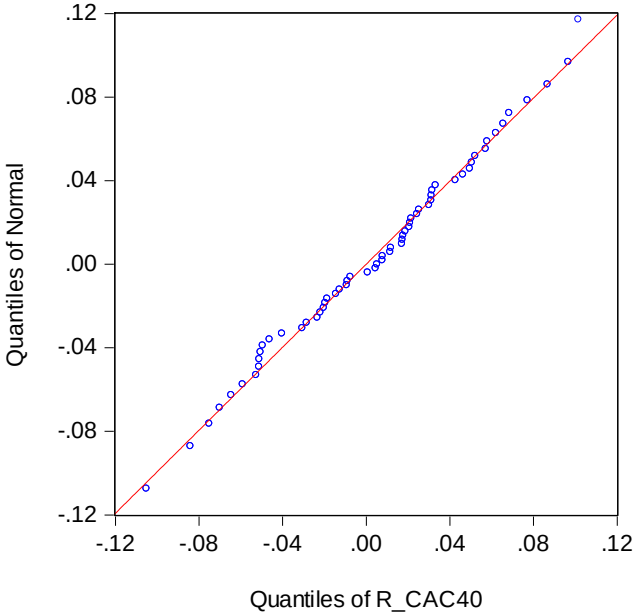
- ➔ Reject H0 if JB > $\chi^2_{2;\alpha}$ or if pvalue < α (the test is statistically significant)

Assessing Normality



Series: R_CAC40
Sample 2012M03 2017M03
Observations 60

Mean	0.004840
Median	0.009694
Maximum	0.101484
Minimum	-0.105081
Std. Dev.	0.046888
Skewness	-0.119506
Kurtosis	2.475716
Excess kurtosis	-0.524284
Jarque-Bera	0.830002
Probability	0.660340



Question 3 –Are the CAC_40 returns normally distributed?

A-No, because the series is left skewed

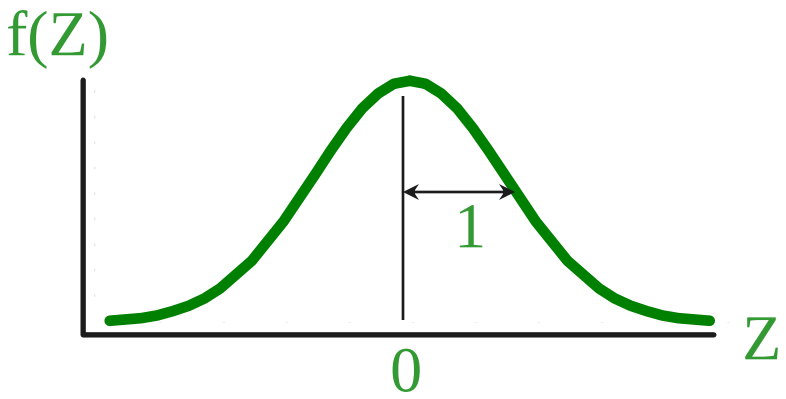
B-Yes, because the Jarque Bera test is not significant

C-No, because the Jarque Bera test is not significant

D-No, because on the QQ plot, points are close to the bissector

The Standardized Normal Distribution

- Also known as the “Z” distribution
- Mean is 0
- Standard Deviation is 1



Values above the mean have positive Z-values

Values below the mean have negative Z-values

- To transform a normally distributed variable into a standard normal: subtract the mean and divide by the st. dev.: $x \sim N(\mu,\sigma) \Rightarrow (x-\mu)/\sigma \sim N(0,1)$

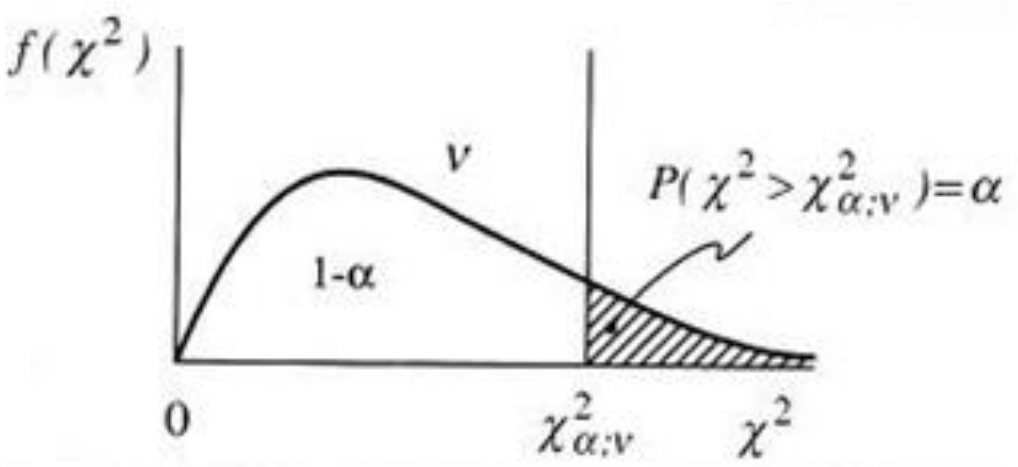
Chi square distribution

Definition:

Let $(X_1, X_2, \ldots, X_n)$ a sample with gaussian distribution N(0;1).

Then:

$$\sum_{i=1}^{n} X_i^{\,2} \quad \text{is} \quad \chi^2(n)$$



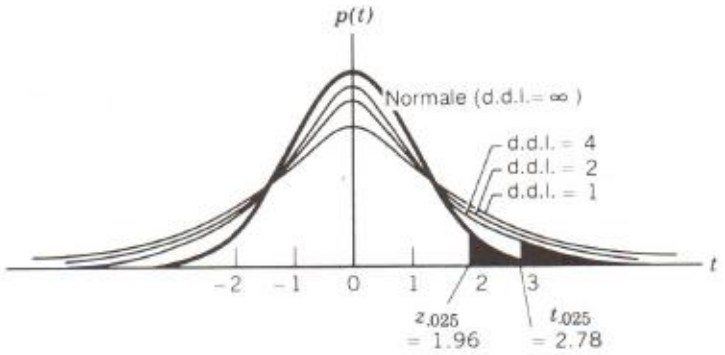
➔ Chi square distributed with n degrees of freedom (dof)

Student Distribution

Definition:

If X~N(0;1) and Y~χ^2(n) with X et Y independent variables, then:

$$t = \frac{X}{\sqrt{Y/n}} \qquad \text{is} \qquad t(n)$$



➔ Student distributed with n degrees of freedom (dof)

➔ When n>30 t(n)~N(0,1)

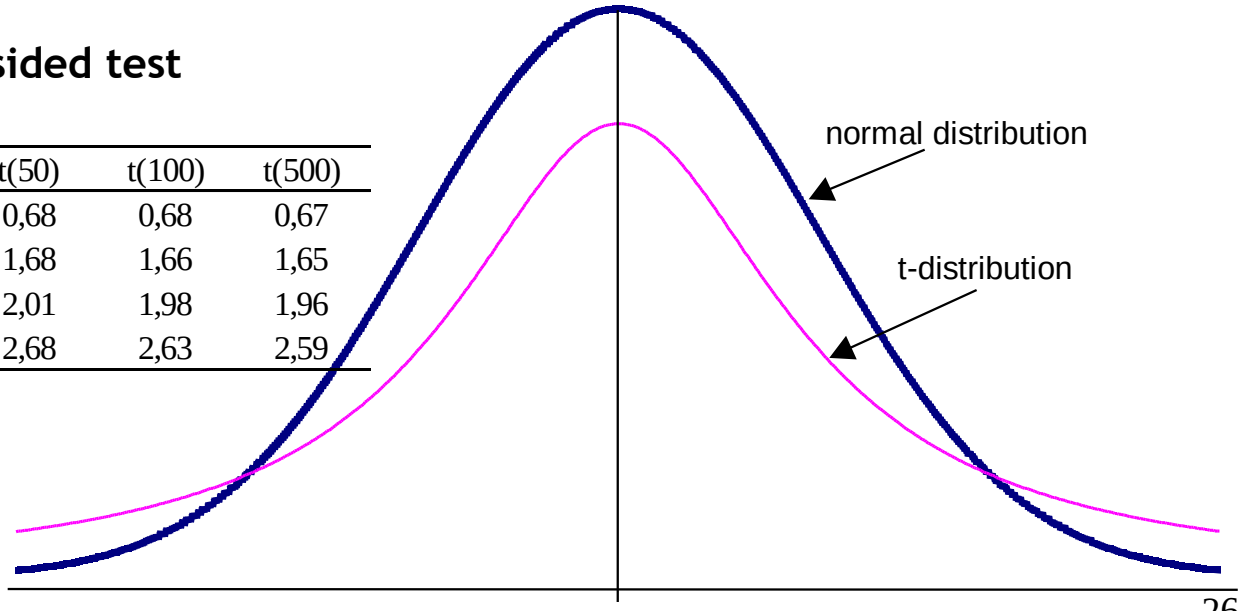
Normal and t-distribution

- *t*-distribution with an infinite number dof ≈ N(0;1)
- t- and the standard normal distribution : both are symmetrical and centred on zero. The t-distribution is characterized by another parameter: its degrees of freedom.

Value for $t_{\alpha/2}$: case of two-sided test

Example from statistical tables

Significance Level	N(0;1)	t(4)	t(50)	t(100)	t(500)
50%	0,67	0,74	0,68	0,68	0,67
10%	1,64	2,13	1,68	1,66	1,65
5%	1,96	2,78	2,01	1,98	1,96
1%	2,58	4,60	2,68	2,63	2,59

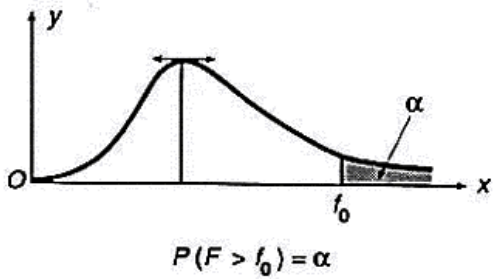


Fisher distribution

Definition:

Let X et Y two independent variables with $X \sim \chi^2(n)$ and $Y \sim \chi^2(p)$, then:

$$\frac{X/n}{Y/p} \quad \text{is} \quad F(n;p)$$



➔ Fisher distributed with n and p degrees of freedom

TUTORIAL XLSTAT

1. Preliminary work on data
2. Descriptive Statistics

Tutorial

- Download data on Excel for the relevant variables:
  - ➔ Stock: Microsoft Corporation
  - ➔ Index: S&P 500
  - ➔ Risk free: Treasury Bill 3 Months
- Create times series of returns and excess returns
- Plot the series of prices and returns

(save as Excel Macro-Enabled Workbook)

Tutorial

- Using software XLSTAT (can be downloaded from Campus):
- **Compute the standard descriptive statistics** (mean, min, max, standard-deviation, skewness, excess kurtosis
- **Histogram**
- **Normality test (JB)**

=> Comments on the returns distributions? Are there outliers (extreme values)? Are the returns normally distributed?

Econometrics & Financial Markets

Linear Regression Model

Toulouse Business School
MSc BIF

Anna CALAMIA
a.calamia@tbs-education.fr

What is regression?

Describing and evaluating the relationship between a given variable (called the dependent variable Y) and one or more other variables (usually known as the independent variable(s), X_1, X_2, …X_k)

$$Y_t = \beta_1 X_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_k X_{kt} + U_t \text{ , t=1, 2, …T}$$

Possible interesting questions:

- Relationship between the expected return of an asset and the market risk premium
- Beta calculation
- Does corporate governance affect firm performance?
- Impact of ad on firm's revenues?
- …

Linear regression Model: Course outline

- Simple linear regression
- Hypothesis and estimation of the coefficients
- Model validation
- Goodness of Fit Statistics
- Generalising to Multiple Linear Regression
- Violation of the assumptions of the CLRM and remedies
- Other problems dealing with CLRM
- Last steps before validating a model

Simple linear regression

Simple regression

- Model :

$$Y_t = \alpha + \beta X_t + U_t$$

One explanatory variable and one constant

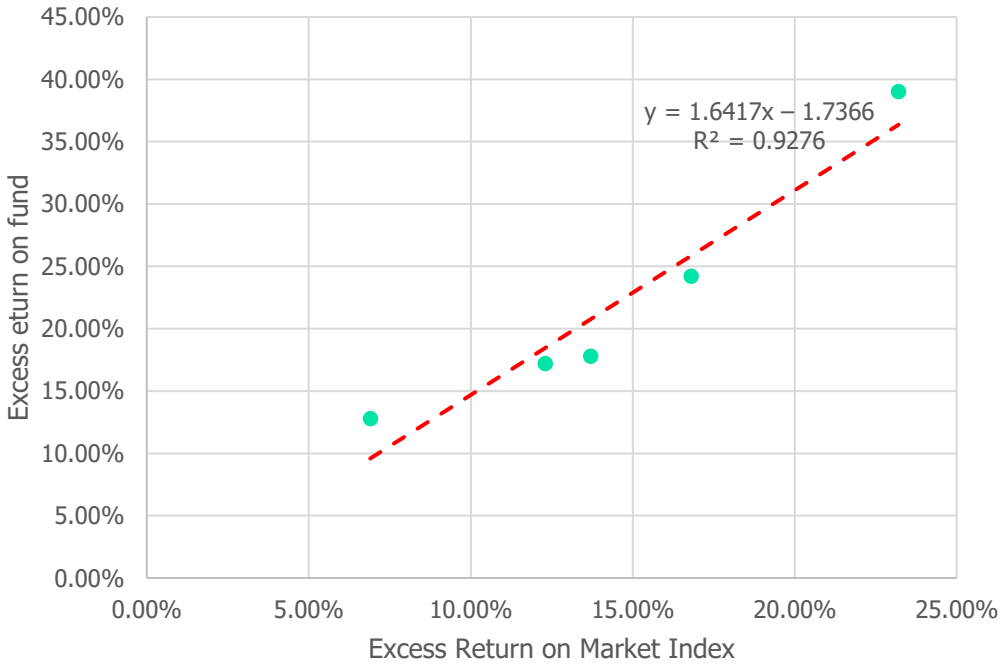
Simple Regression: An Example

- Suppose that we have the following data on the excess returns on a fund manager's portfolio ("fund XXX") together with the excess returns on a market index:

	Y		X
Year, t	Excess return $= r_{XXX,t} - rf_t$		Excess return on market index $= rm_t - rf_t$
1	17.8	$\mathbf{Y} = \beta X + \alpha$ **???**	13.7
2	39.0		23.2
3	12.8		6.9
4	24.2		16.8
5	17.2		12.3

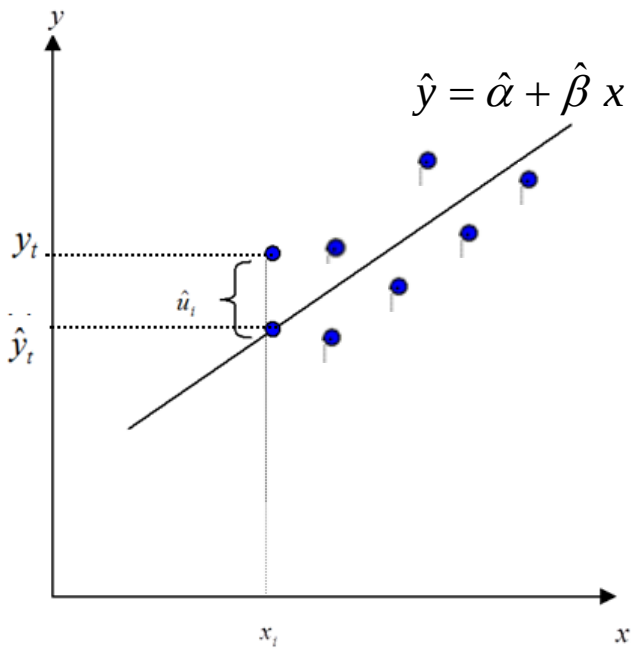
- Does a relationship appear between x and y given the data that we have? ➔ first stage = scatter plot

Simple Regression: Scatter Diagram



Hypothesis and estimation of the coefficients

Ordinary Least Squares



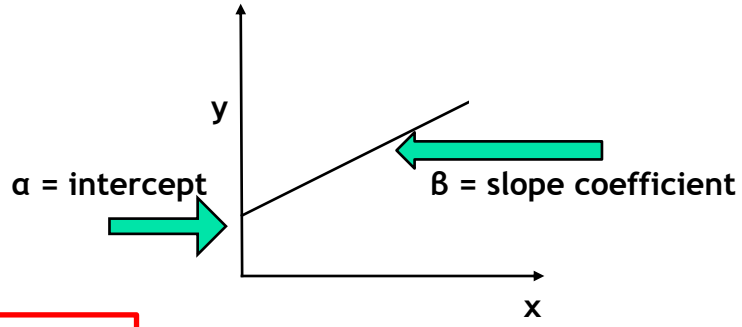
- The most common method used to fit a line to the data is known as **OLS (ordinary least squares).**
- What we actually do is take each distance and square it and **minimize the total sum of the squares** (hence least squares).
- Tightening up the notation, let :
  - ➔ y_t : **actual data**
  - ➔ $\hat{y}_t$: **fitted value** from the regression line
  - ➔ $\hat{u}_t$: **residual**

- So min $\hat{U}_1^2 + \hat{U}_2^2 + \ldots + \hat{U}_T^2$, or minimize $\sum_{t=1}^{T} \hat{U}_t^2$
- This is known as the residual sum of squares, with $\hat{U}_t = Y_t - \hat{Y}_t$

➔ This method of finding the optimum is known as **Ordinary Least Squares (OLS)**

OLS Estimators

Coefficients Estimates



$$\hat{\beta} = \frac{\text{cov}(X;Y)}{\text{var}(X)} \qquad \hat{\alpha} = \overline{Y} - \hat{\beta}\,\overline{X}$$

Calculated by EXCEL, Eviews, SAS, R, ….

$$\text{cov}(X;Y) = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})(Y_t - \overline{Y}) \qquad \overline{X} = \frac{1}{T}\sum_{t=1}^{T}X_t \qquad \overline{Y} = \frac{1}{T}\sum_{t=1}^{T}Y_t$$

$$Var(X) = \frac{1}{T}\sum_{t=1}^{T}(X_t - \overline{X})^2$$

T is the sample size

α And β in the CAPM Example

In the CAPM example used above, the estimates are:

Dependent variable: ER_FUND

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.017366	0.041140	-0.422132	0.7014
ER_MARKET_INDEX	1.641745	0.264778	6.200453	0.0085

R-squared	0.927616	Mean dependent var	0.222000
Adjusted R-squared	0.903488	S.D. dependent var	0.102343
S.E. of regression	0.031794	Akaike info criterion	-3.769896
Sum squared resid	0.003033	Schwarz criterion	-3.926120
Log likelihood	11.42474	Hannan-Quinn criter.	-4.189188
F-statistic	38.44562	Durbin-Watson stat	1.827381
Prob(F-statistic)	0.008452		

Question 4 : Equation of the model

A-ER_MARKET_INDEX=1,64*ER_FUND-0,017

B-ER_FUND=1,64*ER_MARKET_INDEX-0,017

C-I don’t have enough information to conclude

α And β in the CAPM Example

In the CAPM example used above, the estimates are:

Dependent variable: ER_FUND

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.017366	0.041140	-0.422132	0.7014
ER_MARKET_INDEX	1.641745	0.264778	6.200453	0.0085

R-squared	0.927616	Mean dependent var	0.222000
Adjusted R-squared	0.903488	S.D. dependent var	0.102343
S.E. of regression	0.031794	Akaike info criterion	-3.769896
Sum squared resid	0.003033	Schwarz criterion	-3.926120
Log likelihood	11.42474	Hannan-Quinn criter.	-4.189188
F-statistic	38.44562	Durbin-Watson stat	1.827381
Prob(F-statistic)	0.008452		

Question 5 : Interpreting the coefficients

A-When the excess return of the market increases by one, the excess return of the fund is multiplied on average by 1.64

B-When the excess return of the market increases by one, the excess return of the fund increases on average by 1.64

C-When the excess return of the market decreases by one, the excess return of the fund decreases on average by 1.64

Model Validation

-Tests on the coefficients

-R^2

-Analysis of residuals

Coefficients : Precision and Standard Errors

- Regression **estimates** of α and β are **specific to the sample** used in their estimation.
- **Can we rely on these estimates**? Do they vary much from one sample to another? ➔ **measure of the reliability or precision of the estimators**
- **The precision of the estimate is given by its standard error, SE**:

$$SE(\hat{\alpha}) = s\sqrt{\frac{\sum X_t^2}{T\sum (X_t - \overline{X})^2}} \qquad SE(\hat{\beta}) = s\sqrt{\frac{1}{\sum (X_t - \overline{X})^2}}$$

- Where **s is the estimated standard deviation of the residuals**
  - The variance of the random variable *U*, $Var(U) = E[(U)-E(U)]^2 = E(U^2)$ can be estimated by :

$$s^2 = \frac{1}{T-2}\sum \hat{U}_t^2$$

  - $s = \sqrt{s^2}$ **is the standard error of the regression (estimated standard deviation of the residuals)**

Reliability of the coefficients

Reliability ?

- Can we consider that $\hat{\beta}$ is significant ? (statistically different from 0)?
- What about $\hat{\alpha}$?

Coefficients : Hypothesis Testing

- 3 types of tests

$$H0: \beta = \beta_0$$
$$H1: \beta \neq \beta_0$$

← **Two-sided test**

$$H0: \beta = \beta_0$$
$$H1: \beta > \beta_0$$

← **One-sided test (right tail)**

$$H0: \beta = \beta_0$$
$$H1: \beta < \beta_0$$

← **One-sided test (left tail)**

We can use the same type of test for the intercept α

Coefficients : Hypothesis Testing

We assume that $U \sim N(0,\sigma^2)$

- Then the OLS estimators are normally distributed :

$$\hat{\alpha} \sim N(\alpha, \text{Var}(\alpha))$$
$$\hat{\beta} \sim N(\beta, \text{Var}(\beta))$$

- **What if the errors are not normally distributed?**

  The parameter estimates still be normally distributed if the other assumptions of the CLRM hold, and the sample size is sufficiently large.

Coefficients : Hypothesis Testing

- Test Statistics for $\hat{\alpha}$ and $\hat{\beta}$:

$$t = \frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \sim \text{Student distribution (T - 2) degrees of freedom)}$$

$$t = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})} \sim \text{Student distribution (T - 2) degrees of freedom)}$$

- Most commonly used tests :

$H0 : \beta = 0$ $\quad$ $H0 : \alpha = 0$

$H1 : \beta \neq 0$ $\quad$ $H1 : \alpha \neq 0$

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim \text{Student (T - 2) dof} \qquad t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \sim \text{Student (T - 2) dof}$$

- These t-ratio are provided by any econometric software

Coefficients : Hypothesis Testing

- Decision rule to choose between H0 et H1

$$H0: \beta = \beta_0 \qquad H0: \alpha = \alpha_0$$
$$H1: \beta \neq \beta_0 \qquad H1: \alpha \neq \alpha_0$$

1-Use the pvalue of the test (provided by any econometrical software)
pvalue= probability of rejecting H0 given H0 is true

When we take the usual significance level of 5%,

-pvalue < 5% ➔ we reject H0

-pvalue ≥5% ➔ we do not reject H0

2-Compare the t-statistic to a critical value obtained with the Student distribution and a risk level of 5%. When the sample size is large, whatever T, the critical value for a risk level of 5% is around 2 (absolute value).

If we reject the null hypothesis at the 5% level, we say that the result of the test is statistically significant.

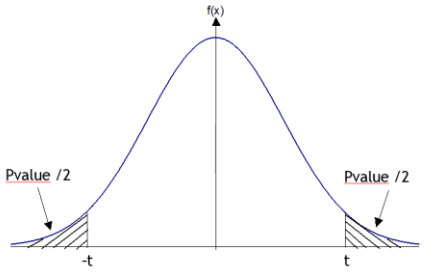
The Test of Significance Approach

- α= 5% determine a rejection region and non-rejection region for a **2-sided test**:

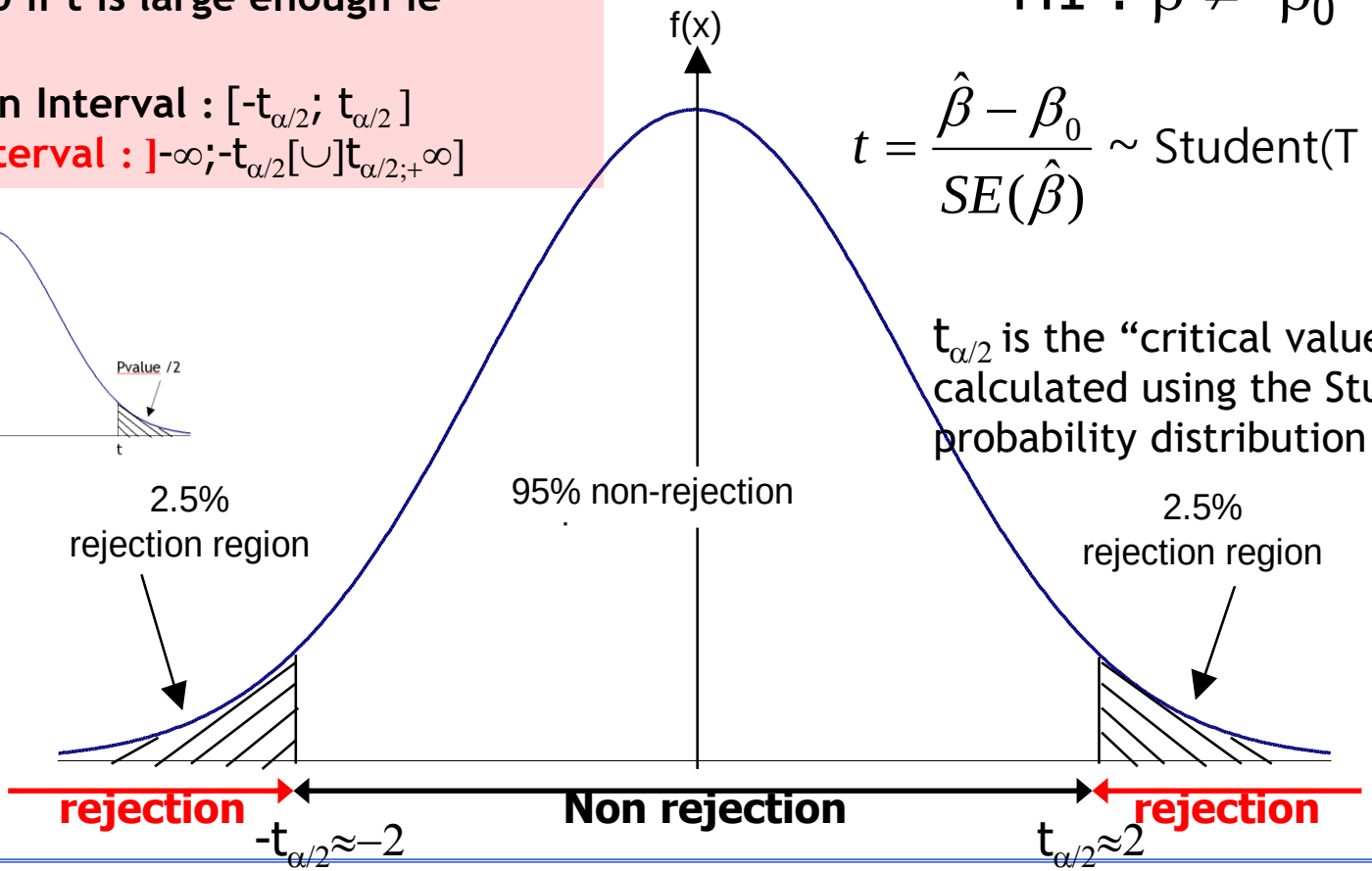
$$H0 : \beta = \beta_0$$
$$H1 : \beta \neq \beta_0$$

We reject H0 if t is large enough ie
$|t| > t_{\alpha/2}$
Non rejection Interval : $[-t_{\alpha/2}; t_{\alpha/2}]$
Rejection Interval : $]-\infty;-t_{\alpha/2}[\cup]t_{\alpha/2;+}\infty]$

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \sim \text{Student}(T - 2)$$



$t_{\alpha/2}$ is the "critical value" and is calculated using the Student probability distribution function



α And β in the CAPM Example

In the CAPM example used above, the estimates are:

Dependent variable: ER_FUND

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.017366	0.041140	-0.422132	0.7014
ER_MARKET_INDEX	1.641745	0.264778	6.200453	0.0085

R-squared	0.927616	Mean dependent var	0.222000
Adjusted R-squared	0.903488	S.D. dependent var	0.102343
S.E. of regression	0.031794	Akaike info criterion	-3.769896
Sum squared resid	0.003033	Schwarz criterion	-3.926120
Log likelihood	11.42474	Hannan-Quinn criter.	-4.189188
F-statistic	38.44562	Durbin-Watson stat	1.827381
Prob(F-statistic)	0.008452		

Question 6 : which affirmation is true?

A- the fund outperforms the market

B- the fund has no residual risk premium

C- the fund underperforms the market

Question 7 : which affirmation is true?

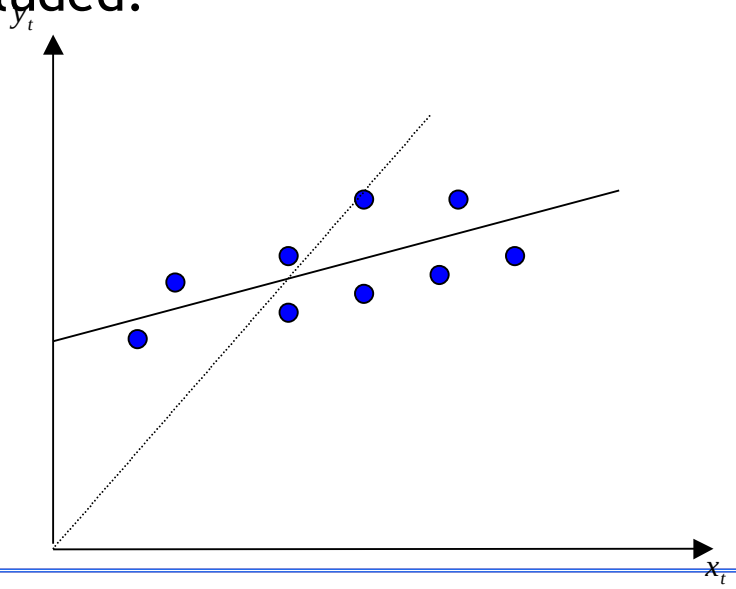
A- the fund excess return is not correlated to the market excess return

B- the fund excess return is correlated to the market excess return

C- the fund excess return is 1.64 times higher than the market excess return

What to do if a coefficient is not significant?

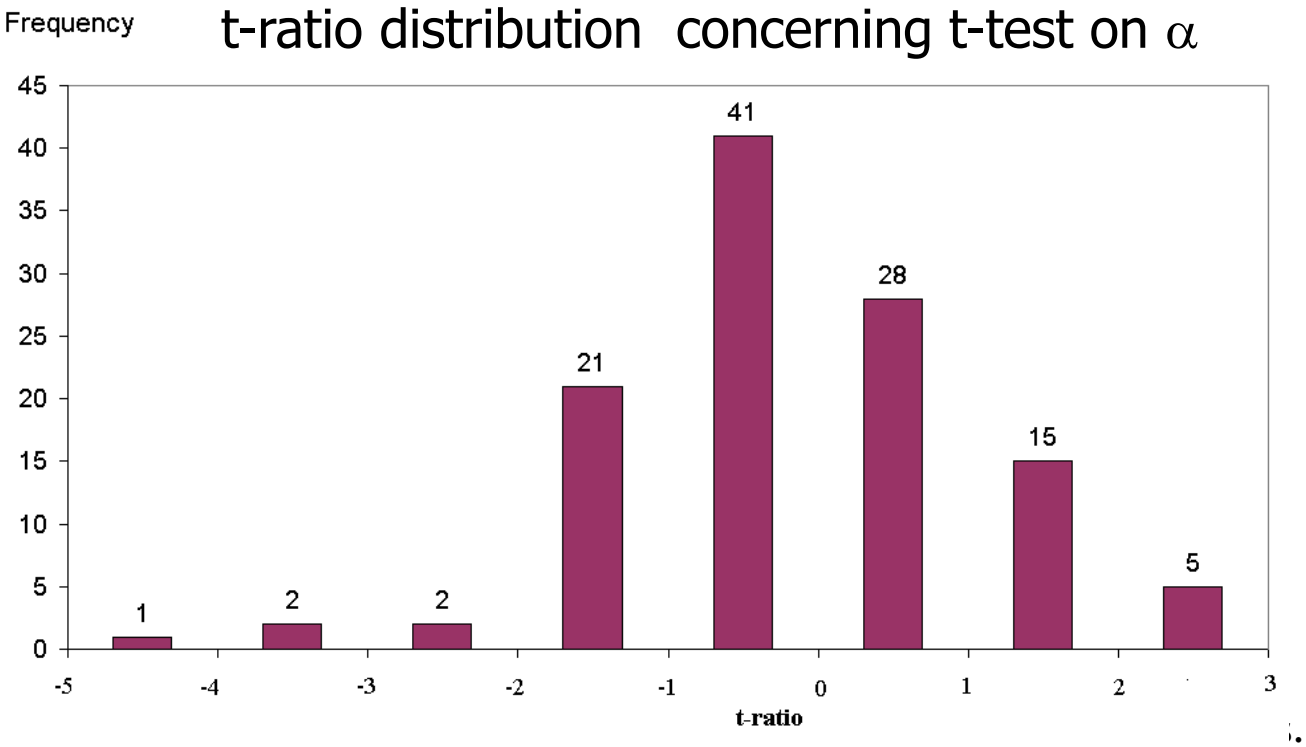
- If we reject H_0, we say that the result is significant. If the coefficient is not "significant" (e.g. the intercept coefficient in the last regression above), then it means that the variable is not helping to explain variations in *y*. Variables that are not significant are usually removed from the regression model.
- In practice there are good statistical reasons for always having a constant even if it is not significant. Look at what happens if no intercept is included:



An Example of the Use of a Simple t- test to Test a Theory in Finance (cf Brooks)

- Testing for the presence and significance of abnormal returns ("Jensen's alpha" - Jensen, 1968).
- The Data: Annual Returns on the portfolios of 115 mutual funds from 1945-1964.
- The model: $R_{jt} - R_{ft} = \alpha_j + \beta_j (R_{mt} - R_{ft}) + u_{jt}$ for j = 1, …, 115
- We are interested in the significance of αj.
- The null hypothesis is H0: $\alpha j = 0$.

Frequency Distribution of t-ratios of Mutual Fund Alphas (**gross of transactions costs**)



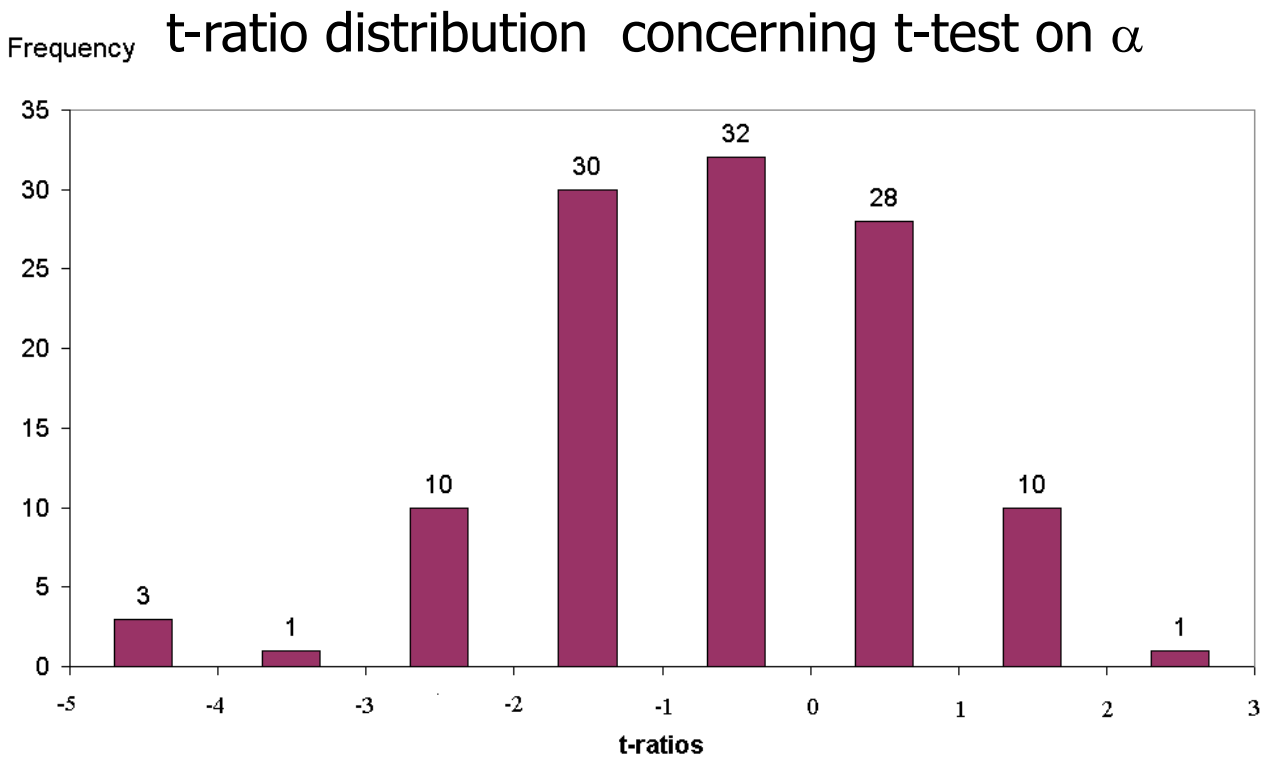
Question 8 : Knowing that the Critical value $t_{\alpha/2}$ for a two-sided test ≈ 2, which affirmation is false?

A- 5 funds underperform the market

B- 5 funds outperform the market

C- no fund has a residual risk premium (not better than the market)

Frequency Distribution of t-ratios of Mutual Fund Alphas (**net of transactions costs**)



Source Jensen (1968). Reprinted with the permission of Blackwell publishers.

Question 9: Knowing that the Critical value $t_{\alpha/2}$ for a two-sided test ≈ 2, what can we conclude ?

Goodness of Fit Statistics

Goodness of Fit Statistics

How well our regression model actually fits the data?

R^2 : proportion of variation in y "explained" by the regressors in the model.

- $R^2 = 1$ ➔ the fitted model explains all variability
- $R^2 = 0$ ➔ no 'linear' relationship (for straight line regression, this means that the straight line model is a constant line (slope=0, intercept= $\bar{y}$) between the response variable and regressors

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

TSS = Variability of Y

ESS = Variability of $\hat{Y}$

RSS = Variability of $\hat{U}$

TSS = ESS + RSS

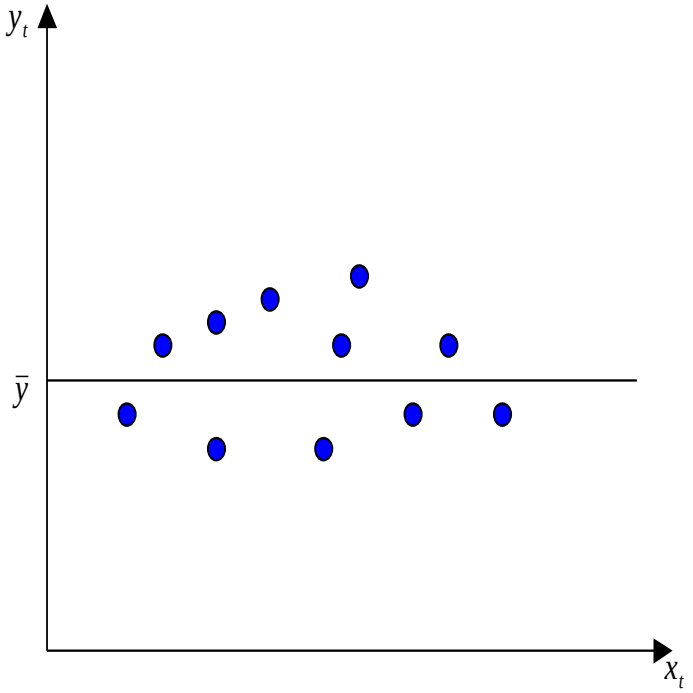
$$\sum_t \left(Y_t - \bar{Y}\right)^2 = \sum_t \left(\hat{Y}_t - \bar{Y}\right)^2 + \sum_t \hat{U}_t^2$$

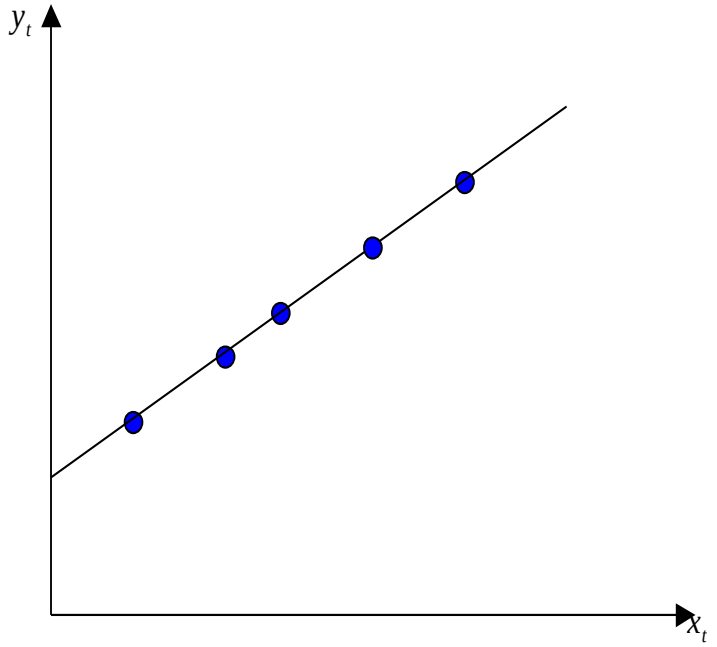
^{t}TSS = Total sum of squares

ESS = Explained sum of squares

^{t}RSS = Residual sum of squares

Illustration of Limit Cases: $R^2 = 0$ and $R^2 = 1$





TUTORIAL XLSTAT

3. Test for CAPM

- Correlation
- Regression
- Tests on coefficients
- Goodness of fit

Tutorial

- XLSTAT – scatter plot
  - ⇒ is there an approximative linear relationship?
  - ⇒ are the variables correlated?

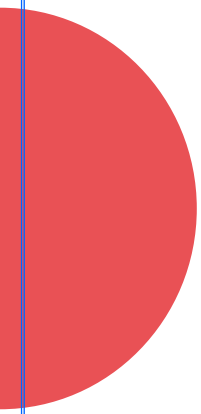
- XLSTAT – linear regression

  ➔ Run the regression: $ER_{msoft,t} = \alpha + \beta(ER_{s\&p,t}) + U_t$

  ➔ Estimate the coefficients of the model: α and β

  ➔ Interpret the significance test for coefficients (t-ratios)
  - ⇒is α significantly different from zero?
  - ⇒ what about β ?

  ➔ Discuss the goodness of fit (R^2)



Generalising to Multiple Linear Regression

Generalising the Simple Model to Multiple Linear Regression

- Before, we have used the model

$$Y_t = \alpha + \beta X_t + U_t \qquad t = 1,2,\ldots,T$$

- If our dependent (Y) variable depends on more than one independent variable?

$$Y_t = \beta_1 X_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \ldots + \beta_k X_{kt} + U_t \quad t = 1,2,\ldots,T$$

Tests on coefficients
T-tests and F-tests

Testing Hypotheses involving only one coefficient : t-test

Hypotheses involving only one coefficient ➔ *t*-test

As seen before the test statistic is :

$$H0: \beta = \beta_0 \qquad H1: \beta \neq \beta_0$$

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \sim \text{Student}(T - k)$$

k = number of regressors
T = sample size

The decision rule remains the same as in the simple regression model

pvalue < 5% ➔ we reject H0 (➔ coefficient different from 0)

Testing Hypotheses involving only one coefficient : t-test

- Relationship between the Malaysian market (RMT) and three others close markets (Indonesia, Singapore and Thailand)

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000378	0.000630	-0.600425	0.5486
R_INDONESIA	0.075668	0.043890	1.724055	0.0855
R_SINGAPORE	0.002118	0.000482	4.392101	0.0000
R_THAILAND	0.092578	0.038079	2.431198	0.0155

R-squared	0.104851	Mean dependent var	-0.000749
Adjusted R-squared	0.097911	S.D. dependent var	0.013048
S.E. of regression	0.012393	Akaike info criterion	-5.933248
Sum squared resid	0.059435	Schwarz criterion	-5.892647
Log likelihood	1163.950	Hannan-Quinn criter.	-5.917155
F-statistic	15.11001	Durbin-Watson stat	1.536228
Prob(F-statistic)	0.000000		

Question 10: Which coefficients are significantly different from 0?

Testing Multiple Hypotheses

Hypothesis involving more than one coefficient simultaneously? ➔ *F*-test

For example H0: $\beta_2 = \beta_3$, H0: $\beta_2 + \beta_3 = 1$, H0: $\beta_1 = 0$ and $\beta_2 = 1$

Remark : We cannot test using this framework nonlinear or multiplicative hypothesis, e.g. H0: $\beta_2 \beta_3 = 2$ or H0: $\beta^2_2 = 1$

The *F*-test involves estimating 2 regressions :

- ➔ The **unrestricted regression** is the one in which the coefficients are freely determined by the data, as we have done before
- ➔ The **restricted regression** is the one in which the coefficients are restricted, i.e. the restrictions are imposed on some βs.
- ➔ Compare the RSS of the 2 regressions to construct the statistics
- ➔ **Test statistic ~ *Fisher* distribution (dof1=m;dof2=T-k)**
- ➔ **reject the null if the test statistic > critical *F*-value or pvalue<5%**

Testing Multiple Hypotheses

A specific F-test : Global Test for Regression Significance

Example

model : $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + U_t,$

then $H_0: \beta_2 = \beta_3 = \beta_4 = 0$

against

H1 : at least one coefficient is significantly different from 0

- test the **global significance of the regression**
- provided automatically by all statistical software
- If pvalue <5%, reject H0 => the regression is globally significant

Testing Multiple Hypotheses

- Example :
  - Write the test for the global significance of the regression (H0 and H1)
  - Conclusion?
  - Are all the coefficient (except the constant) significantly different from 0?

Dependent variable: ER_Microsoft
obs: 63

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.311743	0.669841	0.465399	0.6434
ER_SANDP	0.952967	0.187872	5.072435	0.0000
SMB	-0.135798	0.247568	-0.548526	0.5854
HML	-0.824711	0.336805	-2.448633	0.0173

R-squared	0.421982	Mean dependent var	-0.076405
Adjusted R-squared	0.392591	S.D. dependent var	6.332901
S.E. of regression	4.935638	Akaike info criterion	6.092228
Sum squared resid	1437.271	Schwarz criterion	6.228300
Log likelihood	-187.9052	Hannan-Quinn criter.	6.145746
F-statistic	14.35764	Durbin-Watson stat	2.472399
Prob(F-statistic)	0.000000		

Goodness of Fit Statistics

Goodness of Fit Statistics

How well our regression model actually fits the data?

R^2 : proportion of variation in Y "explained" by the regressors in the model.

- $R^2 = 1$ ➔ the fitted model explains all variability in,
- $R^2 = 0$ ➔ no 'linear' relationship (for straight line regression, this means that the straight line model is a constant line (slope=0, intercept= $\overline{Y}$) between the response variable and regressors

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

TSS = Variability of Y

ESS = Variability of $\hat{Y}$

RSS = Variability of $\hat{U}$

TSS = ESS + RSS

$$\sum_t \left(Y_t - \overline{Y}\right)^2 = \sum_t \left(\hat{Y}_t - \overline{Y}\right)^2 + \sum_t \hat{U}_t^2$$

TSS = Total sum of squares

ESS = Explained sum of squares

RSS = Residual sum of squares

Adjusted R^2

- ***Be careful !*** **R^2 never falls if more regressors are added to the regression**
- to get around these problems : take into account the loss of degrees of freedom associated with adding extra variables

➔ adjusted R^2:

$$\overline{R}^2 = 1 - \left[\frac{T-1}{T-k}(1-R^2)\right]$$

- So if we add an extra regressor, k increases and contrary to the R^2 the $\overline{R}^2$ may decrease.
- As soon as $k \geq 2$, $\overline{R}^2 < R^2$
- While R^2 must be at least zero, $\overline{R}^2$ may take negative values if the model fits the data very poorly.

Adjusted R^2

- Comment ?

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.311743	0.669841	0.465399	0.6434
ERSANDP	0.952967	0.187872	5.072435	0.0000
SMB	-0.135798	0.247568	-0.548526	0.5854
HML	-0.824711	0.336805	-2.448633	0.0173

R-squared	0.421982	Mean dependent var	-0.076405
Adjusted R-squared	0.392591	S.D. dependent var	6.332901
S.E. of regression	4.935638	Akaike info criterion	6.092228
Sum squared resid	1437.271	Schwarz criterion	6.228300
Log likelihood	-187.9052	Hannan-Quinn criter.	6.145746
F-statistic	14.35764	Durbin-Watson stat	2.472399
Prob(F-statistic)	0.000000		

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.263091	0.660063	0.398584	0.6916
ERSANDP	0.934538	0.183763	5.085558	0.0000
HML	-0.833806	0.334431	-2.493212	0.0154

R-squared	0.419034	Mean dependent var	-0.076405
Adjusted R-squared	0.399669	S.D. dependent var	6.332901
S.E. of regression	4.906799	Akaike info criterion	6.065568
Sum squared resid	1444.600	Schwarz criterion	6.167622
Log likelihood	-188.0654	Hannan-Quinn criter.	6.105707
F-statistic	21.63815	Durbin-Watson stat	2.429241
Prob(F-statistic)	0.000000		

CAPM Ford / SP500

- **Comment :** t-statistics? R^2? F-statistic?

Dependent variable: ER_Ford
obs: 63

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.020219	2.801382	0.721151	0.4736
ERSANDP	0.359726	0.794443	0.452803	0.6523

R-squared	0.003350	Mean dependent var	2.097445
Adjusted R-squared	-0.012989	S.D. dependent var	22.05129
S.E. of regression	22.19404	Akaike info criterion	9.068756
Sum squared resid	30047.09	Schwarz criterion	9.136792
Log likelihood	-283.6658	Hannan-Quinn criter.	9.095514
F-statistic	0.205031	Durbin-Watson stat	1.785699
Prob(F-statistic)	0.652297		

CAPM Microsoft / SP500

- **Comment :** t-statistics? R^2? Fstatistic?

Dependent variable: ER_Microsoft
obs: 63

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.108327	0.645998	-0.167690	0.8674
ERSANDP	1.070463	0.183198	5.843195	0.0000

R-squared	0.358859	Mean dependent var	0.121478
Adjusted R-squared	0.348349	S.D. dependent var	6.339973
S.E. of regression	5.117937	Akaike info criterion	6.134611
Sum squared resid	1597.790	Schwarz criterion	6.202647
Log likelihood	-191.2403	Hannan-Quinn criter.	6.161370
F-statistic	34.14293	Durbin-Watson stat	2.208231
Prob(F-statistic)	0.000000		

TUTORIAL XLSTAT

4. Multiple regression

- Global significance
- Tests on coefficients
- Goodness of fit

Tutorial

Extension to multiple regression analysis

Based on the regression:

$$ER_{msoft,t} = \alpha + \beta 1\ (ER_{s\&p,t}) + \beta 2(SIZE) + \beta 2(B/M) + U_t$$

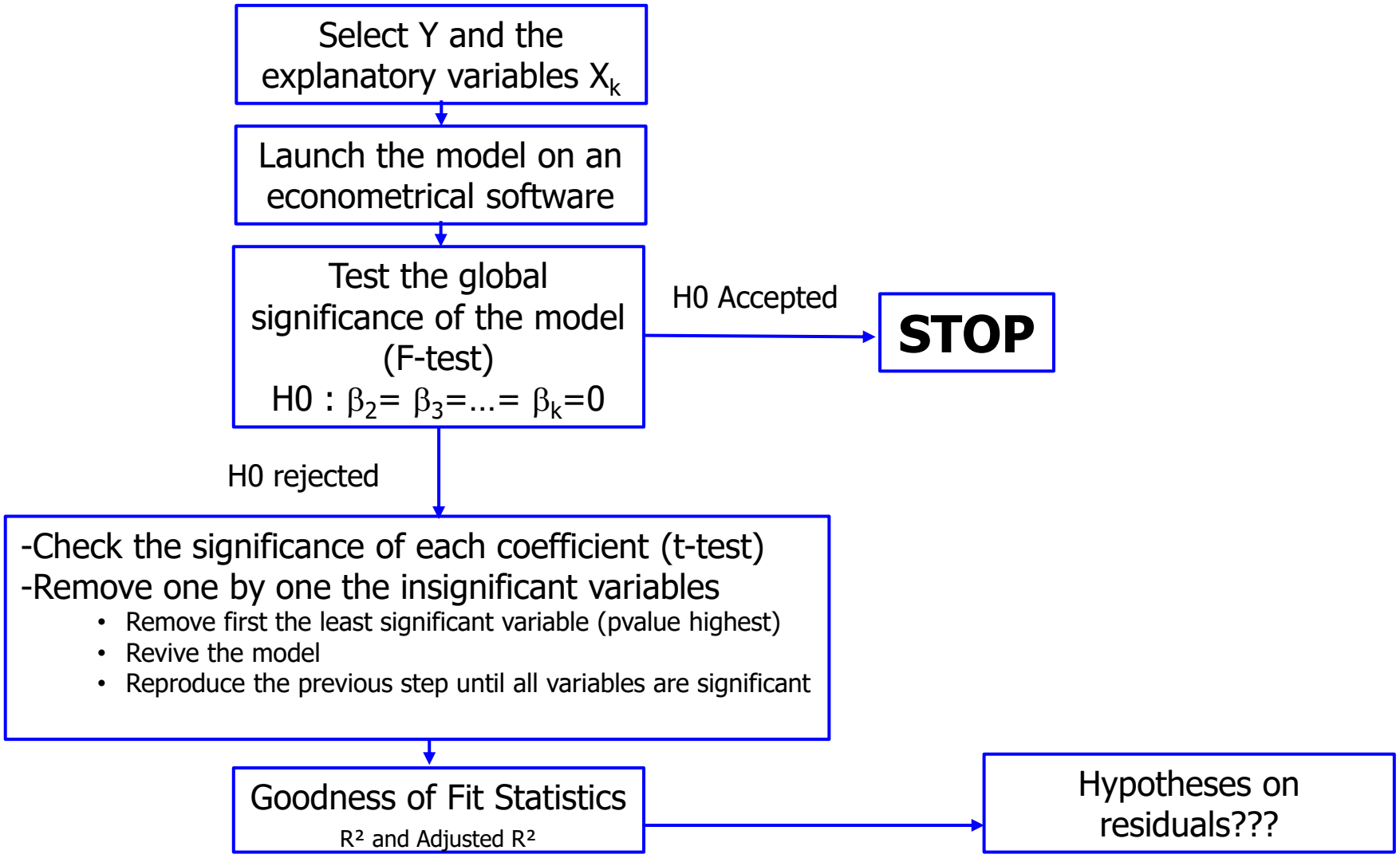
→Interpret the significance test for coefficients (t-ratios)

→If one coefficient is not significant, run again a regression without the corresponding variable (keep the constant even is not significant though)

→Discuss the global significance (F-test) and goodness of fit (adjusted R2)

→Which of the 2 models gives the best estimation?

What you have to retain



Violation of the assumptions of the CLRM and remedies

The Assumptions Underlying the (CLRM)

- First, the CLRM is based on the assumption that the regression model is **linear** in the parameters (model correctly specified)
- We observe data for X_t, but Y_t also depends on U_t. Hence, we usually make the following **assumptions** about the U_t's (the unobserved error terms):

1. $E(U_t) = 0$	The errors have zero mean
2. $U_t \sim N(0, \sigma^2)$	Normally distributed. Useful to make inferences about the population parameters
3. $Var(U_t) = \sigma^2 < \infty$	The variance of the errors is constant and finite over all values of X_t
4. $Cov(U_i, U_j) = 0$	The errors are statistically independent of one another
5. $Cov(U_t, X_t) = 0$	No relationship between the error and corresponding X variate

Violations of the Assumptions of the CLRM

What is the impact on the regression if one or more of these assumptions are not validated?

Violations ➔ pb to infer

- The coefficient estimates are wrong
- The associated standard errors are wrong
- The distribution that we assumed for the test statistics will be inappropriate

Solutions : Operate such that

- The assumptions are no longer violated (clean, transform, use larger sample…)

➔ alternative techniques can be used: alternative regression methods, robust standard errors…

Assumption 1: E(ut) = 0

Assumption that the mean of the disturbances is zero.

- The mean of the residuals will always be zero if there is a constant term included in the regression equation.

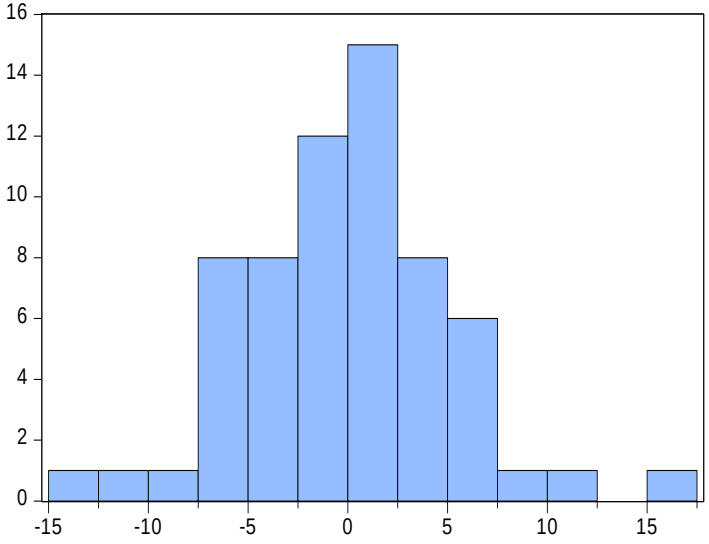
CAPM Microsoft / SP500

Dependent variable: ER_Microsoft
obs: 63

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.108327	0.645998	-0.167690	0.8674
ERSANDP	1.070463	0.183198	5.843195	0.0000

R-squared	0.358859	Mean dependent var	0.121478
Adjusted R-squared	0.348349	S.D. dependent var	6.339973
S.E. of regression	5.117937	Akaike info criterion	6.134611
Sum squared resid	1597.790	Schwarz criterion	6.202647
Log likelihood	-191.2403	Hannan-Quinn criter.	6.161370
F-statistic	34.14293	Durbin-Watson stat	2.208231
Prob(F-statistic)	0.000000		

Comment :
-residuals mean

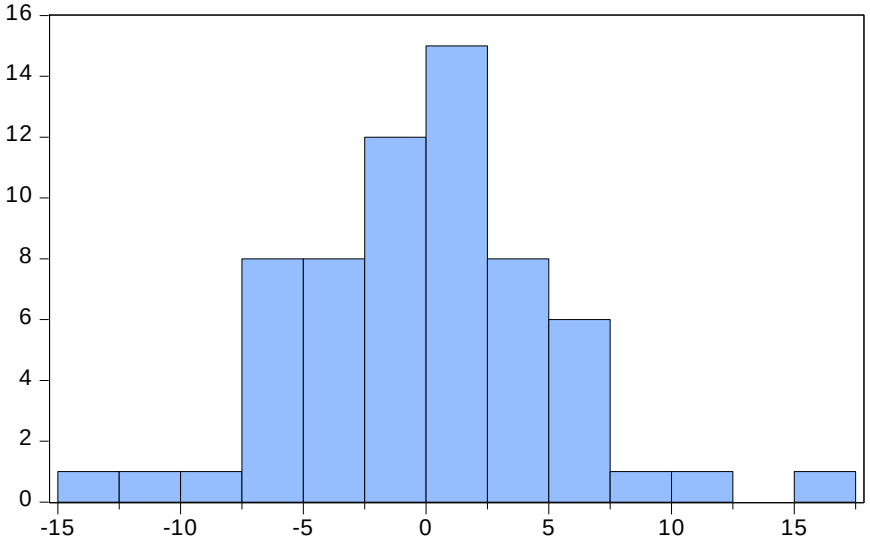


Series: Residuals
Sample 2002M02 2007M04
Observations 63

Mean	-2.04e-16
Median	0.543140
Maximum	15.45907
Minimum	-12.70471
Std. Dev.	5.076496
Skewness	0.172040
Kurtosis	3.726589
Jarque-Bera	1.696599
Probability	0.428142

Assumption 2: Ut ~ N(0,σ^2)

CAPM
(Microsoft/SP500)



Series: Residuals
Sample 2002M02 2007M04
Observations 63

Mean	-2.04e-16
Median	0.543140
Maximum	15.45907
Minimum	-12.70471
Std. Dev.	5.076496
Skewness	0.172040
Kurtosis	3.726589
Excess kurtosis	0.726589
Jarque-Bera	1.696599
Probability	0.428142

Jarque-Bera test:
H0 : the series is normally distributed
H1 : the series is not normally distributed

$$JB = \frac{T-k}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \quad \sim \chi^2 \text{ (2 dof)}$$

T : number of observations; k : number of explanatory variables if the normality of regression residuals is tested, 0 otherwise; S : Skewness; K : Kurtosis; α : risk level

We reject H0 if JB > $\chi^2_{2;\alpha}$ or if pvalue < α

Comment :

- residuals normality?

Residual normality and outliers

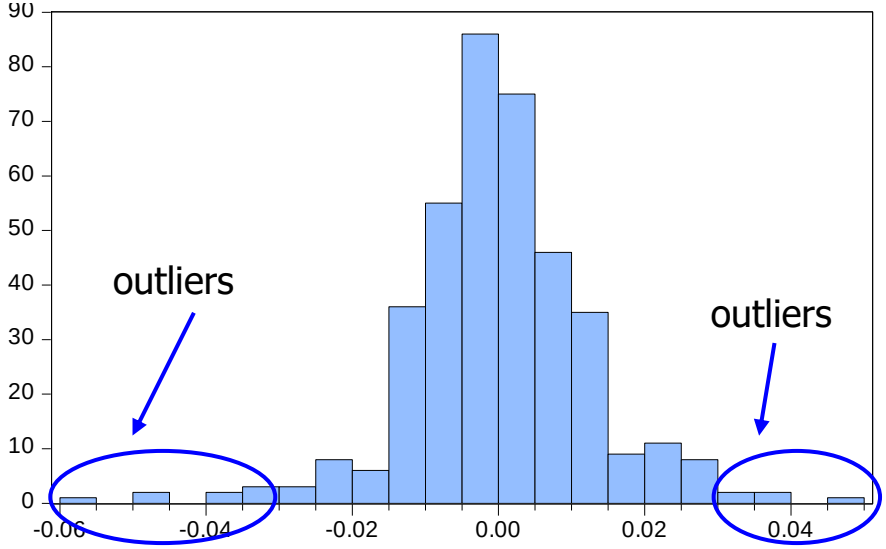
Dependent variable: RMT
obs: 391

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000440	0.000630	-0.698715	0.4851
R_SINGAPORE	0.002254	0.000477	4.727190	0.0000
R_THAILAND	0.096298	0.038114	2.526546	0.0119

R-squared	0.097975	Mean dependent var	-0.000749
Adjusted R-squared	0.093326	S.D. dependent var	0.013048
S.E. of regression	0.012424	Akaike info criterion	-5.930711
Sum squared resid	0.059891	Schwarz criterion	-5.900261
Log likelihood	1162.454	Hannan-Quinn criter.	-5.918642
F-statistic	21.07171	Durbin-Watson stat	1.527428
Prob(F-statistic)	0.000000		

Malaysian Index Market vs Thailand and Singapore

Comment :
-residuals normality?



Series: Residuals
Sample 125 515
Observations 391

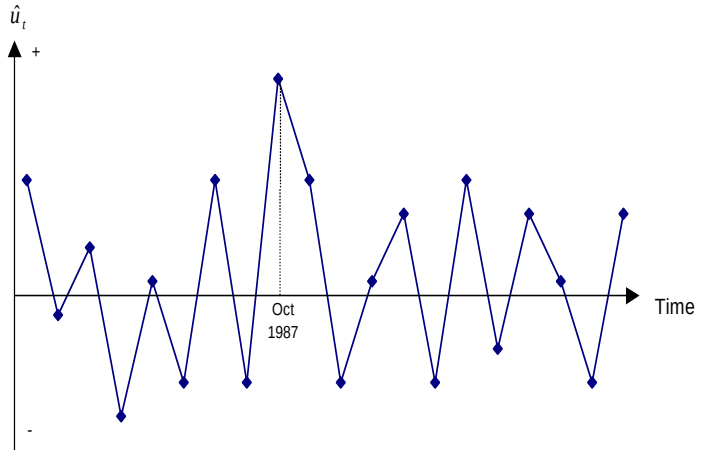
Mean	4.37e-19
Median	-0.000313
Maximum	0.045410
Minimum	-0.056123
Std. Dev.	0.012392
Skewness	-0.247502
Kurtosis	5.714357
Excess kurtosis	2.714357
Jarque-Bera	124.0246
Probability	0.000000

What do we do in case of Non-Normality?

- **Outliers** : one or two very extreme residuals causes us to reject the normality assumption

- Alternative : use dummy variables.

e.g. say we estimate a monthly model of asset returns from 1980-1990, and we plot the residuals, and find a particularly large outlier for October 1987



Create a new variable:
$D87M10_t$ = 1 during October 1987 and zero otherwise.
This effectively knocks out that observation. But we need a theoretical reason for adding dummy variables… (special event …)

Date	dummy
janv-80	0
févr-80	0
mars-80	0
avr-80	0
…	…
juin-87	0
juil-87	0
août-87	0
sept-87	0
oct-87	**1**
nov-87	0
déc-87	0
janv-88	0

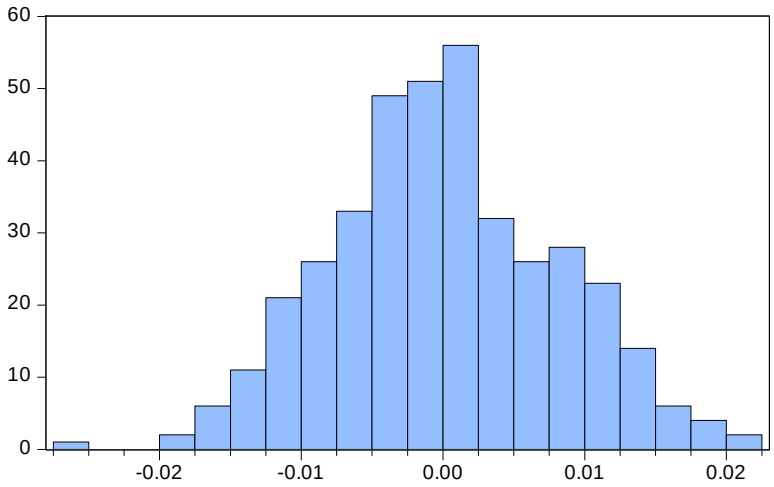
Residual normality and dummies

Dependent variable: RMT
obs: 391

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.000662	0.000429	-1.541863	0.1239
R_SINGAPORE	0.002085	0.000306	6.804805	0.0000
R_THAILAND	0.081598	0.024528	3.326691	0.0010
DUMMYM	-0.030438	0.001881	-16.18540	0.0000
DUMMYP	0.027226	0.001688	16.12586	0.0000

R-squared	0.629828	Mean dependent var	-0.000749
Adjusted R-squared	0.625992	S.D. dependent var	0.013048
S.E. of regression	0.007980	Akaike info criterion	-6.811155
Sum squared resid	0.024578	Schwarz criterion	-6.760405
Log likelihood	1336.581	Hannan-Quinn criter.	-6.791039
F-statistic	164.1896	Durbin-Watson stat	1.794745
Prob(F-statistic)	0.000000		

Comment :
-residuals normality?

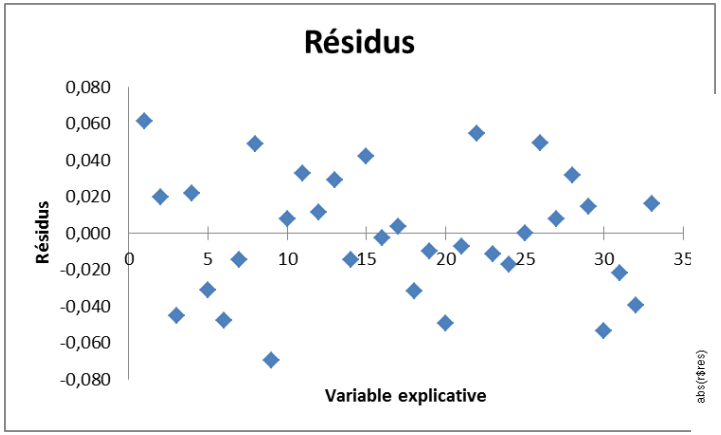


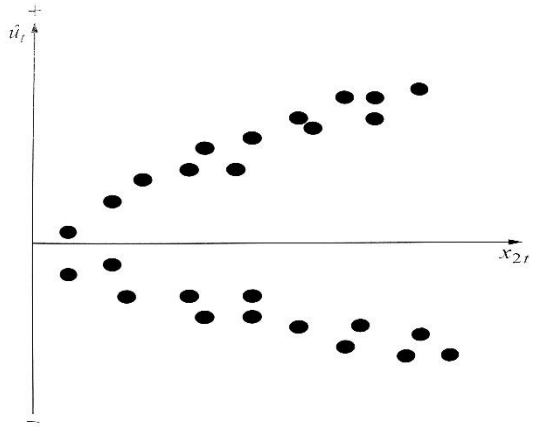
Series: Residuals
Sample 125 515
Observations 391

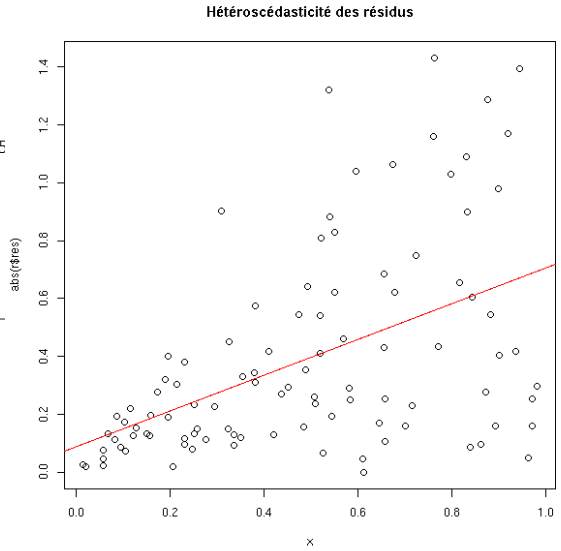
Mean	4.66e-19
Median	-0.000214
Maximum	0.020448
Minimum	-0.025999
Std. Dev.	0.007939
Skewness	0.094913
Kurtosis	2.793170
Excess kurtosis	0.793170
Jarque-Bera	1.283993
Probability	0.526241

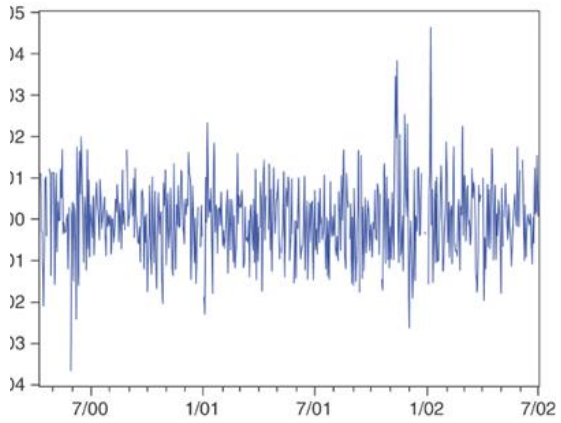
Assumption 3: Var(Ut) = $\sigma^2 < \infty$

- variance of the errors is constant ➔**homoscedasticity**
- variance of the errors is not constant ➔ **heteroscedasticity**









Detection of Heteroscedasticity

- Graphical methods
- Formal tests:

➔ Goldfeld-Quandt test: Split the total sample of length T into two sub-samples of length T_1 and T_2. The regression model is estimated on each sub-sample and the two residual variances are calculated. Test H0: $\sigma_1^2 = \sigma_2^2$ (the variances of the disturbances are equal).

➔ White's test: Check if the variance of the residuals varies systematically with any known variables relevant to the model. Regress $\hat{U}_t^2$ on relevant variables (auxiliary regression). Test statistics based on R^2 of this regression.

Decision rule : $TR^2 > \chi^2_{\alpha,m}$ or pvalue <5% ➔ reject the null hypothesis that the disturbances are homoscedastic.

Model house Price : price = f(rooms, sqfeet)

Comment : Heteroscedasticity?

Dependent variable: Price
obs: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19315.00	31046.62	-0.622129	0.5355
ROOMS	15198.19	9483.517	1.602590	0.1127
SQFEET	128.4362	13.82446	9.290506	0.0000

R-squared	0.631918	Mean dependent var	293546.0
Adjusted R-squared	0.623258	S.D. dependent var	102713.4
S.E. of regression	63044.84	Akaike info criterion	24.97458
Sum squared resid	3.38E+11	Schwarz criterion	25.05903
Log likelihood	-1095.881	Hannan-Quinn criter.	25.00860
F-statistic	72.96353	Durbin-Watson stat	1.757956
Prob(F-statistic)	0.000000		

Heteroskedasticity Test: White

F-statistic	3.991436	Prob. F(5,82)	0.0027
Obs*R-squared	17.22519	Prob. Chi-Square(5)	0.0041
Scaled explained SS	37.67476	Prob. Chi-Square(5)	0.0000

Dependent variable: Resid^2
obs: 88

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.08E+10	1.31E+10	0.822323	0.4133
ROOMS^2	-1.28E+09	8.39E+08	-1.523220	0.1316
ROOMS*SQFEET	1979155.	1819402.	1.087805	0.2799
ROOMS	7.00E+09	5.67E+09	1.234867	0.2204
SQFEET^2	4020.876	2198.691	1.828759	0.0711
SQFEET	-23404693	10076371	-2.322730	0.0227

R-squared	0.195741	Mean dependent var	3.84E+09
Adjusted R-squared	0.146701	S.D. dependent var	8.36E+09
S.E. of regression	7.72E+09	Akaike info criterion	48.43858
Sum squared resid	4.89E+21	Schwarz criterion	48.60749

Question 11 : Which affirmation is true?

A- at 5% risk level we can conclude that the residuals are homoskedastic because of the White's test p-value

B- at 5% risk level we can conclude that the residuals are homoskedastic because the variance of the residuals increases with the SQFEET

C- at 5% risk level we can conclude that the residuals are heteroskedastic because of the White's test p-value

D- I don't know

Assumption 4: Cov (U_t, U_{t-1}) = 0

Cov (U_t , U_s) = 0 for $t \neq s$

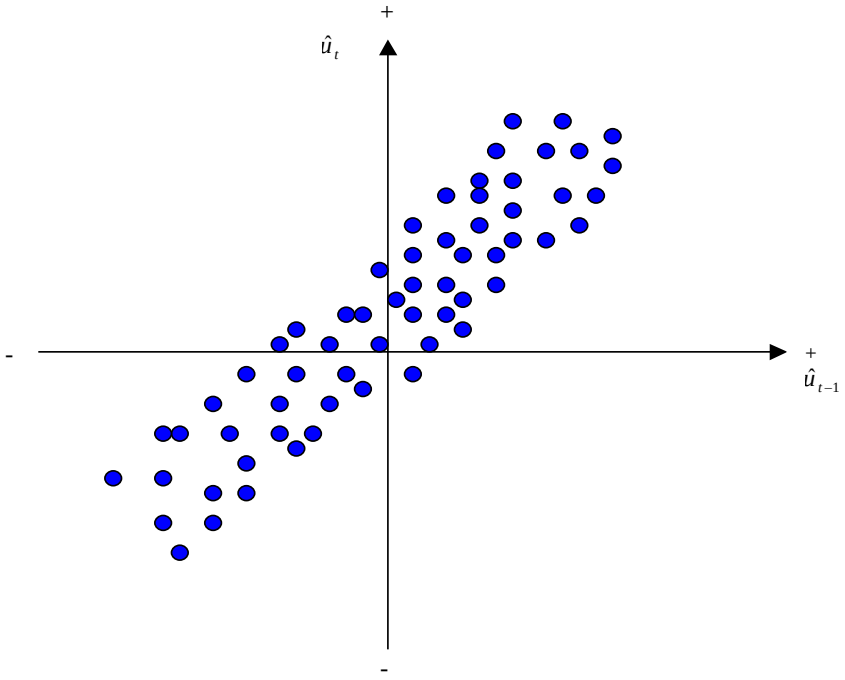
Cov (U_i , U_j) = 0 for $i \neq j$,

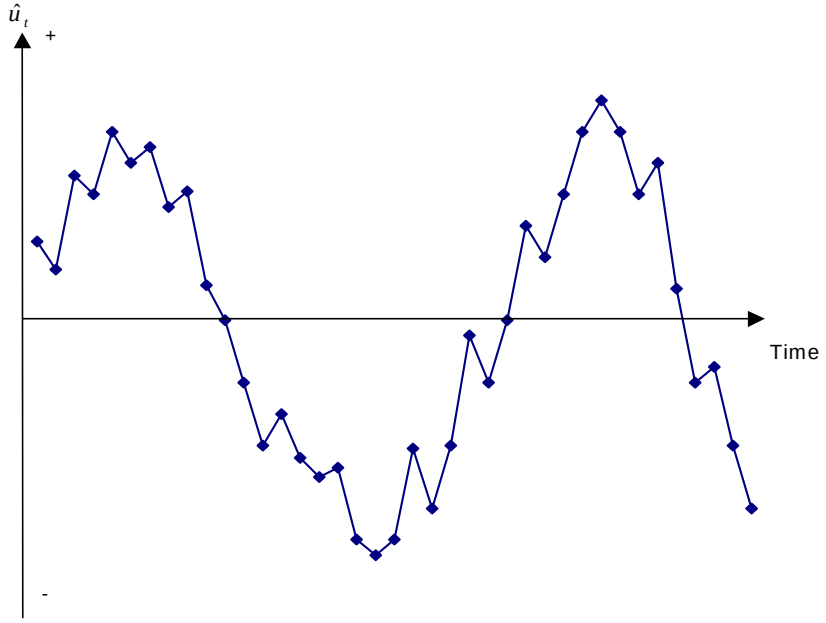
➔ no pattern in the errors.

Background – The Concept of a Lagged Value

t	U_t	U_{t-1}	ΔU_t
1989M09	0.8	-	-
1989M10	1.3	0.8	1.3-0.8=0.5
1989M11	-0.9	1.3	-0.9-1.3=-2.2
1989M12	0.2	-0.9	0.2--0.9=1.1
1990M01	-1.7	0.2	-1.7-0.2=-1.9
1990M02	2.3	-1.7	2.3--1.7=4.0
1990M03	0.1	2.3	0.1-2.3=-2.2
1990M04	0.0	0.1	0.0-0.1=-0.1
.	.	.	.
.	.	.	.
.	.	.	.

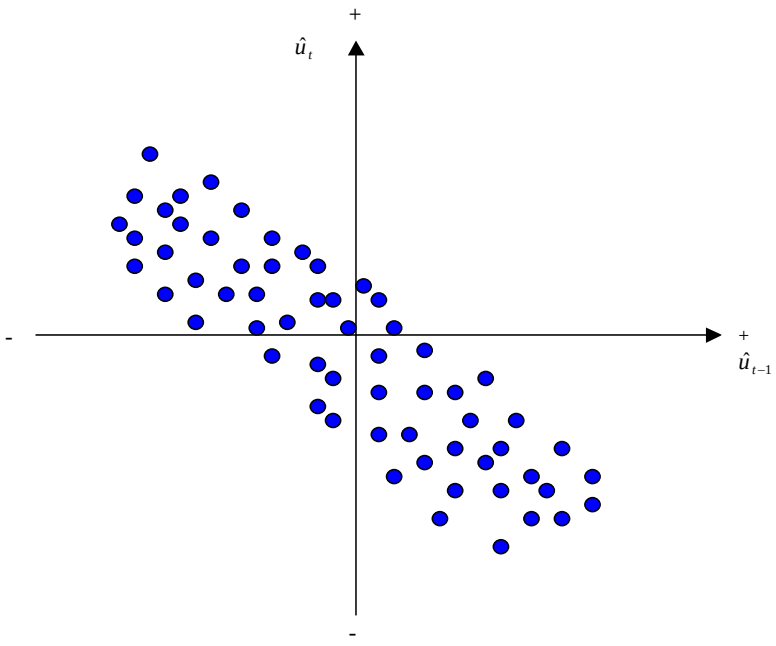
Stereotypical patterns : Positive Autocorrelation

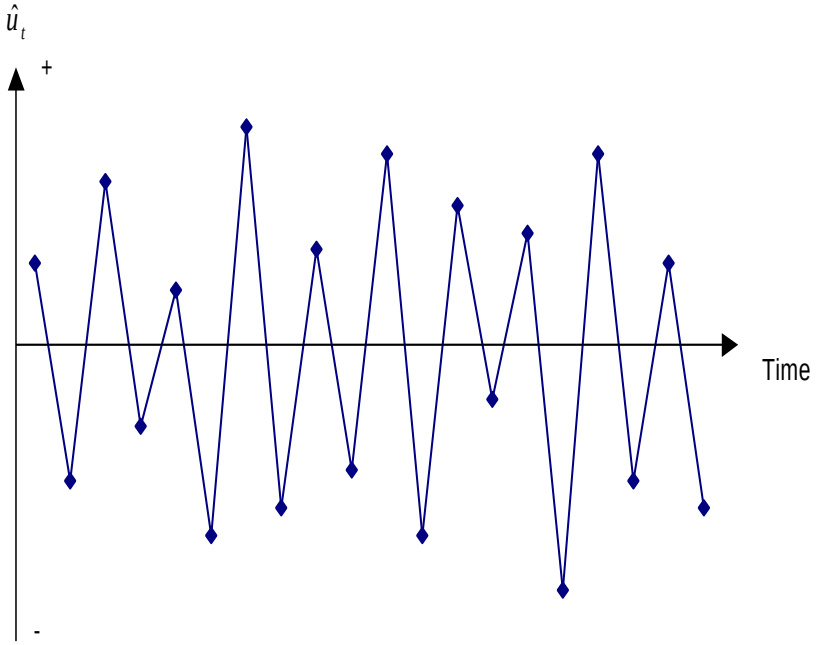




Positive Autocorrelation is indicated by a cyclical residual plot over time.

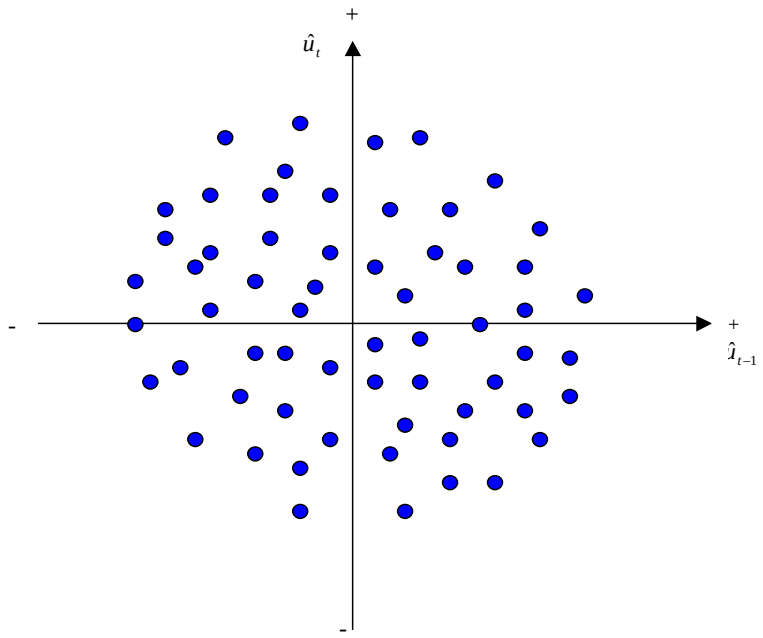
Stereotypical patterns : Negative Autocorrelation

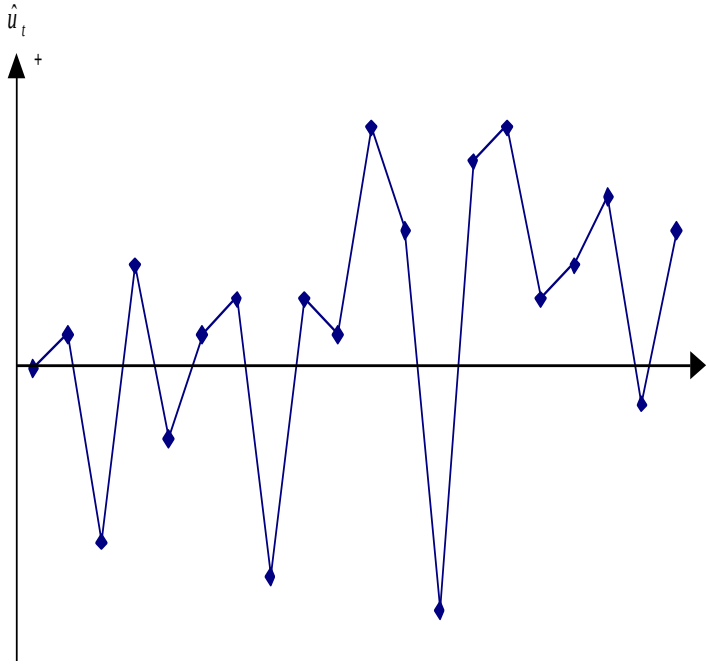




Negative autocorrelation is indicated by an alternating pattern where the residuals cross the time axis more frequently than if they were distributed randomly

No pattern in residuals – No autocorrelation



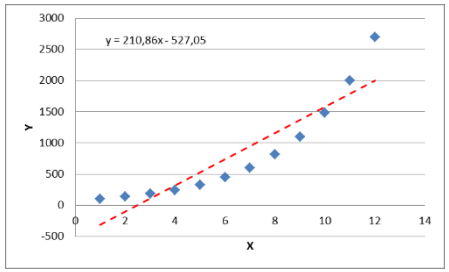


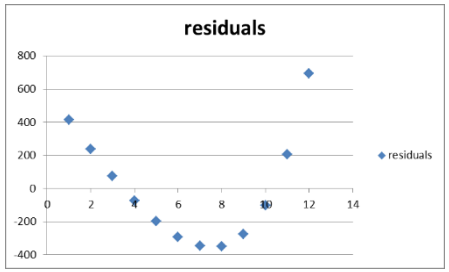
No pattern in residuals at all: this is what we would like to see

What causes autocorrelation?

- **Omitted variables**
  - ➔ Suppose that Y_t is related to $X_{2,t}$ and $X_{3,t}$ but that we do not include $X_{3,t}$ in our model.
  - ➔ The effect of $X_{3,t}$ will be captured by the disturbance U_t. If $X_{3,t}$ as many economic variables depends on $X_{3,t-1}$, $X_{3,t-2}$, … This will lead to unavoidable correlation among U_t, U_{t-1}, U_{t-2}, … and so on.

- **Misspecification in the model**





- **Non stationary variables** **(see Time Series analysis)**

Detecting Autocorrelation: The Durbin-Watson Test

The **Durbin-Watson (DW)** is a test for **first order autocorrelation** - i.e. it tests the relationship between an error and the previous one

$u_t = \rho u_{t-1} + v_t$ where $V_t \sim N(0, \sigma_v^2)$

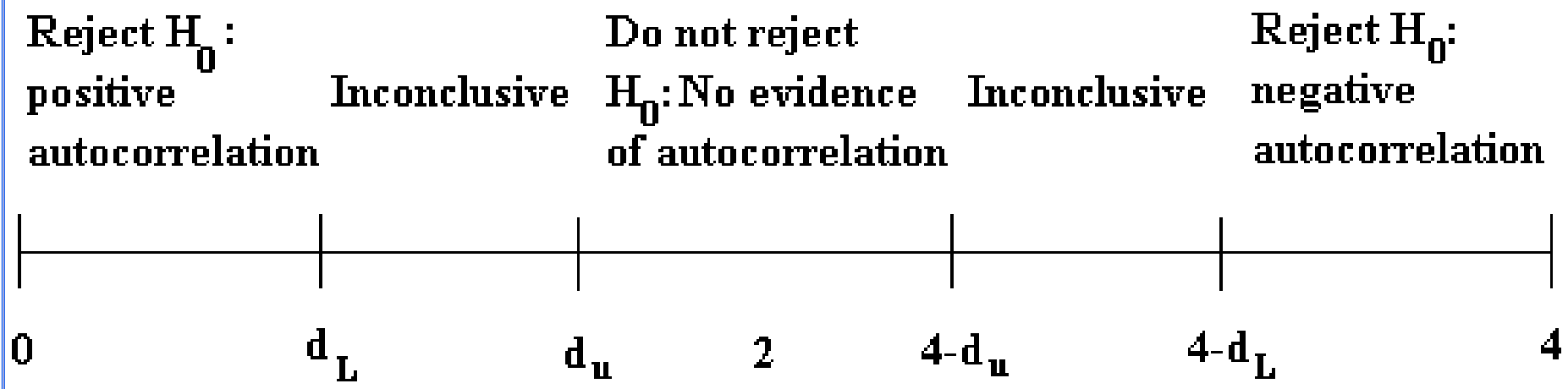
- The DW test statistic : $H_0 : \rho=0$ and $H_1 : \rho\neq0$
- The test statistic is calculated by

$$DW = \frac{\sum_{t=2}^{T} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=2}^{T} \hat{u}_t^2}$$

➔ $DW \approx 2(1-\hat{\rho})$, $-1 \leq \hat{\rho} \leq 1$, where $\hat{\rho}$ is the estimated correlation coefficient

- ➔ $0 \leq DW \leq 4$ If $\hat{\rho} = 0$, $DW = 2$
- ➔ do not reject the null hypothesis if DW is near 2 → i.e. there is little evidence of autocorrelation
- ➔ Refer to DW statistical tables for critical values
- ➔ Low (high) DW indicates positive (negative) autocorrelation

The Durbin-Watson Test: Interpreting the Results



DW has 2 critical values, an upper critical value (d_u) and a lower critical value (d_L), and there is also an intermediate region where we can neither reject nor not reject H_0.

Conditions which must be fulfilled for DW to be a Valid Test

1. Constant term in regression
2. Regressors are non-stochastic
3. No lags of dependent variable

TABLE de DURBIN-WATSON : Test unilatéral de $\rho = 0$ contre $\rho > 0$, au seuil de 5% (test bilatéral : seuil $\alpha = 10\%$)

	k' = 1		k' = 2		k' = 3		k' = 4		k' = 5		k' = 6		k' = 7		k' = 8		k' = 9		k' = 10	
n	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21	0,45	2,47	0,34	2,73	0,25	2,98	0,17	3,22	0,11	3,44
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15	0,50	2,40	0,40	2,62	0,30	2,86	0,22	3,09	0,15	3,30
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10	0,55	2,32	0,45	2,54	0,36	2,76	0,27	2,97	0,20	3,20
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06	0,60	2,26	0,50	2,46	0,41	2,67	0,32	2,87	0,24	3,07
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02	0,65	2,21	0,46	2,40	0,46	2,59	0,37	2,78	0,29	2,97
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99	0,69	2,16	0,60	2,34	0,50	2,52	0,42	2,70	0,34	2,88
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96	0,73	2,12	0,64	2,29	0,55	2,46	0,46	2,63	0,38	2,81
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94	0,77	2,09	0,68	2,25	0,59	2,41	0,50	2,57	0,42	2,73
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92	0,80	2,06	0,71	2,21	0,63	2,36	0,54	2,51	0,46	2,67
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90	0,84	2,03	0,75	2,17	0,67	2,32	0,58	2,46	0,51	2,61
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89	0,87	2,01	0,78	2,14	0,70	2,28	0,62	2,42	0,54	2,56
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88	0,90	1,99	0,82	2,12	0,73	2,25	0,66	2,38	0,58	2,51
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86	0,92	1,97	0,84	2,09	0,77	2,22	0,69	2,34	0,62	2,47
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	0,95	1,96	0,87	2,07	0,80	2,19	0,72	2,31	0,65	2,43
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84	0,97	1,94	0,90	2,05	0,83	2,16	0,75	2,28	0,68	2,40
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83	1,00	1,93	0,93	2,03	0,85	2,14	0,78	2,25	0,71	2,36
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83	1,02	1,92	0,95	2,02	0,88	2,12	0,81	2,23	0,74	2,33
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82	1,04	1,91	0,97	2,00	0,90	2,10	0,84	2,20	0,77	2,31
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81	1,06	1,90	0,99	1,99	0,93	2,08	0,86	2,18	0,79	2,28
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81	1,08	1,89	1,01	1,98	0,95	2,07	0,88	2,16	0,82	2,26
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	1,10	1,88	1,03	1,97	0,97	2,05	0,91	2,14	0,84	2,24
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80	1,11	1,88	1,05	1,96	0,99	2,04	0,93	2,13	0,87	2,22
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80	1,13	1,87	1,07	1,95	1,01	2,03	0,95	2,11	0,89	2,20
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79	1,15	1,86	1,09	1,94	1,03	2,02	0,97	2,10	0,91	2,18
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79	1,16	1,86	1,10	1,93	1,05	2,01	0,99	2,08	0,93	2,16
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,17	1,85	1,12	1,92	1,06	2,00	1,01	2,07	0,95	2,14
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,24	1,84	1,19	1,90	1,14	1,96	1,09	2,00	1,04	2,09
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,29	1,82	1,25	1,87	1,20	1,93	1,16	1,99	1,11	2,04
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77	1,33	1,81	1,29	1,86	1,25	1,91	1,21	1,96	1,17	2,01
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,37	1,81	1,33	1,85	1,30	1,89	1,26	1,94	1,22	1,98
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77	1,40	1,80	1,37	1,84	1,34	1,88	1,30	1,92	1,27	1,96
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77	1,43	1,80	1,40	1,84	1,37	1,87	1,34	1,91	1,30	1,95
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77	1,46	1,80	1,43	1,83	1,40	1,87	1,37	1,90	1,34	1,94
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,48	1,80	1,45	1,83	1,42	1,86	1,40	1,89	1,37	1,92
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77	1,50	1,80	1,47	1,83	1,45	1,86	1,42	1,89	1,40	1,92
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78	1,52	1,80	1,49	1,83	1,47	1,85	1,44	1,88	1,42	1,91
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78	1,54	1,80	1,51	1,83	1,49	1,85	1,46	1,88	1,44	1,90
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,55	1,80	1,53	1,83	1,51	1,85	1,48	1,87	1,46	1,90
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,66	1,80	1,65	1,82	1,64	1,83	1,62	1,85	1,60	1,86	1,59	1,88
200	1,73	1,78	1,75	1,79	1,73	1,80	1,73	1,81	1,72	1,82	1,71	1,83	1,70	1,84	1,69	1,85	1,68	1,86	1,66	1,87

K' is the number of explanatory variables excluding the constant

CAPM Microsoft / SP500

OLS (estimation default)

Dependent variable: ER_Microsoft
obs: 63

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.108327	0.645998	-0.167690	0.8674
ERSANDP	1.070463	0.183198	5.843195	0.0000

R-squared	0.358859	Mean dependent var	0.121478
Adjusted R-squared	0.348349	S.D. dependent var	6.339973
S.E. of regression	5.117937	Akaike info criterion	6.134611
Sum squared resid	1597.790	Schwarz criterion	6.202647
Log likelihood	-191.2403	Hannan-Quinn criter.	6.161370
F-statistic	34.14293	Durbin-Watson stat	2.208231
Prob(F-statistic)	0.000000		

For n=63 obs and k=1,
$[d_l; d_u]$ is = [1,55;1,62]

Question 12: Residuals

A-are autocorrelated

B-are not autocorrelated

C-I have no enough information to answer

Another Test for Autocorrelation: The Breusch-Godfrey Test

- More general test for r^{th} order autocorrelation:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \rho_3 u_{t-3} + ... + \rho_r u_{t-r} + v_t \quad , \quad v_t \sim N(0, \sigma_v^2)$$

- The hypotheses :

$H_0 : \rho_1 = 0$ and $\rho_2 = 0$ and ... and $\rho_r = 0$

$H_1 : \rho_1 \neq 0$ or $\rho_2 \neq 0$ or ... or $\rho_r \neq 0$

- The test :

1. Estimate the linear regression using OLS and obtain the residuals, $\hat{u}_t$
2. Regress $\hat{u}_t$ on all of the regressors from stage 1 (the *x*'s) plus $\hat{u}_{t-1}, \hat{u}_{t-2}, ..., \hat{u}_{t-r}$. Obtain R^2 from this regression.

- Test statistic : $(T-r)R^2 \sim \chi^2(r)$
- Decision rule :

$(T-r)R^2 > \chi^2_{\alpha,r}$ ➔ reject the null hypothesis that there is no autocorrelation (or pvalue <5%)

Consequences of Using OLS in the Presence of Heteroscedasticity and/or autocorrelation

- The coefficient estimates are still **unbiased**
- The associated standard errors are wrong➔ **inferences misleading** because the t-statistic doesn't hold anymore

$$t\text{-statistic}(\hat{\beta}_i) = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Calculated under the hypothesis of homoscedasticity and no autocorrelation

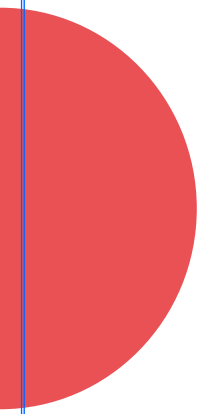
$SE(\hat{\beta})$ is understated
t-statistic is too high and we reject too easily H0

- R^2 likely to be inflated

How Do we Deal with Heteroscedasticity and/or autocorrelation

- Use a specific GLS (generalized least square) procedure
- Transform the variables into logs or reducing by some other measure of "size".
- Use the Cochrane-Orcutt procedure for autocorrelated errors.
- Use **White's heteroscedasticity consistent standard error estimates for** heteroscedastic but serially uncorrelated.
- Use the Newey and West estimator, consistent with both heteroscedasticity and autocorrelation.

Effect of using corrections ➔ in general the standard errors for the slope coefficients are increased relative to the usual OLS standard errors. This makes that we are more "conservative" in hypothesis testing (H0 less easily rejected).



Other problems dealing with CLRM

Assumption 5: Cov (Ut,Xt)=0

All independent variables are uncorrelated with the error term.

Violations: $E(X_{it}u_t) \neq 0$ ➔ Endogeneity of X

➔The coefficient estimates are **biased** and **inconsistent**

Causes:

- ➔ Relevant explanatory variables may be poorly measured
- ➔ Omitted variable
- ➔ Simultaneity => use instrumental variable (IV) and 2SLS to deal with

Parameter Stability

Estimated regressions : $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + U_t$

- **Implicitly assumed that the parameters** ($\beta 1$, $\beta 2$ and $\beta 3$) are **constant** for the entire sample period.

- Test this implicit assumption using parameter stability tests

  H0 : Parameters are constant

➔ **Chow test (analysis of variance test)**

1. Split the data into two sub-period
2. Estimate the regression over the whole period and then for the two sub-periods separately (3 regressions)
3. Obtain the RSS (residuals sum of squares) for each regression
4. Compare the RSS of the whole period regressions with the sum of the 2 sub-periods to construct the statistics
5. Statistics is ~F(k,T-$2k$)
6. Decision rule : If F>Fα(k,T-2k) or pvalue < 5% then reject H0 that parameters stable over time.

Multicollinearity

Multicollinearity : **two or more predictor** variables in a multiple regression model are **highly correlated**, meaning that **one can be linearly predicted from the others**

- Perfect multicollinearity => **Cannot estimate all the coefficients**
- High collinearity

Corr	x_2	x_3	x_4
x_2	-	0.2	0.8
x_3	0.2	-	0.3
x_4	0.8	0.3	-

Measure: Variance Inflation Factor (VIF)

- VIFs ➔ how much of the variance of a coefficient estimate of a regressor has been inflated due to collinearity with the other regressors.

The centered $\mathrm{VIF} = \frac{1}{1-R^2}$

where R^2 is the R^2 from the regression of that regressor on all of the other regressors in the equation.

➔ Multicollinearity if VIF >10

Multicollinearity: Consequences and solutions

Problems if multicollinearity is present but ignored

- The ordinary least-squares estimator does not exist (Predictor matrix is singular and therefore cannot be inverted)
- R^2 high but individual coefficients will have high standard errors.
- Regression becomes very sensitive to small changes in the specification.
- Standard errors for the parameters very high, and significance tests might therefore give inappropriate conclusions.

Solutions

- "Traditional" approaches (e.g. principal component analysis on Xi)
- Some econometricians argue that if the model is otherwise OK, just ignore it
- The easiest ways to "cure" the problems are:
  - ➔ drop one of the collinear variables
  - ➔ transform the highly correlated variables into a ratio
  - ➔ collect more data: longer period or higher frequency

TUTORIAL XLSTAT

5. Check model assumptions

- Normality
- Homoscedasticity
- No Autocorrelation

Tutorial

Based on the regression: $ER_{msoft,t} = \alpha_{msoft} + \beta_{msoft} (ER_{s\&p,t}) + U_t$

- Obtain the residual series
- Plot the residuals over time

- Check for normality :
  - ➔ Histogram
  - ➔ Descriptive statistics
  - ➔ Normality test

- Are the residuals normally distributed?

Tutorial

Based on the regression: $ER_{msoft,t} = \alpha_{msoft} + \beta_{msoft} (ER_{s\&p,t}) + U_t$

- Check for homoscedasticity and no autocorrelation

- Are the residuals homoscedastic ?
- Are the residuals non autocorrelated ?

- If autocorrelation/heteroscedasticity, use appropriate correction

Regression: Global methodology

- Define the variables of interest, based on some theory : Y, X_1, X_2, ..X_p
- Global reliability of the model
- Calculation of model coefficients
- Reliability of each model coefficient
- Goodness of fit
- Assumptions to be checked on residuals of the model
- Conclusion

To validate a model, it should be logically plausible, consistent with underlying financial theory, parsimonious and satisfy the hypothesis on residuals

Econometrics & Financial Markets

Time series analysis

Toulouse Business School
MSc BIF

Anna CALAMIA
a.calamia@tbs-education.fr

Time series data and analysis

Times series data

- Series of data points ordered in time
- E.g. series of index values or stock prices or returns over time (Y_t)
- Multivariate and univariate analysis
- Frequency: daily, weekly, monthly, quarterly, annual…

Date	MSFT_ExR	SP500_ExR
01/01/2020	0.07820	-0.00289
01/02/2020	-0.04931	-0.08514
01/03/2020	-0.02391	-0.12514
01/04/2020	0.13625	0.12677
01/05/2020	0.02244	0.04518
01/06/2020	0.11354	0.01828
01/07/2020	0.00730	0.05503
01/08/2020	0.10001	0.06999
01/09/2020	-0.06521	-0.03930
01/10/2020	-0.03744	-0.02773
01/11/2020	0.05723	0.10748
01/12/2020	0.04167	0.03707
01/01/2021	0.04285	-0.01118
01/02/2021	0.00563	0.05937

Linear regression (e.g. CAPM regression) of Y_t on X_t

- Stationarity assumption
- Non stationary TS ➡ possible spurious regression

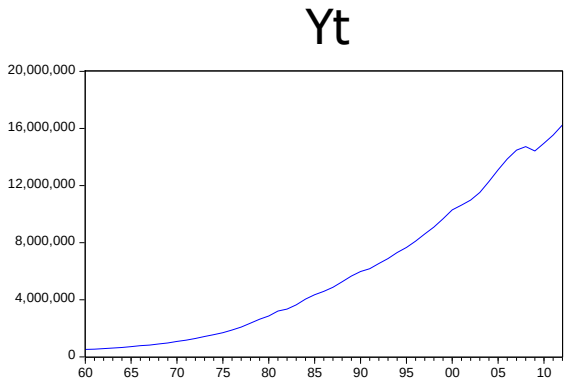
Time series analysis: Outline

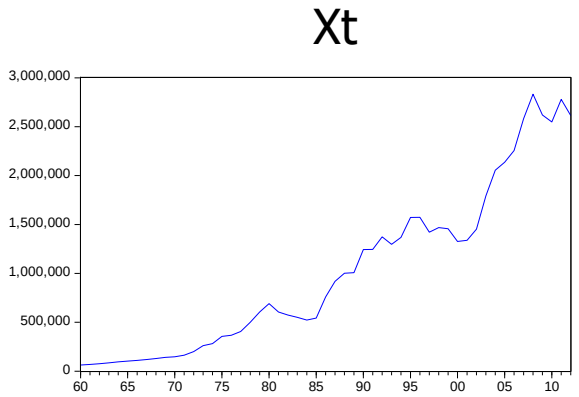
- Spurious Regression
- Stationarity Definition
- From non-stationarity to … Stationarity
- Induce Stationarity
- Modelling and forecasting stationary time series

Spurious Regression

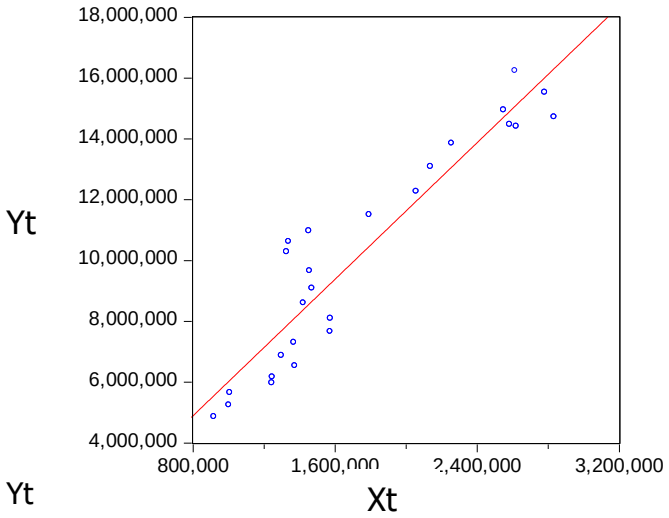
Example of Spurious Regression

- Example : regression of (Yt) to (Xt) and a constant





Comment???



Dependent Variable: Yt
Method: Least Squares
Date: 04/03/14 Time: 15:55
Sample: 1960 2012
Included observations: 53

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	93292.53	204437.6	0.456338	0.6501
Xt	5.764271	0.155177	37.14646	0.0000

R-squared	0.964357	Mean dependent var	5944537.
Adjusted R-squared	0.963658	S.D. dependent var	4976665.
S.E. of regression	948727.8	Akaike info criterion	30.40064
Sum squared resid	4.59E+13	Schwarz criterion	30.47499
Log likelihood	-803.6169	Hannan-Quinn criter.	30.42923
F-statistic	1379.859	Durbin-Watson stat	0.464692
Prob(F-statistic)	0.000000		

Why does stationarity matter?

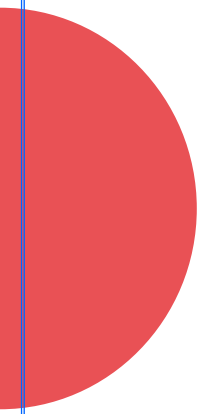
If two variables are trending over time, a regression of one on the other could have a high R^2 even if the two are unrelated

- ➔ consider 2 series Xt and Yt having a similar trend and regress Xt on Yt. Applying the standard asymptotic properties of estimators, you will find that the conjunctions of Mars and Saturn is a powerful predictor of excess returns on NYSE (Novy-Marx 2014). What is the economic justification ?

If variables in the regression model not stationary:

- ➔ usual "*t*-ratios" will not follow a *t*-distribution
- ➔ cannot validly undertake hypothesis tests about the regression parameters.
- ➔ very high R^2 *and* *t*-statistic, but the results may have no economic meaning
- ➔ high level of residuals autocorrelation (DW very low)

➔ **Stationary processes: no spurious regression**



Stationarity Definition

Stationarity

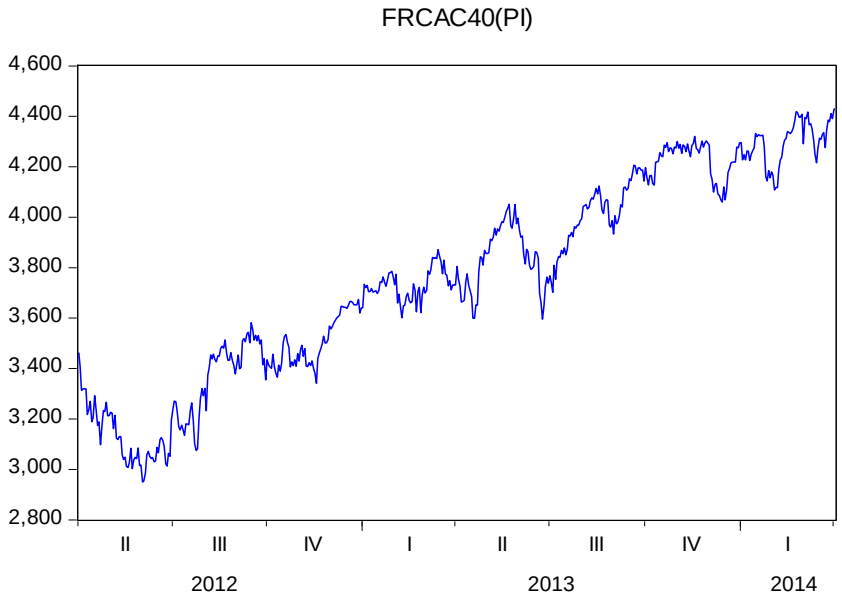
Y_t (with mean μ and variance σ^2) is a Weakly Stationary Process

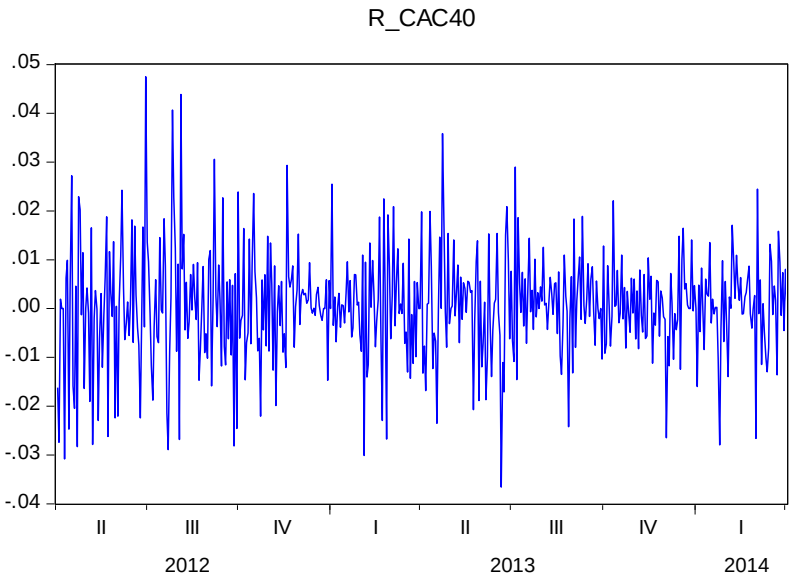
1. $E(Y_t) = \mu$,for all t ⟶ *Mean is constant over time*
2. $Var(Y_t) = \sigma^2$ ⟶ *Variance is constant over time*
3. $Cov(Y_t, Y_s) = \gamma(t-s), t \neq s$ ⟶ *Autocovariances do not depend on time, but only on the difference (t-s)*

Notation : $\gamma(0) = \sigma^2$

Weakly stationary (or covariance stationary) processes have no trend in mean, and no trend in variance, but it does not mean that they have a stable graphic …

Stationarity





Question 13: Which process seems stationary?

A- The series of values of CAC40 because it is mean stationary

B- The series of returns on CAC40 because it is mean stationary

C- The series of values of CAC40 because there is a positive trend

D- The series of returns on CAC40 because it's time trending

Autocorrelation Function

➔ use the **autocorrelations** $\tau(s)$:

$$\tau(s) = \frac{\gamma(s)}{\gamma(0)}$$

s=0,1,2,... and $\gamma(s)$=cov(Yt,Ts)

- Autocorrelation of time series at various lags: Plot $\tau(s)$ against s=0,1,2,...

➔**autocorrelation function (autocorrelogram)**

$$\hat{\tau}(s) = \frac{\sum_{t=s+1}^{T}(Y_t - \bar{Y})(Y_{t-s} - \bar{Y})}{\sum_{t=1}^{T}(Y_t - \bar{Y})^2}, 0 \le s \le T-1$$

- ***Partial Autocorrelation Function***, $\rho(k)$ *is the coefficient of Yt-k in the regression of Yt on Yt-1, Yt-2, …, Yt-k*

  ➔ *measures the correlation between Yt and Yt-k after removing the effects of Yt-k+1 , Yt-k+2 , …, Yt-1 (conditional correlations)*

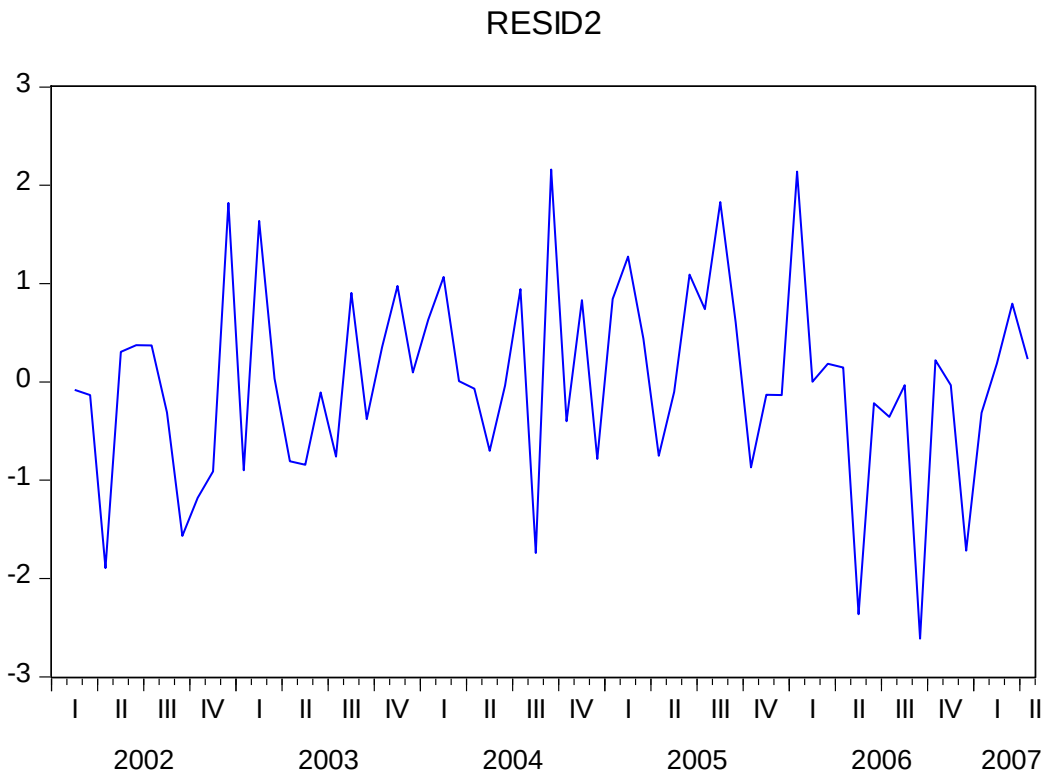
A particular stationary variable : White Noise Process

no discernible structure

$$E(Y_t) = 0$$

$$Var(Y_t) = \sigma^2 \qquad \text{for all t}$$

$$Cov(Y_t, Y_r) = 0 \qquad \text{for } t \neq r$$



Significance tests for the autocorrelation coefficients

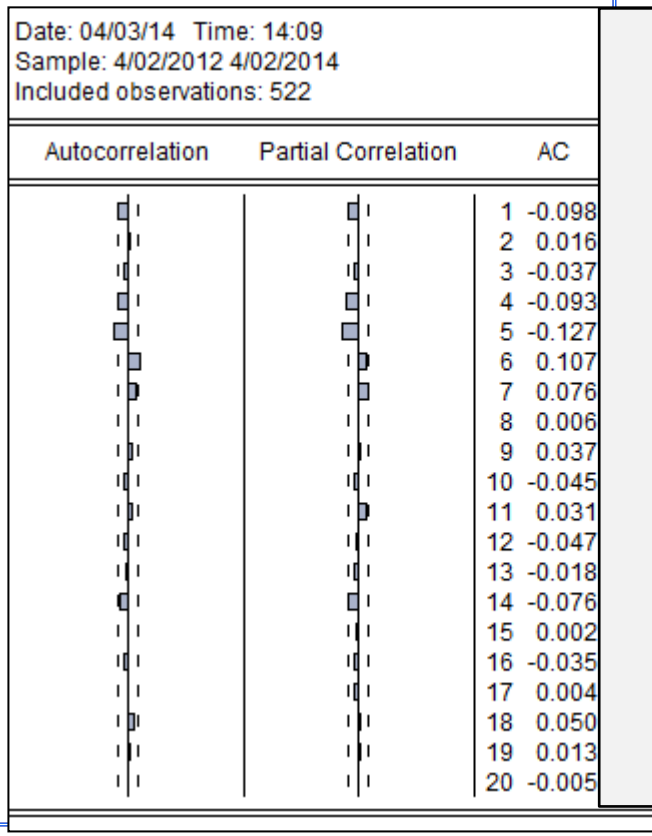
Significance tests for the autocorrelation coefficients at lag s, $\tau(s)$:

1. Compute the t-test of the null hypothesis, H0: $\tau(s)=0$
   - Under HO: $\hat{\tau}(s) \sim$ approximately N(0,1/T)
   - ➔ t-test: $\tau(s)/(1/\sqrt{T})$

     Where $1/\sqrt{T}$ is the standard error of $\tau(s)$ and

     T is the number of observations in the time series
   - Reject if t-test larger >1.96, in absolute value (5% critical value)

2. Equivalently, compute the 95% confidence interval as: $0 \pm 1.96 \times \frac{1}{\sqrt{T}}$ (reject if outside the interval)



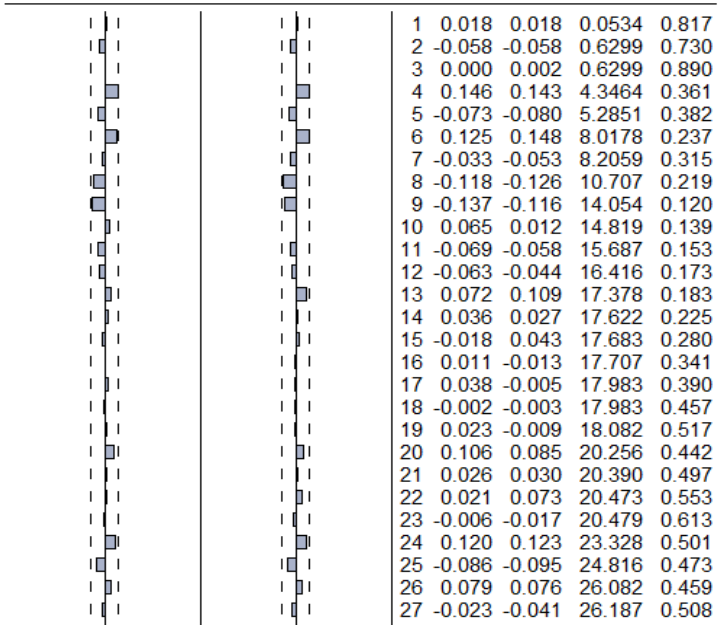
❖ Question 14: is $\tau(1)$ significantly different from 0?

BOX-PIERCE / LJUNG-BOX Q test

➔ **BOX-PIERCE / LJUNG-BOX Q** tests the **joint hypothesis that all correlation coefficients are simultaneously equal to zero.**

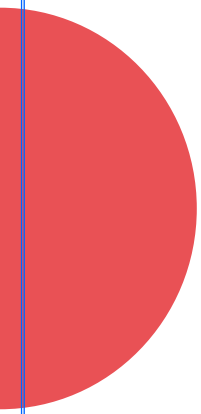
Reject H0 if pvalue <5%: one coefficient is significantly different from zero (the process cannot be approximated by a white noise)

Date: 04/19/17 Time: 21:37
Sample: 1960Q1 2002Q1
Included observations: 168



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.018	0.018	0.0534	0.817
		2	-0.058	-0.058	0.6299	0.730
		3	0.000	0.002	0.6299	0.890
		4	0.146	0.143	4.3464	0.361
		5	-0.073	-0.080	5.2851	0.382
		6	0.125	0.148	8.0178	0.237
		7	-0.033	-0.053	8.2059	0.315
		8	-0.118	-0.126	10.707	0.219
		9	-0.137	-0.116	14.054	0.120
		10	0.065	0.012	14.819	0.139
		11	-0.069	-0.058	15.687	0.153
		12	-0.063	-0.044	16.416	0.173
		13	0.072	0.109	17.378	0.183
		14	0.036	0.027	17.622	0.225
		15	-0.018	0.043	17.683	0.280
		16	0.011	-0.013	17.707	0.341
		17	0.038	-0.005	17.983	0.390
		18	-0.002	-0.003	17.983	0.457
		19	0.023	-0.009	18.082	0.517
		20	0.106	0.085	20.256	0.442
		21	0.026	0.030	20.390	0.497
		22	0.021	0.073	20.473	0.553
		23	-0.006	-0.017	20.479	0.613
		24	0.120	0.123	23.328	0.501
		25	-0.086	-0.095	24.816	0.473
		26	0.079	0.076	26.082	0.459
		27	-0.023	-0.041	26.187	0.508

Question 15: All the pvalues >5%, which implies that ….
A- none of the coefficients are significant
B- all the coefficients are significant
C- the process is non stationary
D- The process is non gaussian



From non-stationarity to … Stationarity

3 types of Non-Stationarity

Various illustrations of non-stationarity :

- (1) the **random walk model with drift**
  (Difference Stationary with drift):

$$Y_t = \mu + Y_{t-1} + U_t \quad , U_t \text{ WN} \quad (1)$$

- (2) the **random walk model without drift**
  ($\mu = 0$, DS without drift) :

$$Y_t = Y_{t-1} + U_t \quad , U_t \text{ WN} \quad (2)$$

- (3) the **deterministic trend process** (Trend stationary):

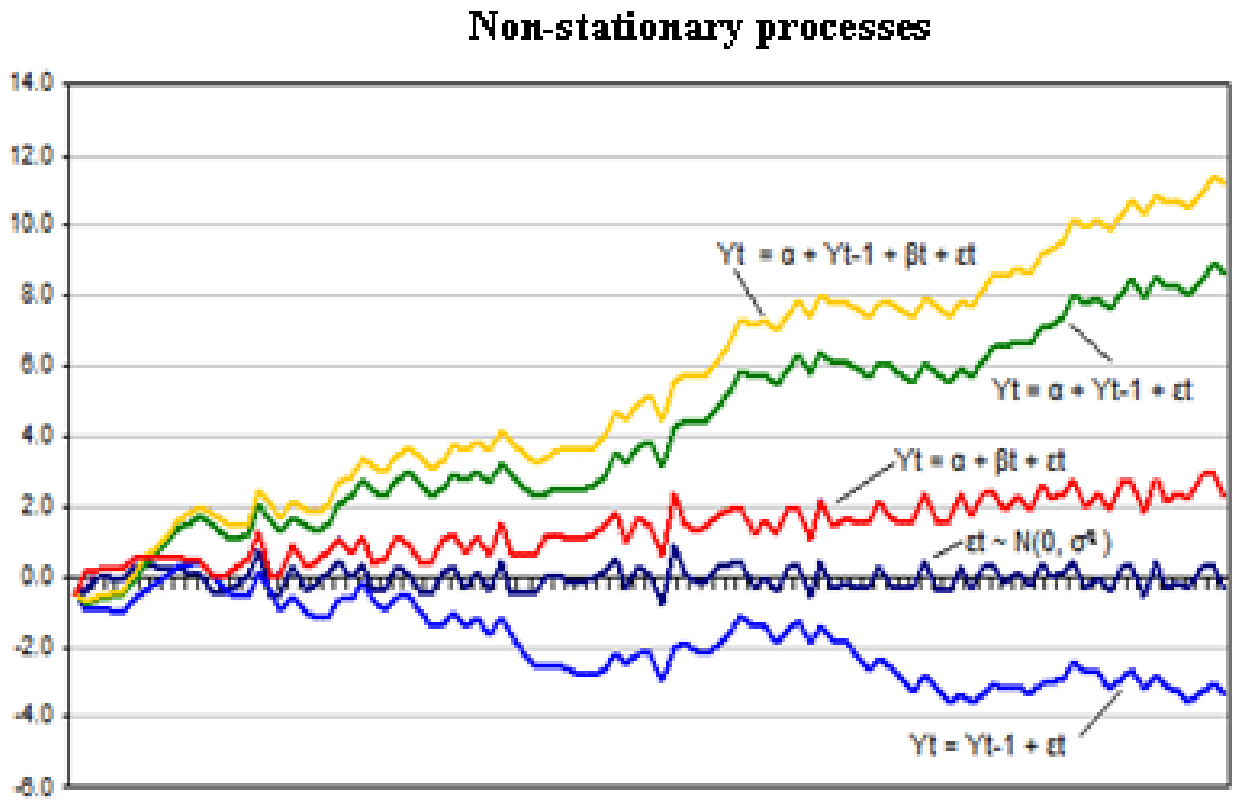
$$Y_t = \alpha + \beta t + U_t \quad , U_t \text{ WN} \quad (3)$$

Variable at date t depends on the value of the previous period : shock at a period will have permanent rather than transitory effects (shocks persist in the system)

STOCHASTIC NON STATIONARITY

- The series moves linearly in time
- ut is stationary : shocks have no impact on the later evolution, the series always returns to its long term trend

Non-Stationarity



Non-stationarity: study of shocks

- Generalization: consider the process defined by

$$Y_t = \phi Y_{t-1} + U_t \text{ , with } \phi \text{ a generic parameter}$$

- By T successive substitutions (of Yt-1…) we get:

$$Y_T = U_T + \phi U_{T-1} + \phi^2 U_{T-2} + \phi^3 U_{T-3} + \ldots + \phi^T y_0$$

1. $\phi < 1 \Rightarrow \phi^T \to 0$ as $T \to \infty$

   shocks to the system gradually die away➔ **stationarity**

2. $\phi = 1 \Rightarrow \phi^T = 1 \forall\ T$

   shocks persist in the system ➔ **random walk, non stationarity**

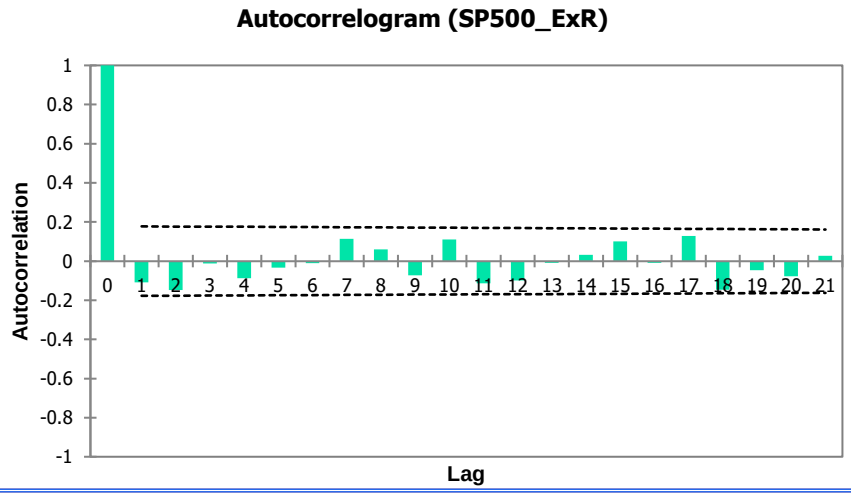
3. $\phi > 1$ shocks propagate and become more influential as time goes on➔ **explosive case, non stationarity**

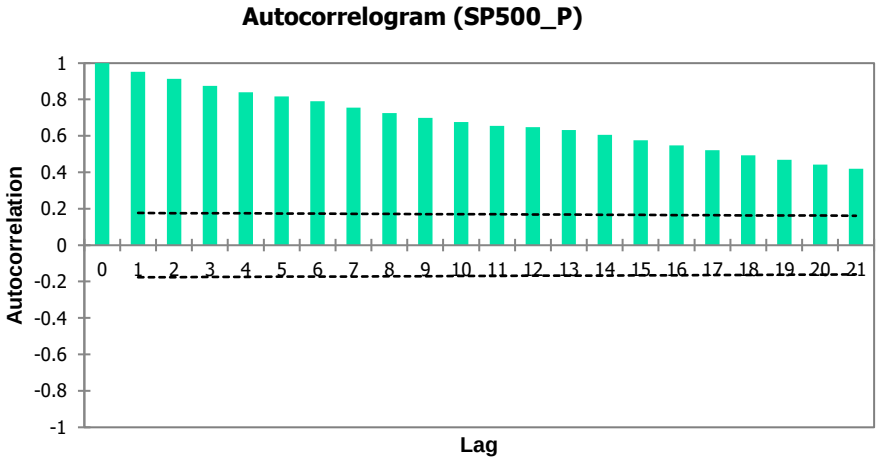
   ➔ Explosive case does not describe many data series in economics and finance.

we use $\phi = 1$ to characterise the non-stationarity

Test for non stationarity

- Autocorrelogram :
  - ➔ For a stationary time series, either autocorrelations at all lags are statistically indistinguishable from zero, or the autocorrelations drop off rapidly to zero as the number of lags becomes large.
  - ➔ The autocorrelation function of a nonstationary process decreases very slowly even at very high lags.





Test for non stationarity

- Unit root test: If $\phi = 1 \Rightarrow$ the series has a unit root, it is a random walk and is not covariance stationary.

  - Dickey Fuller test based on a transformed version of the model:

$$Y_t = \mu + \phi Y_{t-1} + U_t$$
$$Y_t - Y_{t-1} = \mu + (\phi - 1) Y_{t-1} + U_t$$

  - The null hypothesis of the Dickey-Fuller test is H0: $\phi - 1 = 0$ and the alternative hypothesis is Ha: $\phi - 1 < 0$ (stationary).
  - Specific (larger) critical values (Dickey Fuller statistical tables)
  - Possible to add an intercept and a deterministic trend
  - If Yt serially correlated, may include lags of Yt (Augmented DF test)

Induce Stationarity

Induce Stationarity for Random Walk Process : Difference-Stationary series

Random walk with (or without) drift:

$$Y_t = \mu + Y_{t-1} + U_t, \qquad U_t \sim WN \qquad (1)$$

If we take (1) and subtract Y_{t-1} from both sides:

$$Y_t - Y_{t-1} = \mu + U_t$$

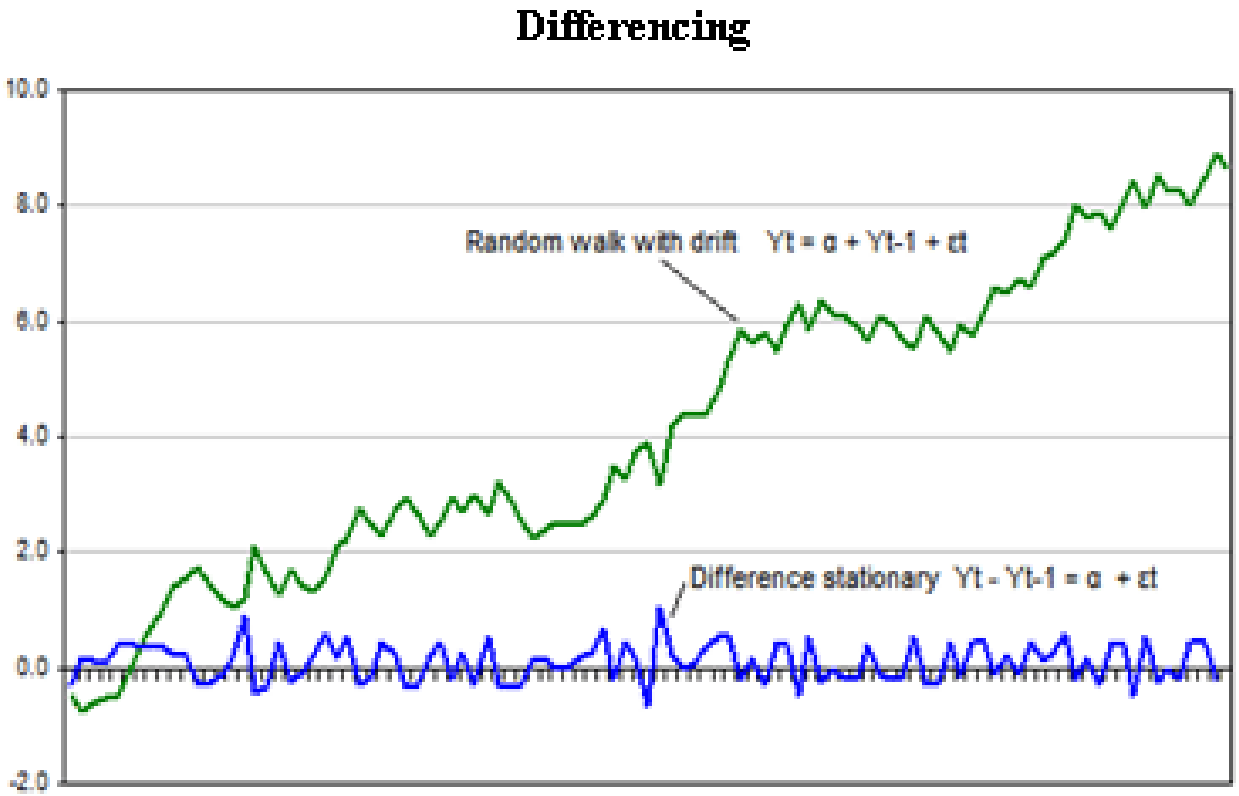
$$\Delta Y_t = \mu + U_t$$

We say that we have induced stationarity by "differencing once".

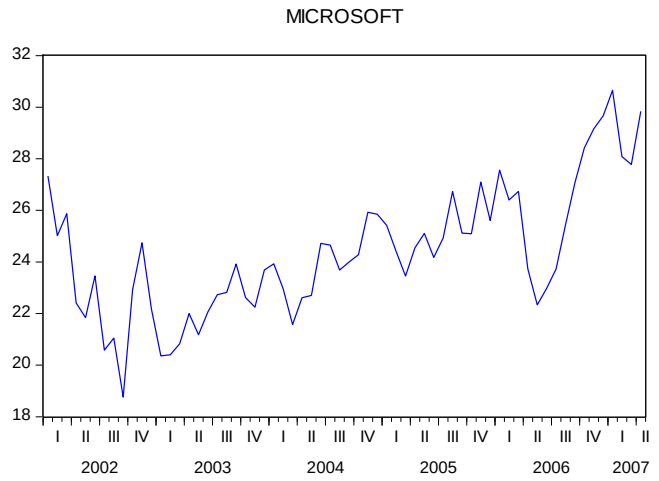
- A series is integrated of order 1 ($Y_t \sim I(1)$) if **Y_t is non-stationary but ΔY_t is stationary** (The series contains a unit-root)
- A series is integrated of order d ($Y_t \sim I(d)$) if **Y_t is non-stationary but $\Delta^d Y_t$ is stationary** (The series contains d unit-root)

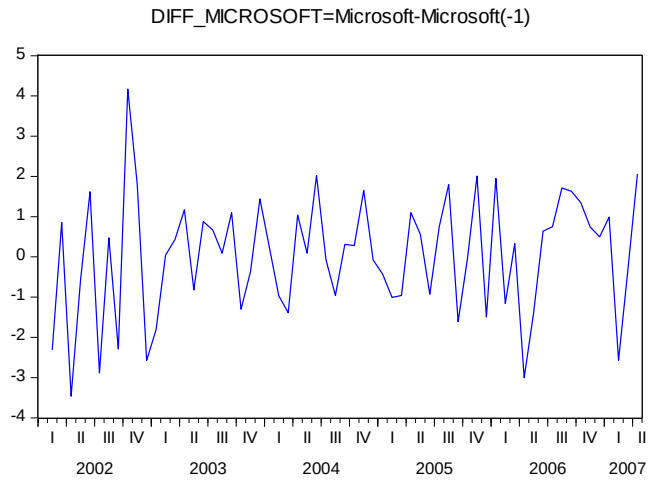
Most economic and financial series contain a single unit root

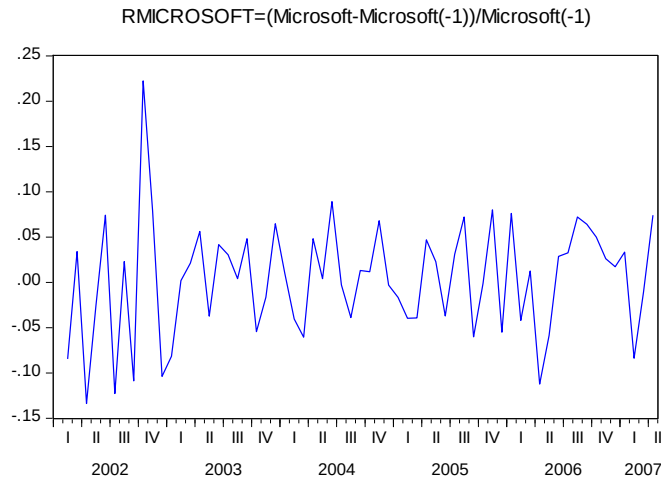
Induce Stationarity for Random Walk Process : Difference-Stationary series

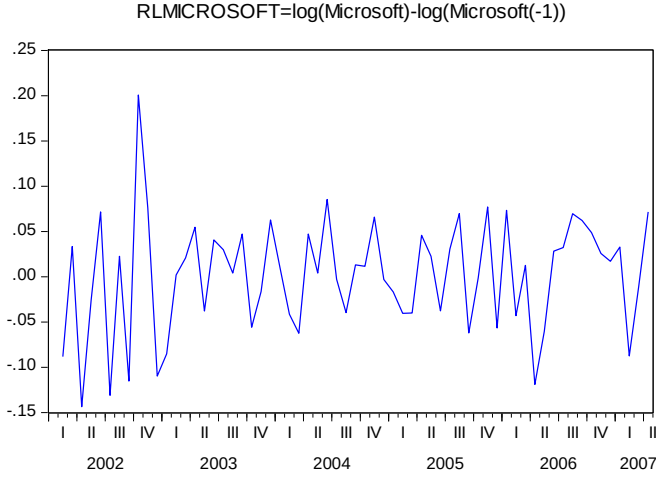


Example Price/Return









Induce Stationarity for a Deterministic Trend Process: Detrending

The trend-stationary process

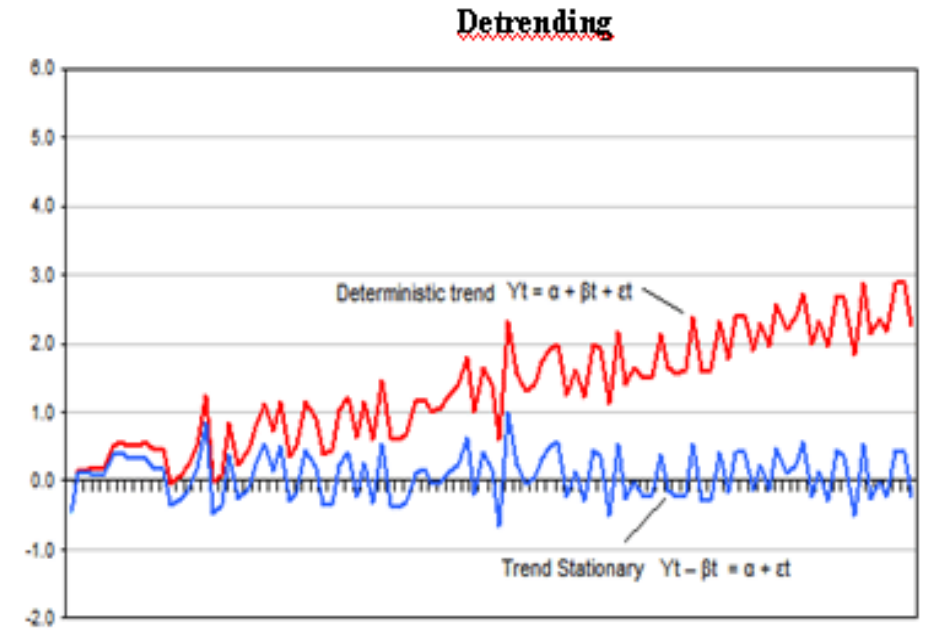
$$Y_t = \alpha + \beta t + U_t$$

➔ **deterministic non-stationarity**

- Subtracting the trend βt:

=> Y_t - βt = α + U_t, is stationary

Detrending : run a regression of the form Yt=α+βt+Ut and fit a model on the residuals from Yt-βt=α+Ut (from which the linear trend has been removed)



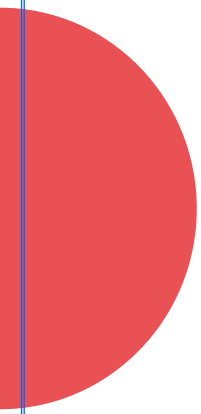
Seasonality: regular patterns of movement within the year => include seasonal lags in an autoregressive model. Then, Y_t is non stationary but Y_{t-s} is stationary (quarterly (s=4) or annual(s=12) seasonality)

TUTORIAL XLSTAT

6. Time series: Test for stationarity

Tutorial

- XLSTAT – Time series analysis
- Plot the series as a function of time
  - ➔ Prices and returns
- Autocorrelation function and partial autocorrelation function
- Are the return series stationary? What about the price series?



Modelling stationary time series

Univariate Time Series Modelling and Forecasting

- Time-series models: to explain the past and to predict the future of a time series, Yt (stock price, return)
- Predicting or forecasting the future behaviour of financial variables :
  - ➔ Times series models use the information in past values of the same variable
  - ➔ Alternatively, regression models based on hypothesized causal relationships with other variables
- Objective:
  - ➔ We observe values of Yt, for t=1, … T.
  - ➔ We want to model this series and forecast futures values (T+1, T+2…) given its past values
    - One period ahead forecast
    - Multi period forecast (chain-rule)

Classical models

- Different models to represent univariate time series
- Classical TS models:

  Assumptions:
  - (Yt)t is a (weakly) stationary process
  - Ut is a white noise WN(0,σ^2)

  - Autoregressive model: Yt is explained by its own past values
  - Moving average: Yt explained by a moving average of current and past WN errors
  - ARMA(p,q): Generalization of the first two models
  - ARIMA Model: non stationary processes
  - SARIMA(p,d,q): processes with seasonality s
  - ARCH, GARCH: Autoregressive conditional heteroscedasticity

Classical models

- ***Autoregressive model of order p, AR(p)***

$Y_t = \mu + \phi_1 Y_{t-1} + U_t$ — *AR(1)*

$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + U_t$ — *AR(2)*

$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + U_t$ — *AR(p)*

- ***Moving average of order q, MA(q)***

$Y_t = \mu + U_t + \theta_1 U_{t-1}$ — *MA(1)*

$Y_t = \mu + U_t + \theta_1 U_{t-1} + \theta_2 U_{t-2}$ — *MA(2)*

$Y_t = \mu + U_t + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \ldots + \theta_q U_{t-q}$ — *MA(q)*

- ***ARMA(p,q)***

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + U_t + \theta_1 U_{t-1} + \theta_2 U_{t-2} + \ldots + \theta_q U_{t-q}$$

- **ARIMA(p,d,q) and SARIMA(pdq)**

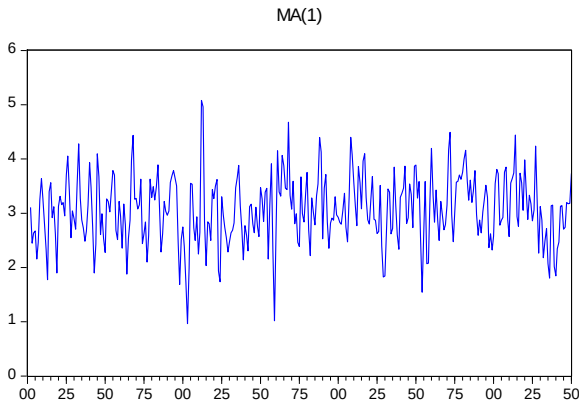
Y_t is I(1) if $\Delta Y_t = Y_t - Y_{t-1}$ is stationary

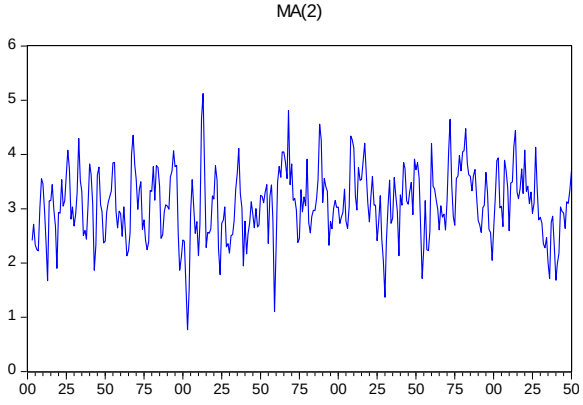
➔ Y_t is an ARIMA(p,1,q) if ΔY_t is an ARMA(p,q) process

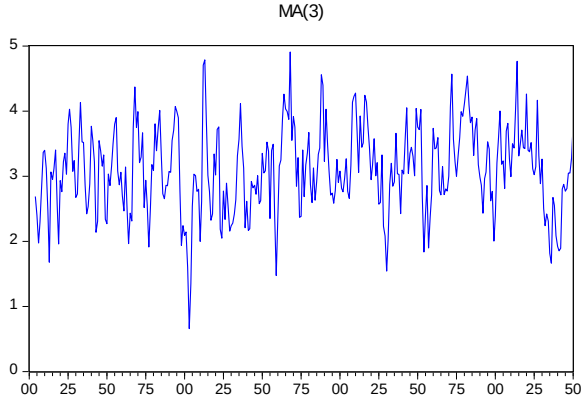
➔ Y_t is an ARIMA(p,d,q) if $\Delta^d Y_t$ is an ARMA(p,q) process

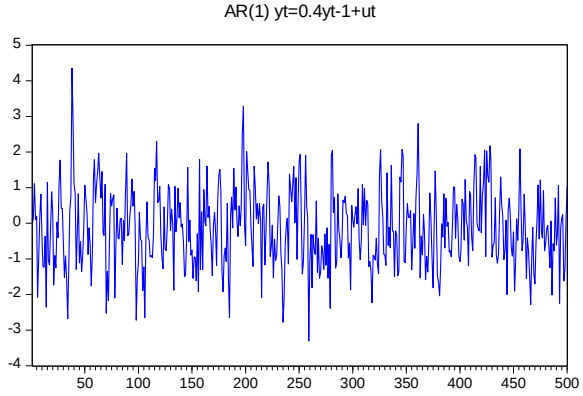
Y_t is a SARIMA(p,d,q) process with seasonality s if Y_t is nonstationary but $Y_t - Y_{t-s}$ is stationary

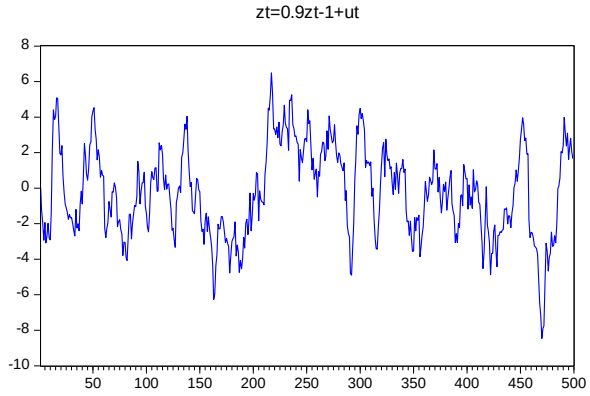
Example of MA and AR Processes

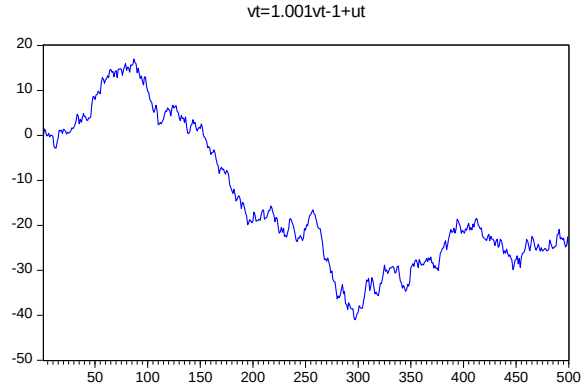












Classical models

How to choose among these models?

How to select the parameters p, q?

Use the properties of the autocorrelation function to identify AR and MA processes and the order p,q.

ACF and PACF

- *If Yt is stationary AR(p), The autocorrelation function decays exponentially to zero (autocorrelations start large and decline gradually) while the PACF drops to zero after p:*
  - ➔ $\tau(k)$ ➔ *0* at exponential decay
  - ➔ $\rho(k)=0$ *for* $k>p$
  - ➔ $\rho(p) \neq 0$

- *If Yt is MA(q), it is weekly stationary and the autocorrelations drop to zero after first q autocorrelations, while The PACF decays to zero at an exponential decay:*
  - ➔ $\tau(k)=0$ *for* $k>q$
  - ➔ $\tau(q)\neq 0$
  - ➔ $\rho(k)$ ➔ *0* at exponential decay

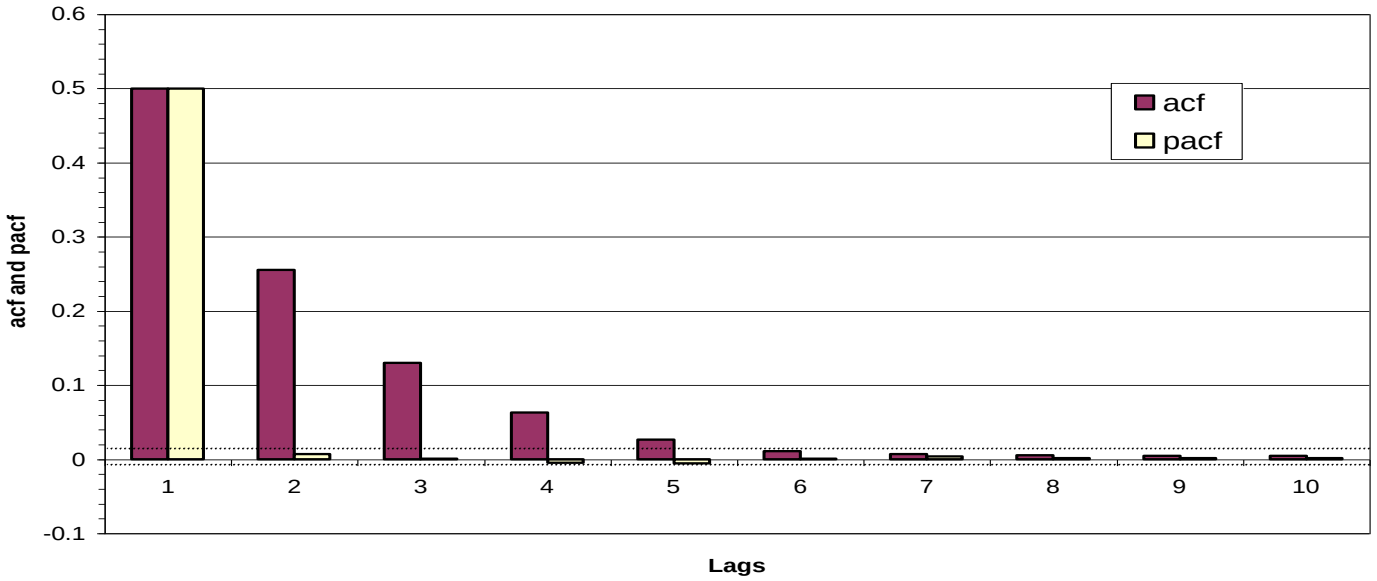
- *For a stationary (and invertible) ARMA process :*
  - ➔ *both acf and pacf are geometrically decaying*

Exemples of AR(1) Process

Question 16: Is $Y_t = 0.5Y_{t-1} + U_t$ stationary?
A- Yes because $\phi = 0.5$ is strictly less than 1
B- No because Y_t is an AR(1) which is always non stationary
C-Yes because Y_t is an AR(1) which is always stationary
D- No because Yt has a unit root

Question 17: Is $Y_t = 1.0Y_{t-1} + U_t$ stationary?
A- Yes because $\phi = 1$ is strictly larger than 1
B- No because Y_t is an AR(1) which is always non stationary
C-Yes because Y_t is an AR(1) which is always stationary
D- No because Yt is a Random Walk

Sample acf and pacf plots for standard processes



Question 18: Which model could fit the best?

A- an AR(5)

B- an AR(1)

C- a white moise

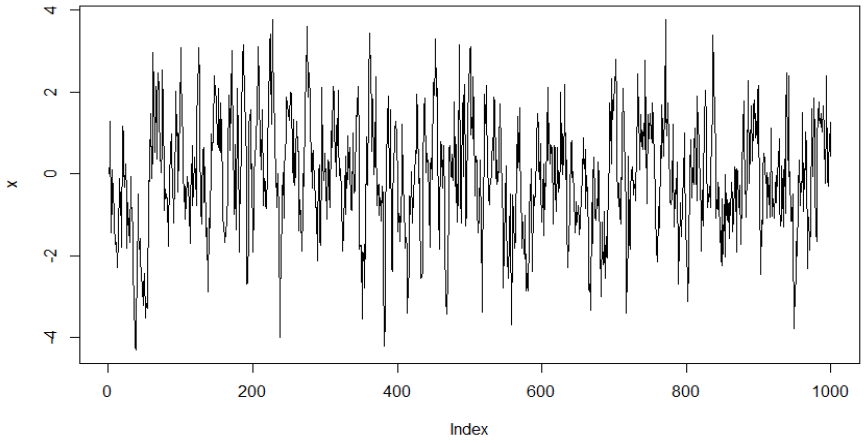
D- This series is non stationary

Model selection

- Check for stationarity of the series (trend, ACF and PACF, unit root)
- Transform if necessary (difference, detrend…)
- ACF, PACF to **choose the model AR or MA** and determine the order
  - ➔ choose the **parameter q** such that the autocorrelation values are not significant for any lag greater than q
  - ➔ choose the **parameter p** such that the partial autocorrelation values are not significant for any lag greater than p
- Estimate the model parameters and the residuals
- Model check:
  - ➔ Keep lag only if coefficient is significant
  - ➔ Residual diagnostics (residuals autocorrelation,…)
  - ➔ Goodness of fit $\overline{R}^2$ and information criteria for model selection:
    - AIC, SBC: based on RSS + correction for the number of parameters. The smaller the AIC or SBC the better the fit of the model.
  - ➔ The model should be parsimonious and plausible

Example: AR(1) model

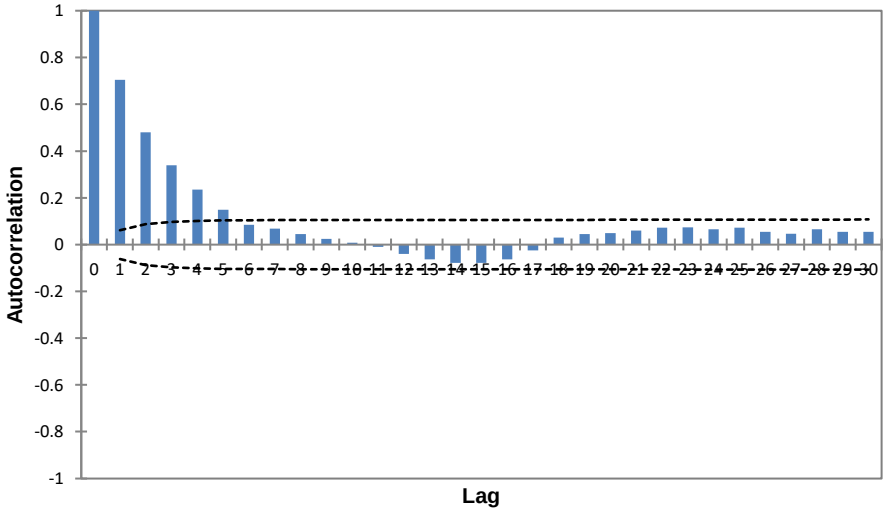
$$Y_t = 0.7\,Y_{t-1} + U_t$$



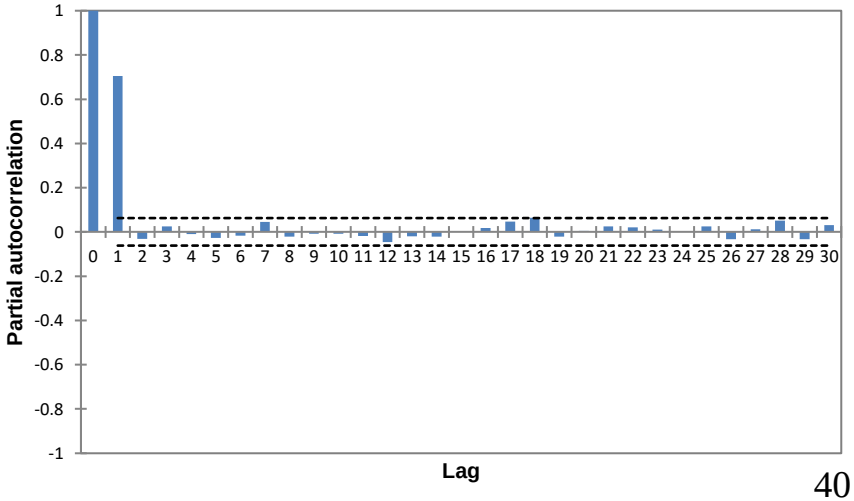
Dickey-Fuller test (x):	
Tau (Observed value)	-8,428
Tau (Critical value)	-3,393
p-value (one-tailed)	< 0,0001
alpha	0,05

Statistic	DF	Value	p-value
Jarque-Bera	2	1,089	0,580
Box-Pierce	6	927,599	< 0,0001
Ljung-Box	6	931,142	< 0,0001
Box-Pierce	12	936,692	< 0,0001
Ljung-Box	12	940,329	< 0,0001

Autocorrelogramx



Partial autocorrelogramx



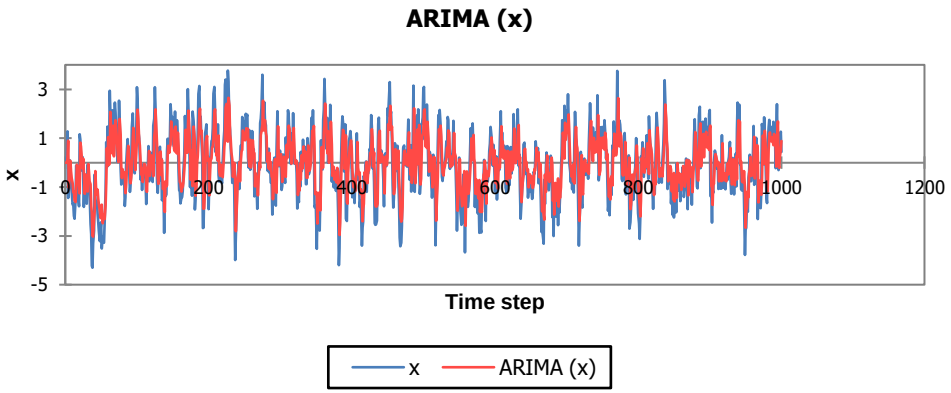
Example: AR(1) model

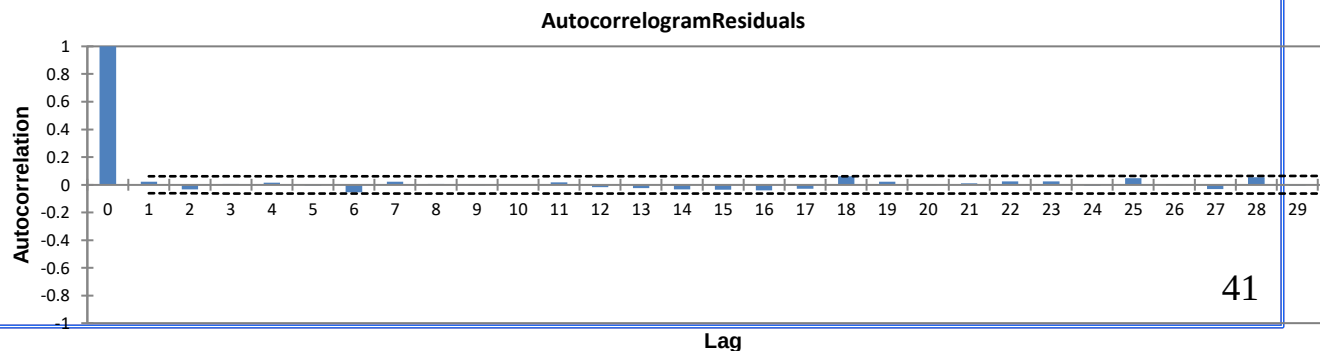
On XLSTAT (or R…) fit an Arima model with autoregressive order 1, 0 degrees of differencing, and an MA order of 0.

Goodness of fit statistics:	
Observations	1000
DF	998
SSE	992,8727718
MSE	0,992872772
RMSE	0,996430014
WN Variance	0,992872772
MAPE(Diff)	244,6768992
MAPE	244,6768992
-2Log(Like.)	2831,411794
FPE	0,994860505
AIC	2835,411794
AICC	2835,42383
SBC	2845,227304
Iterations	8

Model parameters:

Parameter	Value	standard error	Lower bound (95%)	Upper bound (95%)
Constant	0,000	0,107	-0,209	0,209
AR(1)	0,705	0,022	0,661	0,749





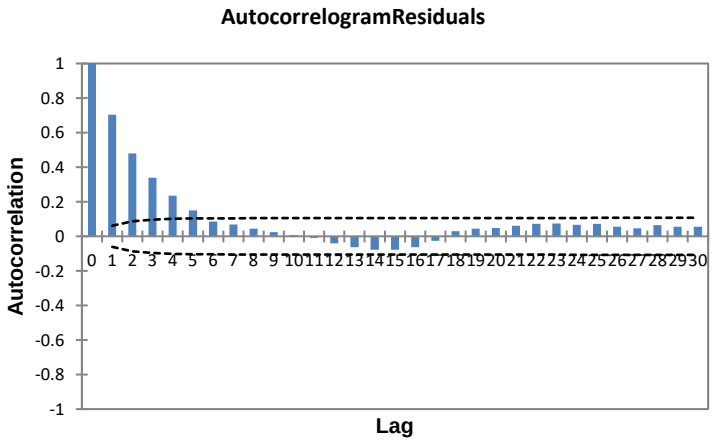
Example: AR(1) model

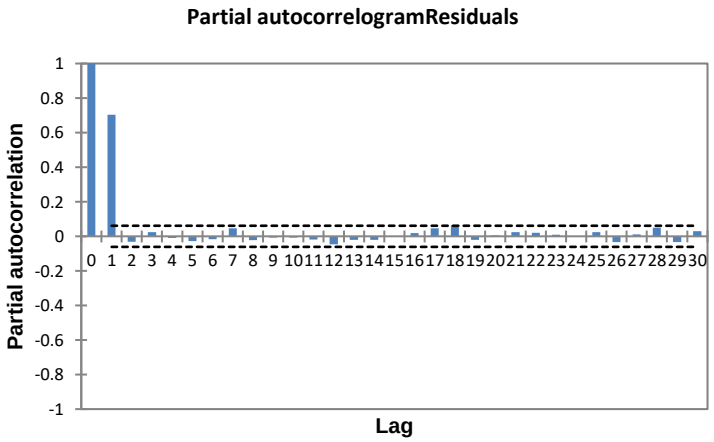
Compare with an ARIMA(2,0,0)

Model parameters:				
Parameter	Value	standard error	Lower bound (95%)	Upper bound (95%)
Constant	0,000	0,103	-0,203	0,203
Parameter	Value	standard error	Lower bound (95%)	Upper bound (95%)
AR(1)	0,727	0,032	0,665	0,789
AR(2)	-0,030	0,032	-0,092	0,031

Goodness of fit statistics:	
Observations	1000
DF	997
SSE	991,9509
MSE	0,991951
RMSE	0,995967
WN Variance	0,991951
MAPE(Diff)	240,7977
MAPE	240,7977
-2Log(Like.)	2830,485
FPE	0,995927
AIC	2836,485
AICC	2836,509
SBC	2851,208
Iterations	57

Compare with an ARIMA(0,0,0)





Forecasting with ARMA Models

- We have estimated an AR(2)
- We are at time *t* and we want to forecast 1,2,..., *s* steps ahead.
- We know Y_t , Y_{t-1}, ..., and U_t , U_{t-1}, ...

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + U_t$$
$$Y_{t+1} = \mu + \phi_1 Y_t + \phi_2 Y_{t-1} + U_{t+1}$$
$$Y_{t+2} = \mu + \phi_1 Y_{t+1} + \phi_2 Y_t + U_{t+2}$$
$$Y_{t+3} = \mu + \phi_1 Y_{t+2} + \phi_2 Y_{t+1} + U_{t+3}$$

$$f_{t,1} = \mathrm{E}(Y_{t+1 \mid t}) = \mathrm{E}_t(\mu + \phi_1 Y_t + \phi_2 Y_{t-1} + U_{t+1}) = \mu + \phi_1 Y_t + \phi_2 Y_{t-1}$$
$$f_{t,2} = \mathrm{E}(Y_{t+2 \mid t}) = \mathrm{E}_t(\mu + \phi_1 Y_{t+1} + \phi_2 Y_t + U_{t+2}) = \mu + \phi_1 f_{t,1} + \phi_2 Y_t$$
...
$$f_{t,s} = \mu + \phi_1 f_{t,s-1} + \phi_2 f_{t,s-2}$$

- Similarly, we can generate forecasts for a MA(q) and for ARMA(p,q)

In-Sample and Out-of-Sample

- In-sample forecasts: predicted values from the estimated time-series model (generated for the same set of data used to estimate the model's parameters).
- Out-of-sample forecasts: forecasts made from the estimated time-series model for a time period different from the one for which the model was estimated.
- Holdout sample: last observations of the sample used to construct out-of-sample forecasts and test the model performance.

→ Ability of the forecast:

➔ Measures of out of sample forecast accuracy: RMSE (root mean square error) measures of the difference between values predicted by a model and the actual values:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (Y_{obs,t} - \hat{Y}_{model,t})^2}{n}}$$

➔ Other measures:

MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error)

➔ The model with the smallest values for RMSE provides the most accurate forecasts

Econometrics & Financial Markets

Other tools and methods :
ANOVA, ANCOVA
Logit model
Panel data
Diff-in-diff
Event study

Toulouse Business School

MSc BIF

Anna CALAMIA

a.calamia@tbs-education.fr

ANOVA, ANCOVA

Categorical or qualitative variables

- ANOVA (analysis of variance):
  - ➔ Compares means among groups, based on a numerical response variable (dependent variable) and qualitative explanatory variable (*factors*).
  - ➔ Seeks to identify sources of variation in the response variable: Variation in DV about its mean is explained by one or more categorical independent variables or is unexplained (random error).
  - ➔ Assumptions on errors as in linear regression
  - ➔ One-way and Two-way ANOVA

- ANCOVA (analysis of covariance):
  - ➔ Similar to ANOVA but uses both qualitative (*factor*) and quantitative (*covariate*) explanatory variables. The variance of the dependent variable is decomposed in variance explained by the covariates, by the factors and the residual variance.

TUTORIAL XLSTAT

7. ANCOVA

- Regression with quantitative and qualitative variables

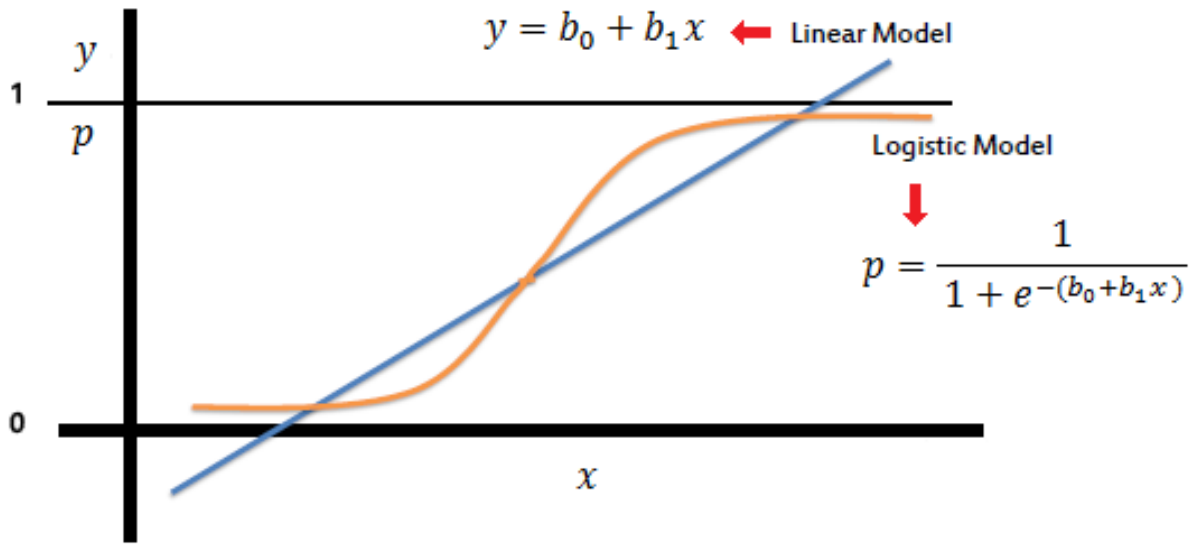
Logit model

Regressions with qualitative dependent variables

- **Qualitative dependent variables** (or categorical dependent variables) are dummy variables used as dependent variables.
- **Probit and logit** regression models are used to model the effect of a series of variables on a **binary response variable** (with two possible values, such as pass/fail)
- They are based on the estimation of the probability of a discrete outcome given the values of the independent variables used to explain that outcome: probability that Y = 1 (a condition is fulfilled) given the values of the independent variables.
- **Probit model** is based on the normal distribution
- **Logit model** is based on the logistic distribution

Logit model

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \varepsilon$$



TUTORIAL XLSTAT

8. LOGIT

- Regression with binary response variables

Panel data analysis

Panel data

Panel Data: observations about different cross sections over time (2 dimensions)

- Y_{it} is observed for all individuals i=1,...,N across all time periods t=1,...,T
- E.g., relation between returns and earnings for several stocks over time
- Panel data allows to control for unobservable variables and heterogeneity

Quarter	Stock	EBIT	Return
2017-Q1	AIRBUS	533	-0.00433
2017-Q2	AIRBUS	529	-0.00305
2017-Q3	AIRBUS	414	0.005125
2017-Q4	AIRBUS	878	-0.01073
2018-Q1	AIRBUS	168	-0.00192
2018-Q2	AIRBUS	871	0.02181
2018-Q3	AIRBUS	1524	-0.00661
2018-Q4	AIRBUS	2155	0.002388
2017-Q1	CARREFOUR	1385	0.005002
2017-Q2	CARREFOUR	546	0.00317
2017-Q3	CARREFOUR	546	0.008257
2017-Q4	CARREFOUR	427	0.001944
2018-Q1	CARREFOUR	427	0.003275
2018-Q2	CARREFOUR	-169	-0.00431
2018-Q3	CARREFOUR	-169	-0.00151
2018-Q4	CARREFOUR	931	0.005734

- **Pooled OLS**
  - homogeneous panel data
  - model parameters are common across individuals: $y_{it}=\alpha+\beta X_{it}+\epsilon_{it}$

Heterogeneous panel: FE and RE

- **Fixed Effects**

  Includes unobservable individual-specific and/or time-specific effects, possibly correlated with the observed explanatory variable:

  $Y_{it} = \beta X_{it} + \alpha_i + \epsilon_{it}$

  $Y_{it} = \beta X_{it} + \alpha_i + \tau_t + \epsilon_{it}$

  where αi is the unknown intercept for each entity (i=1….N), composed of a constant intercept and an individual-specific term; τt captures any unobservable time-specific effects.

  → Within estimator or least squares dummy variable(LSVD)

- **Random Effects**

  Includes unobservable time-specific and/or individual-specific effects which act like individual-specific stochastic error terms, uncorrelated with the regressors.

  → GLS with appropriate error structure (accounting for individual-specific error)

- **Hausman test** to chose between fixed or random effects.

  H0: the preferred model is random effects; H1: fixed effects

  Tests whether the unique errors are correlated with the regressors (H0: they are not)

Panel Data - example

firm	year	inv	value	capital
1	1935	317.6	3078.5	2.8
1	1936	391.8	4661.7	52.6
1	1937	410.6	5387.1	156.9
1	1938	257.7	2792.2	209.2
1	1939	330.8	4313.2	203.4
1	1940	461.2	4643.9	207.2
1	1941	512	4551.2	255.2
1	1942	448	3244.1	303.7
1	1943	499.6	4053.7	264.1
1	1944	547.5	4379.3	201.6
1	1945	561.2	4840.9	265
1	1946	688.1	4900.9	402.2
1	1947	568.9	3526.5	761.5
1	1948	529.2	3254.7	922.4
1	1949	555.1	3700.2	1020.1
1	1950	642.9	3755.6	1099
1	1951	755.9	4833	1207.7
1	1952	891.2	4924.9	1430.5
1	1953	1304.4	6241.7	1777.3
1	1954	1486.7	5593.6	2226.3
2	1935	209.9	1362.4	53.8
2	1936	355.3	1807.1	50.5
2	1937	469.9	2676.3	118.1
2	1938	262.3	1801.9	260.2
2	1939	230.4	1957.3	312.7
2	1940	361.6	2202.9	254.2
2	1941	472.8	2380.5	261.4
2	1942	445.6	2168.6	298.7
2	1943	361.6	1985.1	301.8
2	1944	288.2	1813.9	279.1
2	1945	258.7	1850.2	213.8
2	1946	420.3	2067.7	132.6
2	1947	420.5	1796.7	264.8
2	1948	494.5	1625.8	306.9
2	1949	405.1	1667	351.1
2	1950	418.8	1677.4	357.8
2	1951	588.2	2289.5	342.1
2	1952	645.5	2159.4	444.2
2	1953	641	2031.3	623.6
2	1954	459.3	2115.5	669.7
....	....	....	....	....

Summary statistics:

Variable	Observations	Minimum	Maximum	Mean	Std. deviation
inv	200	0.930	1486.700	145.958	216.875
value	200	58.120	6241.700	1081.681	1314.470
capital	200	0.800	2226.300	276.017	301.104

Results for variable inv:

Goodness of fit statistics:

rsq	0.769
adjrsq	0.767

Joint test of significance (F or Chi-square test):

statistic.Chisq	parameter.df	p.value.Chisq
657.295	2	1.8634E-143

Coefficients:

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	-57.865	29.393	-1.969	0.049
value	0.110	0.011	10.429	<0.0001
capital	0.308	0.017	17.948	<0.0001

R-sq:
within = 0.7668
between = 0.8196
overall = 0.8061

hausman fixed random

Test: Ho: difference in coefficients not systematic

Prob>chi2 = 0.3119

Difference-in-differences

Difference-in-differences

Used to estimate the effects of a sudden change in economic environment, policy, or general treatment on a population

- **Treatment group**: subject to the change i.e. to the **treatment (**sudden exogenous source of variation)
- **Control group**: similar in characteristic to the treatment group but not subject to the change
- Quantifiable and measurable **outcome**
- Measure of treatment effects based on **between-group cross-sectional differences** and **within-group time-series differences**
- **Parallel trend assumption**: in the absence of treatment, the difference between the groups is constant over time

Example: Card and Krueger(AER,1994)

- Does an increase in minimum wage have a negative impact on employment?

- Study the evolution of the number of employees in fast-food restaurants in New Jersey (NJ) following the increase in minimum wage from $4.25 (Feb. 1992) to $5.05 (Nov. 1992)
- Comparison with the evolution of employment in Pennsylvania(PA), a neighbouring state

Diff-in-diff: basic principle

- The average (expected) number y of employees, in state s at time t:

$$E(y|s,t) = \gamma_s + \lambda_t$$

- Where γs is a constant specific to state s and λt is a constant specific to time period t
- A change (treatment) on minimum wage occurs in state s at period t and creates a shock on employment equal to β
- For a given restaurant i operating in state s at date t, the number of employees will be equal to

$$y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + \epsilon_{ist}$$

- Where ϵist is the error term and Dst is a dummy variable that takes on the value 1 for observations from the treated group (NJ) after the treatment (post), 0 otherwise

Diff-in-diff: basic principle

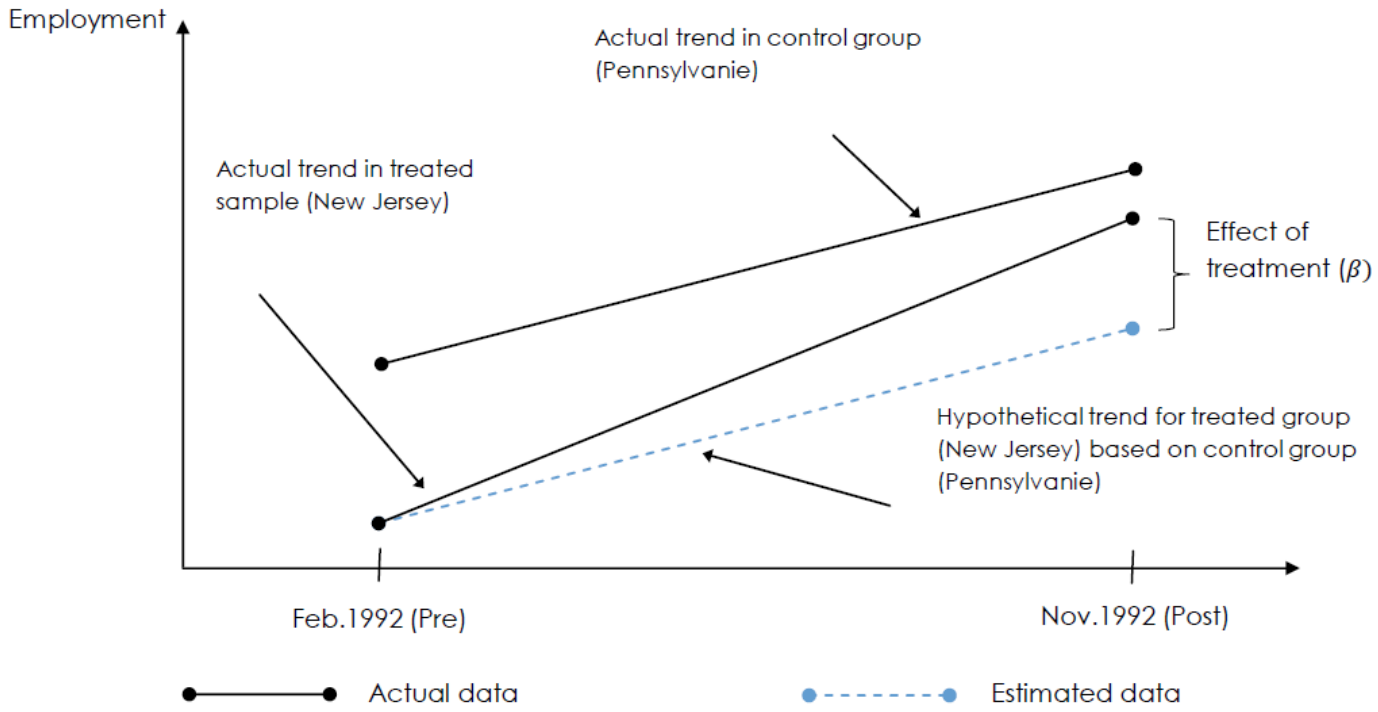
- The question is: How can we measure β ?
- Compute first the difference between the average number of employees in PA, after and before the treatment
  - ➔ This difference allows to remove differences across states
- Compute next the difference between the average number of employees in NJ, after and before the treatment
- Finally, compute second difference minus first difference:

$$E(y_i|s = NJ, t = Post) - E(y_i|s = NJ, t = Pre) \; - \; [E(y_i|s = PA, t = Post) - E(y_i|s = PA, t = Pre)]$$

  - ➔ This is equal to : β

    We have eliminated the common trend between the groups, λt, and the permanent differences between the groups, leaving a very simple estimate of the treatment effect, ß.

Diff-in-diff: basic principle



Diff-in-diff: regression specification

- Simple DID does not compute the statistical significance of the shock β
- Regression specification allows to overcome this problem:

$$Y_{ist} = \alpha + \gamma D_s + \lambda D_t + \beta(D_{st}) + \varepsilon_{ist}$$

where: Yist is the number of employees in restaurant i in state s at period t; Ds is a dummy variable that takes on the value 1 for restaurants in NJ (i.e. treated group) and 0 otherwise; Dt is a dummy variable that takes on the value 1 for observations made after the wage increase (treatment); Dst is a dummy (interaction) variable that takes on the value 1 for observations in NJ after the treatment; εist is the error term

What is the interpretation of the regression coefficients ?

α (intercept) is the average number of employees in restaurants operating in PA (during the Pre period)
γ is the difference between the average number of employees in NJ and PA
λ is the difference between the average number of employees working in restaurants in the $Post$ and the Pre periods
β is the DiD estimator, Average differential change in y from the first to the second time period of the treatment group relative to the control group

Event studies

Event studies: what for?

Event studies aim at quantifying the effects of an (unexpected) economic event on the value of firms

- Financial economics: corporate events, market efficiency
- Macroeconomic policy: fed rates, trade deficits
- Accounting: earning announcements
- Law and economics: changes in legal environment and regulation
- Marketing: brand strategy announcements
- ...

How asset prices react to a given event:

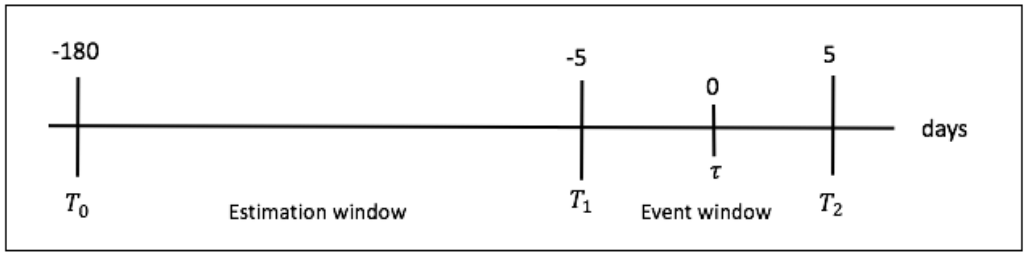
- events are reflected in asset prices (assuming markets are informationally efficient)
- prices are easily observed
- well-performing models are available to isolate the impact of a given event on asset prices

Test of market efficiency: Is the relevant information impounded into prices immediately or with delay?

Event studies design

Step 1: Definition of the event and event window

- ➔ Unexpected event
- ➔ Exact date (date 0)
- ➔ Short vs long horizon
- ➔ Define an event window (period over which prices are examined) that is larger than the exact period of interest
  - Inclusion of the days prior to the event (-1, -2, ...) aims at accounting for possible anticipation of the event as well as information leakage
  - Inclusion of the days after the event (+1, +2, ...) aims at capturing posterior abnormal movements that occur after market close



Step 2: Selection criteria (Which firms to be included in the study?)

- ➔ Restrictions imposed by data availability and reliability
- ➔ Restrictions imposed by representativeness issues
- ➔ Some summary statistics (market capitalisation, average return, industry representation, distribution of events through time...) might prove useful to identify potential biases in the initial sample as well as outliers

Step 3: Normal and abnormal returns

- Problem: how to isolate price movements induced by the event of interest from contemporaneous movements, unrelated to the event?

➔ We need a measure of abnormal returns:

Abnormal returns are computed as the difference between observed returns and normal returns (predicted returns):

$$ARit = R^*it - E(Rit|Xt)$$

Normal returns correspond to the expected returns if the event had not taken place

➔ market model or other asset pricing models (CAPM, FF 3 factors, etc.) to estimate normal returns: $Ri,t = \alpha i + \beta i Rm,t + \xi i,t$

➔ Cumulative Abnormal Returns (CAR), computed as the cumulative sum of abnormal return over the event window (for a given security)

➔ AR averaged in the cross section of sample stocks to compute AARs on each event date: time series of average abnormal returns (AAR)

➔ AAR across securities can also be aggregated over time to compute Cumulated Average Abnormal Returns over the event window (CAAR)

Step 4: Estimation procedure

- Estimate the parameters of the model that is used to generate normal returns over the event window
  - ➔ The estimation is performed on the **estimation window**

- Check that the estimation window is not contaminated by events that are likely to impact the parameters of the model that generates normal returns
- The **event window** (or part of it) should not be included in the estimation window (when feasible). May also introduce a buffer zone between the estimation window and the event window.
- Common choices for the length of the estimation window are 120 days or 250 days
- The estimation of the parameters can be made through **OLS**

Step 5: Test procedure

Once abnormal returns are computed and aggregated, the objective is to test their significance:

- Test of the null hypothesis: Event has no impact on returns, i.e., no abnormal returns

  - Comparison of the distribution of actual returns with the distribution of predicted returns
  - Typically, the specific null hypothesis to be tested is whether the mean abnormal return in the event window is equal to 0
  - Occasionally, other parameters of the cross-sectional variation in abnormal returns can be used, such as the median or variance
  - Parametric tests, such as t-test (based on normality assumption) and non-parametric tests

Event study results
Interpretation and conclusions

Question the reliability of results:

- Interpretation of results
- Robustness tests using various sub-samples
- Incidence of outliers?
- Sensitivity to the choice of the estimation window?
- Sensitivity to the normal-return generating model?
- Other issues (clustering, event induced variance, partially anticipated events, event-date uncertainty, short vs long horizon…)