Introduction to Financial Econometrics

Chapter 3: Assessing the Multiple Linear Model

Overall fitting, significance and misspecification tests

Christophe Hurlin

Univ Orléans

May, 2019

1. Introduction

Three questions

1. **Question 1**: How well does the fitted line describe the data? or How much variation of the dependent variable is explained by the model?
2. **Question 2**: How informative is (resp., are) an explanatory variable (resp., the independent variables) for the dependent variable?
3. **Question 3**: Are assumptions satisfied?

1. Introduction

To answer...

1. **Overall fitting criterion of the fitted model:**
    - Coefficient of determination
    - Adjusted coefficient of determination
2. **Proceed with individual and global significance tests:**
    - Student test
    - Fisher test
3. **Misspecification tests**
    - Residuals: outliers and qq-plot
    - Functional form
    - Heteroskedasticity
    - Autocorrelation
    - Non-normality of the error terms

1. Introduction

The outline of this chapter is the following:

1. Introduction

References

- Brooks, C., Introductory Econometrics for finance, Cambridge University Press, 2nd edition.
- Campbell, J.,Y., A.W. Lo and A.C. MacKinlay, The Econometrics of financial markets, Princeton University Press, 1997.
- Greene W. (2007), Econometric Analysis, sixth edition, Pearson - Prentice Hil
- Tsay, R., 2002, Analysis of Financial Time Series, Wiley Series

Section 2

Overall adjustment

2. Overall adjustment

Objectives

1. Define the standard error (SE) of the regression.
2. Introduce the coefficient of determination R^2.
3. Define the adjusted R^2.

2. Overall adjustment

Reminder (Chapter 2)

Suppose that there is a **constant term** in the linear model.

$$y_t = \beta_1 + \sum_{k=2}^{K} \beta_k x_{t,k} + \varepsilon_t = \sum_{k=1}^{K} \beta_k x_{t,k} + \varepsilon_t$$

with $x_{t,1} = 1$, $\forall t$ and denote by $\widehat{y}_t$ the fitted or predicted values

$$\widehat{y}_t = \sum_{k=1}^{K} \widehat{\beta}_k x_{t,k}$$

The mean of the fitted (adjusted) values of y equals the mean of the actual values of y:

$$\overline{\widehat{y}}_T = \overline{y}_T$$

2. Overall adjustment

Definition (standard error of the regression)

A first measure of the overall fitting is the **standard error (SE)** of the regression

$$\widehat{\sigma}_{\varepsilon} = \sqrt{\frac{1}{T-K}\sum_{t=1}^{T}\widehat{\varepsilon}_t^2} = \sqrt{\frac{SSR}{T-K}}$$

Note: The quantity $\widehat{\sigma}_{\varepsilon}$ is also called the **Root Mean Squared Error (RMSE)**.

2. Overall adjustment

Definition (Coefficient of determination)

The **coefficient of determination** of the multiple linear regression model (with a constant term) is the ratio of the (empirical) variance explained by model to the (empirical) variance of y_t:

$$\mathrm{R}^2 = \frac{\sum_{t=1}^{T} \left(\widehat{y}_t - \overline{y}_T\right)^2}{\sum_{t=1}^{T} \left(y_t - \overline{y}_T\right)^2} = 1 - \frac{\sum_{t=1}^{T} \widehat{\varepsilon}_t^2}{\sum_{t=1}^{T} \left(y_t - \overline{y}_T\right)^2}$$

Interpretation

The coefficient of determination measures the proportion of the total variance (or variability) in y that is accounted for by variation in the regressors (or the model).

$$0 \leq \mathrm{R}^2 \leq 1$$

2. Overall adjustment

Properties

1. R^2 is a measure of the extent to which x and y are **linearly related**;
2. $\mathrm{R}^2 \longrightarrow 1$ means that there is a strong positive or negative relationship between the two variables;
3. $\mathrm{R}^2 \longrightarrow 0$ does not rule out the existence of some nonlinear relationships;
4. R^2 depends on the nature of data (micro/macro data, aggregate/disaggregate data);
5. R^2 is meaningless in the absence of an intercept.

Problem

- The R^2 automatically and spuriously increases when extra explanatory variables are added to the model.

2. Overall adjustment

Definition (Adjusted R-squared)

The **adjusted R-squared** coefficient is defined to be:

$$\overline{\mathrm{R}}^2 = 1 - \frac{T-1}{T-p-1}\left(1-\mathrm{R}^2\right)$$

where p denotes the number of regressors (not counting the constant term, i.e., $p = K - 1$ if there is a constant or $p = K$ otherwise).

2. Overall adjustment

Properties

1. $\overline{\mathrm{R}}^2 \leq \mathrm{R}^2$
2. if T is large $\overline{\mathrm{R}}^2 \simeq \mathrm{R}^2$
3. The adjusted R-squared $\overline{\mathrm{R}}^2$ can be **negative**.

Example (CAPM)

We consider the CAPM model

$$z_{\texttt{intel},t} = \alpha + \beta z_{\texttt{market},t} + \varepsilon_t$$

where $z_{\texttt{intel},t}$ is the excess (log-) return for Intel Corp and $z_{\texttt{market},t}$ is the excess (log-) return for the S&P500 observed from August 22, 2017 to August 17, 2018 (250 observations). We know that

$$SSR = 0.0509 \qquad \sum_{t=1}^{T} \left(y_t - \overline{y}_T \right)^2 = 0.0880$$

What are the values of the R^2 and the adjusted R^2?

Note: the data are available within the file `Data_CAPM_returns.xlsx`.

Solution

$$\begin{aligned}
\mathrm{R}^2 &= 1 - \frac{SSR}{\sum_{t=1}^{T}\left(y_t - \overline{y}_T\right)^2} \\
&= 1 - \frac{0.0509}{0.0880} \\
&= 0.4216
\end{aligned}$$

$$\begin{aligned}
\overline{\mathrm{R}}^2 &= 1 - \frac{T-1}{T-p-1}\left(1 - \mathrm{R}^2\right) \\
&= 1 - \frac{250-1}{250-1-1}\left(1 - 0.4216\right) \\
&= 0.4193
\end{aligned}$$

2. Overall adjustment

Figure: Matlab output (R^2 and adjusted R^2)

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE          tStat       pValue
                  __________    __________    _______    __________

    (Intercept)   0.00022961    0.00090928    0.25252       0.80085
    x1                1.5452       0.11505     13.431    2.8468e-31


Number of observations: 250, Error degrees of freedom: 248
Root Mean Squared Error: 0.0143
R-squared: 0.421,  Adjusted R-Squared 0.419
F-statistic vs. constant model: 180, p-value = 2.85e-31
```

2. Overall adjustment

Figure: Excel output (R^2 and adjusted R^2)

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*								
4	Coefficient de détermination multiple	0,6489074							
5	Coefficient de détermination R^2	0,42108081							
6	Coefficient de détermination R^2	0,41874646							
7	Erreur-type	0,01433559							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de liberté*	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*			
12	Régression	1	0,03707066	0,03707066	180,3844885	2,8468E-31			
13	Résidus	248	0,05096626	0,00020551					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95,0%*	*Limite supérieure pour seuil de confiance = 95,0%*
17	Constante	0,00022961	0,00090928	0,25251848	0,800849498	-0,00156128	0,0020205	-0,00156128	0,0020205
18	Variable X 1	1,54515835	0,1150465	13,4307293	2,84683E-31	1,31856557	1,77175113	1,31856557	1,77175113

2. Overall adjustment

Example (CAPM)

We now consider the CAPM model for Microsoft:

$$z_{i,t} = \alpha + \beta z_{m,t} + \varepsilon_t$$

where $z_{i,t}$ is excess return of Microsoft at time t and $z_{i,t}$ is the market excess return. The data for Microsoft, S&P500 and Tbill (closing prices) are available from 11/1/1993 to 04/03/2003 **Question:** write a Matlab code in order to compute (1) the estimates of α and β for Microsoft, (2) the corresponding standard errors, (3) the SE of the regression, (4) the SSR, (5) the coefficient of determination and (6) the adjusted R^2.

Note: the data are available within the file `Capm_Microsoft.xls`.

2. Overall adjustment

Figure: Matlab code: residuals, SSR, R^2 and adjusted R^2

```
clear all ; clc ; close all ; format long

data=xlsread('capm.xls');
r_tbill=data(2:end,9);              % Return on the Tbill
r_msft=data(2:end,10);              % Return on MSFT
r_sp500=data(2:end,11);             % Return on the SP500
y=r_msft-r_tbill;                   % Excess return on MSFT
x=r_sp500-r_tbill;                  % Excess return on MSFT

T=length(y);                        % Sample size
X=[ones(T,1) x];                    % Matrix X (explicative variables)
beta=X\y;                           % Beta estimates

res=y-X*beta;                       % Residuals
SSR=sum(res.^2);                    % SSR
SSE=sqrt(1/(T-2)*SSR);              % Standard error (SE) of regression
R2=1-SSR/sum((y-mean(y)).^2);       % R2
adjR2=1-(T-1)/(T-2)*(1-R2);         % Adjusted R2

V=SSE^2*inv(X'*X);
std=sqrt(diag(V));
disp('beta and std ')
disp([beta std])
```

2. Overall adjustment

Figure: Matlab output: residuals, SSR, R^2 and adjusted R^2

```
SSR =

   0.177900069479780


SSE =

   0.008682245410767


R2 =

   0.454512697248921


adjR2 =

   0.454281558561315

beta and std
   0.000274089254513   0.000178812721819
   1.125056007502154   0.025370992355959
```

2. Overall adjustment

Figure: Eviews output: CAPM model for Microsoft

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/09/13 Time: 21:53
Sample(adjusted): 2 2363
Included observations: 2362 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.000274	0.000179	1.532829	0.1255
RSP500	1.125056	0.025371	44.34419	0.0000

R-squared	0.454513	Mean dependent var	0.000617
Adjusted R-squared	0.454282	S.D. dependent var	0.011753
S.E. of regression	0.008682	Akaike info criterion	-6.654227
Sum squared resid	0.177900	Schwarz criterion	-6.649343
Log likelihood	7860.642	F-statistic	1966.407
Durbin-Watson stat	2.028898	Prob(F-statistic)	0.000000

Key Concepts

1. Standard Error (SE) of the regression
2. Root Mean Squared Error (RMSE)
3. Coefficient of determination: R^2
4. Adjusted R^2

Section 3

Statistical Hypothesis Testing

Objectives

The objective of this section is to define the following concepts:

1. Null and alternative hypotheses
2. One-sided and two-sided tests
3. Rejection region, test statistic and critical value
4. Size, power and power function
5. p-value

Introduction

1. A statistical hypothesis **test** is a method of **making** a **rule of decision** (as concerned a statement about a **population parameter**) using the data of **sample**.

2. Statistical hypothesis tests define a procedure that **controls (fixes) the probability of incorrectly deciding** that a default position (null hypothesis) is incorrect.

Introduction (cont'd)

In general we distinguish two types of tests:

1. The **parametric tests** assume that the data have come from a type of probability distribution and makes inferences about the parameters of the distribution.

2. The **non-parametric tests** refer to tests that do not assume the data or population have any characteristic structure or parameters.

In this course, we only consider the parametric tests.

Introduction (cont'd)

A statistical test is based on three elements:

1. A null hypothesis and an alternative hypothesis
2. A rejection region based on a test statistic and a critical value
3. A type I error and a type II error

Introduction (cont'd)

A statistical test is based on three elements:

1. **A null hypothesis and an alternative hypothesis**
2. A rejection region based on a test statistic and a critical value
3. A type I error and a type II error

3. Statistical hypothesis testing

Definition (Hypothesis)

A **hypothesis** is a statement about a population parameter. The formal testing procedure involves a statement of the hypothesis, usually in terms of a "**null**" and an "**alternative**," conventionally denoted H_0 and H_1, respectively.

Definition (One-sided test)

A **one-sided test** has the general form:

$$\begin{array}{lll} \mathrm{H}_0 & : & \theta = \theta_0 \\ \mathrm{H}_1 & : & \theta < \theta_0 \end{array} \quad \text{or} \quad \begin{array}{l} \mathrm{H}_0 : \theta = \theta_0 \\ \mathrm{H}_1 : \theta > \theta_0 \end{array}$$

Definition (Two-sided test)

A **two-sided test** has the general form:

$$\begin{aligned} \mathrm{H}_0 &: \quad \theta = \theta_0 \\ \mathrm{H}_1 &: \quad \theta \neq \theta_0 \end{aligned}$$

Introduction (cont'd)

A statistical test is based on three elements:

1. A null hypothesis and an alternative hypothesis
2. **A rejection region based on a test statistic and a critical value**
3. A type I error and a type II error

Definition (Rejection region)

The **rejection region** is the set of values of the test statistic (or equivalently the set of samples) for which the null hypothesis is rejected. The rejection region is denoted W. For example, a standard rejection region W is of the form:

$$\mathrm{W} = \{x : T(x) \lesseqgtr c\}$$

where x denotes a sample $\{x_1, .., x_N\}$, $T(x)$ the realization of a **test statistic** (random variable) and c the **critical value**.

Remarks

1. A (hypothesis) test is thus a rule that specifies:
    1. For which sample values the decision is made to "**fail to reject $\mathbf{H_0}$**" as true;
    2. For which sample values the decision is made to "**reject $\mathbf{H_0}$**".
    3. **Never say "Accept H1", "fail to reject H1" etc..**
2. The rejection region is also called the **critical region**.
3. The complement of the rejection (critical) region is the **non-rejection region**, denoted $\overline{W}$.

Remark

The rejection region is defined as to be:

$$W = \{x : \underbrace{T(x)}_{\text{test statistic}} \lesseqgtr \underbrace{c}_{\text{critical value}}\}$$

$T(x)$ is the realization of the statistic (random variable):

$$T(X) = T(X_1, .., X_N)$$

The test statistic $T(X)$ has an exact or an asymptotic distribution D under the null H_0.

$$T(X) \underset{H_0}{\sim} D \quad \text{or} \quad T(X) \underset{H_0}{\overset{d}{\rightarrow}} D$$

3. Statistical hypothesis testing

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of i.i.d. normal random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$, $N = 100$ and $\sigma^2 = 1$. We want to test

$$\mathrm{H}_0 : m = 1.2 \qquad \mathrm{H}_1 : m = 1$$

An econometrician proposes the following decision rule:

$$\mathrm{W} = \{x : \overline{x}_N < c\}$$

Under the null, the test statistic $\overline{X}_N = N^{-1} \sum_{i=1}^{N} X_i$ has a normal distribution

$$\overline{X}_N \underset{\mathrm{H}_0}{\sim} \mathcal{N}\left(1.2, \frac{\sigma^2}{N}\right)$$

Introduction (cont'd)

A statistical test is based on three elements:

1. A null hypothesis and an alternative hypothesis
2. A rejection region based on a test statistic and a critical value
3. **A type I error and a type II error**

		Decision	
		Fail to reject H_0	Reject H_0
Truth	H_0	Correct decision	**Type I error**
	H_1	**Type II error**	Correct decision

3. Statistical hypothesis testing

Definition (Size)

The probability of a type I error is the (nominal) **size** of the test. This is conventionally denoted α and is also called the **significance level**.

$$\alpha = \Pr\left(W \middle| H_0 \right)$$

Definition (Power)

The **power** of a test is the probability that it will correctly lead to rejection of a false null hypothesis:

$$\text{power} = \Pr\left(\mathrm{W} \middle| \mathrm{H}_1\right) = 1 - \beta$$

where β denotes the probability of type II error, i.e. $\beta = \Pr\left(\overline{\mathrm{W}} \middle| \mathrm{H}_1\right)$ and $\overline{\mathrm{W}}$ denotes the non-rejection region.

There is a **trade-off** between the size and the probability of type II error:

- Decreasing the size (trough a variation of the critical value) induces an increase of the probability of type II error, and thus a decrease of the power.
- Increasing the size (trough a variation of the critical value) induces an decrease of the probability of type II error, and thus an increase of the power.

Solution (critical value)

The (nominal) **size** α *is* **fixed by the analyst** *and the critical value is deduced from* α.

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of i.i.d. continuous random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$, $N = 100$ and $\sigma^2 = 1$. We want to test

$$\mathrm{H}_0 : m = m_0 = 1.2 \qquad \mathrm{H}_1 : m = m_1 = 1$$

An econometrician proposes the following rule of decision:

$$\mathrm{W} = \{x : \overline{x}_N < c\}$$

Questions: What is the value of the critical value for a level $\alpha = 5\%$?

3. Statistical hypothesis testing

Solution

The rejection region is W$= \{x : \overline{x}_N < c\}$. Under the null $\text{H}_0 : m = m_0$:

$$\overline{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i \underset{\text{H}_0}{\sim} \mathcal{N}\left(m_0, \frac{\sigma^2}{N}\right)$$

So, by definition:

$$\begin{aligned}
\alpha &= \Pr\left(\text{W}|\,\text{H}_0\right) \\
&= \Pr\left(\overline{X}_N < c\,\middle|\,\text{H}_0\right) \\
&= \Pr\left(\frac{\overline{X}_N - m_0}{\sigma/\sqrt{N}} < \frac{c - m_0}{\sigma/\sqrt{N}}\,\middle|\,\text{H}_0\right) \\
&= \Phi\left(\frac{c - m_0}{\sigma/\sqrt{N}}\right)
\end{aligned}$$

where $\Phi(.)$ is the cdf of the standard normal distribution.

Solution (cont'd)

We know that:

$$\alpha = \Phi\left(\frac{c - m_0}{\sigma/\sqrt{N}}\right) \Leftrightarrow \Phi^{-1}(\alpha) = \frac{c - m_0}{\sigma/\sqrt{N}}$$

So, the critical value that corresponds to a significance level of α is:

$$c = m_0 + \frac{\sigma}{\sqrt{N}}\Phi^{-1}(\alpha)$$

NA: if $m_0 = 1.2$, $m_1 = 1$, $N = 100$, $\sigma^2 = 1$ and $\alpha = 5\%$ with $\Phi^{-1}(0.05) = -1.6449$ then the rejection region is

$$\mathrm{W} = \{x : \overline{x}_N < 1.0355\}$$

Summary

1. Fix the **significance level** of the test and compute the **critical value** of the test
2. If the realization of the test statistic belongs to the rejection region W, **reject** H_0.

$$T(x) \in \text{W} \quad => \text{Reject } \text{H}_0 \text{ for a significance level } \alpha\%$$

If the realization of the test statistic does not belongs to the rejection region W, **do not reject** H_0.

$$T(x) \notin \text{W} \quad => \text{ Fail to reject } \text{H}_0 \text{ for a significance level } \alpha\%$$

Remark

1. Changing the significance level induces a change in the critical value, in hence in your conclusion.

2. That is why it is particularly important to mention the choice of the **significance level** α in your decision...

For a 5% significance level, we reject the null hypothesis....

For a 5% significance level, we fail to reject the null hypothesis....

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of *i.i.d.* continuous random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$ with $\sigma^2 = 1$ and $N = 100$. We want to test

$$\mathrm{H}_0 : m = 1.2 \qquad \mathrm{H}_1 : m = 1$$

The rejection region for a significance level is:

$$\mathrm{W} = \left\{ x : \overline{x}_N < m_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}\left(\alpha\right) \right\}$$

Question: if the realization of the sample mean is equal to $\overline{x}_N = 1.13$, what is the conclusion of the test for a significance level of 5% and 30%?

Solution

For a significance level $\alpha = 5\%$ we have

$$\begin{aligned} m_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha) &= 1.2 + \frac{1}{10} \times \Phi^{-1}(0.05) \\ &= 1.2 + \frac{1}{10} \times (-1.6449) \\ &= 1.0355 \end{aligned}$$

$$\mathrm{W} = \{x : \overline{x}_N < 1.0355\}$$

If we observe $\overline{x}_N = 1.13$, this realization does not belong to the rejection region:

$$\overline{x}_N \notin \mathrm{W}$$

For a significance level of 5%, we **do not reject the null** hypothesis $\mathrm{H}_0 : m = 1.2$

Solution (con'td)

For a significance level $\alpha = 30\%$ we have

$$\begin{aligned} m_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha) &= 1.2 + \frac{1}{10} \times \Phi^{-1}(0.30) \\ &= 1.2 + \frac{1}{10} \times (-0.5244) \\ &= 1.1476 \end{aligned}$$

$$\mathrm{W} = \{x : \overline{x}_N < 1.1476\}$$

If we observe $\overline{x}_N = 1.13$, this realization belongs to the rejection region:

$$\overline{x}_N \in \mathrm{W}$$

For a significance level of 30%, we **reject the null** hypothesis $\mathrm{H}_0 : m = 1.2$

3. Statistical hypothesis testing

Fact (two-sided tests)

If one considers a **two-sided test** *and if the distribution of the test statistics is* **symmetric** *(Normal, Student, etc.), the critical value has to be computed with a risk level of $\alpha/2$ to ensure that the size of the test is equal to α.*

3. Statistical hypothesis testing

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of $i.i.d.$ continuous random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$ where σ^2 is known. We want to test

$$\mathrm{H}_0 : m = m_0 \quad \text{vs} \quad \mathrm{H}_1 : m \neq m_0$$

The test statistic $\overline{X}_N$ has a normal distribution (symmetric) under the null H_0.The rejection region of the two-sided test of size α is then defined by:

$$\mathrm{W} = \left\{ x : \left| \frac{\overline{x}_N - m_0}{\sigma / \sqrt{N}} \right| > \underbrace{\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)}_{\text{critical value based on } \alpha/2} \right\}$$

Solution

- The decision "Reject H0" or "fail to reject H0" is not so informative!
- Indeed, there is some "arbitrariness" to the choice of α (level).
- Another strategy is to ask, for every α, whether the test rejects at that level.
- Another alternative is to use the so-called **p-value**—the smallest level of significance at which H_0 would be rejected given the value of the test-statistic.

Definition (p-value)

Suppose that for every $\alpha \in [0,1]$, one has a size α test with rejection region W_{α}. Then, the **p-value** is defined to be:

$$\text{p-value} = \inf \left\{ \alpha : T\left(y\right) \in W_{\alpha} \right\}$$

The p-value is the smallest level at which one can reject H_0.

3. Statistical hypothesis testing

The p-value is a **measure of evidence against** H_0:

p-value	evidence
< 0.01	Very strong evidence against H_0
$0.01 - 0.05$	Strong evidence against H_0
$0.05 - 0.10$	Weak evidence against H_0
> 0.10	Little or no evidence against H_0

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of *i.i.d.* continuous random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$ with $\sigma^2 = 1$ and $N = 100$. We want to test

$$\mathrm{H}_0 : m = 1.2 \qquad \mathrm{H}_1 : m = 1$$

The rejection region for a significance level is:

$$\mathrm{W} = \left\{ x : \overline{x}_N < m_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}\left(\alpha\right) \right\}$$

Question: if the realization of the sample mean is equal to 1.13, determine the rejection regions for different levels between 1% and 30%, and compute an approximation of the p-value.

3. Statistical hypothesis testing

Solution

$$\mathrm{W} = \left\{ x : \overline{x}_N < m_0 + \frac{\sigma}{\sqrt{N}} \Phi^{-1}(\alpha) \right\}$$

The realization of the sample mean is equal to 1.13, so the p-value ranges between 0.24 and 0.25.

α	$\Phi^{-1}(\alpha)$	$m_0 + \frac{\sigma}{\sqrt{n}} \Phi^{-1}(\alpha)$	Conclusion
0.01	-2.3263	0.9674	No rejection of H_0
0.05	-1.6449	1.0355	No rejection of H_0
0.10	-1.2816	1.0718	No rejection of H_0
0.15	-1.0364	1.0964	No rejection of H_0
0.20	-0.8416	1.1158	No rejection of H_0
0.24	-0.7063	1.1294	No rejection of H_0
0.25	-0.6745	1.1326	Rejection of H_0
0.30	-0.5244	1.1476	Rejection of H_0

Remarks

1. For each type of test (one-sided, two-sided) and given the properties of the distribution of the test statistic, there exist different formula to compute the p-value (without computing the critical values for different levels).

2. For a significance level α fixed by the analyst, the decision rule is

$$\text{p-value} < \alpha \implies \text{Reject } H_0 \text{ for a significance level } \alpha\%$$

$$\text{p-value} > \alpha \implies \text{Do not reject } H_0 \text{ for a significance level } \alpha\%$$

Example (Test on the mean)

Consider a sequence $X_1, .., X_N$ of $i.i.d.$ continuous random variables with $X_i \sim \mathcal{N}\left(m, \sigma^2\right)$ with $\sigma^2 = 1$ and $N = 100$. We want to test

$$\mathrm{H}_0 : m = 1.2 \qquad \mathrm{H}_1 : m = 1$$

The p-value of the test is equal to 0.2420. Question: what is your conclusion for a significance level $\alpha = 5\%$?

Solution

The p-value is larger than the significance level $\alpha = 5\%$. For a significance level of 5%, we do not reject the null $\mathrm{H}_0 : m = 1.2$.

Summary: For a test, you have always two ways to proceed

1. **Critical value approach:** Compare the realization of the test statistic to the rejection region for a fixed level of risk

$$T(x) \in \mathrm{W} \;\; => \text{Reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

$$T(x) \notin \mathrm{W} \;\; => \;\; \text{Do not reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

2. **p-value approach:** Compare the p-value to the significance level $\alpha\%$

$$\text{p-value} < \alpha \;\; => \text{Reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

$$\text{p-value} > \alpha => \;\; \text{Do not reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

Key concepts

1. Null and alternative hypotheses
2. One-sided and two-sided tests
3. Rejection region, test statistic and critical value
4. Type I and type II errors
5. Size and power
6. p-value
7. Critical value and p-value approaches

Section 4

Individual and Global Significance Tests

Objectives

1. To test the value of an individual parameter in a linear model
2. To compute a Student t-test statistic, its p-value and its critical region
3. To test some global restrictions on the parameters
4. To compute a Fisher test statistic, its p-value and its critical region
5. To compute the p-value associated to a Fisher test statistic
6. To introduce the global F test
7. To apply all these tests to the CAPM model.

Model

Consider the **multiple linear regression model** (cf. Chapter 2):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\mathbf{y}$ is a $T \times 1$ vector of observations y_i for $i = 1, .., T$
- $\mathbf{X}$ is a $T \times K$ matrix of K explicative variables $\mathbf{x}_{t,k}$ for $k = 1, ., K$ and $t = 1, .., T$
- $\boldsymbol{\varepsilon}$ is a $T \times 1$ vector of error terms ε_t.
- $\boldsymbol{\beta} = (\beta_1..\beta_K)^\top$ is a $K \times 1$ vector of parameters

4. Individual and global significance tests

Fact (Assumptions)

We assume that the multiple linear regression model satisfy the **assumptions A1-A6**

A1: linearity	The model is linear with β
A2: identification	$\mathbf{X}$ is an $T \times K$ matrix with rank K
A3: exogeneity	$\mathbb{E}\left(\varepsilon \middle| \mathbf{X} \right) = \mathbf{0}_{T \times 1}$
A4: spherical error terms	$\mathbb{V}\left(\varepsilon \middle| \mathbf{X} \right) = \sigma^2 \mathbf{I}_T$
A5: data generation	$\mathbf{X}$ may be fixed or random
A6: normal distribution	$\varepsilon \vert \mathbf{X} \sim \mathcal{N}\left(\mathbf{0}_{T \times 1}, \sigma^2 \mathbf{I}_T \right)$

4. Individual and global significance tests

Parametric tests

The β_k are unknown features of the population, but:

1. One can formulate a **hypothesis about their value**;
2. One can construct a test statistic with a known **distribution** under H_0;
3. One can take a "decision" meaning "reject H0" if the value of the test statistic is too unlikely.

4. Individual and global significance tests

For that, we introduce two types of tests

1. Individual significance tests: the **Student** tests or t-tests:

2. Global significance tests: the **Fisher** tests of F-tests

Subsection 4.1

Individual Significance Tests: Student Tests

Two types of tests concerning one individual parameter

1. **Two-sided test**

$$\begin{aligned} \mathrm{H}_0 &: \beta_k = a_k \\ \mathrm{H}_1 &: \beta_k \neq a_k \end{aligned}$$

where a_k is some reference value.

2. **One-sided test**

$$\begin{aligned} \mathrm{H}_0 &: \beta_k = a_k \\ \mathrm{H}_1 &: \beta_k < a_k \end{aligned} \quad \text{or} \quad \begin{aligned} \mathrm{H}_0 : \beta_k = a_k \\ \mathrm{H}_1 : \beta_k > a_k \end{aligned}$$

4.1. The Student test

Why is this important?

- Because $a_k = 0$ means that there is no merit to using the k^{th} independent variable as a predictor (regressor);
- Because it allows testing the statistical significance of each estimate and to select some of the explanatory variables.

Reminder (cf. chapter 1)

Fact (Linear regression model)

Under the **assumption A6 (normality)**, *the estimators* $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\sigma}^2$ *have a finite sample distribution given by:*

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\right)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2}\left(T-K\right) \sim \chi^2\left(T-K\right)$$

Moreover, $\widehat{\boldsymbol{\beta}}$ *and* $\widehat{\sigma}^2$ *are independent. This result holds whether or not the matrix* $\mathbf{X}$ *is considered as random. In this last case, the distribution of* $\widehat{\boldsymbol{\beta}}$ *is conditional to* $\mathbf{X}$.

4.1. The Student test

Property: Any subset of $\widehat{\boldsymbol{\beta}}$ has a joint normal distribution.

$$\widehat{\beta}_k \sim \mathcal{N}\left(\beta_k, \sigma^2 m_{kk}\right)$$

where m_{kk} is k^{th} diagonal element of $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.

4.1. The Student test

Reminder

If X and Y are two independent random variables such that

$$X \sim \mathcal{N}(0,1)$$

$$Y \sim \chi^2(\theta)$$

then the variable Z defined as to be

$$Z = \frac{X}{\sqrt{Y/\theta}}$$

has a Student's t-distribution with θ degrees of freedom

$$Z \sim t_{(\theta)}$$

Student test statistic

Consider a test with the null:

$$\mathrm{H}_0 : \beta_k = a_k$$

Under the null H_0 :

$$\frac{\widehat{\beta}_k - a_k}{\sigma\sqrt{m_{kk}}} \underset{\mathrm{H}_0}{\sim} \mathcal{N}(0,1)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2}(T-K) \underset{\mathrm{H}_0}{\sim} \chi^2(T-K)$$

and these two variables are independent...

Student test statistic (cont'd)

$$\frac{\widehat{\beta}_k - a_k}{\sigma\sqrt{m_{kk}}} \underset{\mathrm{H}_0}{\sim} \mathcal{N}(0,1)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2}(T-K) \underset{\mathrm{H}_0}{\sim} \chi^2(T-K)$$

So, under the null H_0 we have:

$$\frac{\frac{\widehat{\beta}_k - a_k}{\sigma\sqrt{m_{kk}}}}{\sqrt{\frac{\widehat{\sigma}^2}{\sigma^2}\frac{(T-K)}{(T-K)}}} = \frac{\widehat{\beta}_k - a_k}{\widehat{\sigma}\sqrt{m_{kk}}} \underset{\mathrm{H}_0}{\sim} t_{(T-K)}$$

4.1. The Student test

Definition (Student t-statistic)

Under the null $\text{H}_0 : \beta_k = a_k$, the **Student test-statistic** or **t-statistic** is defined to be:

$$\text{T}_k = \frac{\widehat{\beta}_k - a_k}{\widehat{\text{se}}\left(\widehat{\beta}_k\right)} \underset{\text{H}_0}{\sim} t_{(T-K)}$$

where T is the number of observations, K is the number of explanatory variables (including the constant term), $t_{(T-K)}$ is the Student t-distribution with $T - K$ degrees of freedom and

$$\widehat{\text{se}}\left(\widehat{\beta}_k\right) = \widehat{\sigma}\sqrt{m_{kk}}$$

with m_{kk} is k^{th} diagonal element of $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$.

4.1. The Student test

Remarks

1. Under the assumption A6 (normality) and under the null $H_0 : \beta_k = a_k$, the Student test-statistic has an **exact (finite sample)** distribution.

$$\mathrm{T}_k \underset{H_0}{\sim} t_{(T-K)}$$

2. The term $\widehat{\mathrm{se}}\left(\widehat{\beta}_k\right)$ denotes the estimator of the SE of the OLS estimator $\widehat{\beta}_k$ and it corresponds to the square root of the k^{th} diagonal element of $\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\beta}}\right)$ (cf. Chapter 2):

$$\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\beta}}\right) = \widehat{\sigma}^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

4.1. The Student test

Fact

All the regression analysis tools (R, Matlab, Python, Excel, Stata, etc.) report the t-statistics and their p-value associated to the **test of nullity** *of the individual parameters*

$$H_0 : \beta_k = 0 \qquad H_1 : \beta_k \neq 0$$

4.1. The Student test

Figure: Matlab output (t-stats and p-values)

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE           tStat      pValue
                   __________     __________   _______    __________

    (Intercept)    0.00022961     0.00090928   0.25252       0.80085
    x1                 1.5452        0.11505    13.431    2.8468e-31

Number of observations: 250, Error degrees of freedom: 248
Root Mean Squared Error: 0.0143
R-squared: 0.421,  Adjusted R-Squared 0.419
F-statistic vs. constant model: 180, p-value = 2.85e-31
```

4.1. The Student test

Figure: Excel output (tstats and p-values)

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*								
4	Coefficient de détermination multiple	0,6489074							
5	Coefficient de détermination R^2	0,42108081							
6	Coefficient de détermination R^2	0,41874646							
7	Erreur-type	0,01433559							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de liberté*	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*			
12	Régression	1	0,03707066	0,03707066	180,384488	2,8468E-31			
13	Résidus	248	0,05096626	0,00020551					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	***Statistique t***	***Probabilité***	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95,0%*	*Limite supérieure pour seuil de confiance = 95,0%*
17	Constante	0,00022961	0,00090928	**0,25251848**	**0,8008495**	-0,0015613	0,0020205	-0,0015613	0,0020205
18	Variable X 1	1,54515835	0,1150465	**13,4307293**	**2,8468E-31**	1,31856557	1,77175113	1,31856557	1,77175113
19									

Example (CAPM)

We consider the CAPM model

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \varepsilon_t$$

Question: Given the elements reported by Excel (cf. previous slide), compute the realizations of the Student test-statistics for the following hypotheses

$$\begin{array}{lllll} \mathrm{H}_0 : \beta_2 = 0 & \quad & \mathrm{H}_0 : \beta_2 = 0 & \quad \mathrm{H}_0 : \beta_2 = 0 & \quad \mathrm{H}_0 : \beta_2 = 1 \\ \mathrm{H}_1 : \beta_2 < 0 & \quad & \mathrm{H}_1 : \beta_2 > 0 & \quad \mathrm{H}_0 : \beta_2 \neq 0 & \quad \mathrm{H}_0 : \beta_2 \neq 1 \end{array}$$

Note: the data are available within the file `Data_CAPM_returns.xlsx`.

4.1. The Student test

Solution

The three tests

$$\begin{array}{lcl} H_0 & : & \beta_2 = 0 \\ H_1 & : & \beta_2 < 0 \end{array} \qquad \begin{array}{l} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 > 0 \end{array} \qquad \begin{array}{l} H_0 : \beta_2 = 0 \\ H_0 : \beta_2 \neq 0 \end{array}$$

have the same t-statistic defined as

$$T_{\beta_2=0}\left(y\right) = \frac{\widehat{\beta}_2 - 0}{\widehat{se}\left(\widehat{\beta}_2\right)} = \frac{\widehat{\beta}_2}{\widehat{se}\left(\widehat{\beta}_2\right)} = \frac{1.5451}{0.1150} = 13.4307$$

Note: be careful, the p-values of unilateral and bilateral tests are not identical

4.1. The Student test

Solution (cont'd)

Concerning the latest test

$$\begin{aligned} \mathrm{H}_0 &: \quad \beta_2 = 1 \\ \mathrm{H}_0 &: \quad \beta_2 \neq 1 \end{aligned}$$

The t-statistic is defined as

$$\mathrm{T}_{\beta_2=1}\left(y\right) = \frac{\widehat{\beta}_2 - 1}{\widehat{\mathrm{se}}\left(\widehat{\beta}_2\right)} = \frac{1.5451 - 1}{0.1150} = 4.7386$$

4.1. The Student test

As for any test, you two equivalent decision rules for the t-test

1. **Critical value approach:** Compare the realization of the test statistic to the rejection region for a fixed level of risk

$$\text{if } \mathrm{T}_k \in \mathrm{W} \quad => \text{Reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

2. **p-value approach:** Compare the p-value to the significance level $\alpha\%$

$$\text{if p-value} < \alpha \quad => \text{Reject } \mathrm{H}_0 \text{ for a significance level } \alpha\%$$

Let us define rejection region and the p-values...

Definition (rejection region - left tailed test)

Consider the **one-sided (left tailed) test**:

$$\mathrm{H}_0 : \beta_k = a_k \quad \mathrm{H}_1 : \beta_k < a_k$$

The **rejection region** is defined as to be:

$$\mathrm{W} = \{y : \mathrm{T}_k(y) < c_\alpha\}$$

where c_α is the α-quantile of the Student's t-distribution with $T - K$ degrees of freedom and $\mathrm{T}_k(y)$ is the realization of the Student test-statistic.

Example (One-sided test)

Consider the CAPM model for Microsoft (cf. Chapter 2) and the following results (Eviews). We want to test the beta of MSFT as

$$H_0 : \beta_{MSFT} = 1 \quad \text{against} \quad H_1 : \beta_{MSFT} < 1$$

Question: give a conclusion for a nominal size of 5%.

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/30/13 Time: 17:15
Sample: 2 21
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001205	0.986860	0.3368
RSP500	1.989787	0.314210	6.332664	0.0000

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

4.1. The Student test

Solution

Step 1: compute the t-statistic

$$\mathrm{T}_{MSFT}(y) = \frac{\widehat{\beta}_{MSFT} - 1}{\widehat{\mathrm{se}}\left(\widehat{\beta}_{MSFT}\right)} = \frac{1.9898 - 1}{0.3142} = 3.1501$$

Step 2: Determine the rejection region for a nominal size $\alpha = 5\%$ and $T = 20$:

$$\mathrm{T}_{MSFT} \underset{\mathrm{H}_0}{\sim} t_{(20-2)} \qquad c_{0.05} = -1.7341$$

$$\mathrm{W} = \{y : \mathrm{T}_k(y) < -1.7341\}$$

Conclusion: for a significance level of 5%, we fail to reject the null $\mathrm{H}_0 : \beta_{MSFT} = 1$ against $\mathrm{H}_1 : \beta_{MSFT} < 1$.

Definition (rejection region - right tailed test)

Consider the **one-sided (right tailed) test**:

$$\mathrm{H}_0 : \beta_k = a_k \qquad \mathrm{H}_1 : \beta_k > a_k$$

The **rejection region** is defined as to be:

$$\mathrm{W} = \{ y : \mathrm{T}_k (y) > c_{1-\alpha} \}$$

where $c_{1-\alpha}$ is the $1-\alpha$-quantile of the Student's t-distribution with $T-K$ degrees of freedom and $\mathrm{T}_k (y)$ is the realization of the Student test-statistic.

Example (One-sided test)

Consider the CAPM model for Microsoft (cf. Chapter 2) and the following results (Eviews). We want to test the beta of MSFT as

$$H_0 : \beta_{MSFT} = 1 \quad \text{against} \quad H_1 : \beta_{MSFT} > 1$$

Question: give a conclusion for a nominal size of 5%.

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/30/13 Time: 17:15
Sample: 2 21
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001205	0.986860	0.3368
RSP500	1.989787	0.314210	6.332664	0.0000

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

Solution

Step 1: compute the t-statistic

$$\mathrm{T}_{MSFT}(y) = \frac{\widehat{\beta}_{MSFT} - 1}{\widehat{\mathrm{se}}\left(\widehat{\beta}_{MSFT}\right)} = \frac{1.9898 - 1}{0.3142} = 3.1501$$

Step 2: Determine the rejection region for a nominal size $\alpha = 5\%$ and a sample size $T = 20$.

$$\mathrm{T}_{MSFT} \underset{\mathrm{H}_0}{\sim} t_{(20-2)} \qquad c_{0.95} = 1.7341$$

$$\mathrm{W} = \{y : \mathrm{T}_k(y) > 1.7341\}$$

Conclusion: for a significance level of 5%, we reject the null $\mathrm{H}_0 : \beta_{MSFT} = 1$ against $\mathrm{H}_1 : \beta_{MSFT} > 1$ □

Definition (rejection region for two-sided or bilateral test)

Consider the **two-sided test**

$$\mathrm{H}_0 : \beta_k = a_k \qquad \mathrm{H}_1 : \beta_k \neq a_k$$

The **rejection region** for a significance level of $\alpha\%$ (say, 5%) is

$$\mathrm{W} = \left\{ y : \left| \mathrm{T}_k \left(y \right) \right| > c_{1-\alpha/2} \right\}$$

where $c_{1-\alpha/2}$ is the $1-\alpha/2$ (say, 97.5%) critical value of a Student t-distribution with $T-K$ degrees of freedom and $\mathrm{T}_k \left(y \right)$ is the realization of the Student test-statistic.

4.1. The Student test

Example (Two-sided test)

Consider the CAPM model (cf. Chapter 2) and the following results (Eviews). We want to test the beta of MSFT as

$$H_0 : \beta_{MSFT} = 1 \quad \text{against} \quad H_1 : \beta_{MSFT} \neq 1$$

Question: give a conclusion for a nominal size of 5%.

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/30/13 Time: 17:15
Sample: 2 21
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001205	0.986860	0.3368
RSP500	1.989787	0.314210	6.332664	0.0000

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

Solution

Step 1: compute the t-statistic

$$\mathrm{T}_{MSFT}(y) = \frac{\widehat{\beta}_{MSFT} - 1}{\widehat{\mathrm{se}}\left(\widehat{\beta}_{MSFT}\right)} = \frac{1.9898 - 1}{0.3142} = 3.1501$$

Step 2: Determine the rejection region for a nominal size $\alpha = 5\%$ and a sample size $T = 20$.

$$\mathrm{T}_{MSFT} \underset{\mathrm{H}_0}{\sim} t_{(20-2)} \qquad c_{0.975} = 2.1009$$

$$\mathrm{W} = \{y : |\mathrm{T}_k(y)| > 2.1009\}$$

Conclusion: for a significance level of 5%, we reject the null $\mathrm{H}_0 : \beta_{MSFT} = 1$ against $\mathrm{H}_1 : \beta_{MSFT} \neq 1$ $\square$

Summary: rejection regions for a Student test

H_0	H_1	**Rejection region**
$\beta_k = a_k$	$\beta_k > a_k$	$W = \{y : T_k(y) > c_{1-\alpha}\}$
$\beta_k = a_k$	$\beta_k < a_k$	$W = \{y : T_k(y) < c_{\alpha}\}$
$\beta_k = a_k$	$\beta_k \neq a_k$	$W = \{y : \lvert T_k(y) \rvert > c_{1-\alpha/2}\}$

Note: c_{β} denotes the β-quantile (critical value) of the Student t-distribution with $T - K$ degrees of freedom.

4.1. The Student test

Definition (P-values)

The **p-values** of Student tests are equal to:

$$\text{Two-sided test: } \text{p-value} = 2 \times F_{T-K}\left(-\left|\mathrm{T}_k\left(y\right)\right|\right)$$

$$\text{Right tailed test: } \text{p-value} = 1 - F_{T-K}\left(\mathrm{T}_k\left(y\right)\right)$$

$$\text{Left tailed test: } \text{p-value} = F_{T-K}\left(-\mathrm{T}_k\left(y\right)\right)$$

where $\mathrm{T}_k\left(y\right)$ is the realization of the Student test-statistic and $F_{T-K}\left(.\right)$ the cdf of the Student's t-distribution with $T-K$ degrees of freedom.

4.1. The Student test

Example (One-sided test)

Consider the CAPM model for Microsoft. We want to test:

$$\mathrm{H}_0 : c = 0 \quad \text{against} \quad \mathrm{H}_1 : c \neq 0$$

$$\mathrm{H}_0 : \beta_{MSFT} = 0 \quad \text{against} \quad \mathrm{H}_1 : \beta_{MSFT} \neq 0$$

Question: find the corresponding p-values.

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/30/13 Time: 18:45
Sample: 2 21
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001205	0.986860	
RSP500	1.989787	0.314210	6.332664	

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

4.1. The Student test

Solution

Since we consider two-sided tests with $T = 20$ and $K = 2$:

$$\text{p-value}_c = 2 \times F_{18}\left(-\left|T_c\left(y\right)\right|\right) = 2 \times F_{18}\left(-0.9868\right) = 0.3368$$

$$\text{p-value}_c = 2 \times F_{18}\left(-\left|T_{MSFT}\left(y\right)\right|\right) = 2 \times F_{18}\left(-6.3326\right) = 5.7e^{-006}$$

Dependent Variable: RMSFT
Method: Least Squares
Date: 11/30/13 Time: 18:51
Sample: 2 21
Included observations: 20

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001205	0.986860	0.3368
RSP500	1.989787	0.314210	6.332664	0.0000

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

2. Overall adjustment

Example (CAPM)

Consider the CAPM model for Intel Corp (8/2/2017-8/17/2018)

$$z_{\texttt{intel},t} = \alpha + \beta z_{\texttt{market},t} + \varepsilon_t$$

Question: when the CAPM is expressed with excess returns (as in our case), we can test the mean-variance efficiency by testing the nullity of the intercept

$$\mathrm{H}_0 : \alpha = 0 \quad \mathrm{H}_1 : \alpha \neq 0$$

Given the Excel results (next slide), what is your conclusion for a significance level $\alpha = 10\%$? Please use both the **critical value** and the **p-value approaches**.

Note: the data are available within the file "Data_CAPM_returns.xlsx".

4.1. The Student test

Figure: CAPM model for Intel Corp.

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*								
4	Coefficient de détermination multiple	0,6489074							
5	Coefficient de détermination R^2	0,42108081							
6	Coefficient de détermination R^2	0,41874646							
7	Erreur-type	0,01433559							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de liberté*	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*			
12	Régression	1	0,03707066	0,03707066	180,384488	2,8468E-31			
13	Résidus	248	0,05096626	0,00020551					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95,0%*	*Limite supérieure pour seuil de confiance = 95,0%*
17	Constante	0,00022961	0,00090928	0,25251848	0,8008495	-0,0015613	0,0020205	-0,0015613	0,0020205
18	Variable X 1	1,54515835	0,1150465	13,4307293	2,8468E-31	1,31856557	1,77175113	1,31856557	1,77175113

4.1. The Student test

Solution

p-value approach: The p-value computed by Excel is equal to 0.8000, so we have

$$\text{p-value} = 0.8000 \; > \; \text{significance level} \; = 10\%$$

Conclusion: for a significance level of 10%, we do no reject the null $\text{H}_0 : \alpha = 0$ against $\text{H}_1 : \alpha \neq 0$.

Solution (cont'd)

Critical value approach: The realization of the t-statistic (computed by Excel) is equal to

$$T_\alpha = 0.2525$$

The critical region is defined by

$$W = \{y : |T_k(y)| > 1.6510\}$$

$$T_\alpha \underset{H_0}{\sim} t_{(250-2)} \qquad c_{1-\frac{0.10}{2}} = c_{0.95} = 1.6510$$

So, the t-statistic does not belongs to the critical region for $\alpha = 10\%$.

Conclusion: for a significance level of 10%, we do no reject the null $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$.

Fact (Student test with large sample)

For a **large sample** *size* T

$$T_k \underset{H_0}{\sim} t_{(T-K)} \approx \mathcal{N}(0,1)$$

Then, the rejection region for a Student two-sided test becomes

$$W = \left\{ y : |T_k(y)| > \Phi^{-1}(1-\alpha/2) \right\}$$

where $\Phi(.)$ *denotes the cdf of the standard normal distribution. For* $\alpha = 5\%$, $\Phi^{-1}(0.975) = 1.96$, *so we have:*

$$W = \{ y : |T_k(y)| > 1.96 \}$$

4.1. The Student test

Non-normality of the error term

- Consider a case of a semi-parametric model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon}|\,\mathbf{X} \sim \text{Unknown distribution}$$

$$\mathbb{E}\left(\boldsymbol{\varepsilon}|\,\mathbf{X}\right) = \mathbf{0}_{T\times 1} \quad \mathbb{V}\left(\boldsymbol{\varepsilon}|\,\mathbf{X}\right) = \sigma^2 \mathbf{I}_T$$

- Without the normality assumption, the two following results **do not hold**

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta},\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2}\left(T-K\right) \sim \chi^2\left(T-K\right)$$

4.1. The Student test

Non-normality of the error term

- So, the t-test statistic **does not have** a Student distribution

$$T_k = \frac{\widehat{\beta}_k - a_k}{\widehat{\text{se}}\left(\widehat{\beta}_k\right)} \underset{H_0}{\sim} \text{Unknown distribution}$$

- However, we know that

$$\sqrt{T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}\right)$$

where

$$\mathbf{Q} = \text{plim} \frac{1}{T} \mathbf{X}^\top \mathbf{X} = \mathbb{E}_X \left(\mathbf{x}_i^\top \mathbf{x}_i\right)$$

or equivalently

$$\widehat{\boldsymbol{\beta}} \overset{asy}{\approx} \mathcal{N}\left(\boldsymbol{\beta}_0, \frac{\sigma^2}{T} \mathbf{Q}^{-1}\right)$$

4.1. The Student test

4.1. The Student test

Subsection 4.2

Global Significance Tests:

Fisher Tests

4.2. The Fisher test

Consider the two-sided test associated to p **linear constraints** on the parameters β_k :

$$\begin{aligned} \mathrm{H}_0 &: \quad \mathbf{R}\beta = \mathbf{q} \\ \mathrm{H}_1 &: \quad \mathbf{R}\beta \neq \mathbf{q} \end{aligned}$$

where $\mathbf{R}$ is a $p \times K$ matrix and $\mathbf{q}$ is a $p \times 1$ vector.

Example (Linear constraints)

If $K = 4$ and if we want to test $\text{H}_0 : \beta_1 + \beta_2 = 0$ and $\beta_2 - 3\beta_3 = 4$, then we have $p = 2$ linear constraints with:

$$\underset{(2\times4)}{\mathbf{R}}\ \underset{(4,1)}{\boldsymbol{\beta}} = \underset{(2\times1)}{\mathbf{q}}$$

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -3 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 4 \end{pmatrix}$$

Example (Linear constraints)

If $K = 4$ and if we want to test $\text{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$, then we have $p = 3$ linear constraints with:

$$\underset{(3\times4)}{\mathbf{R}} \underset{(4,1)}{\boldsymbol{\beta}} = \underset{(3\times1)}{\mathbf{q}}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

4.2. The Fisher test

Definition (Fisher test-statistic)

Under assumptions A1-A6 (cf. Chapter 2), the **Fisher test-statistic** is defined as to be:

$$\mathrm{F} = \frac{1}{p}\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{q}\right)^{\top}\left(\widehat{\sigma}^{2}\mathbf{R}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{R}^{\top}\right)^{-1}\left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{q}\right)$$

where $\widehat{\boldsymbol{\beta}}$ denotes the OLS estimator. Under the null $\mathrm{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{q}$**,** the F-statistic has a Fisher exact (finite sample) distribution

$$\mathrm{F} \underset{\mathrm{H}_0}{\sim} F_{(p, T-K)}$$

4.2. The Fisher test

Reminder

If X and Y are two independent random variables such that

$$X \sim \chi^2(\theta_1)$$

$$Y \sim \chi^2(\theta_2)$$

then the variable Z defined by

$$Z = \frac{X/\theta_1}{Y/\theta_2}$$

has a Fisher distribution with θ_1 and θ_2 degrees of freedom

$$Z \sim F_{(\theta_1,\theta_2)}$$

Intuition of the Proof

We can base the test of H_0 on the **Wald criterion**:

$$\underset{(1\times 1)}{W} = \underset{(1\times p)}{\mathbf{m}^{\top}} \left(\mathbb{V} \underset{p\times p}{(\mathbf{m})} \right)^{-1} \underset{p\times 1}{\mathbf{m}}$$

$$= \left(\mathbf{R}\widehat{\beta} - \mathbf{q} \right)^{\top} \left(\sigma^2 \mathbf{R} \left(\mathbf{X}^{\top}\mathbf{X} \right)^{-1} \mathbf{R}^{\top} \right)^{-1} \left(\mathbf{R}\widehat{\beta} - \mathbf{q} \right)$$

Under assumption A6 (normality)

$$W \underset{\text{H}_0}{\sim} \chi^2 (p)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2} (T - K) \sim \chi^2 (T - K)$$

These two variables are independent.

Intuition of the Proof (cont'd)

$$W \underset{\mathrm{H}_0}{\sim} \chi^2(p)$$

$$\frac{\widehat{\sigma}^2}{\sigma^2}(T-K) \sim \chi^2(T-K)$$

These two variables are independent. So, their ratio has a Fisher distribution

$$\mathrm{F} = \frac{\frac{W}{p}}{\frac{\widehat{\sigma}^2}{\sigma^2}\frac{(T-K)}{(T-K)}} \underset{\mathrm{H}_0}{\sim} F_{(p,T-K)}$$

$$\mathrm{F} = \frac{1}{p}\left(\mathbf{R}\widehat{\beta}-\mathbf{q}\right)^{\top}\left(\widehat{\sigma}^2\mathbf{R}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{R}^{\top}\right)^{-1}\left(\mathbf{R}\widehat{\beta}-\mathbf{q}\right)$$

4.2. The Fisher test

Definition (Fisher test-statistic)

Under assumptions A1-A6 (cf. chapter 3), the **Fisher test-statistic** can be defined as a function of the SSR of the constrained (H_0) and unconstrained model (H_1):

$$\text{F} = \left(\frac{SSR_0 - SSR_1}{SSR_1} \right) \left(\frac{T-K}{p} \right)$$

where SSR_0 denotes the sum of squared residuals of the constrained model estimated under H_0 and SSR_1 denotes the sum of squared residuals of the unconstrained model estimated under H_1.

Example (Fisher test and CAPM model)

Consider the extended CAPM model for Microsoft:

$$r_{MSFT,t} = \beta_1 + \beta_2 r_{SP500,t} + \beta_3 r_{Ford,t} + \beta_4 r_{GE,t} + \varepsilon_t$$

where $r_{MSFT,t}$ is the excess return for Microsoft, $r_{SP500,t}$ for the SP500, $r_{Ford,t}$ for Ford and $r_{GE,t}$ for general electric. We want to test the following linear constraints:

$$H_0 : \beta_2 = 1 \text{ and } \beta_3 = \beta_4$$

Question: write a Matlab code to compute the F-statistic.

Note: the data are available in `Data_CAPM_Microsoft.xlsx`.

4.2. The Fisher test

Solution

In this problem, the null $H_0 : \beta_2 = 1$ and $\beta_3 = \beta_4$ can be written as:

$$\underset{(2\times 4)}{\mathbf{R}} \underset{(4,1)}{\boldsymbol{\beta}} = \underset{(2\times 1)}{\mathbf{q}}$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

4.2. The Fisher test

Figure: Matlab code to compute the F-statistic

```
clear ; clc ; close all

data=xlsread('Data_CAPM_Microsoft.xls');
r_MSFT=data(:,1);                     % Excess return for MSFT
r_SP500=data(:,2);                    % Excess return for SP500
r_Ford=data(:,3);                     % Excess return for Ford
r_GE=data(:,4);                       % Excess return for GE
T=length(r_MSFT);                     % Sample size

% Estimation under H1
X=[ones(T,1) r_SP500 r_Ford r_GE];    % Matrix of explicative variables
y=r_MSFT;                             % Dependent variable
beta=pinv(X'*X)*X'*y;                  % OLS estimator (H1)
res=y-X*beta;                         % Residuals
SSR1=sum(res.^2);                     % SSR of unconstrained model
var_eps=SSR1/(T-4);                   % Estimated variance
disp(' '),disp('beta under H1'),disp(beta')

% Fisher test statistic
R=[0 1 0 0 ; 0 0 1 -1];               % Matrix R
q=[1 ; 0];                            % Vector q
disp('R'),disp(R)
disp('q'),disp(q)

F=(1/2)*(R*beta-q)'*pinv(var_eps*R*pinv(X'*X)*R')*(R*beta-q);
disp('Fisher test statistics')
disp(F)
```

4.2. The Fisher test

Figure: Matlab output (F-statistic)

```
beta under H1
    0.0012    2.7619    0.3131   -0.1391

R
     0     1     0     0
     0     0     1    -1

q
     1
     0

Fisher test statistics
    4.3406
```

4.2. The Fisher test

Definition (Rejection region of a Fisher test)

The **critical region** of the Fisher test $\mathrm{H}_0 : \mathbf{R}\beta = \mathbf{q}$ against $\mathrm{H}_1 : \mathbf{R}\beta \neq \mathbf{q}$ at the $\alpha\%$ (say, 5%) is defined as:

$$\mathrm{W} = \{y : \mathrm{F}(y) > d_{1-\alpha}\}$$

where $d_{1-\alpha}$ is the $1-\alpha$ critical value (say 95%) of the Fisher distribution with p and $T-K$ degrees of freedom and $\mathrm{F}_k(y)$ is the realization of the Fisher test-statistic.

4.2. The Fisher test

Example (Fisher test and CAPM model)

Consider the extended CAPM model for Microsoft:

$$r_{MSFT,t} = \beta_1 + \beta_2 r_{SP500,t} + \beta_3 r_{Ford,t} + \beta_4 r_{GE,t} + \varepsilon_t \quad t = 1, \ldots, 24$$

where $r_{MSFT,t}$ is the excess return for Microsoft, $r_{SP500,t}$ for the SP500, $r_{Ford,t}$ for Ford and $r_{GE,t}$ for general electric. We want to test the following linear constraints:

$$H_0 : \beta_2 = 1 \text{ and } \beta_3 = \beta_4$$

Question: given the realization of the Fisher test-statistic (cf. previous example), conclude for a significance level $\alpha = 5\%$.

Note: the data are available in `Data_CAPM_Microsoft.xlsx`.

Solution

Step 1: compute the F-statistic (cf. Matlab code)

$$\mathrm{F}(y) = 4.3406$$

Step 2: Determine the rejection region for a nominal size $\alpha = 5\%$ for $T = 24$, $K = 4$ and $p = 2$

$$\mathrm{F} \underset{\mathrm{H}_0}{\sim} F_{(2,20)} \quad c_{0.95} = 3.4928$$

$$\mathrm{W} = \{y : \mathrm{F}(y) > 3.4928\}$$

Conclusion: for a significance level of 5%, we reject the null $\mathrm{H}_0 : \mathbf{R}\beta = \mathbf{q}$ against $\mathrm{H}_1 : \mathbf{R}\beta \neq \mathbf{q}$ $\square$

4.2. The Fisher test

Corollary (Student test-statistic and Fisher test-statistic)

Consider the test

$$H_0 : \beta_k = a_k \quad \text{versus} \quad H_1 : \beta_k \neq a_k$$

the **Fisher test-statistic** *corresponds to the squared of the corresponding* **Student's test-statistic**

$$F = T_k^2$$

4.2. The Fisher test

Definition (p-value)

The **p-value** of the F-test is equal to:

$$\text{p-value} = 1 - F_{p,T-K}\left(\mathrm{F}\left(y\right)\right)$$

where $\mathrm{F}(y)$ is the realization of the F-statistic and $F_{p,T-K}\left(.\right)$ the cdf of the Fisher distribution with p and $T-K$ degrees of freedom.

Definition (Global F-test)

In a multiple linear regression model with a constant term

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ik} + \varepsilon_i$$

the **global F-test** corresponds to the test of significance of all the explicative variables:

$$\text{H}_0 : \beta_2 = \ldots = \beta_K = 0$$

Under the assumption A6 (normality), the global F-test-statistic satisfies:

$$\text{global-F} \underset{\text{H}_0}{\sim} F_{(K-1,T-K)}$$

Remarks

1. The global F-test is a test designed to see if the model is useful overall. This test statistic is displayed in all the regression analysis tools (R, Stata, Matlab, Python, Excel, etc.)

2. The null $\mathrm{H}_0 : \beta_2 = \ldots = \beta_K = 0$ can be written as:

$$\underset{(K-1\times K)}{\mathbf{R}} \underset{(K,1)}{\boldsymbol{\beta}} = \underset{(K-1\times 1)}{\mathbf{q}}$$

$$\left(\begin{array}{cccccc} 0 & 1 & 0 & 0 & .. & 0 \\ 0 & 0 & 1 & 0 & .. & 0 \\ .. & .. & 0 & 1 & .. & .. \\ .. & .. & .. & .. & .. & .. \\ 0 & .. & 0 & 0 & .. & 1 \end{array} \right) \left(\begin{array}{c} \beta_1 \\ .. \\ .. \\ \beta_K \end{array} \right) = \left(\begin{array}{c} 0 \\ .. \\ .. \\ 0 \end{array} \right)$$

4.2. The Fisher test

Figure: Excel output (global F-test and corresponding p-value)

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*								
4	Coefficient de détermination multiple	0,6489074							
5	Coefficient de détermination R^2	0,42108081							
6	Coefficient de détermination R^2	0,41874646							
7	Erreur-type	0,01433559							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de liberté*	*Somme des carrés*	*Moyenne des carrés*	***F***	***Valeur critique de F***			
12	Régression	1	0,03707066	0,03707066	**180,384488**	**2,8468E-31**			
13	Résidus	248	0,05096626	0,00020551					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95,0%*	*Limite supérieure pour seuil de confiance = 95,0%*
17	Constante	0,00022961	0,00090928	0,25251848	0,8008495	-0,0015613	0,0020205	-0,0015613	0,0020205
18	Variable X 1	1,54515835	0,1150465	13,4307293	2,8468E-31	1,31856557	1,77175113	1,31856557	1,77175113

4.2. The Fisher test

Corollary (Global F-test)

In a multiple linear regression model with a constant term

$$y_i = \beta_1 + \sum_{k=2}^{K} \beta_k x_{ik} + \varepsilon_i$$

the global F-test-statistic can also be defined as:

$$F = \left(\frac{R^2}{1 - R^2} \right) \left(\frac{T - K}{K - 1} \right)$$

where R^2 denotes the (unadjusted) coefficient of determination.

4.2. The Fisher test

Example (Global F-test and CAPM model)

Consider the extended CAPM model for Microsoft:

$$r_{MSFT,t} = \beta_1 + \beta_2 r_{SP500,t} + \beta_3 r_{Ford,t} + \beta_4 r_{GE,t} + \varepsilon_t$$

Question: Given the elements reported by Eviews, compute the global F-test, its critical region for $\alpha = 5\%$, the p-value and conclude about the overall significance of the model.

Figure: Eviews estimation output

Dependent Variable: R_MSFT
Method: Least Squares
Date: 11/30/13 Time: 22:37
Sample: 2 25
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001219	0.000974	1.250453	0.2256
R_SP500	2.761927	0.629752	4.385734	0.0003
R_FORD	0.313054	0.174803	1.790895	0.0885
R_GE	-0.139065	0.287520	-0.483672	0.6339

R-squared	0.722707	Mean dependent var	0.000978
Adjusted R-squared	0.681113	S.D. dependent var	0.008312
S.E. of regression	0.004694	Akaike info criterion	-7.734203
Sum squared resid	0.000441	Schwarz criterion	-7.537861
Log likelihood	96.81044	F-statistic	
Durbin-Watson stat	2.036200	Prob(F-statistic)	

4.2. The Fisher test

Solution: critical value approach

$$r_{MSFT,t} = \beta_1 + \beta_2 r_{SP500,t} + \beta_3 r_{Ford,t} + \beta_4 r_{GE,t} + \varepsilon_t$$

The null hypothesis considered for global F-test is ($p = 3$):

$$\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

The realization of the global F-test statistic is equal to:

$$\begin{aligned} \text{global-F} &= \left(\frac{\mathrm{R}^2}{1-\mathrm{R}^2}\right)\left(\frac{T-K}{K-1}\right) \\ &= \left(\frac{0.722707}{1-0.722707}\right) \times \left(\frac{24-4}{4-1}\right) \\ &= 17.3753 \end{aligned}$$

Solution: critical value approach (cont'd)

The critical rejection region for a nominal size $\alpha = 5\%$, $T = 24$, $K = 4$ and $p = 3$ is defined as

$$\text{global-F} \underset{H_0}{\sim} F_{(3,20)} \quad c_{0.95} = 3.0984$$

$$\mathrm{W} = \{y : \text{global-F}\,(y) > 3.0984\}$$

So, the realization of global-F test statistic belongs to the critical region.

$$\text{global-F}\,(y) = 17.3753 \in \mathrm{W}$$

Conclusion: for a significance level of 5%, we reject the null $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. All the variables of the model are significant.

4.2. The Fisher test

Solution: p-value approach

The realization of the global F-test statistic is equal to (cf. infra) :

$$\text{global-F} = 17.3753$$

The p-value is equal to

$$\text{p-value} = 1 - F_{p,T-K}\left(\mathrm{F}\left(y\right)\right) = 1 - F_{3,20}\left(17.3753\right) = 8.5997e^{-06}$$

where $F_{3,20}\left(.\right)$ the cdf of the Fisher distribution with $p = 3$ and $T - K = 20$ degrees of freedom. The p-value is smaller than the significance level $\alpha = 5\%$.

Conclusion: for a significance level of 5%, we reject the null $\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$. The model is "significant".

4.2. The Fisher test

Figure: Eviews output (F-test statistic and p-value)

Dependent Variable: R_MSFT
Method: Least Squares
Date: 12/01/13 Time: 00:03
Sample: 2 25
Included observations: 24

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001219	0.000974	1.250453	0.2256
R_SP500	2.761927	0.629752	4.385734	0.0003
R_FORD	0.313054	0.174803	1.790895	0.0885
R_GE	-0.139065	0.287520	-0.483672	0.6339

R-squared	0.722707	Mean dependent var	0.000978
Adjusted R-squared	0.681113	S.D. dependent var	0.008312
S.E. of regression	0.004694	Akaike info criterion	-7.734203
Sum squared resid	0.000441	Schwarz criterion	-7.537861
Log likelihood	96.81044	F-statistic	17.37532
Durbin-Watson stat	2.036200	Prob(F-statistic)	0.000009

4. Individual and global significance tests

Key Concepts

1. One-sided and two-sided tests
2. Student t-test statistic
3. Linear restrictions on the parameters
4. Fisher F-test statistic
5. Critical value or p-value based decision rules
6. Global F test

Section 5

Misspecification Tests

5. Misspecification tests

Objectives

1. Testing the main assumptions A1-A6 of the multiple linear model
2. Introducing the main visual procedures and testing procedures
3. Understand the consequences of the perfect and imperfect multi-collinearity
4. Understand the consequences of the heteroscedasticity and autocorrelation
5. Introduce the White, Breusch-Pagan, Jarque Bera and Lilliefors tests
6. Introduce the QQ-plots

Overview

- If the model is well-specified: there should be no information contained in the error term (residuals).
- If this condition is not satisfied, this does represent a violation of assumptions underlying the MLR model (Assumptions A1-A6).
- This leads to different problems: inappropriateness of the OLS estimator, misinterpretation of statistical tests, misinterpretation of results...
- This also means that there exist some "arbitrage opportunities" that can be used in order to improve the knowledge (inference, predictions) of the dependent variable.

To avoid such situations you have to apply **misspecification tests**

1. **Visual procedures** using residuals: outliers, qq-plots, residual vs fitted values, and residuals vs explanatory variables.

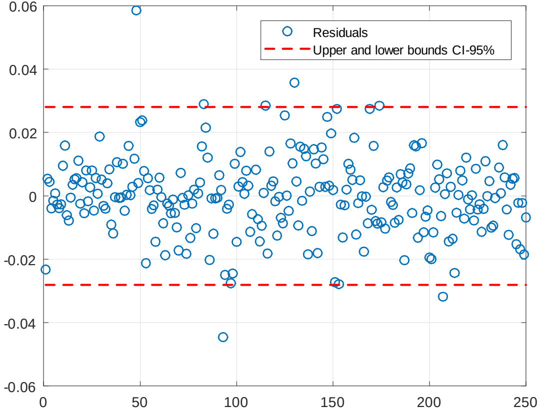
2. **Testing procedures** using residuals (misspecification tests): Among others,
   - Heteroscedasticity
   - Autocorrelation
   - Non-normality of the error terms
   - Parameter instability

General methodology of misspecification tests:

Starting point is to plot residuals!

5. Misspecification tests

Figure: OLS-based residuals of CAPM model for Intel Corp (Aug 2017 - Aug 2018)



General methodology of misspecification tests (cont'd):

1. **Step 1:** OLS estimation of the initial model and computation of residuals $\widehat{\varepsilon}_t = y_t - \hat{y}_t$, $t = 1, \cdots, T$.
2. **Step 2:** OLS estimation of an **auxiliary** regression for the residuals (or a transformation of the residuals).
3. **Step 3:** Computation of the coefficient of determination, R^2, and then of the test statistic.
4. **Step 4:** Decision using a critical value approach or the p-value approach.

Note: These misspecification tests are generally upper-tailed tests.

Subsection 5.1

Identification Assumption

5.2. Identification assumption

Reminder (ch. chapter 1)

Definition (Assumption 2: Full column rank)

$\mathbf{X}$ is an $T \times K$ matrix with rank K.

1. **Perfect multicollinearity:** The correlation coefficient between two explanatory variables is -1 or $+1$.
2. **Imperfect or near multicollinearity:** It occurs when high positive or negative (linear) correlation coefficients among the explanatory variables.

5.2. Identification assumption

Remarks

1. Perfect multi-collinearity is generally not difficult to spot and is signalled by most statistical software.

2. Imperfect multi-collinearity is a more serious issue.

Example (Perfect multi-collinearity)

We consider the following extended CAPM model

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \beta_3 x_t + \varepsilon_t$$

where the variable x_t is defined as

$$x_t = 2 \times z_{\texttt{market},t}$$

Question: What is output from Excel's Regression command when you try to estimate β_1, β_2, and β_3?

Note: the data are available within the file `Data_CAPM_returns.xlsx`.

5.2. Identification assumption

Figure: Data exemple (perfect multi-collinearity)

	A	B	C	D
1	date	r_Intel_ex	r_SP500_ex	r_SP500*2
2	22-Aug-17	-0,00778902	0,00986457	0,01972914
3	23-Aug-17	0,00026178	-0,00348629	-0,00697258
4	24-Aug-17	0,00141431	-0,00210382	-0,00420765
5	25-Aug-17	-0,00118022	0,00164435	0,00328869
6	28-Aug-17	-0,00060432	0,00045956	0,00091911
7	29-Aug-17	0,00227855	0,00081493	0,00162987
8	30-Aug-17	0,00456915	0,00457732	0,00915465
9	31-Aug-17	0,00511917	0,00567801	0,01135602
10	1-Sep-17	0,00054327	0,00195372	0,00390745
11	5-Sep-17	-0,00202366	-0,00760626	-0,01521252
12	6-Sep-17	0,02088228	0,00309555	0,0061911
13	7-Sep-17	-0,00619893	-0,00020634	-0,00041267
14	8-Sep-17	-0,00992473	-0,00151776	-0,00303553
15	11-Sep-17	0,01631983	0,01075317	0,02150634
16	12-Sep-17	0,00887852	0,00333055	0,00666111
17	13-Sep-17	0,00660033	0,00072908	0,00145817
18	14-Sep-17	0,00409224	-0,00112934	-0,00225869
19	15-Sep-17	0,01412621	0,00181794	0,00363589
20	18-Sep-17	-2,7262E-05	0,0014276	0,0028552
21	19-Sep-17	0,00616923	0,00108183	0,00216366
22	20-Sep-17	-0,00433467	0,00060634	0,00121269
23	21-Sep-17	0,00347329	-0,00307804	-0,00615609
24	22-Sep-17	-0,00056515	0,00062024	0,00124048
25	25-Sep-17	-0,00056528	-0,00225173	-0,00450346

5.2. Identification assumption

Figure: Excel output in case of perfect multi-collinearity

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*								
4	Coefficient de détermination multiple	0,6489074							
5	Coefficient de détermination R^2	0,42108081							
6	Coefficient de détermination R^2	0,4147142							
7	Erreur-type	0,01433559							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de liberté*	*Somme des carrés*	*Moyenne des carrés*	*F*	*Valeur critique de F*			
12	Régression	2	0,03707066	0,01853533	180,384488	5,0859E-49			
13	Résidus	248	0,05096626	0,00020551					
14	Total	250	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*pour seuil de*	*pour seuil de*	*our seuil de c*	*our seuil de c*
17	Constante	0,00022961	0,00090928	0,25251848	0,8008495	-0,00156128	0,0020205	-0,00156128	0,0020205
18	Variable X 1	0	0	65535	#NOMBRE!	0	0	0	0
19	Variable X 2	0,77257917	0,05752325	13,4307293	#NOMBRE!	0,65928279	0,88587556	0,65928279	0,88587556

Patterns of near multicollinearity

1. The coefficient of determination is large;
2. Individual coefficients have high standard errors;
3. Confidence intervals are quite wide;
4. The regression is quite sensitive to small changes in the specification.

5.2. Identification assumption

Example (Perfect multi-collinearity)

We consider the following extended CAPM model

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \beta_3 x_t + \varepsilon_t$$

where the variable x_t is defined as

$$x_t = I_t \times z_{\texttt{market},t}$$

$$I_t \sim U_{[a,b]}$$

Question: What is output from Excel's Regression command when $[a, b] = [0.990; 1.01]$ and $[a, b] = [0.999; 1.001]$?

Note: the data are available within the file "Chapter2_Exercice3.xlsx".

5.2. Identification assumption

Figure: Near multi-collinearity (case $[a, b] = [0.990; 1.1010]$)

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*				a	0,990			
4	Coefficient de détermination multiple	0,64922717			b	1,010			
5	Coefficient de détermination R^2	0,42149592			corr(X1,X2)	0,99996571			
6	Coefficient de détermination R^2	0,41681168							
7	Erreur-type	0,01435943							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de libert*	*mme des car*	*yenne des ca*	*F*	*'eur critique de F*			
12	Régression	2	0,0371072	0,0185536	89,9816418	4,4141E-30			
13	Résidus	247	0,05092972	0,00020619					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*pour seuil de*	*pour seuil de*	*our seuil de c*	*our seuil de c*
17	Constante	0,00021425	0,00091152	0,23504691	0,81436707	-0,0015811	0,0020096	-0,0015811	0,0020096
18	Variable X 1	-4,57139541	14,5292442	-0,31463408	0,75330511	-33,1884091	24,0456183	-33,1884091	24,0456183
19	Variable X 2	61,2849704	145,571563	0,42099548	0,67412447	-225,434928	348,004869	-225,434928	348,004869

5.2. Identification assumption

Figure: Near multi-collinearity (case $[a, b] = [0.999; 1.1001]$)

	A	B	C	D	E	F	G	H	I
1	RAPPORT DÉTAILLÉ								
2									
3	*Statistiques de la régression*				a	0,999			
4	Coefficient de détermination multiple	0,65083346			b	1,001			
5	Coefficient de détermination R^2	0,42358419			corr(X1,X2)	0,99999963			
6	Coefficient de détermination R^2	0,41891685							
7	Erreur-type	0,01433349							
8	Observations	250							
9									
10	ANALYSE DE VARIANCE								
11		*Degré de libert*	*mme des carı*	*yenne des ca*	*F*	*leur critique de F*			
12	Régression	2	0,03729105	0,01864552	90,7550522	2,8241E-30			
13	Résidus	247	0,05074587	0,00020545					
14	Total	249	0,08803692						
15									
16		*Coefficients*	*Erreur-type*	*Statistique t*	*Probabilité*	*pour seuil de*	*pour seuil de*	*our seuil de c*	*our seuil de cc*
17	Constante	0,00024272	0,00090923	0,26695269	0,78972833	-0,00154812	0,00203356	-0,00154812	0,00203356
18	Variable X 1	151,600909	144,88029	1,04638739	0,29640511	-133,757448	436,959266	-133,757448	436,959266
19	Variable X 2	-150,038311	144,863407	-1,03572265	0,30134448	-435,363414	135,286792	-435,363414	135,286792

Remedies: near multicollinearity can be tackled (to some extent) by

1. Penalized regressions (e.g., Ridge, Lasso, Elastic-net regressions),
2. Principal components analysis;
3. General-to-specific or specific-to-general approach.

Subsection 5.2

Homoscedasticity Assumption

Definition (Heteroscedasticity)

Disturbances ε_t are **heteroscedastic** when they have different (conditional) variances:

$$\mathbb{V}\left(\varepsilon_i \middle| \mathbf{X}\right) \neq \mathbb{V}\left(\varepsilon_j \middle| \mathbf{X}\right) \quad \text{for } i \neq j$$

or equivalently

$$\sigma_i^2 \neq \sigma_j^2 \quad \text{for } i \neq j$$

Example (Heteroscedasticity)

If the disturbances are **heteroscedastic** but they are still assumed to be uncorrelated across observations, so $\mathbb{V}\left(\left.\boldsymbol{\varepsilon}\right|\mathbf{X}\right)$ is defined as:

$$\mathbb{V}\left(\left.\boldsymbol{\varepsilon}\right|\mathbf{X}\right)=\begin{pmatrix} \sigma_1^2 & 0 & .. & 0 \\ 0 & \sigma_2^2 & .. & 0 \\ .. & .. & .. & .. \\ 0 & .. & .. & \sigma_T^2 \end{pmatrix}$$

Remarks

1. Heteroscedasticity often arises in volatile high-frequency time-series data such as daily observations in **financial markets.**

2. Heteroscedasticity often arises in **cross-section data** where the scale of the dependent variable and the explanatory power of the model tend to vary across observations. Microeconomic data such as expenditure surveys are typical

What are the consequences of the heteroscedasticity?

1. The OLS estimator is **unbiased**

2. The OLS estimator is (weakly) **consistent**

3. The OLS estimator is **asymptotically normally distributed**

But...

1. The inference based on the estimator $\widehat{\mathbb{V}}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)=\widehat{\sigma}^2\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}$ is **misleading**.

2. The OLS is **inefficient**.

$$\mathbb{V}\left(\widehat{\boldsymbol{\beta}}_{OLS}\right)-I_T^{-1}\left(\boldsymbol{\beta}_0\right) \text{ is a positive definite matrix}$$

Consequence: standard errors are generally inappropriate and thus inference and interpretation might be misleading.

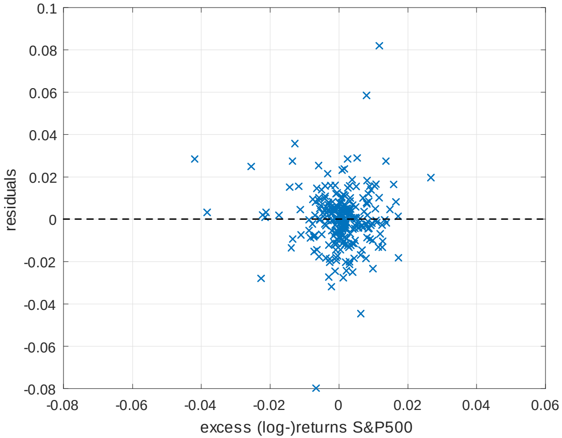
How to detect the heteroscedasticity?

1. **Visual procedure:** look at residuals vs each explanatory variable and detect if there any specific pattern.

2. **Testing procedure:** use a White test to detect the heteroscedasticity

5.3. Homoscedasticity assumption

Figure: CAPM-based residuals vs S&P500



Definition (White test for heteroscedasticity)

The **White test for heteroscedasticity** is based on the auxiliary regression:

$$\widehat{\varepsilon}_t^2 = \gamma_0 + \sum_{k=2}^{K} \gamma_k x_{t,k} + \sum_{k=2}^{K} \sum_{j \geq k}^{K} \gamma_{k,j} x_{t,k} x_{t,j} + v_t$$

where $\widehat{\varepsilon}_t$ denotes the OLS residual at time t. The test is

$$\mathrm{H}_0 : \gamma_k = 0 \text{ and } \gamma_{k,j} = 0 \ \forall \, (k,j) \qquad => \ \text{Homoscedasticity}$$

$$\mathrm{H}_1 : \exists k : \gamma_k \neq 0 \ \text{ or } \gamma_{k,j} \neq 0 => \ \text{Heteroscedasticity}$$

White (1980) proposes the following procedure and test-statistic:

1. **Step 1:** Estimation of the model using the OLS estimator of $\boldsymbol{\beta}$.
2. **Step 2:** Determine the residuals $\widehat{\varepsilon}_i = y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_{OLS}$.
3. **Step 3 (auxiliary regression):** Regress $\widehat{\varepsilon}_i^2$ on a constant and all unique columns vectors contained in $\mathbf{X}$ and all the squares and cross-products of the column vectors in $\mathbf{X}$.
4. **Step 4:** Determine the coefficient of determination, R^2, of the previous regression. The **White test-statistic** is defined as

$$T_{\mathrm{W}} = T \times R^2$$

Definition (White test for heteroscedasticity)

Under the null, the **White test-statistic** converges:

$$T_{\mathrm{W}} = T \times \mathrm{R}^2 \underset{\mathrm{H}_0}{\overset{d}{\to}} \chi^2 \left(m - 1\right)$$

where m is the number of explanatory variables in the regression of $\widehat{\varepsilon}_i^2$. The critical region of size α is

$$\mathrm{W} = \left\{ y : T_{\mathrm{W}} \left(y\right) > \chi^2_{1-\alpha} \right\}$$

where $\chi^2_{1-\alpha}$ denotes the 1-α critical value of the chi-squared distribution $\chi^2 \left(m - 1\right)$.

Example (extended CAPM)

We want to estimate by OLS the parameters of the extended CAPM model given by

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \beta_3 \text{inflation}_t + \varepsilon_t$$

Question: Compute the white test statistic and conclude about the homoscedasticity of ε_t at a 5% significance level.

Note: the data are available in `Data_CAPM_extended.xlsx`.

5.3. Homoscedasticity assumption

Solution

Step 1: Estimation of the model using the OLS estimator of $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \beta_3 \text{inflation}_t + \varepsilon_t$$

Step 2: Determine the residuals

$$\widehat{\varepsilon}_t = z_{\texttt{intel},t} - \widehat{\beta}_1 - \widehat{\beta}_2 z_{\texttt{market},t} - \widehat{\beta}_3 \text{inflation}_t$$

Step 3 (auxiliary regression): Regress $\widehat{\varepsilon}_i^2$ on a constant and all unique columns vectors contained in $\mathbf{X}$ and all the squares and cross-products of the column vectors in $\mathbf{X}$.

$$\begin{aligned} \widehat{\varepsilon}_t^2 \quad = \quad & \gamma_1 + \gamma_2 z_{\texttt{market},t} + \gamma_3 \text{inflation}_t \\ & + \gamma_4 z_{\texttt{market},t}^2 + \gamma_5 \text{inflation}_t^2 + \gamma_6 z_{\texttt{market},t} \times \text{inflation}_t + v_t \end{aligned}$$

Step 4: Determine the coefficient of determination, R^2, of the previous regression.

5.3. Homoscedasticity assumption

Table 3: OLS estimation of the auxiliary regression (White test)

Variable	Estimate	Standard error	t-stat	p-value
Intercept	-0.011784	0.011377	-1.0358	0.3025
SP500	0.15217	0.18193	0.83642	0.40467
Inflation	-0.09595	1.8622	-0.051525	0.959
SP500$\times$Inflation	-44.587	37.024	-1.2043	0.23097
SP500^2	3.7512	2.397	1.5649	0.12037
Inflation^2	169.58	214.55	0.79042	0.43093

Note: R-squared is **0.0262** and the number of observations is **120**.

Solution (cont'd)

The realization of White test statistic is equal to:

$$T_{\mathrm{W}} = T \times \mathrm{R}^2 = 120 \times 0.0262 = 3.1391$$

Critical value approach: The critical region is defined by

$$T_{\mathrm{W}} \underset{\mathrm{H}_0}{\overset{d}{\rightarrow}} \chi^2 (6-1) \qquad \chi^2_{0.95,5} \simeq 11.07$$

$$\mathrm{W} = \{y : T_{\mathrm{W}}(y) > 11.07\}$$

So, the White test statistic does not belongs to the critical region for a level $\alpha = 5\%$.

Conclusion: For a significance level of 5%, we fail to reject the null $\mathrm{H}_0 : \gamma_2 = \ldots = \gamma_6$, i.e. the homoscedasticity assumption.

Solution (cont'd)

p-value approach: the p-value is equal to

$$\text{p-value} = 1 - G_{m-1}\left(T_{\mathrm{W}}\left(y\right)\right)$$

where $G_{m-1}\left(.\right)$ is the cdf of the chi-squared distribution $\chi^2\left(m-1\right)$. Since $T_{\mathrm{W}} = 3.1391$ and $m = 6$, we have

$$\text{p-value} = 1 - G_5\left(3.1391\right) = 1 - 0.3214 = 0.6786$$

The p-value is larger than the risk level $\alpha = 5\%$.

Conclusion: For a significance level of 5%, we fail to reject the null $\mathrm{H}_0 : \sigma_t^2 = \sigma^2$, i.e. the homoscedasticity assumption.

Remark: All the econometric software (R, Stata, Excel) have a command to compute the White test statistic and its p-value.

Example (White's (1980) test for heteroscedasticity)

Consider the generalized linear regression model:

$$\text{AVGEXP}_i = \beta_1 + \beta_2 \text{AGE}_i + \beta_3 \text{Ownrent}_i + \beta_4 \text{Income}_i + \beta_5 \text{Income}_i^2 + \varepsilon_i$$

where AVGEXP denotes the Avg. monthly credit card expenditure, Ownrent denotes a binary variable (individual owns (1) or rents (0) home), Age denotes the age in years, Income denotes the income divided by 10,000. Eviews computes the White test statistic and reports the results of the auxiliary regression.

Note: the data are available in file `Data_White_Test.xls`.

6. Testing for heteroscedasticity

Figure: White test with Eviews

White Heteroskedasticity Test:

F-statistic	1.244819	Probability	0.266541
Obs*R-squared	14.65386	Probability	0.260914

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 12/14/13 Time: 21:00
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	876511.9	913863.8	0.959128	0.3402
AGE	28775.90	31660.00	0.908904	0.3659
AGE^2	-644.2271	425.9743	-1.512361	0.1341
AGE*OWNRENT	5681.491	8776.134	0.647380	0.5191
AGE*INCOME	6853.915	11227.53	0.610456	0.5432
AGE*INCOME2	-647.8628	1274.148	-0.508467	0.6124
OWNRENT	195763.1	474111.1	0.412905	0.6807
OWNRENT*INCOME	-177650.5	199416.6	-0.890851	0.3755
OWNRENT*INCOME2	11325.35	21530.66	0.526010	0.6002
INCOME	-1509045.	778264.9	-1.938986	0.0557
INCOME^2	498964.2	253154.3	1.970989	0.0519
INCOME*INCOME2	-63934.08	34454.00	-1.855636	0.0669
INCOME2^2	2820.726	1630.189	1.730306	0.0871

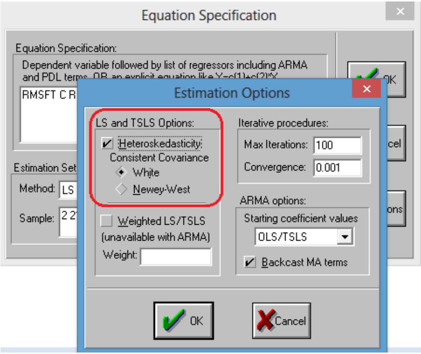
R-squared	0.146539	Mean dependent var	70384.57
Adjusted R-squared	0.028820	S.D. dependent var	287729.4
S.E. of regression	283552.9	Akaike info criterion	28.06892
Sum squared resid	6.99E+12	Schwarz criterion	28.40759
Log likelihood	-1390.446	F-statistic	1.244819
Durbin-Watson stat	1.745177	Prob(F-statistic)	0.266541

To go further...

In the presence of heteroscedasticity,

- Transform the variables into logs or reduce by some other measures of size;
- Compute the White's **heteroscedasticity consistent** standard error estimates: these standard errors are generally increased with respect to the OLS ones.
- Compute estimators that have better properties (at least, asymptotically) as for instance the generalized least squares estimator (in the presence of known heteroscedasticity).

Example: Eviews procedure to compute the heteroscedasticity consistent standard error estimates



Dependent Variable: RMSFT
Method: Least Squares
Date: 12/14/13 Time: 16:12
Sample: 2 21
Included observations: 20
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001189	0.001160	1.025585	0.3187
RSP500	1.989787	0.311130	6.395357	0.0000

R-squared	0.690203	Mean dependent var	-0.000180
Adjusted R-squared	0.672992	S.D. dependent var	0.009272
S.E. of regression	0.005302	Akaike info criterion	-7.546873
Sum squared resid	0.000506	Schwarz criterion	-7.447300
Log likelihood	77.46873	F-statistic	40.10263
Durbin-Watson stat	1.955366	Prob(F-statistic)	0.000006

Subsection 5.3

Autocorrelation Assumption

5.4. Autocorrelation assumption

Definition (Autocorrelation)

Disturbances are **autocorrelated (or correlated)** when:

$$\mathbb{C}ov\left(\varepsilon_i, \varepsilon_j \middle| \mathbf{X}\right) \neq 0 \quad \text{for } i \neq j$$

Example (Autocorrelation)

For instance, **time-series data** are usually homoscedastic, but autocorrelated, so $\mathbb{V}\left(\varepsilon | \mathbf{X}\right)$ would be:

$$\mathbb{V}\left(\varepsilon | \mathbf{X}\right) = \begin{pmatrix} \sigma^2 & \sigma_{12} & .. & \sigma_{1N} \\ \sigma_{21} & \sigma^2 & .. & \sigma_{2N} \\ .. & .. & .. & .. \\ \sigma_{N1} & .. & .. & \sigma^2 \end{pmatrix}$$

Causes of autocorrelated (or spatially correlated) errors:

1. "Business cycle inertia" generally causes positive autocorrelation in macroeconomic time series;
2. Overlapping effect of shocks: the shock has an effect at time t but also persists at some other periods;
3. Model misspecification: Omitted variables which are correlated across time (inertia).

What are the consequences of the autocorrelation?

They are similar to those of the heteroscedasticity

1. The OLS estimator is **unbiased,** (weakly) **consistent** and **asymptotically normally distributed**
2. But, the OLS is **inefficient** and the inference based on the estimator $\widehat{\mathbb{V}}\left(\widehat{\beta}_{OLS}\right)=\widehat{\sigma}^{2}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}$ is **misleading**

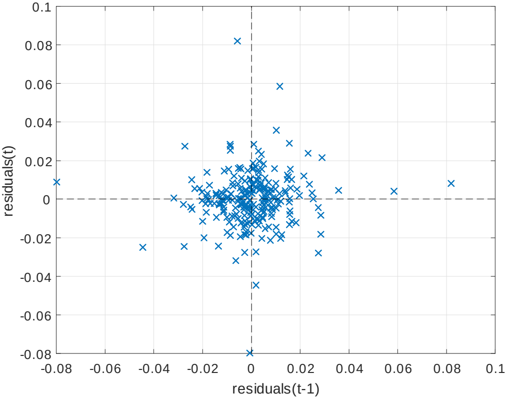
Consequence: standard errors are generally inappropriate and thus inference and interpretation might be misleading.

How to detect the autocorrelation?

1. **Visual procedure:** look at residuals vs lagged residuals (e.g., $\widehat{\varepsilon}_t$ vs $\widehat{\varepsilon}_{t-1}$)
2. **Basic statistics:** compute the autocorrelations of the residuals
3. **Testing procedure:** use a Breush-Pagan test to detect the autocorrelation

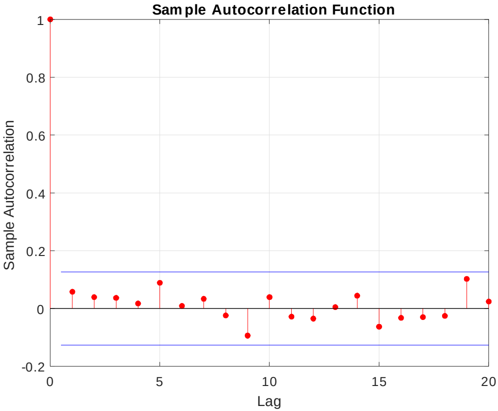
5.4. Autocorrelation assumption

Figure: CAPM-based residuals vs rst lag of residuals



5.4. Autocorrelation assumption

Figure: Autocorrelation function fo the CAPM-based residuals (Matlab function autocorr)



Definition (Breusch Pagan test for autocorrelation)

The **Breusch Pagan test for autocorrelation** is based on the auxiliary regression:

$$\widehat{\varepsilon}_t = \gamma_0 + \gamma_1 \widehat{\varepsilon}_{t-1} + \cdots + \gamma_p \widehat{\varepsilon}_{t-p} + v_t$$

where $\widehat{\varepsilon}_t$ denotes the OLS residual at time t. The test is

$$\mathrm{H}_0 : \gamma_1 = \ldots = \gamma_p \quad => \text{ Non autocorrelation}$$

$$\mathrm{H}_1 : \exists k : \gamma_k \neq 0 \quad => \text{ Autocorrelation}$$

Definition (Breusch Pagan test for autocorrelation)

The **Breusch Pagan test** statistic is equal to

$$T_{\mathrm{BP}} = T \times \mathrm{R}^2 \underset{\mathrm{H}_0}{\overset{d}{\to}} \chi^2\left(p\right)$$

where p is the number of lagged residuals introduced in the auxiliary regression. The critical region of size α is

$$\mathrm{W} = \left\{ y : T_{\mathrm{W}}\left(y\right) > \chi^2_{1-\alpha} \right\}$$

where $\chi^2_{1-\alpha}$ denotes the $1-\alpha$ critical value of the $\chi^2\left(p\right)$ distribution.

Example (CAPM and Breusch Pagan test)

Consider the CAPM model for Intel Corp.

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \varepsilon_t$$

Question: Test the autocorrelation of the residuals with a maximum lag order $p = 5$ and a significance level $\alpha = 5\%$.

Note: the data are available within the file `Data_CAPM_returns.xlsx`.

Solution

Step 1: Estimate the parameters of the CAPM model for Intel Corp.

$$z_{\texttt{intel},t} = \beta_1 + \beta_2 z_{\texttt{market},t} + \varepsilon_t$$

and compute the residuals

$$\widehat{\varepsilon}_t = z_{\texttt{intel},t} - \widehat{\beta}_1 - \widehat{\beta}_2 z_{\texttt{market},t} - \widehat{\beta}_3 \text{inflation}_t$$

Step 2: Consider the auxiliary regression with $p = 5$ lags:

$$\widehat{\varepsilon}_t = \gamma_0 + \gamma_1 \widehat{\varepsilon}_{t-1} + \cdots + \gamma_5 \widehat{\varepsilon}_{t-5} + v_t$$

5.4. Autocorrelation assumption

Figure: Auxiliary regression for the Breush Pagan test

```
Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	4.2403e-05	0.00092132	0.046024	0.96333
x1	0.055857	0.064459	0.86656	0.38705
x2	0.034393	0.064748	0.53118	0.59579
x3	0.028113	0.064747	0.43419	0.66454
x4	0.0076935	0.064895	0.11855	0.90573
x5	0.086472	0.06445	1.3417	0.18097

```
Number of observations: 245, Error degrees of freedom: 239
Root Mean Squared Error: 0.0144
R-squared: 0.0137, Adjusted R-Squared -0.0069
F-statistic vs. constant model: 0.666, p-value = 0.65
```

Solution (cont'd)

The realization of Breusch Pagan test statistic is equal to:

$$T_{\mathrm{BP}} = T \times \mathrm{R}^2 = 245 \times 0.0137 = 3.3565$$

Critical value approach: The critical region is defined by

$$T_{\mathrm{BP}} \underset{\mathrm{H}_0}{\overset{d}{\rightarrow}} \chi^2(5) \qquad \chi^2_{0.95,5} \simeq 11.07$$

$$\mathrm{W} = \{y : T_{\mathrm{W}}(y) > 11.07\}$$

So, the Breusch Pagan test statistic does not belongs to the critical region for a level $\alpha = 5\%$.

Conclusion: For a significance level of 5%, we fail to reject the null $\mathrm{H}_0 : \gamma_1 = \ldots = \gamma_5$, i.e. the no-autocorrelation assumption.

Solution (cont'd)

p-value approach: the p-value is equal to

$$\text{p-value} = 1 - G_p\left(T_{\text{BP}}\left(y\right)\right)$$

where $G_p(.)$ denotes the cdf of the chi-squared distribution $\chi^2(p)$. Since $T_{\text{BP}} = 3.3565$ and $p = 5$, we have

$$\text{p-value} = 1 - G_5\left(3.3565\right) = 1 - 0.3548 = 0.6452$$

The p-value is larger than the risk level $\alpha = 5\%$.

Conclusion: For a significance level of 5%, we fail to reject the null $\text{H}_0 : \gamma_1 = \ldots = \gamma_5$, i.e. the no-autocorrelation assumption.

Subsection 5.4

Normality Assumption

5.5. Normality assumption

Reminder (ch. Chapter 2)

Definition (Assumption 6: Normal distribution)

The disturbances are normally distributed.

$$\varepsilon_i | \mathbf{X} \sim \mathcal{N}\left(0, \sigma^2\right)$$

or equivalently

$$\boldsymbol{\varepsilon} | \mathbf{X} \sim \mathcal{N}\left(\mathbf{0}_{T \times 1}, \sigma^2 \mathbf{I}_T\right)$$

5.3. Homoscedasticity assumption

How to asses the normality of the residuals?

1. **Visual procedure:** QQ-plot of the residuals $\widehat{\varepsilon}_t$
2. **Testing procedure:** parametric tests (Jarque-Bera test) or distribution tests (Lilliefors or Anderson-Darling tests)

Visual inspection: QQ-plot

Definition (QQ-plot)

A **Q–Q (quantile-quantile) plot** is a graphical method for comparing two probability distributions by plotting their quantiles against each other. It can be used to test the normal distribution of the residuals, and hence of the error terms.

Visual inspection: QQ-plot

Step 1: From the sequence $\{\widehat{\varepsilon}_t, \ t = 1, \cdots, T\}$, compute the **standardized** residuals

$$\widetilde{\varepsilon}_t = \frac{\widehat{\varepsilon}_t}{\text{se}\left(\widehat{\varepsilon}\right)}, \quad \text{for} \ \ t = 1, \cdots, T$$

Step 2: Rank the standardized series by increasing order

$$\left\{\widetilde{u}_t^{\star}, t = 1, \cdots, T \ \right\} \text{ with } \widetilde{u}_t^{\star} < \widetilde{u}_{t'}^{\star} \text{ for } t < t'$$

Visual inspection: QQ-plot (cont'd)

Step 3: Plot the points for $t = 1, \cdots, T$

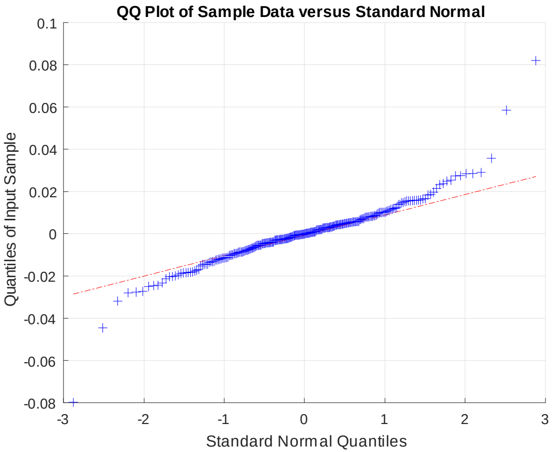
$$\left(\Phi^{-1}\left(\frac{t-0.5}{T} \right), \widetilde{u}_t^{\star} \right)$$

where $\Phi(.)$ is the cdf of the standard normal distribution. Remark: the choice of quantiles for the theoretical distribution depends upon context and purpose. We can also consider the values t/T as these are the quantiles that the sampling distribution realizes. Other choices are the use of $(t-0.5)/T$, or instead to space the points evenly in the uniform distribution.

Step 4: By visual inspection (this is not a formal testing procedure), the empirical distribution of the standardized series, $\{\widetilde{u}_t^{\star}, t = 1, \cdots, T\}$, is "close" to the (standard) normal distribution if the points $\left(q_{\frac{t-0.5}{T}}, \widetilde{u}_t^{\star} \right)$ are "close" to the 45-degree line.

5.5. Normality assumption

Figure: QQ-plot of the residuals



A parametric test: The Jarque-Bera test

Definition (Jarque-Bera test)

The **Jarque-Bera** test for the residuals $\widehat{\varepsilon}_t$ is defined as

$$\begin{aligned} \mathrm{H}_0 &: \mathbb{S}\left(\widehat{\varepsilon}_t\right) = 0 \quad \text{and} \quad \mathbb{K}\left(\widehat{\varepsilon}_t\right) = 3 \\ \mathrm{H}_1 &: \mathbb{S}\left(\widehat{\varepsilon}_t\right) \neq 0 \quad \text{and/or} \quad \mathbb{K}\left(\widehat{\varepsilon}_t\right) \neq 3 \end{aligned}$$

Note: If the null is rejected, the distribution of $\widehat{\varepsilon}_t$ is not normal. If the null is not rejected, the distribution of $\widehat{\varepsilon}_t$ is mesokurtic, but not necessarily normal.

5.5. Normality assumption

Definition (Jarque-Bera test statistic)

The **Jarque-Bera test statistic** for residuals $\widehat{\varepsilon}_t$ is defined as

$$\mathrm{JB} = \left(\frac{T-K}{6}\right)\left(\widehat{S}^2 + \frac{1}{4}\left(\widehat{K}-3\right)^2\right) \underset{H_0}{\sim} \chi^2(2)$$

where $\widehat{S}$ and $\widehat{K}$ respectively denote the sample skewness and kurtosis. The critical region of size α is

$$\mathrm{W} = \left\{ y : \mathrm{JB}\left(y\right) > \chi^2_{1-\alpha,2} \right\}$$

where $\chi^2_{1-\alpha,2}$ is the $1-\alpha$ critical value of the $\chi^2\left(2\right)$ distribution.

A parametric test: The Jarque-Bera test

Step 1: Compute the observed test statistic

$$\text{JB} = \left(\frac{T-K}{6}\right)\left(\widehat{S}^2 + \frac{1}{4}\left(\widehat{K}-3\right)^2\right)$$

Step 1: Using a critical value approach and taking the significance level α (e.g., $\alpha = 0.01, 0.05$ or 0.1), reject the null hypothesis if

$$\text{JB} \geq \chi^2_{1-\alpha,2}$$

where $\chi^2_{1-\alpha,2}$ denotes the $1-\alpha$ critical value of the $\chi^2(2)$ distribution.

Distribution tests

Compare the empirical cumulative distribution, G_T, with the cdf $\Phi(.)$ of a standard normal distribution. The empirical cdf is defined by:

$$G_T(x) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}_{\widetilde{u}_t^\star \leq x}$$

where

$$\mathbb{I}_{\widetilde{u}_t^\star \leq x} = \begin{cases} 1 & \text{if } \widetilde{u}_t^\star \leq x \\ 0 & \text{otherwise.} \end{cases}$$

Interpretation: The value $G_T(x)$ is the proportion (relative to T) of observations $\widetilde{u}_t^\star$ below x, e.g.

$$G_T(\widetilde{u}_t^\star) = \frac{t}{T}$$

Definition (Distribution tests)

The **Lilliefors** test statistic is defined as

$$\mathrm{T}_L = \max_{t=1,\cdots,T} \left| G_T\left(\widetilde{u}_t^{\star}\right) - \Phi(\widetilde{u}_t^{\star}) \right| = \max_{t=1,\cdots,T} \left| \frac{t}{T} - \Phi(\widetilde{u}_t^{\star}) \right|$$

The **Anderson-Darling test** statistic is defined as

$$\mathrm{AD} = -T - \frac{1}{T} \sum_{t=1}^{T} \left[(2t-1) \ln \Phi(\widetilde{u}_t^{\star}) + (2T+1-2t) \ln\left(1 - \Phi(\widetilde{u}_t^{\star})\right) \right]$$

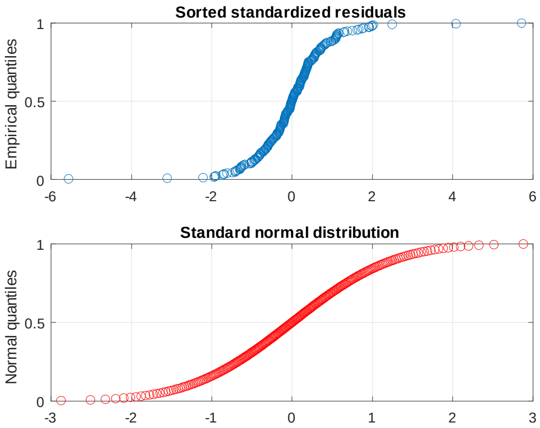
where $\Phi\left(.\right)$ is the cdf of a standard normal distribution

Decision rule

- In both cases, reject the null hypothesis of normality if the observed test-statistic is greater than the (tabulated or simulated) critical value(s)
- Lilliefors test: $0.805/\sqrt{T}$ (10%), $0.886/\sqrt{T}$ (5%), and $1.031/\sqrt{T}$ (1%)
- Anderson-Darling: 0.631 (10%), 0.752 (5%), and 1.035 (1%).

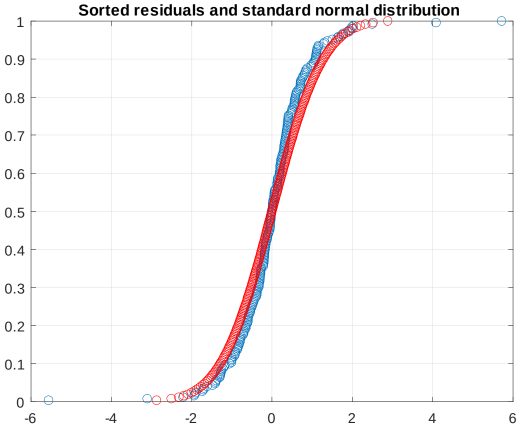
5.5. Normality assumption

Figure: Empirical cdf vs Standard normal cdf



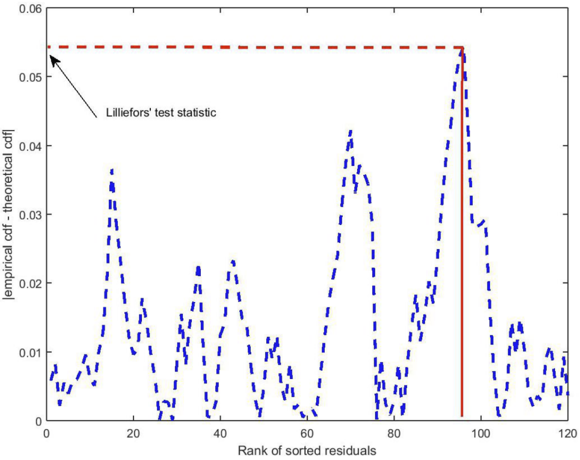
5.5. Normality assumption

Figure: Empirical cdf vs Standard normal cdf



5.5. Normality assumption

Figure: Lilliefors test statistic



Key concepts

1. Visual procedures and testing procedures
2. Misspecification tests
3. Auxiliary regression
4. Perfect and imperfect (near) multi-collinearity
5. Heteroscedasticity and autocorrelation
6. White and Breusch-Pagan tests
7. White's heteroscedasticity consistent standard errors
8. QQ-plots, Jarque Bera and Lilliefors tests

End of Chapter 3

Christophe Hurlin