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)

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

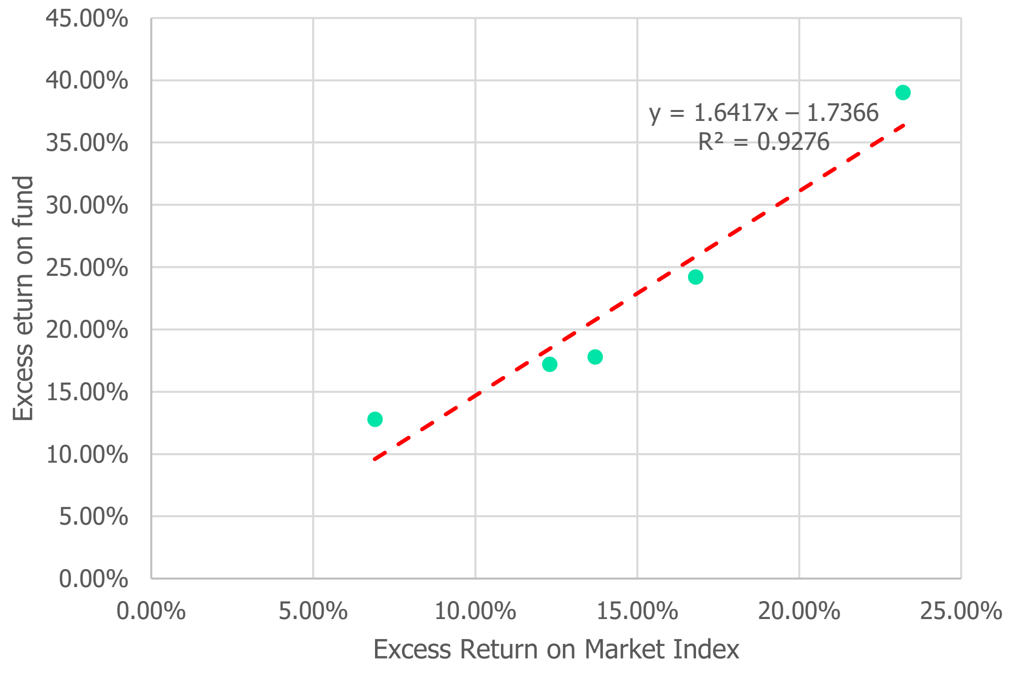
Simple Regression: An Example

- 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	**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

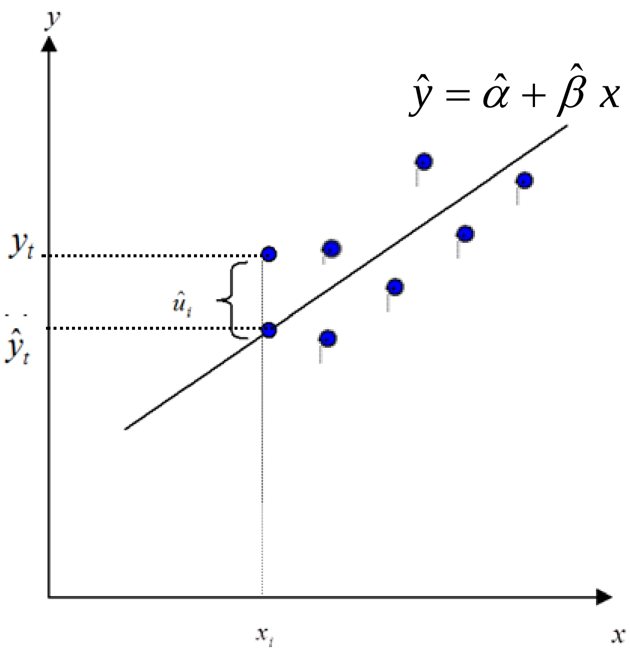
- Does a relationship appear between x and y given the data that we have? ➔ first stage = scatter plot

Simple Regression: Scatter Diagram



Hypothesis and estimation of the coefficients

Ordinary Least Squares



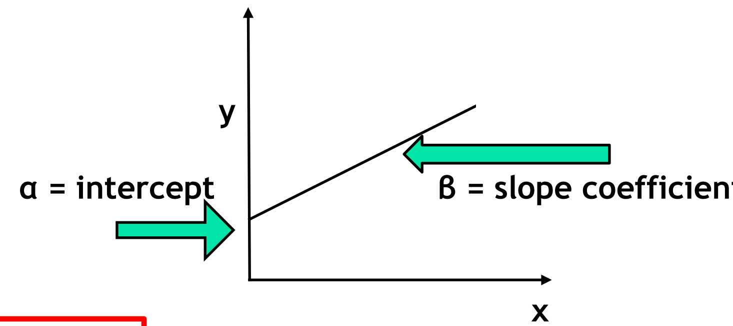
- The most common method used to fit a line to the data is known as **OLS (ordinary least squares).**
- What we actually do is take each distance and square it and **minimize the total sum of the squares** (hence least squares).
- Tightening up the notation, let :
  - ➔ y_t : **actual data**
  - ➔ $\hat{y}_t$: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



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

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	$\boxed{-0.017366}$	0.041140	-0.422132	0.7014
ER_MARKET_INDEX	$\boxed{1.641745}$	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 :

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

  - $\boxed{s = \sqrt{s^2}}$ **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 :

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

- **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 $\hat{\beta}$:

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

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

- Most commonly used tests :

$$\text{H0} : \beta = 0 \qquad \text{H0} : \alpha = 0$$

$$\text{H1} : \beta \neq 0 \qquad \text{H1} : \alpha \neq 0$$

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

- 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).

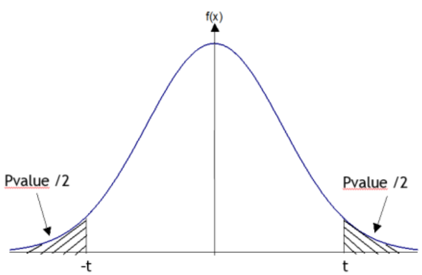
If we reject the null hypothesis at the 5% level, we say that the result of the test is statistically significant.

The Test of Significance Approach

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

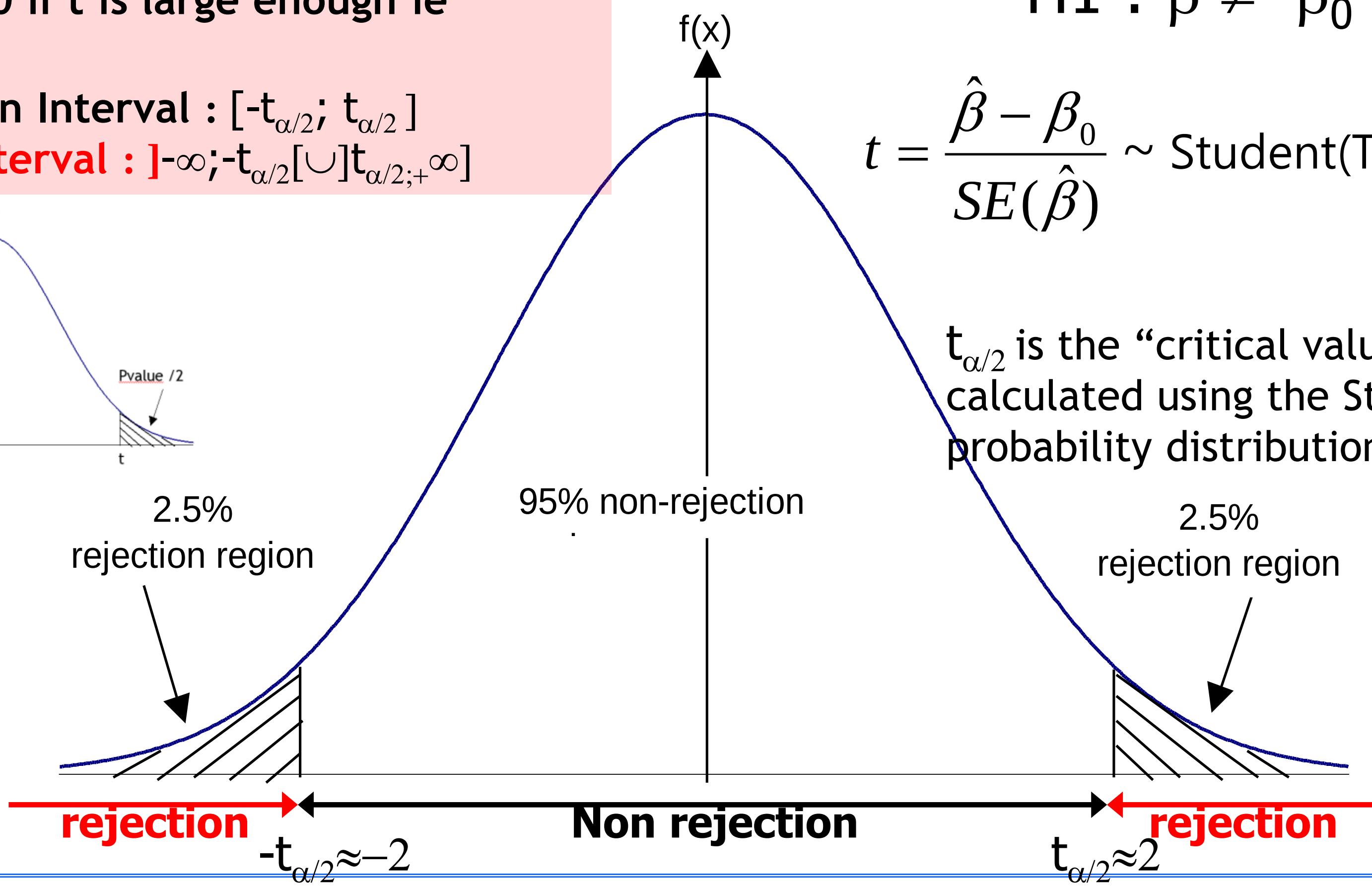
$$H0 : \beta = \beta_0$$
$$H1 : \beta \neq \beta_0$$

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?

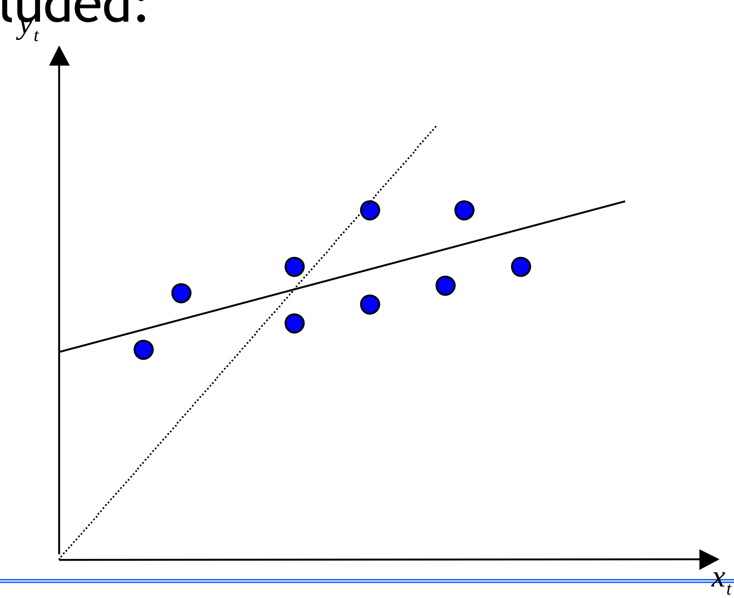
A- the fund excess return is not correlated to the market excess return

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).

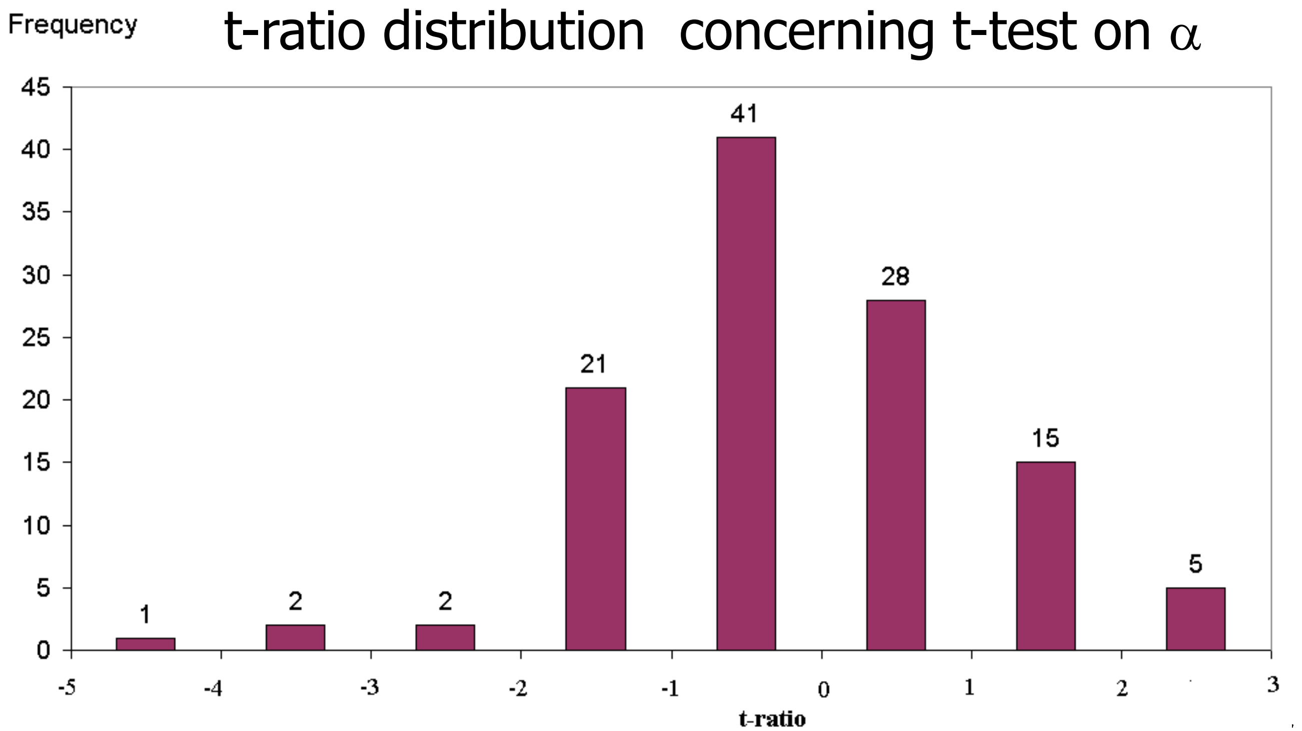
- The Data: Annual Returns on the portfolios of 115 mutual funds from 1945-1964.

- 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 α_j.

- The null hypothesis is H0: $\alpha_j = 0$.

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



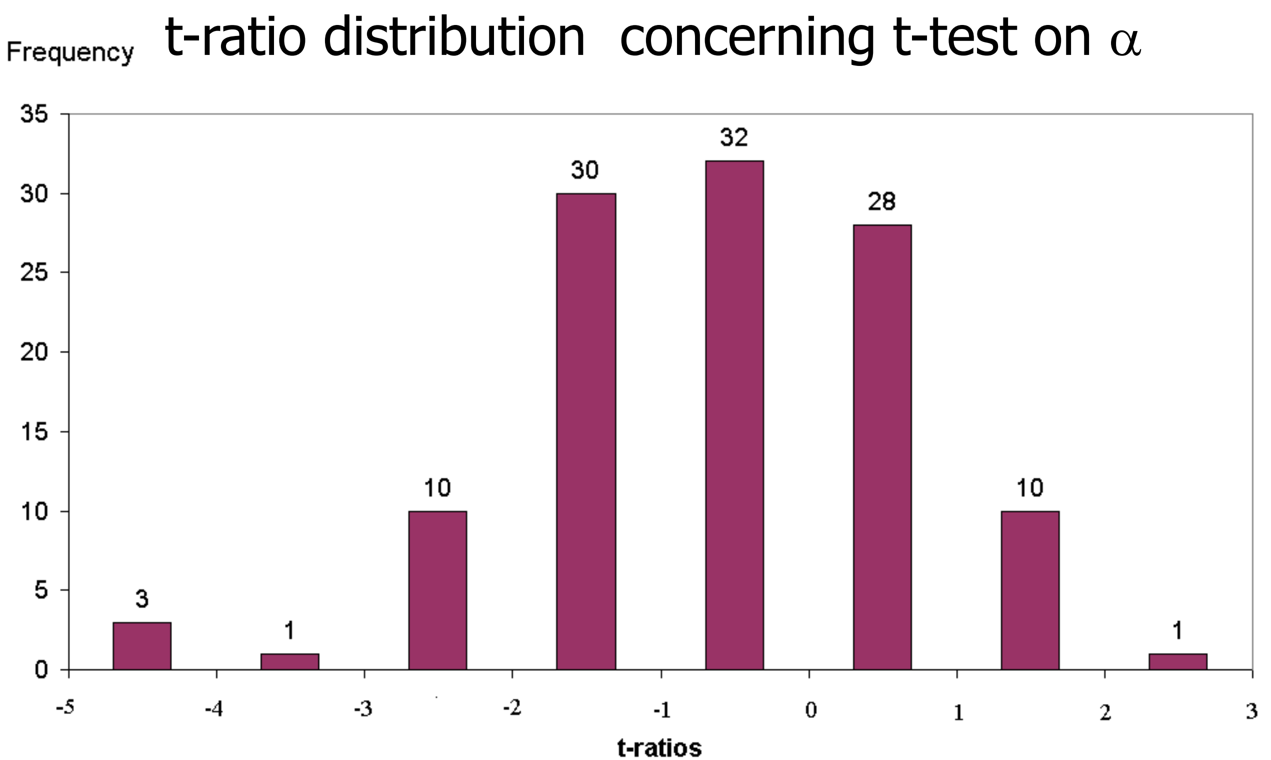
Question 8 : Knowing that the Critical value $t_{\alpha/2}$ for a two-sided test ≈ 2, which affirmation is false?

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

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

TSS = Variability of Y

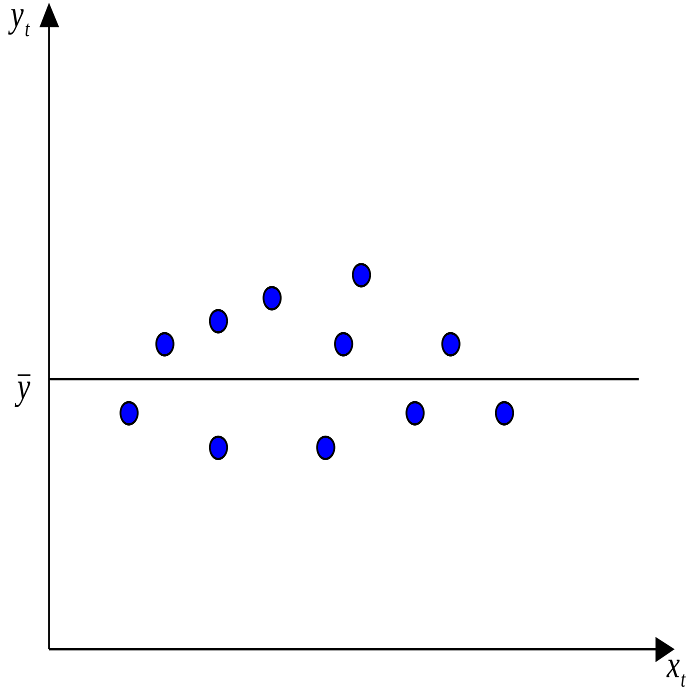
ESS = Variability of $\hat{Y}$

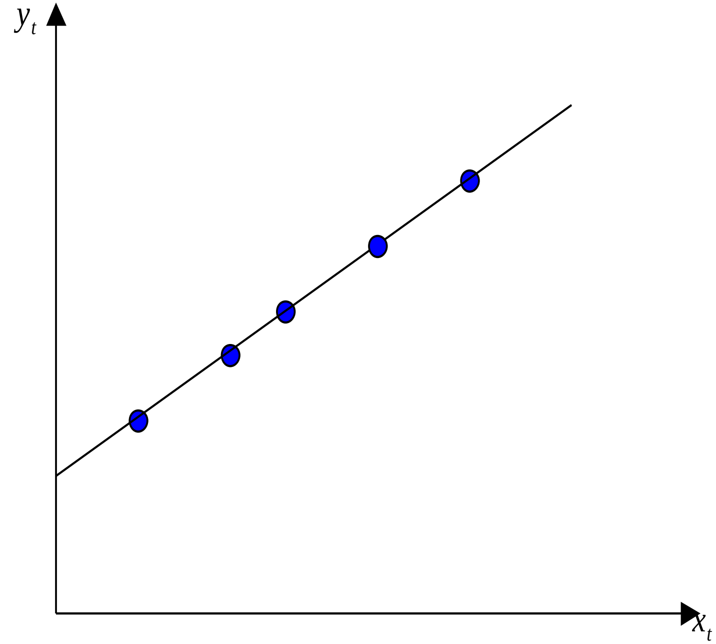
RSS = Variability of $\hat{U}$

TSS = ESS + RSS

$$\underbrace{\sum_t \left(Y_t - \bar{Y}\right)^2}_{\text{TSS = Total sum of squares}} = \underbrace{\sum_t \left(\hat{Y}_t - \bar{Y}\right)^2}_{\text{ESS = Explained sum of squares}} + \underbrace{\sum_t \hat{U}_t^2}_{\text{RSS = Residual sum of squares}}$$

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 :

$$H0: \beta = \beta_0$$

$$H1: \beta \neq \beta_0$$

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

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

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

TSS = Variability of Y

ESS = Variability of $\hat{Y}$

RSS = Variability of $\hat{U}$

TSS = ESS + RSS

$$\underbrace{\sum_t \left(Y_t - \overline{Y}\right)^2}_{\text{TSS = Total sum of squares}} = \underbrace{\sum_t \left(\hat{Y}_t - \overline{Y}\right)^2}_{\text{ESS = Explained sum of squares}} + \underbrace{\sum_t \hat{U}_t^2}_{\text{RSS = Residual sum of squares}}$$

Adjusted R^2

- ***Be careful ! R²*** **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:

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

- So if we add an extra regressor, k increases and contrary to the R^2 the $\bar{R}^2$ may decrease.
- As soon as $k \geq 2$, $\bar{R}^2 < R^2$
- While R^2 must be at least zero, $\bar{R}^2$ 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

Select Y and the explanatory variables X_k

↓

Launch the model on an econometrical software

↓

Test the global significance of the model (F-test)

$H0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$

H0 Accepted → **STOP**

H0 rejected ↓

-Check the significance of each coefficient (t-test)

-Remove one by one the insignificant variables

- Remove first the least significant variable (pvalue highest)
- Revive the model
- Reproduce the previous step until all variables are significant

↓

Goodness of Fit Statistics

R^2 and Adjusted R^2

→ Hypotheses on residuals???

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: $E(u_t) = 0$

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.

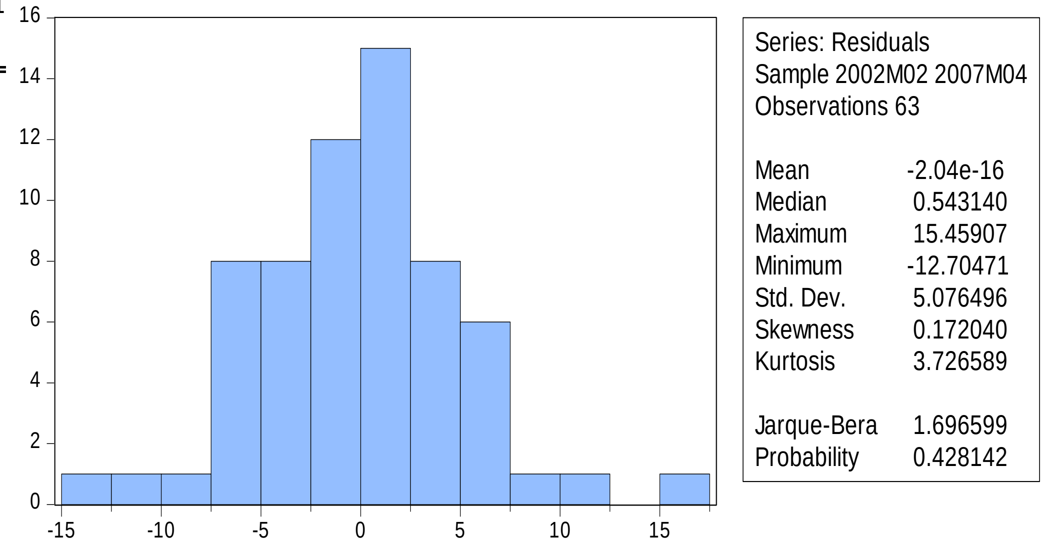
CAPM Microsoft / SP500

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

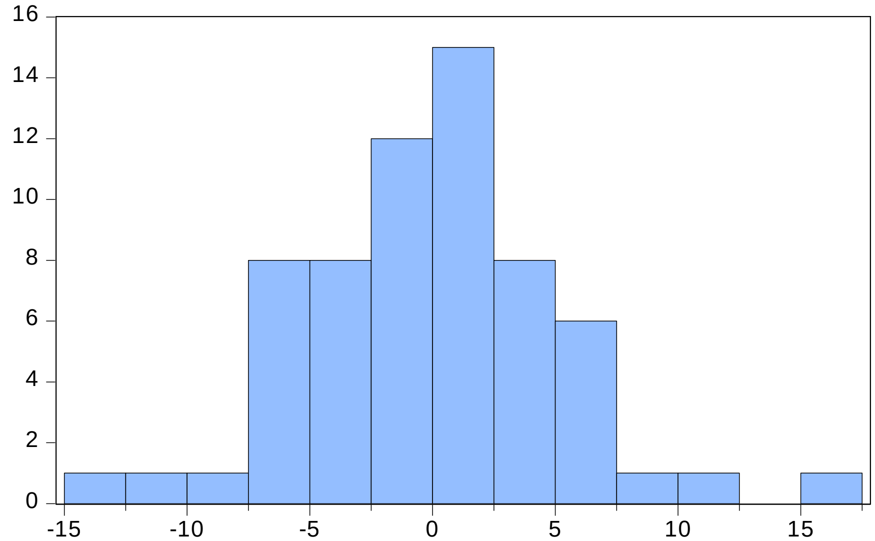
R-squared	0.358859	Mean dependent var	0.121478
Adjusted R-squared	0.348349	S.D. dependent var	6.339973
S.E. of regression	5.117937	Akaike info criterion	6.134611
Sum squared resid	1597.790	Schwarz criterion	6.202647
Log likelihood	-191.2403	Hannan-Quinn criter.	6.161370
F-statistic	34.14293	Durbin-Watson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

Comment :
-residuals mean



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

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

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?

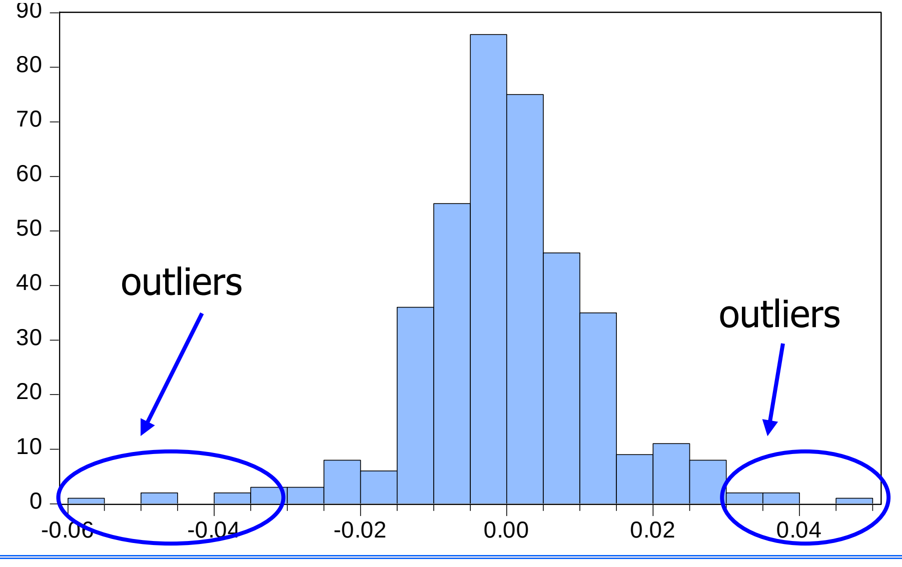
Residual normality and outliers

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



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

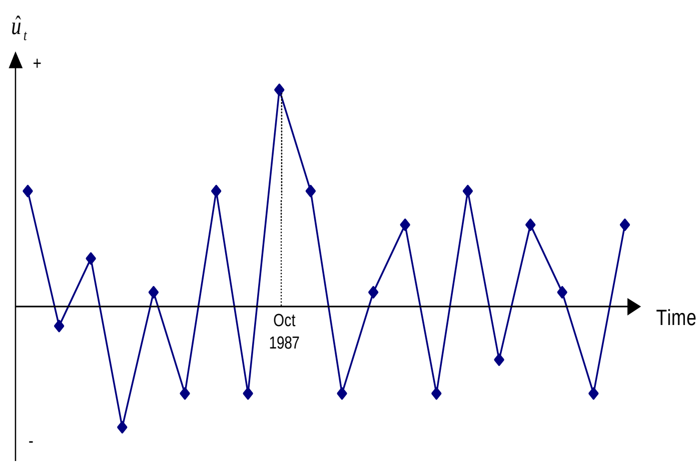
Comment :
-residuals normality?

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

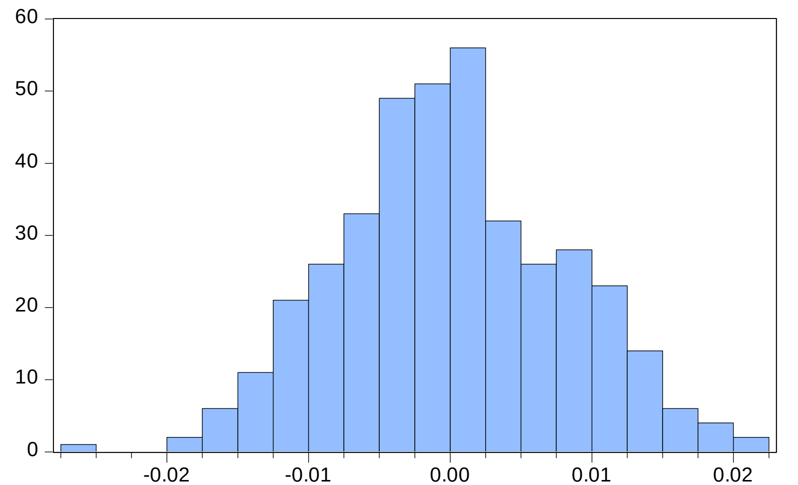
Residual normality and dummies

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		

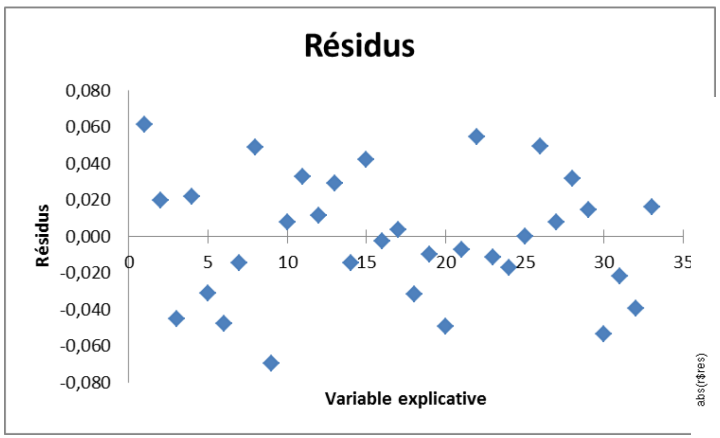
Comment :
-residuals normality?

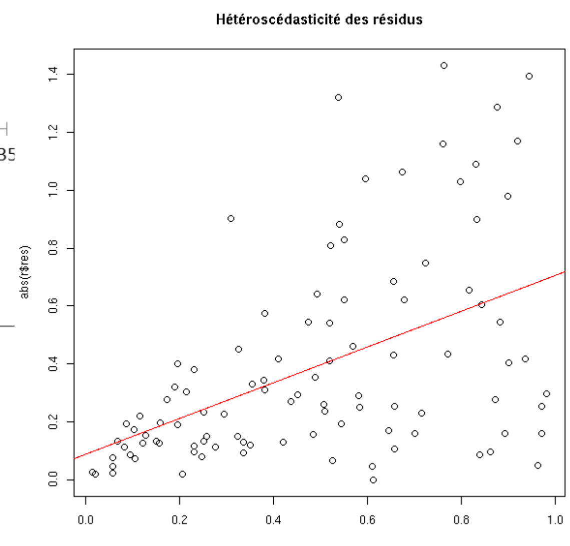


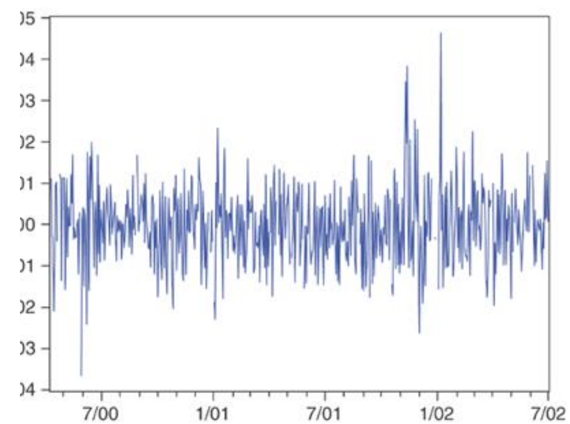
Series: Residuals	
Sample 125 515	
Observations 391	
Mean	4.66e-19
Median	-0.000214
Maximum	0.020448
Minimum	-0.025999
Std. Dev.	0.007939
Skewness	0.094913
Kurtosis	2.793170
Excess kurtosis	0.793170
Jarque-Bera	1.283993
Probability	0.526241

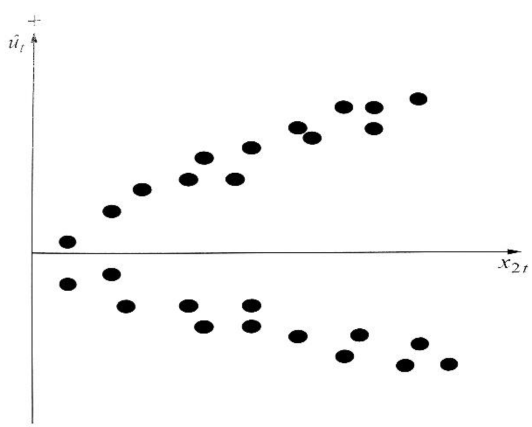
Assumption 3: Var(Ut) = $\sigma^2 < \infty$

- 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 $t \neq s$

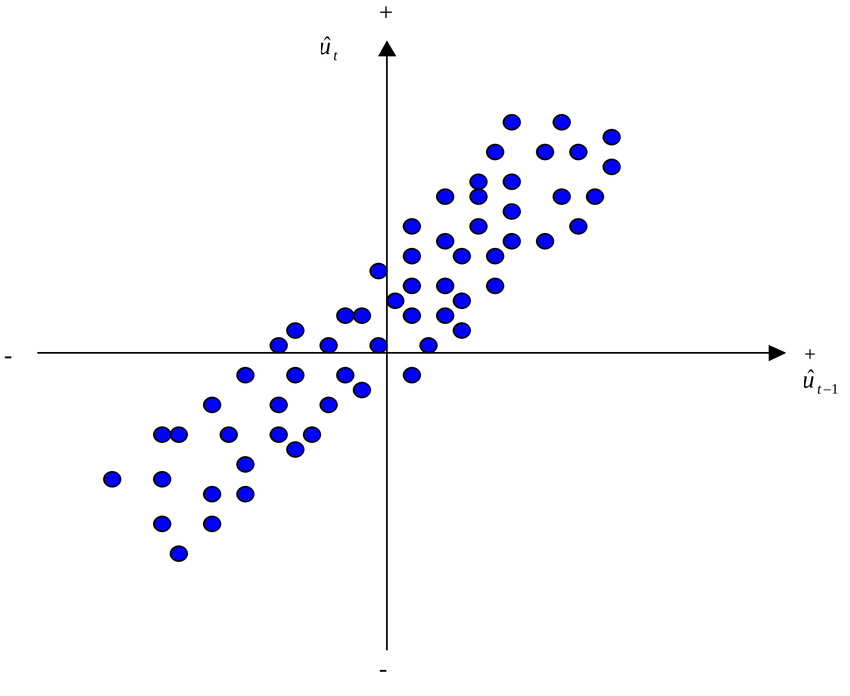
Cov $(U_i\ , U_j) = 0$ for $i \neq j$,

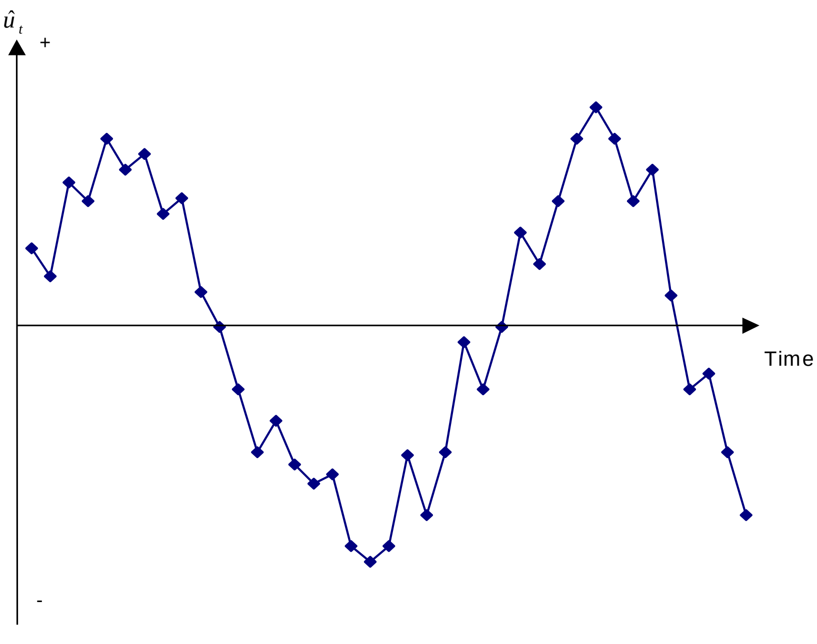
→ no pattern in the errors.

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
.	.	.	.
.	.	.	.
.	.	.	.

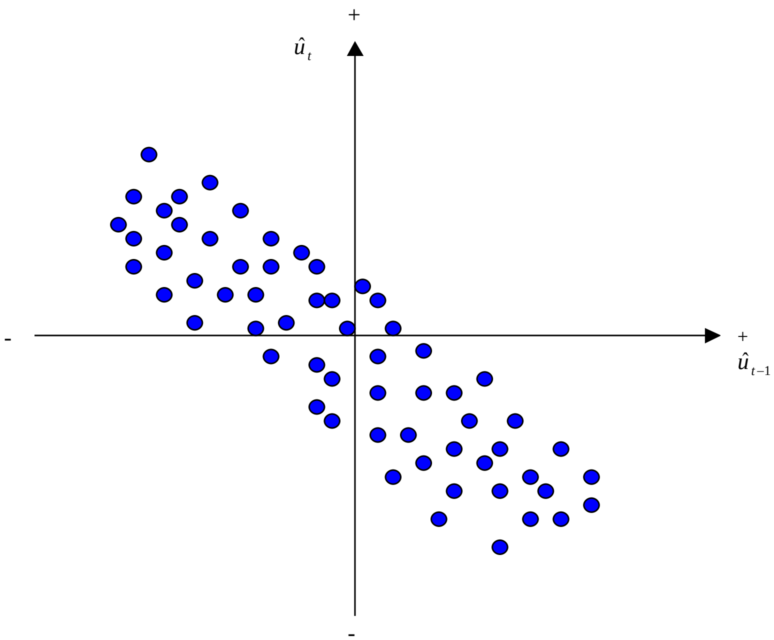
Stereotypical patterns : Positive Autocorrelation

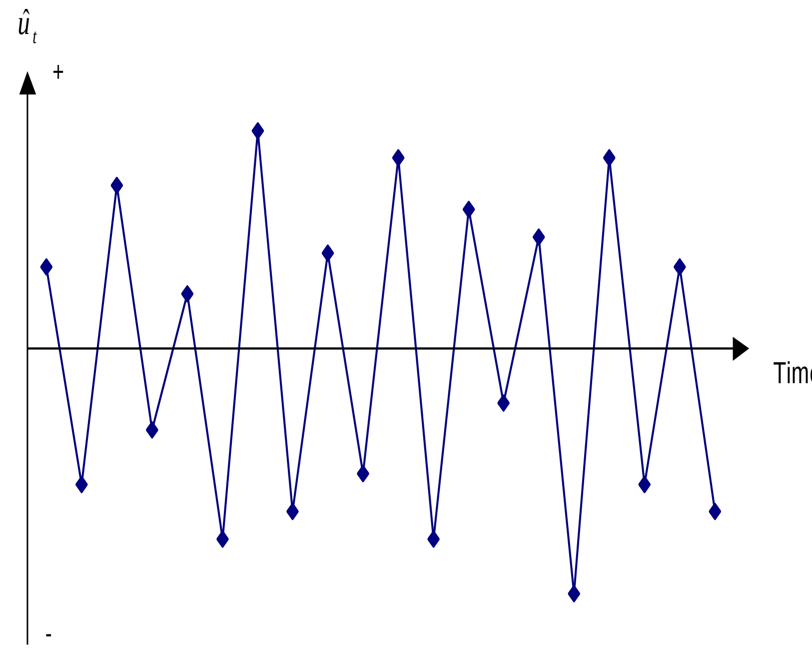




Positive Autocorrelation is indicated by a cyclical residual plot over time.

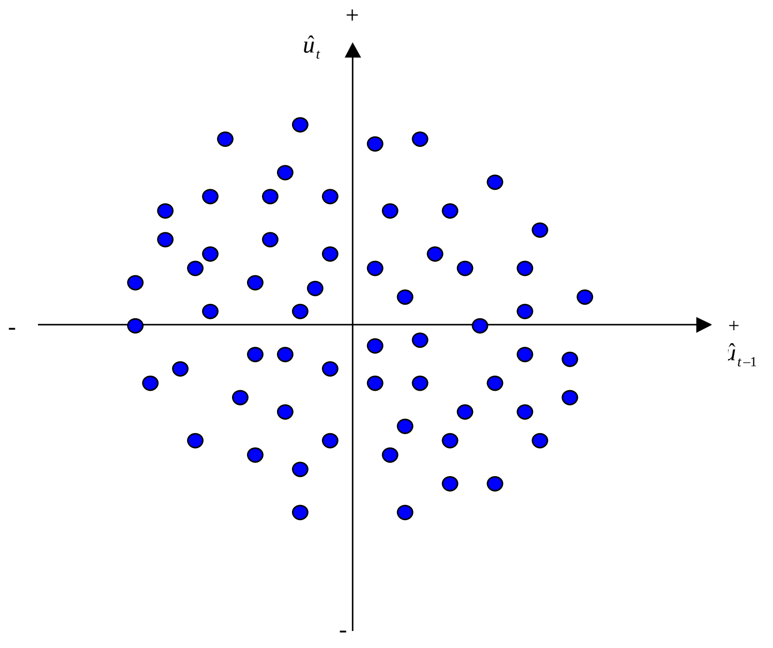
Stereotypical patterns : Negative Autocorrelation

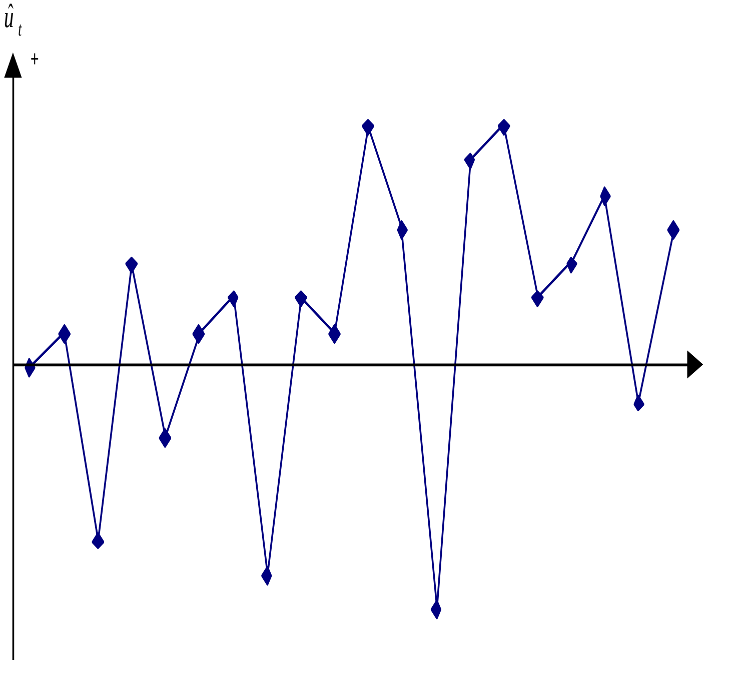




Negative autocorrelation is indicated by an alternating pattern where the residuals cross the time axis more frequently than if they were distributed randomly

No pattern in residuals – No autocorrelation



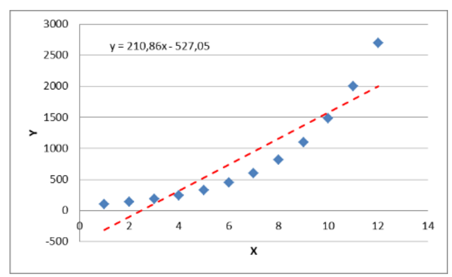


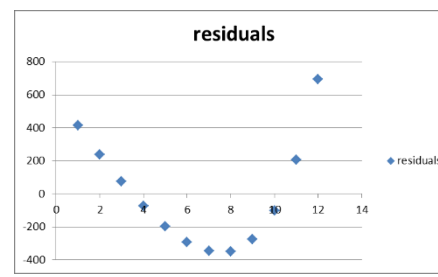
No pattern in residuals at all: this is what we would like to see

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 $V_t \sim N(0, \sigma_v^2)$

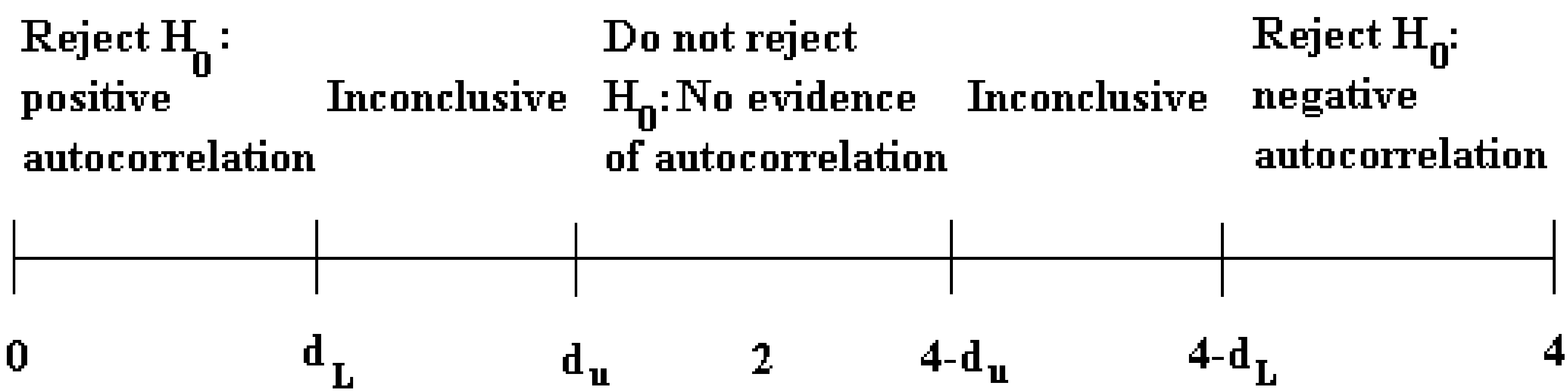
- The DW test statistic : $H_0 : \rho=0$ and $H_1 : \rho\neq 0$
- The test statistic is calculated by

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

➔ $DW \approx 2(1 - \hat{\rho})$, $-1 \leq \hat{\rho} \leq 1$, 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 $\rightarrow$ 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.

<u>Conditions which must be fulfilled for DW to be a Valid Test</u>

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:

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

- The hypotheses :

$$\boxed{\begin{array}{l} \mathrm{H}_0 : \rho_1 = 0 \text{ and } \rho_2 = 0 \text{ and } \ldots \text{ and } \rho_r = 0 \\ \mathrm{H}_1 : \rho_1 \neq 0 \text{ or } \rho_2 \neq 0 \text{ or } \ldots \text{ or } \rho_r \neq 0 \end{array}}$$

- 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 $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-r}$. Obtain R^2 from this regression.

- Test statistic : $\boxed{(T-r)R^2 \sim \chi^2(r)}$
- 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

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

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 $F > F_\alpha(k, T-2k)$ 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	<u>0.8</u>
x_3	0.2	-	0.3
x_4	<u>0.8</u>	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 $\mathrm{VIF} = \dfrac{1}{1-R^2}$ 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