UNIVERSITY of DEBRECEN

SZÉCHENYI 2020

Introduction to Econometrics for Engineers and Engineering Managers
Part I

Editor:
Judit T. Kiss

The course material/work/publication is supported by the EFOP-3.4.3-16-2016-00021 "Development of the University of Debrecen for the Simultaneous Improvement of Higher Education and its Accessibility" project. The project is supported by the European Union and co-financed by the European Social Fund.

SZÉCHENYI 2020

HUNGARIAN GOVERNMENT

European Union
European Social Fund

INVESTING IN YOUR FUTURE

Author:
Judit T. Kiss

Proofreading:

George Seel

Manuscript closed: 01/03/2018

ISBN 978-963-490-018-4

1. An Introduction to Econometrics

"There are two things you are better off not watching in the making: sausages and econometric estimates" (Leamer, 1983)

In the following chapters, we try to understand the concept of econometrics, the main steps in empirical analysis, and types of data.

1.1. What is Econometrics? What are the goals of econometrics?

Econometrics is the application of statistical methods for estimating economic relationships, creating new economic models, testing existing theories, and evaluating economic policies.

Econometrics is about how we can use theory and data from economics, business, and the social sciences, along with tools from statistics, to answer ''how much'' questions (Hill et al., 2011:3).

By using econometrics we try to understand, characterize and measure economic and socioeconomic phenomena. What kind of questions can arise in this field? Among others:

- What are the main factors influencing the gross domestic product?
- How should we describe the demand for a given product?
- What are the key determinants of consumer choice?
- What is the relationship between inflation and unemployment?
- What is the effect of increasing price on total revenue?

Some previous questions belong to the field of microeconomics, and some to macroeconomics. It is therefore important to distinguish between microeconometrics and macroeconometrics. *Microeconometrics focuses on the behaviour of individual decision-making units such as the decisions of consumers, firms, and workers, based on the analysis of different data (cross section and panel data).* In the field of microeconometrics, professionals generally use the main tools of microeconomics:

- Constrained optimization,
- Equilibrium analysis,
- Comparative statistics.

Constrained optimization is an analytical tool for making the optimal choice, taking into account limitations or restrictions on choice (Besanko-Braeutigam, 2011: 6).
Equilibrium analysis is used to analyse and describe a condition or state that could continue indefinitely in a system, or at least until there is a change in some exogenous variable (Besanko-Braeutigam, 2011: 12).

Comparative statistics: Analysis used to examine how a change in a certain exogenous variable affects the level of a certain endogenous variable in an economic system (Besanko-Braeutigam, 20).

Macroeconomics deals with the operation of an economy as a whole. Macroeconomics treats issues and problems related to economic growth, sustainable development, employment, unemployment, inflation, economic and social institutional systems, and the business cycle, among others. Certain elements of macroeconometrics rely on the analysis of time-series data and panel data.

However, we cannot establish that econometrics can be divided only into micro- and macroeconometrics, because there are no clear dividing lines between the categories, and we cannot forget the other fields of economics, for example the financial element of econometrics.

1.2. Types of data and economic data

In order to make empirical analysis and to make inferences we must have data. How can we collect and derive data?

Types of data

We can distinguish between experimental data and non-experimental data.

Experimental data
Experimental data are often collected under controlled circumstances, where the analyst is able to fix the values of independent variables at a predetermined value, and the experiment can be repeated several times. Using experimental data is very rare in social sciences; however these data are often collected in natural sciences. If we would like to analyse the impact of unemployment benefits on the level of unemployment, or we would like to know the main determinants of Gross Domestic Product (GDP), we can use data from previous years and from different countries. We are not easily able to change the number of employees or the amount of human capital in an economy in order to analyse a change in the unemployment rate and GDP. It is almost impossible to create laboratory environments to analyse the market mechanism and the operation of an economy.

Non-experimental data

Analysts do not have absolute control over the conditions under which data are collected. Analysts are not able to select for given values of independent variables, and the experiment cannot be repeated several times. Non-experimental data are population surveys or other sample surveys, and administrative records, among others. Non-experimental data are sometimes called observational data (or retrospective data).

Economic data

Economic data can be:

- Cross-sectional data,
- Time-series data,
- Panel data.

Cross-sectional data

Cross-sectional data are collected across sample units at a given point of time or a particular time period. Sample units may be individuals, consumers, households, firms, employees, countries, and other economic units. For example, firms draw up and submit to the National Statistical Office in Hungary, on a monthly basis, a definitive statistical report of their activity; we would view these data for one year as a cross-sectional data set. Another example: assume that we would like to analyse the internal rate of return to education. Therefore, we collect data on wages, education, experience, gender and other characteristics by randomly drawing 1 000 individuals from the active population at a given time period. The collected data set can be seen as a cross-sectional data set.

Time-series data

"A time-series data set is collected over discrete intervals of time.
Macroeconomic data are usually reported in monthly, quarterly, or annual terms. Financial data, such as stock prices, can be recorded daily, or at even higher frequencies. The key feature of time-series data is that the same economic quantity is recorded at a regular time interval". (Hill, et al., 2011)

Time series data in Table 1.1 shows the Hungarian GDP data between 2005 and 2016.

Table 1.1 Gross Domestic Product per capita in Hungary between 2000 and 2015.

Year	Gross Domestic Product per capita (HUF)
2000	1 304 629.2
2001	1 510 019.9
2002	1 714 957.1
2003	1 883 335.8
2004	2 080 075.9
2005	2 227 684.9
2006	2 398 186.5
2007	2 541 859.6
2008	2 696 887.9
2009	2 623 798.4
2010	2 708 583.8
2011	2 824 597.6
2012	2 889 059.8
2013	3 045 294.7
2014	3 283 864.9
2015	3 454 121.2

Source: Hungarian Central Statistical Office (2017a)

Time series data are data collected over several time periods.

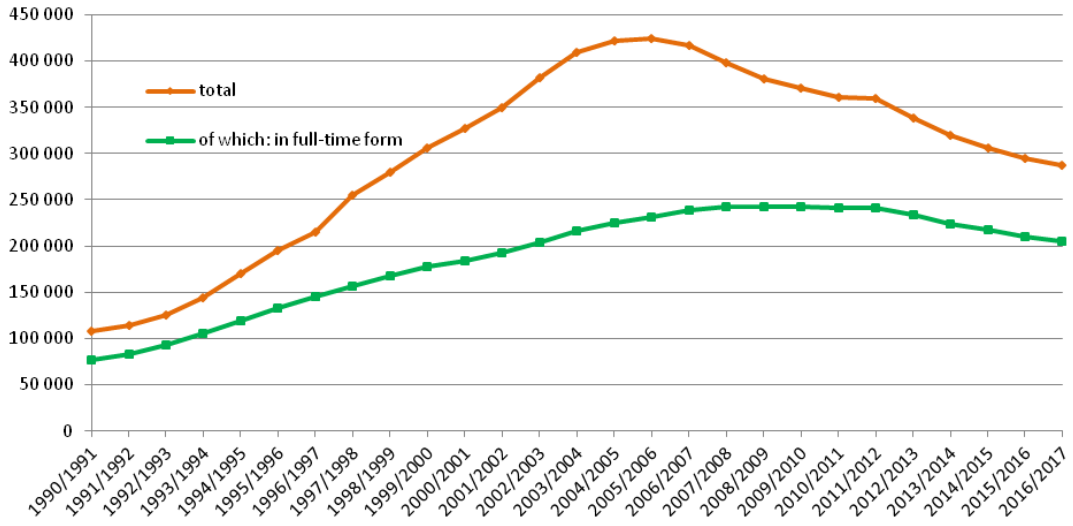
The number of college and university students in Hungary between 1990 and 2017 can be seen as another time-series data set (Table 1.2; Fig. 1.1).

Table 1.2 The number of college and university students in Hungary between 1990 and 2017

School year	Number of students		School year	Number of students	
	total	*of which on full-time courses*		*total*	*of which on full-time courses*
1990/1991	108 376	76 601	**2004/2005**	421 520	225 512
1991/1992	114 690	83 191	**2005/2006**	424 161	231 482
1992/1993	125 874	92 382	**2006/2007**	416 348	238 674
1993/1994	144 560	105 240	**2007/2008**	397 704	242 893
1994/1995	169 940	118 847	**2008/2009**	381 033	242 928
1995/1996	195 586	132 997	**2009/2010**	370 331	242 701
1996/1997	215 115	145 843	**2010/2011**	361 347	240 727
1997/1998	254 693	156 904	**2011/2012**	359 824	241 614
1998/1999	279 397	168 183	**2012/2013**	338 467	233 678
1999/2000	305 702	177 654	**2013/2014**	320 124	223 604
2000/2001	327 289	183 876	**2014/2015**	306 524	217 248
2001/2002	349 301	192 974	**2015/2016**	295 316	210 103
2002/2003	381 560	203 379	**2016/2017**	287 018	205 560
2003/2004	409 075	216 296			

Source: Hungarian Central Statistical Office (2017b)

Figure 1.1 The annual change in the number of college and university students in Hungary between 1990 and 2017



Source: Hungarian Central Statistical Office (2017b)

Graphs similar to Fig 1.1 help researchers to understand the change in time and what happened in the past, to analyse any trends over time; researchers may forecast future values of data by using different econometrics methods.

Panel data

Panel data or longitudinal data represent observations of individual economic-units which are followed over time. Panel data includes the corresponding data for the same cross-sectional units, for example the same set of countries, firms or employees.

If we collect data on the number of students in tertiary education and the amount of GDP for different countries and years, we get a panel data set. Table 1.3 shows the first part (educational data) of the mentioned panel data.

Table 1.3 Students in tertiary education as a percentage of those aged 20-24 years in the population (%)

Country	2013	2014	2015	Country	2013	2014	2015
	(%)	(%)	(%)		(%)	(%)	(%)
Belgium	69.2	70.4	72.2	**Netherlands**	63.9	66.1	78.8
Bulgaria	62.1	65.4	68.9	**Austria**	78.5	77.6	78.6
Czech Republic	64.8	64.9	63.4	**Poland**	71.2	68.0	66.4
Denmark	81.0	81.8	83.2	**Portugal**	64,9	64.3	61.0
Germany (until 1990 former territory of the FRG)	56.9	62.5	64.9	**Romania**	48.5	48.5	48.0
Estonia	70.3	70.0	70.3	**Slovenia**	83.2	83.1	79.5
Ireland	75.8	82.3	90.9	**Slovakia**	54.0	52.1	50.2
Greece	106.8	-	117.8	**Finland**	90.9	89.8	88.4
Spain	80.6	835	84.7	**Sweden**	65.4	63.9	63.8
France	59.3	61.6	63.4	**United Kingdom**	55.2	54.6	54.0
Croatia	64.9	66.5	65.7	**Iceland**	78.4	80.4	75.6
Italy	60.7	59.5	59.3	**Liechtenstein**	37.1	36.8	33.6
Cyprus	45.0	49.9	56.0	**Norway**	75.3	77.3	77.9
Latvia	65.8	66.8	69.0	**Switzerland**	56.2	58.1	59.1
Lithuania	74.3	69.6	-	**Former Yugoslav Republic of Macedonia, the**	38.2	38.9	41.3
Luxembourg	19.9	20.1	19.9	**Serbia**	55.2	57.3	57.6
Hungary	57.0	52.5	49.2	**Turkey**	-	88.1	96.8
Malta	41.6	41.9	44.8				

Source: Eurostat (2017a)

Let assume that the personal income tax system will change at the beginning of the next year. We would like to analyse the effect of the change in personal income tax on household consumption. We survey a random sample this year and another random sample two years later following the introduction of the new personal income tax system, in order to implement the examination. The data set collected from the two surveys will result in a pooled cross section instead of panel data, since the two data sets are taken from different random samples. As we have mentioned, in the case of panel data the sample units are the same.

1.3. The main application areas of econometrics, and the main elements of empirical analysis

The main application areas of econometrics are the following, among others:

- estimating economic links, examining relationship among variables, specifying models,
- testing hypothesis,
- forecasting variables.

Examination of economic relationships

Economists generally want to know whether relationships between certain economic variables might exist. For example,

- What is the relationship between working experience and wages?
- What is the effect of increasing price on total revenue?
- What is the relationship between tax rates and tax revenues?
- Is there any relationship between the tax burden and GDP?
- What is the relationship between supply chain and firm performance?

Important information to enable firm managers to understand how changes in price affect the total revenue includes the relationship between price and total revenue. From our previous microeconomic studies we know that according to the law of demand, there is a negative relationship between price and quantity demanded for a product, when all other influencing factors remain the same. This means that if the price increases the quantity demanded decreases; however, the total revenue might increase and decrease. The change in total revenue depends on the degree of changes in quantity demanded caused by the change in price. The price elasticity shows how the quantity demanded for a product changes if the price increases or decreases.

Testing hypothesis

In the field of economics, there are many theoretical model and different theories. However, we can see rapid economic change over time, during which many theories and models can remain the same or change. As a result of the changing economic relationship and environment, hypothesis testing is very important, despite existing evidence. Under hypothesis testing we examine whether the given sample provides support for the examined theory or provides evidence against the theory.

Forecasting of variables

If we have information about the relationship between variables, we would like to know the future values of the examined variables. We try to estimate the expected values of the examined variables based on their values in the past.

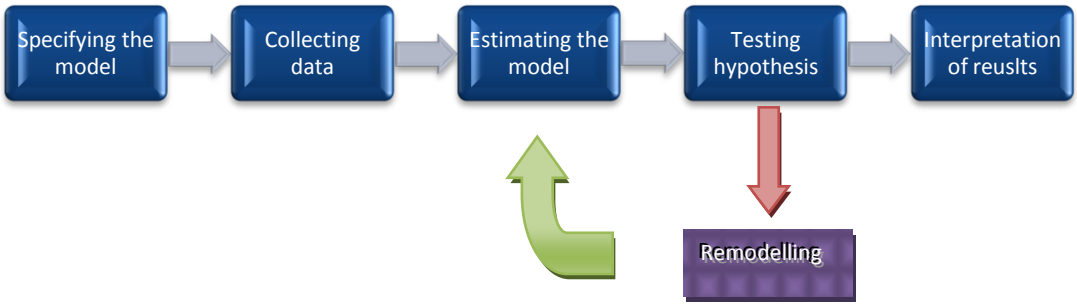
Firm managers would like to know the expected demand for the firm's products, its production costs, and its expected share price. Useful information for the government may be the expected level of unemployment or the forecasted value of Gross Domestic Product.

The main elements of empirical analysis

The main steps of an empirical analysis are the following (Fig. 1):

1. Specifying the model,
2. Collecting data,
3. Estimating the model,
4. Testing the hypothesis,
5. Evaluating and interpreting the results.

Figure 1.2 The main steps of an empirical analysis



1. *Specifying the model*

The model is specified by using an equation or equations which describe(s) the behaviour of economic units or variables, and the relationship between economic variables. The development of the initial model is based on economic theories, studies, or previous empirical results. Let us assume we would like to analyse our students' performance in econometrics. First of all, we choose the main influencing factors, so we try to describe the Student's PERFormance in ECONometrics (sperfecon) as a function of various different factors:

$$\text{Sperfecon} = \text{f(lect, stud, abil, prev, gend)}, \qquad (1.1)$$

where lect is the number of econometrics lectures and seminars attended by the students, stud is the number of hours they have studied, abil characterizes the students' ability, prev represents the previous econometrics knowledge, and gend is the students' gender. The specification of the function (1.1) is based on our intuition, because we do not have knowledge of any similar research. We have to give the form of our function before carrying out the econometrics analysis. Let us assume that the function form (1.1) is linear; in this case we can give the function as the following:

$$\text{Sperfecon} = \beta_0 + \beta_1 \cdot \text{lect} + \beta_2 \cdot \text{stud} + \beta_3 \cdot \text{abil} + \beta_4 \cdot \text{prev} + \beta_5 \cdot \text{gend} + \text{u}, \quad (1.2)$$

where u is the error term, which includes the unobserved factors.

2. *Collecting data*

In the next step, we collect data in order to test our model. In the previous subchapter, we have mentioned that the data can be non-experimental or experimental data. We can use data from previous surveys, or we can make surveys in different ways ourselves. In this section, it is important to decide how to manage the unobservable variables. For example, a student's ability is difficult to observe; in this case we can choose other variables to describe a student's ability, such as mathematics test scores.

3. *Estimating the model and testing the model*

In this step, we try to quantify our model. We estimate the unknown parameters (β_0, $\beta_1, \dots, \beta_5$) of the econometric model. By estimating the unknown parameters we can quantify the relationship between the dependent ($Sperfecon$) and independent variables($lect, stud, abil, prev, gend$). We test our model to check the underlying assumptions and the validity of the chosen functional form and parameters. It may be necessary to modify our model in the light of the test results.

4. *Evaluation and interpretation of results*

In the last step, we evaluate and interpret our model, before finally drawing conclusions. We can verify previous economic statements, or we can explore new relationships between economic variables, among others.

Example 1.1

Let us assume that we would like to analyse the quantity of ice cream demanded. Give the main variables of the quantity of ice cream demanded; explain your reasoning and give the direction of the relationship.

From our previous microeconomics studies we know that the demand for a good depends on the price of the good, the price of substitute goods, the consumers' tastes, the consumers' income, and the price of complementary goods. The quantity of ice cream demanded as a dependent variable can be given as a function of different independent variables:

$$\text{Quantity of ice cream demanded} = \text{f(price ic, price sg, tastes, income, price cg)},$$

where

price ic = the price of ice cream,

price sg = price of substitute goods,

tastes = consumers' tastes,

income = consumers' income,

price cg = price of complementary goods.

According to the law of demand, there is a negative relationship between the price of ice cream and the quantity demanded. This means that if the price of ice cream increases, consumers buy less ice cream, all other things being equal.
Let us assume that the price of a substitute good for ice cream decreases. According to the law of demand, the demand for the substitute good (frozen yoghurt) increases, and the demand for the ice cream probably decreases, all other things being equal.
If the consumers' income decreases, they have less to spend on different kind of goods, and the quantity of ice cream demanded decreases, if all other influencing factors remain the same. We must mention that the relationship between quantity demanded and consumers' income is not the same for all goods. It depends on the measure of the income elasticity of a good. Goods can be normal, inferior and superior (luxury) goods. If ice cream can be seen as a normal good, the quantity demanded is positively related to the consumers' income.
Finally, suppose that the price of ice cream cones increases. What is the effect of the change in the price of ice cream cones on the quantity of ice cream demanded? The price of ice cream increases as the price of ice cream cones increases, which means that the quantity of ice cream demanded decreases. Ice cream and ice cream cones are complementary goods; the price of ice cream cones and the quantity demanded move in the opposite direction, ceteris paribus.
The model of the quantity of ice cream demanded can be specified as the following:

$$\text{Sperfecon} = \beta_0 + \beta_1 \cdot \text{price ic} + \beta_2 \cdot \text{price sg} + \beta_3 \cdot \text{tastes} + \beta_4 \cdot \text{income} + \beta_5 \cdot \text{price cg} + \\ + u,$$

where u is the error term, and β_i are the parameters of the independent variables.

We have repeatedly mentioned the following expressions: all other things being equal, ceteris paribus, all other influencing factors remain the same. Ceteris paribus means that other things are equal or other influencing factors are equal. If we would like to analyse the effect of a change in an independent variable on the dependent variable, it is necessary to assume that only the given independent variable changes and other factors remain the same.

1.4. Terms and Questions

ceteris paribus,
complementary goods,
cross-sectional data set,
econometrics,
experimental data,
forecasting of variables,
inferior goods,
law of demand,
luxury goods,
macroeconometrics,
microeconometrics,
non-experimental data,
normal goods,
observational data,
panel data,
pooled cross sectional data set,
specifying the model,
substitute goods
superior goods,
testing hypothesis,
time series data set.

Problems

Theoretical questions

1. What is the difference between experimental data and non-experimental data?

2. What is econometrics?

3. Give the definition of a cross sectional data set.

4. How can we specify a model?

5. Explain what the relationship is between specifying the model and testing the model.

6. What is the difference between a cross-sectional data set and a pooled cross sectional data set? Give an example of each.

7. Give the main steps of empirical research.

8. Explain what the difference is between microeconometrics and macroeconometrics.

9. Give an example of a time series data set.

10. Give the definition of a panel data set.

11. Give an example of an econometric model.

12. What is the difference between statistics and econometrics?

Calculation exercise

1.

Suppose that you would like to analyse students' height at Debrecen University. Give the height as a function of independent variables. List the main independent variables.

2.

Which of the following pairs can be seen as dependent and independent variables (separately)?

a) The growth rate of the price level and the real interest rate.
b) The amount of GDP in an economy and the level of human capital in the same economy.
c) Nationality of students and their height.
d) The aggregate net investment and the amount of aggregate income.
e) The quantity of pencils demanded and the number of primary school pupils in a country.
f) The quality of the lecture and the number of students.

3.

Give an example of a dependent variable and at least two independent variables which affect the dependent variable. Give the direction of the relationship.

4.

You would like to analyse the price of flats in Debrecen. The price is the dependent variable of your econometric model. Give the list of independent variables and the direction of the relationship separately.

5.

Assume that your lecturer has collected data on the performance of students. These data are summarized in the Table below. The required minimum level of the econometrics test was 100 points, and the maximum score was 200 points.

Students' performance		
Observation	Test points	Time spent on econometrics studies (minutes)
1	195	684
2	200	720
3	58	144
4	82	216
5	74	192
6	200	960
7	32	66
8	194	552
9	200	714
10	200	746.4
11	80	294
12	180	576
13	94	300
14	52	144
15	88	343.2

a) Find the smallest and largest values of the test points.
b) Find the average of the test points in the sample.
c) Find the average of the time spent on studying in the sample.
d) How many students have passed?
e) Which one is the dependent variable?
f) Does it make sense to think that the students' performance depends on the quality of the teacher's lectures?
g) Give the direction of the relationship between the two variables (dependent variable and independent variable).

h) List other independent variables which can have an impact on the dependent variables. Explain your reasoning and the give direction of the relationship.

6.

An estate agent collected data on 20 flats in Debrecen in 2017.

Data of flats			
Number	Price	Floor area	Age
	Million HUF	Square metres	years
1	15.6	48	2
2	26.4	55	2
3	13.6	71	3
4	26.8	82	0
5	28.8	100	3
6	38.4	85	1
7	19.2	70	8
8	17.6	73	9
9	17.2	74	10
10	15.6	66	7
11	7.6	35	47
12	13.6	53	18
13	15.2	73	18
14	8	39	63
15	12.8	67	23
16	8.8	48	63
17	10.8	51	31
18	21.6	61	5
19	10	53	40
20	9.2	54	78

a) Give the types of the given data set.
b) Find the smallest and largest values of the prices.
c) Find the average of the prices in the sample.
d) Which one is the dependent variable?
e) What is the relationship between the price and the floor area?
f) Is the relationship between price and age positive or negative?
g) Give an independent variable which has a positive effect on the price.
h) Give an independent variable which has a negative effect on the price.

7.

Assume that you built a model of the price of the *ith* flat as a function of the floor area (FLA), and the result of your estimation is the following:

$$\widehat{Price}_i = 3.2 + 0.4 \cdot \mathrm{FLA_i}.$$

a) Give the graph of the estimated function. (Excel)
b) Give the mathematical meaning of the estimated coefficients.
c) Interpret the economic meaning of the estimated coefficients.
d) You have possibility to add another independent variable to the equation. Give your choice and explain it.

8.

Assume that you built a model of price of the *ith* flat as a function of floor area (FLA), age (AGE), and the number of rooms (ROOM):

$$Price_i = \beta_0 + \beta_1 \cdot \mathrm{FLA_i} + \beta_2 \cdot \mathrm{AGE_i} + \beta_3 \cdot \mathrm{ROOM_i} + \mathrm{u_i}.$$

a) What is the meaning of the β_1 coefficient?
b) What is the meaning of the β_2 coefficient?
c) What is the meaning of the β_3 coefficient?
d) You have the possibility to add another independent variable to the equation. Give your choice and explain it.

2. Simple Linear Regression Analysis

In the subsequent sections we will build a model in order to study the relationship between only two variables, and we will try to interpret the estimated model. This is the first step in an econometric analysis. The first part of this chapter deals with the linear regression model with only two variables, and in the subsequent sections we will study how to test the goodness of the model and evaluate the results.

2.1. An Introduction – correlation versus regression

In this section, we introduce the concept of a linear regression model, show several varieties of such a model, and explain the estimation method (least squares) that is generally applied with regression models. Regression Analysis is the process used to describe the relationship between two or more variables. If we would like to analyse only the strength of the relationship between two variables, we can use a correlation coefficient, such as the Pearson correlation coefficient.

Correlation
The correlation between variables is a measure of the nature and degree of association between the variables.

According to our previous studies, we know that the Pearson correlation coefficient (linear correlation coefficient) can be given by the following equation:

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}, \qquad (2.1)$$

where X and Y are the two variables, $\bar{X}$ $(\bar{Y})$ is the mean of variable X (Y), and n is the number of observations. The value of the linear correlation coefficient can be found between -1 and 1:

$$-1 \leq r_{xy} \leq 1.$$

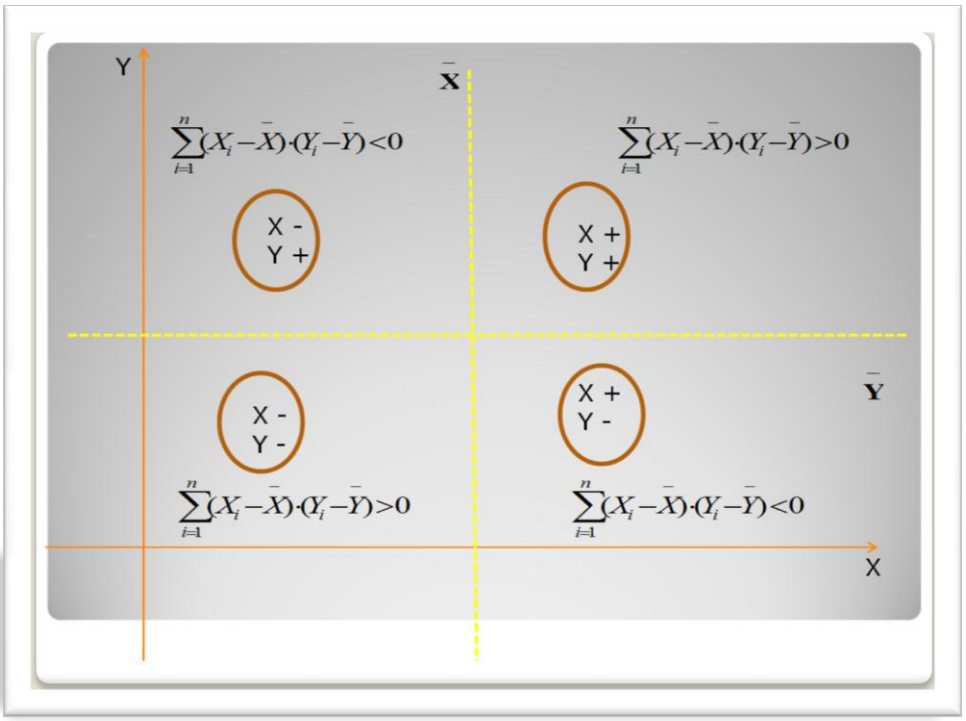
There is a positive correlation between two variables if the two variables change in the same direction. This means that if one variable increases (decreases), the other variable also increases (decreases). In the case of a positive association, the coefficient is positive:

$$0 \leq r_{xy}.$$

A perfect positive (negative) correlation exists when the value of the indicator is equal to 1 (-1). If there is no correlation between the two variables, the correlation coefficient is zero. We can say that two variables are independent: the change in one variable has no effect on the other variable.
The Pearson correlation coefficient measures the linear correlation between two variables.

Pearson's correlation coefficient is equal to the ratio of the *covariance* of the two variables and their *standard deviations*.
The correlation measures the nature and degree of association between the variables; however it does not show the functional relationship between the variables and the causality.

Figure 2.1 The Pearson correlation coefficient



Example 2.1

Consider the following time-series data on the total expenditure for advertising and the total revenue of a given firm. Examine the association between the two variables.

Years	**Total expenditure on advertising** (million HUF)	**Total revenue** (million HUF)
1.	8	20
2.	7	16
3.	4	15
4.	3	14
5.	5	19
6.	4	12
7.	5	18
8.	7	24
9.	3	16
10.	5	22
11.	9	28
12.	6	25
SUM	66	229
Mean	5.5	19.08

To examine the association between total expenditure on advertising and the total revenue, we calculate the Pearson correlation coefficient:

Years	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$
1.	2.5	0.92	2.29
2.	1.5	-3.08	-4.63
3.	-1.5	-4.08	6.13
4.	-2.5	-5.08	12.71
5.	-0.5	-0.08	0.04
6.	-1.5	-7.08	10.63
7.	-0.5	-1.08	0.54
8.	1.5	4.92	7.38
9.	-2.5	-3.08	7.71
10.	-0.5	2.92	-1.46
11.	3.5	8.92	31.21
12.	0.5	5.92	2.96
Sum			75.5
Sum of Squares	41	260.92	

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{75.5}{\sqrt{41 \cdot 206.92}} = \frac{75.5}{\sqrt{10\,697.72}} = 0.73$$

According to the result, we can see that there is a relatively strong association between the two variables. The correlation coefficient is positive; this refers to the positive correlation between total expenditure on advertising and the total revenue. However, there is no information on the causation. We do not know whether the change in expenditure on advertising causes the change in total revenue, or vice versa.

Example 2.2

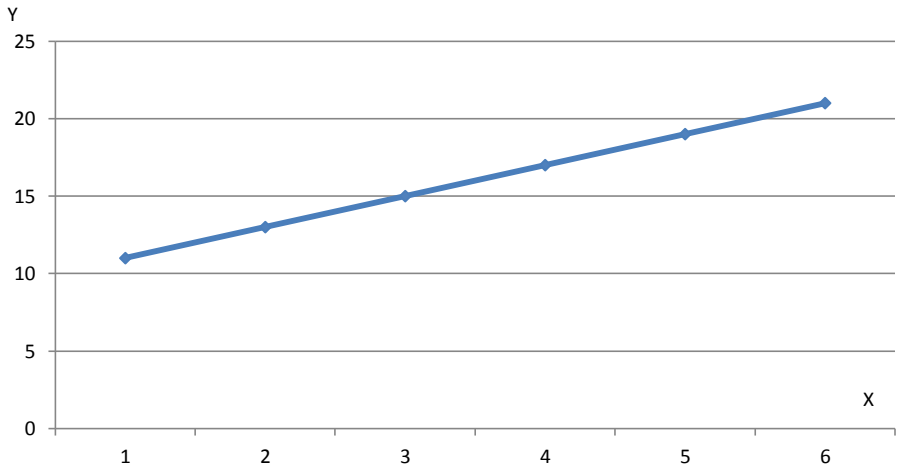
Examine the Pearson correlation coefficient for the following data:

Observations	X_i	Y_i
1.	1	11
2.	2	13
3.	3	15
4.	4	17
5.	5	19
6.	6	21
Mean	3.5	16

Years	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$
1.	2.5	0.92	2.29
2.	1.5	-3.08	-4.63
3.	-1.5	-4.08	6.13
4.	-2.5	-5.08	12.71
5.	-0.5	-0.08	0.04
6.	-1.5	-7.08	10.63
Sum			35
Sum of Squares	17.5	70	

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{35}{\sqrt{17.5 \cdot 70}} = \frac{35}{\sqrt{1225}} = 1.$$

There is a perfect positive correlation between the two variables. If we illustrate the given data, we can see the perfect linear association:



The equation of the function is:

$$f(x) = 2 \cdot x + 9.$$

Example 2.3

Given the following function:

$$f(x) = x^2 + 3.$$

The following table contains y values when x goes from 1 to 6.

Observations	X_i	Y_i
1.	1	4
2.	2	7
3.	3	12
4.	4	19
5.	5	28
6.	6	39
Mean	3.5	18.17

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{122.5}{\sqrt{17.5 \cdot 894.8}} = 0.9789.$$

We might expect a perfect correlation, because there is a functional relationship between *X* and *Y*; however, the value of the coefficient is less than 1. The reason for the result comes from the main characteristic of the indicator. As we mentioned earlier, the Pearson coefficient measures the degree of linear association. The relationship between the dependent (*y*) and independent variables (*x*) is other than linear, since the dependent variable is given by a parabolic function of the independent variable:

$$y = x^2 + 3.$$

2.2. Simple Regression Model

2.2.1. The theoretical regression equation

As we have seen in the previous section, the correlation coefficient shows us whether two variables are associated with one another, but it does not provide information about the kind of relationship.

The starting point is the simple linear regression model, and we assume that there are two variables. It is further assumed that the relationship is linear between the two variables, and the theoretical regression equation is:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i, \qquad (2.2)$$

where *Y* is the dependent variable (or explained variable) and *X* is the independent variable (or explanatory variable). Y_i (X_i) shows the *ith* observation of the dependent variable (independent variable). *Y* is also called the response variable, while X is called the controlled variable (or regressor). Finally *u* is the error term or disturbance; it (*u*) represents other omitted independent variables which may explain the behaviour of the dependent variable *Y*. The error term represents all of the variation in *Y* that cannot be explained by *X*:

- measurement error (for example sampling error),

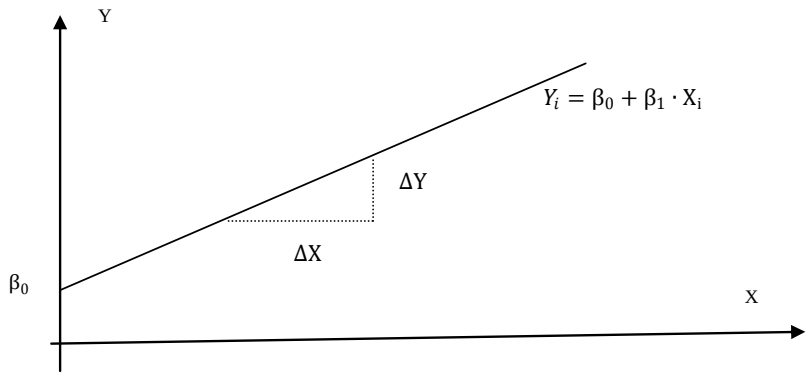
- inadequate function specification,
- omitted independent (explanatory) variables,
- omitted data at the estimation, because the data are not available.

There are two coefficients (parameters) in equation (2.2); β_0 is the constant term, and β_1 is the slope coefficient. The coefficient β_0 shows the intercept value where the linear function intercepts the vertical axis, when the value of *X* is zero. The coefficient β_1 shows the amount that Y changes as X increases (or decreases) by one unit. The coefficient β_1 is the slope of the line:

$$\beta_1 = \frac{\Delta Y}{\Delta X}.$$

The slope β_1 is constant over the entire function (Fig. 2.2).

Figure 2.2 The regression line



Equation (2.2) represents a simple regression model, because it contains only one independent (explanatory) variable. The model is called a *multiple regression model* if two or more independent variables are included in the model.
In the regression equation (2.2), the first part ($\beta_0 + \beta_1 \cdot X_i$) is the *deterministic term* or non-random component, because it represents the value of the dependent variable (Y) at the given value of X. The second part of the regression equation (2.2) is the *stochastic or random component*. The regression line illustrated in Fig. 2.2 is called the theoretical regression line. It is only theoretical because it can never be observed; we can only estimate the parameters of the line.

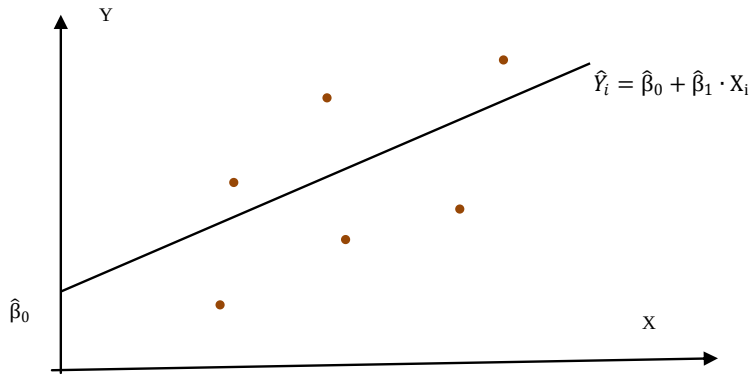
2.2.2. The estimated regression equation

Let us assume you have six observations of the independent variable and its dependent variable. You would like to give the best estimation of the parameters (β_0, β_1) in equation (2.2). You illustrate the estimated line in Figure 2.3. The intersection of the fitted line with the vertical axis is an estimation of β_0, and it is denoted by $\hat{\beta}_0$. Similarly the estimation of the slope β_1 is $\hat{\beta}_1$. The estimated equation is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i,$$

where $\hat{Y}_i$ ("*Y hat*") is the estimated (or fitted) value of Y_i.

Figure 2.3 The estimated line

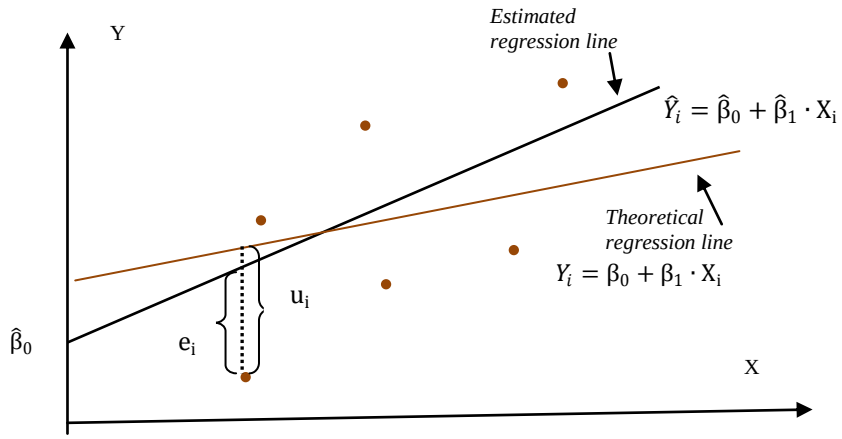


The difference between the actual value of the dependent variable and the estimated value of the dependent variable is the $residual$ (e_i):

$$e_i = Y_i - \hat{Y}_i.$$

The smaller the residual, the better is the estimation, because the difference between Y_i and $\hat{Y}_i$ is smaller. The residual generally is denoted by $\hat{u}_i$. It should be noted that the residual and the error term are not identical concepts; their values are different. The error term is the difference between the observed value of the dependent variable and the theoretical regression line (Fig.2.4).

Figure 2.4 The estimated and the theoretical regression line



2.2.3. The estimation of the regression coefficients of the econometric model

We estimate the regression parameters by using the OLS (Ordinary Least Squares) method. The main technique of the method is to minimize the sum of the squared residual:

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \quad (2.3)$$

We minimize the sum of the squared difference between the actual value and the estimated value of the Ys, over all the observation data. The first step is to take the difference between the actual value of Y and the estimated value of Y. In the next step we take the squares of the differences; in other cases some of the sums of the differences may eliminate each other. Let us substitute the function of the estimated regression line $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i$ instead of $\hat{Y}_i$:

$$F = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i)^2.$$

Finally, we find the values of the regression coefficient $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimize the sum of the squared function (F):

We take the partial derivatives of the F function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and we get two equations:

$$\frac{\partial F}{\partial \hat{\beta}_0} = 2 \cdot \left(\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i) \cdot (-1) \right) = (-2 \cdot) \sum_{i=1}^{n} Y_i + 2 \cdot n \cdot \hat{\beta}_0 + 2 \cdot \hat{\beta}_1 \sum_{i=1}^{n} X_i,$$

$$\frac{\partial F}{\partial \hat{\beta}_1} = 2 \cdot \left(\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot X_i \right) \cdot (-X_i)) =$$

$$= (-2 \cdot) \sum_{i=1}^{n} (Y_i \cdot X_i) + 2 \cdot \hat{\beta}_0 \cdot \sum_{i=1}^{n} X_i + 2 \cdot \hat{\beta}_1 \sum_{i=1}^{n} X_i^2.$$

The first-order condition for the minimum is $\frac{\partial F}{\partial \hat{\beta}_0} = 0$ and $\frac{\partial F}{\partial \hat{\beta}_1}$. To determine the minimum point, the two equations must be set to equal zero:

$$(-2 \cdot) \sum_{i=1}^{n} Y_i + 2 \cdot n \cdot \hat{\beta}_0 + 2 \cdot \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i = 0, \quad (2.4)$$

$$(-2 \cdot) \sum_{i=1}^{n} (Y_i \cdot X_i) + 2 \cdot \hat{\beta}_0 \cdot \sum_{i=1}^{n} X_i + 2 \cdot \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i^2 = 0. \quad (2.5)$$

$$\sum_{i=1}^{n} Y_i = n \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i, \quad (2.4a)$$

$$\sum_{i=1}^{n} (Y_i \cdot X_i) = \hat{\beta}_0 \cdot \sum_{i=1}^{n} X_i + \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i^2 \quad (2.5a)$$

The equations (2.4a) and (2.5a) are called normal equations of the regression model.

Rewrite the equation (2.4a) by using the arithmetical average of *X* and *Y*:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n},$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n},$$

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{X}.$$

We solve the two equations (2.4a) and (2.5a) for $\hat{\beta}_0$ and $\hat{\beta}_1$; substituting $\hat{\beta}_1$in equation (2.5a) we obtain:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X},$$

$$\sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) = (\bar{Y} - \hat{\beta}_1 \cdot \bar{X}) \cdot \sum_{i=1}^{n}\mathrm{X_i} + \hat{\beta}_1 \cdot \sum_{i=1}^{n}\mathrm{X_i}^2. \quad (2.5b)$$

$$\sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) = n \cdot \bar{Y} \cdot \bar{X} - \hat{\beta}_1 \cdot n \cdot \bar{X}^2 + \hat{\beta}_1 \cdot \sum_{i=1}^{n}\mathrm{X_i}^2. \quad (2.5c)$$

Let us separate the terms involving $\hat{\beta}_1$ and put them on one side of the equation (2.5c):

$$\hat{\beta}_1 \cdot \left(\sum_{i=1}^{n}\mathrm{X_i}^2 - n \cdot \bar{X}^2\right) = \sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - n \cdot \bar{Y} \cdot \bar{X}.$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - n \cdot \bar{Y} \cdot \bar{X}}{\sum_{i=1}^{n}\mathrm{X_i}^2 - n \cdot \bar{X}^2}. \quad (2.6)$$

The next formula is also generally used in OLS estimations. Prove it:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(\mathrm{X_i} - \bar{X})^2}. \quad (2.7)$$

We modify the formula (2.6) by introducing the average of X and Y ($\bar{X}$, $\bar{Y}$) where it is possible, in order to derive the formula (2.7):

$$\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}}) \cdot (Y_i - \bar{Y}) = \sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - \bar{Y} \cdot \sum_{i=1}^{n}X_i - \bar{X} \cdot \sum_{i=1}^{n}Y_i + n \cdot \bar{X} \cdot \bar{Y},$$

$$\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}}) \cdot (Y_i - \bar{Y}) = \sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - n \cdot \bar{Y} \cdot \bar{X} - n \cdot \bar{Y} \cdot \bar{X} + n \cdot \bar{X} \cdot \bar{Y},$$

$$\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}}) \cdot (Y_i - \bar{Y}) = \sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - n \cdot \bar{Y} \cdot \bar{X}.$$

The denominator can be written as follows:

$$\sum_{i=1}^{n}(\mathrm{X_i} - \bar{X})^2 = \sum_{i=1}^{n}\mathrm{X_i}^2 - 2 \cdot \bar{X} \cdot \sum_{i=1}^{n}X_i + n \cdot \bar{X}^2 = \sum_{i=1}^{n}\mathrm{X_i}^2 - 2 \cdot n \cdot \bar{X}^2 + n \cdot \bar{X}^2,$$

$$\sum_{i=1}^{n}(\mathrm{X_i} - \bar{X})^2 = \sum_{i=1}^{n}\mathrm{X_i}^2 - n \cdot \bar{X}^2.$$

The least squares estimators – The estimated values of the regression parameters

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} \quad (2.8)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{\mathrm{X}}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(\mathrm{X_i} - \bar{X})^2} \quad (2.9)$$

Example 2.4

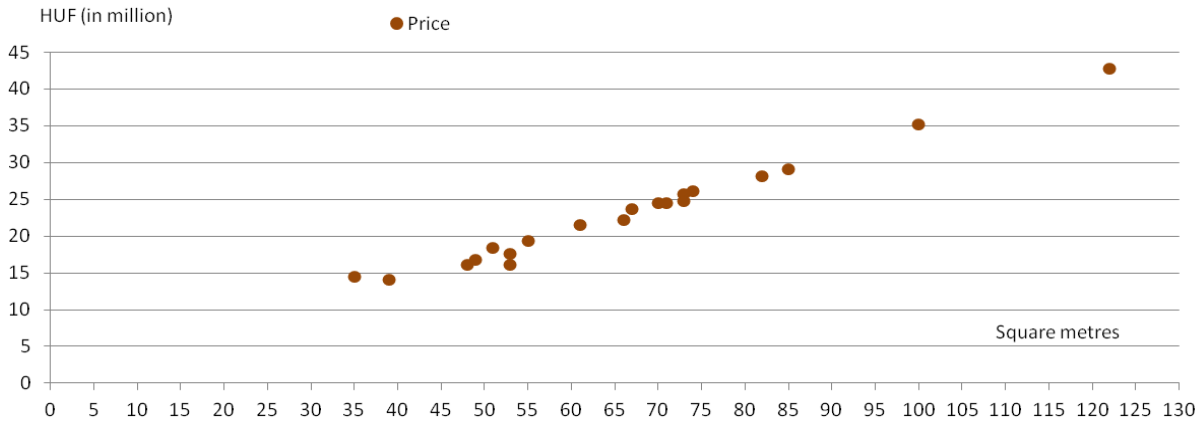
An estate agent collected data on the price and floor area of 20 flats in Debrecen in September 2017.

Data of flats		
Number	Price	Floor area
	Million HUF	Square metres
1	16.2	48
2	19.4	55
3	24.5	71
4	28.2	82
5	35.2	100
6	29.1	85
7	24.6	70
8	25.8	73
9	26.1	74
10	22.2	66
11	14.5	35
12	17.6	53
13	24.8	73
14	14.1	39
15	23.75	67
16	16.8	49
17	18.4	51
18	21.5	61
19	16.2	53
20	43	122

Given are twenty observations for two variables, X and Y. Let us illustrate the data on a scatter diagram.

A scatter diagram illustrates the relationship between two variables (Fig. 2.5). The first variable is the dependent (the price of the flat) and the second one is the independent variable (the floor area of the flat).

Figure 2.5 Scatter diagram – Prices of flats and floor area of flats



Let us determine the estimated regression equation by computing the parameters β_0 and β_1. The determination of β_0 requires the means of X and Y; compute the value of the two means:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{461.75}{20} = 23.0975,$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{1327}{20} = 66.35.$$

Develop calculations for the regression analysis in the next table.

Number	Price Million HUF	Floor area Square metres	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	16.20	48.00	-6.90	-18.35	126.57	336.72
2	19.40	55.00	-3.70	-11.35	41.97	128.82
3	24.50	71.00	1.40	4.65	6.52	21.62
4	28.20	82.00	5.10	15.65	79.85	244.92
5	35.20	100.00	12.10	33.65	407.25	1132.32
6	29.10	85.00	6.00	18.65	111.95	347.82
7	24.60	70.00	1.50	3.65	5.48	13.32
8	25.80	73.00	2.70	6.65	17.97	44.22
9	26.10	74.00	3.00	7.65	22.97	58.52
10	22.20	66.00	-0.90	-0.35	0.31	0.12
11	14.50	35.00	-8.60	-31.35	269.53	982.82
12	17.60	53.00	-5.50	-13.35	73.39	178.22
13	24.80	73.00	1.70	6.65	11.32	44.22
14	14.10	39.00	-9.00	-27.35	246.08	748.02
15	23.75	67.00	0.65	0.65	0.42	0.42
16	16.80	49.00	-6.30	-17.35	109.26	301.02
17	18.40	51.00	-4.70	-15.35	72.11	235.62
18	21.50	61.00	-1.60	-5.35	8.55	28.62
19	16.20	53.00	-6.90	-13.35	92.08	178.22
20	43.00	122.00	19.90	55.65	1107.57	3096.92
Sum	461.95	1327.00			2811.17	8122.55
Mean	23.0975	66.35				406.13

Using equation (2.9) and the information in Table, we can calculate the slope of the regression function:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{2811.17}{8122.55} = 0.34609421,$$

or

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) - n \cdot \bar{Y} \cdot \bar{X}}{\sum_{i=1}^{n} \mathrm{X_i}^2 - n \cdot \bar{X}^2} = \frac{33461.55 - 20 \cdot 23.0975 \cdot 66.35}{96169 - 20 \cdot 66.35^2} = \frac{2811.17}{8122.55} =$$
$$= 0.34609421.$$

Using equation (2.8), the intercept of the regression function can be calculated:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} = 23.0975 - 0.34609421 \cdot 66.35 = 0.134149.$$

The normal equations (2.4a) and (2.5a) are as follows:

$$\sum_{i=1}^{n} \mathrm{Y_i} = \mathrm{n} \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot \sum_{i=1}^{n} \mathrm{X_i},$$

$$\sum_{i=1}^{n}(Y_i \cdot \mathrm{X_i}) = \hat{\beta}_0 \cdot \sum_{i=1}^{n} \mathrm{X_i} + \hat{\beta}_1 \cdot \sum_{i=1}^{n} \mathrm{X_i}^2$$

$$461.95 = 20 \cdot \hat{\beta}_0 + 1\,327 \cdot \hat{\beta}_1,$$
$$33\,461.55 = 1\,327 \cdot \hat{\beta}_0 + \hat{\beta}_1 \cdot 96\,169.$$

Solving the two normal equation for $\hat{\beta}_0$ and $\hat{\beta}_1$, we get the parameters, which are:

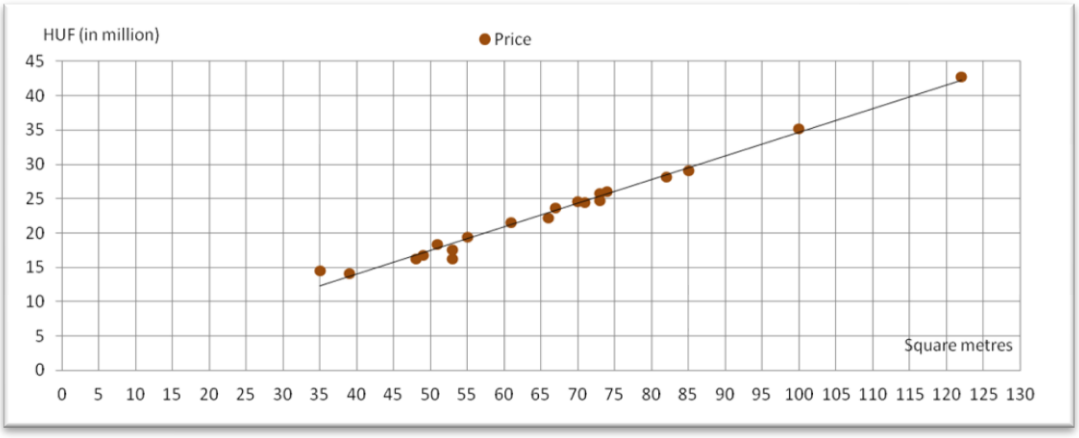
$$\hat{\beta}_1 = 0.34609.$$

$$\hat{\beta}_0 = 0.134.$$

The estimated regression function is (Fig. 2.6):

$$\hat{Y} = 0.134149 + 0.34609 \cdot \mathrm{X}.$$

Figure 2.6 The estimated and the theoretical regression line



The slope of the regression line is positive, indicating that as the square metre area increases, the price also increases. The value of the slope is $\hat{\beta}_1 = 0.34609 \approx 0.346$. This represents that an increase (decrease) in the floor area of a flat by one square metre results in an increase (decrease) in price of approximately 346 000 HUF, other influencing factor remaining the same.

2.3. Total sum of squares, explained sum of squares, and residual sum of squares

In the previous subchapter we minimized the sum of squared residuals:

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

The sum of squared residuals is denoted by SSE (sum of squares due to error). The variation of the dependent variable around the sample mean can be divided into parts. Let us start with the formula of the residual:

$$e_i = Y_i - \hat{Y}_i,$$
$$Y_i = \hat{Y}_i + e_i.$$

Subtracting the sample mean from both sides of the equation, we obtain:

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + e_i,$$

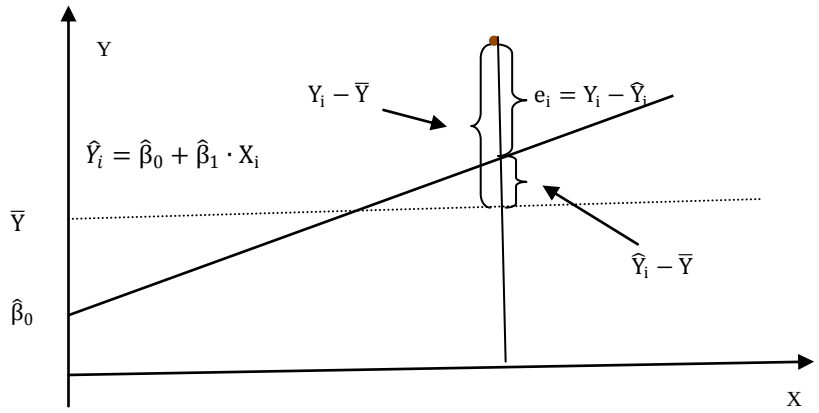
Substituting $Y_i - \hat{Y}_i$ instead of e_i, we get:

$$Y_i - \bar{Y} = \hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i.$$

The difference between the actual value of Y and the mean value of Y has two parts:

- $(\hat{Y}_i - \bar{Y})$ the difference between the estimated value of Y and the mean value of Y, and
- $(Y_i - \hat{Y}_i)$ the difference between the actual value of Y and the estimated value of Y (Fig. 2.7).

Figure 2.7 The explained and unexplained parts of Y



We take the squared sum of both sides of the equation, and we obtain:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}) \cdot \left(Y_i - \hat{Y}_i\right),$$

$$2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}) \cdot \left(Y_i - \hat{Y}_i\right) = 0.$$

From the first normal equation (2.4a), we get:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right) = \sum_{i=1}^{n} Y_i - n \cdot \hat{\beta}_0 - \hat{\beta}_1 \cdot \sum_{i=1}^{n} X_i = 0,$$

$$2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}) \cdot \left(Y_i - \hat{Y}_i\right) = 2 \cdot \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}) \cdot e_i = 2 \cdot \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 \cdot X_i - \bar{Y}) \cdot e_i,$$

$$2 \cdot \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 \cdot \mathrm{X_i} - \bar{Y}) \cdot e_i = 2 \cdot (\hat{\beta}_0 \cdot \sum_{i=1}^{n} e_i + \hat{\beta}_1 \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i} \cdot e_i - \bar{Y} \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} e_i),$$

$$\sum_{i=1}^{n} e_i = 0$$

From the first normal equation (2.5a), we get:

$$\hat{\beta}_1 \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i} \cdot e_i = \hat{\beta}_1 \cdot \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{X_i} \cdot (\mathrm{Y_i} - \mathrm{n} \cdot \hat{\beta}_0 - \hat{\beta}_1 \cdot \mathrm{X_i}) = 0.$$

Finally we get:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, \qquad (2.10)$$

If we would like to analyse the total sample variability, we can examine the variation of the dependent variable about the sample mean $\bar{Y}$:

$$\mathrm{SST} = \sum_{\mathrm{i=1}}^{\mathrm{n}}(Y_i - \bar{Y})^2$$

Variance Analysis

The total sum of squares shows the total variations of Y around the sample mean. The total sum of squares consists of two parts: the explained sum of squares (SSR) and the residual sum of squares (SSE):

$$SSR = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2,$$

$$SSE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

$$\mathrm{SST} = SSR + SSE.$$

The decomposition of the variance is:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2.$$

The variation of the estimated values of Y around the sample mean is the explained sum of squares. It is the sum of squares due to the regression. SSR is the component of total variation in Y around its mean that is explained by the regression. SSE is the component of total variation in Y around its mean that is not explained by the regression.

> *"Looking at the overall fit of an estimated model is useful not only for evaluating the quality of the regression, but also for comparing models that have different data sets or combinations of independent variables. We can never be sure that one estimated model represents the truth any more than another, but evaluating the quality of the fit of the equation is one ingredient in a choice between different formulations of a regression model. Be careful, however! The quality of the fit is a minor ingredient in this choice, and many beginning researchers allow themselves to be overly influenced by it."*
>
> Studenmund (2014)

Example 2.5

Let us give the decomposition of the variance for the exercise 2.4.

Number	Price Million HUF	Floor area Square metres	$(Y_i - \bar{Y})$	$\hat{Y}_i$	$(\hat{Y}_i - \bar{Y})$	$(Y_i - \hat{Y}_i)$
1	16.20	48.00	-6.90	16.75	-6.351	-0.55
2	19.40	55.00	-3.70	19.17	-3.928	0.23
3	24.50	71.00	1.40	24.71	1.609	-0.21
4	28.20	82.00	5.10	28.51	5.416	-0.31
5	35.20	100.00	12.10	34.74	11.646	0.46
6	29.10	85.00	6.00	29.55	6.455	-0.45
7	24.60	70.00	1.50	24.36	1.263	0.24
8	25.80	73.00	2.70	25.40	2.302	0.40
9	26.10	74.00	3.00	25.75	2.648	0.35
10	22.20	66.00	-0.90	22.98	-0.121	-0.78
11	14.50	35.00	-8.60	12.25	-10.850	2.25
12	17.60	53.00	-5.50	18.48	-4.620	-0.88
13	24.80	73.00	1.70	25.40	2.302	-0.60
14	14.10	39.00	-9.00	13.63	-9.466	0.47
15	23.75	67.00	0.65	23.32	0.225	0.43
16	16.80	49.00	-6.30	17.09	-6.005	-0.29
17	18.40	51.00	-4.70	17.78	-5.313	0.62
18	21.50	61.00	-1.60	21.25	-1.852	0.25
19	16.20	53.00	-6.90	18.48	-4.620	-2.28
20	43.00	122.00	19.90	42.36	19.260	0.64
Sum	461.95	1327.00				
Mean	23.0975	66.35				
Sum of squares			987.512375		972.9288	14.5836

$$\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 987.5124$$

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = 972.9288,$$

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = 14.5836.$$

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE},$$

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 987.5124 = 972.9288 + 14.5836.$$

2.4. Goodness of fit

The main purpose of the regression analysis to provide a good estimation, and explain the variation of the dependent variable Y. Using the R-squared (coefficient of determination) indicator we can measure what proportion of the total sum squares is explained by the regression. R-squared is the coefficient of determination providing information about how well the independent variable explains the dependent variable. R-squared is equal to the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

The indicator shows the proportion of the sample variation in Y that is explained by the independent variable (X):

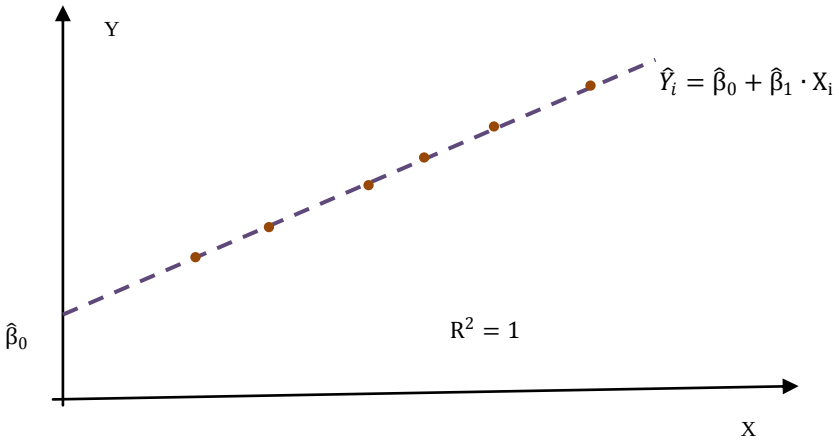
$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

The value of the coefficient of determination can be between zero and 1:

$$0 \leq R^2 \leq 1.$$

The coefficient is equal to one, if the data points fit on the regression line.

Figure 2.8 The perfect fit



In this case the fit is perfect, and the residuals are zero (Fig. 2.8). Indeed, the minimum criterion of the residual sum of squares is equivalent to the maximum criterion of the coefficient of determination:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}.$$

R^2 is equal to zero, if the dependent and independent variable are not correlated:

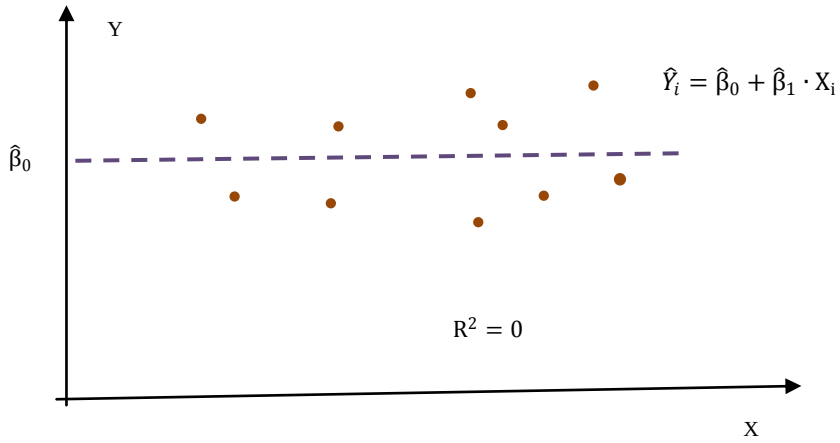
$$R^2 = \frac{SSR}{SST} = 0,$$

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = 0.$$

The regression line is a horizontal (regression) line, if the value of x changes, the value of Y remains the same; i.e. it does not change (Fig. 2.9):

$$\hat{Y}_i = \bar{Y}.$$

Figure 2.9 Zero coefficient of determination



If the value of the coefficient (R^2) is greater, the regression line fits the sample data better.

What is the acceptable value of (R^2)? There is no general rule to evaluate R^2; however, a coefficient above 0.5 may represent a good fit. It is very important to discover the variables that significantly influence the dependent variable. Our estimation can result in a high determination of the coefficient (R^2); however the independent variable chosen and the dependent variable can, in fact, be influenced by another variable.

Example 2.6

Try to answer the following question:

How well does the estimated regression function fit the data?

Let us compute the coefficient of determination for the example 2.4.

$$R^2 = \frac{SSR}{SST} = \frac{972.9288}{987.5124} = 0.985232.$$

On the basis of the result of R^2, we can state that 98.5232% of the total sum of squares is explained by the regression equation.

We started the second chapter by introducing the linear correlation coefficient. Finally, at the end of the chapter, we return to a discussion of the correlation coefficient. The main reason for this is that there is a relationship between the coefficient of determination (R^2) and the linear correlation coefficient. In the last chapter we discuss the association between the two indicators.

2.5. The relationship between R^2 and the linear correlation coefficient

In the previous chapter, we have found that the coefficient of determination is as follows:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

Let us give the correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

The square of the linear correlation coefficient is equal to the coefficient of determination (R^2):

$$R^2 = \frac{(\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y}))^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2} = {r_{xy}}^2.$$

Let us prove the statement:

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot X_i - \bar{Y}\right)^2. \qquad (2.11)$$

Substituting the regression parameters into the equation (2.11):

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X},$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

$$\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot X_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left(\bar{Y} - \hat{\beta}_1 \cdot \bar{X} + \hat{\beta}_1 \cdot X_i - \bar{Y}\right)^2 = \cdot,$$

$$= \sum_{i=1}^{n}\left(-\hat{\beta}_1 \cdot \bar{X} + \hat{\beta}_1 \cdot X_i\right)^2 =$$

$$= {\hat{\beta}_1}^2 \cdot \sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y}))^2}{(\sum_{i=1}^{n}(X_i - \bar{X})^2)^2} \sum_{i=1}^{n}(X_i - \bar{X})^2 =$$

$$= \frac{(\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y}))^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Insert the given formula instead of $\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2$ into the equation for R^2:

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(X_i-\bar{X})\cdot(Y_i-\bar{Y})\right)^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = r_{xy}{}^2. \quad (2.12)$$

At this moment, we know that the association between the two indicators is valid if there is only one independent variable and one dependent variable. Later (in Chapter 5), we will examine the validity of (2.12) in multiple regression models. Do not forget that the Pearson correlation coefficient measures the strength and direction of the ***linear association*** between two variables.

Example 2.7

Let us compute the linear correlation coefficient for exercise 2.4.

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2\cdot\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} = \frac{2811.1675}{\sqrt{8122.55\cdot 987.512375}} = 0.9925885,$$

$$r_{xy}{}^2 = 0.985232 = R^2.$$

There is a strong linear association between the dependent and independent variable.

2.6. Terms and Questions

coefficient of determination,
correlation,
dependent variable,
deterministic term,
disturbance term,
error term,
explained part,
explained sum of squares,
explained variable,
explanatory variable,
estimated regression line,
goodness of fit,
independent variable,
linear correlation coefficient,
multiple regression model,
normal equations,
ordinary least square method,
Pearson correlation coefficient,
perfect fit,
random component,
regression analysis,
regression coefficient,
regression equation,
regression parameter,
residual,
residual sum of squares,
R-squared,
simple regression model,
standard deviations,
stochastic component,
theoretical regression line,
total sum of squares,
unexplained part,
zero coefficient of determination.

Theoretical questions

1. What is the difference between dependent and independent variables?

2. What is the simple linear regression model?

3. How can we analyse the goodness of fit? How well does the estimated regression function fit the data?

4. Give the definition of the linear correlation coefficient.

5. What is the difference between an estimated regression line and a theoretical regression line?

6. What does a perfect fit mean?

7. How can we interpret the zero coefficient of determination?

8. How can we specify a linear regression model? Give an example of a linear regression model?

9. What does the intercept mean in the linear regression equation?

10. Explain how the R-squared is calculated.

11. What does the slope mean in the linear regression equation?

12. Explain what the difference is between residual and error terms.

13. How can we calculate the total sum of squares?

14. Give the definition of the disturbance term.

15. How can we derive normal equations? Give an example of each.

16. How can we compute the residual sum of squares?

17. What is the relationship between the total sum of squares and the explained sum of squares?

18. Explain what the difference is between the correlation analysis and the regression analysis.

Calculation exercise

1.

You have the following information about a sample survey:

X_i	Y_i
1	8
6	19
10	30
15	44
20	54
22	60
30	82

a) Illustrate these data on a scatter diagram.
b) Find the mean of data *X*.
c) Find the mean of data *Y*.
d) Compute the linear correlation coefficient.
e) Evaluate the result of the correlation coefficient.
f) Write the normal equations.
g) Compute the regression parameters.
h) Predict the value of Y when $X = 15$, and $X = 18$.

2.

Assume that the regression model is $Y = \beta_0 + \beta_1 \cdot X$. Prove that if Y is changed to $Y^* = \alpha_0 + \alpha_1 \cdot Y$, the changed intercept will be given by ${\beta_0}^* = \alpha_0 + \alpha_1 \cdot \beta_0$, where $Y^* = {\beta_0}^* + {\beta_1}^* \cdot Y$.

3.

You have the following information about a sample survey:

X_i	Y_i
2	5
5	15
12	39
14	33
17	40
21	70
25	62

a) Illustrate these data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram. Is there an association between the two variables?
c) Find the mean of data *X*.
d) Find the mean of data *Y*.
e) Compute the linear correlation coefficient.
f) Evaluate the result of the correlation coefficient.
g) Write the normal equations.
h) Compute the regression parameters.
i) Predict the value of Y when $X = 5$, and determine the value of the residual.
j) Predict the value of Y when $X = 30$.

4.

You have the following information about a sample survey:

X_i	Y_i
5	186
14	156
24	148
32	125
48	78
60	55
68	30

a) Illustrate these data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram. Is there a relationship between the two variables?
c) Find the mean of data *X*.
d) Find the mean of data *Y*.
e) Compute the linear correlation coefficient.
f) Evaluate the result of the correlation coefficient.
g) Compute SST, SSR and SSE.

h) Compute the coefficient of determination.
i) How well does the estimated regression function fit the data?

5.

You have the following information about a sample survey:

X_i	Y_i
5	186
14	156
24	148
32	125
48	78
60	55
68	30

a) Illustrate these data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram. Is there a positive or negative association between the two variables?
c) Compute the linear correlation coefficient.
d) Evaluate the result of the correlation coefficient.
e) Compute SST, SSR and SSE.
f) Compute the coefficient of determination.
g) How well does the estimated regression function fit the data?

6.

Suppose that you would like to analyse the height of students at Debrecen University. You have the following information about their height (centimetres) and weight (kilograms):

Students' data		
Observation	Height	Weight
1	162	60
2	158	55
3	175	72
4	182	75
5	164	61
6	195	104
7	201	112
8	180	102
9	178	75
10	163	58
11	155	49
12	160	56
13	169	64
14	171	68
15	194	104

a) Illustrate the data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram. Is there a positive or negative association between the two variables?
c) Find the mean of data *X*.
d) Find the mean of data *Y*.
e) Compute the linear correlation coefficient.
f) Evaluate the result of the correlation coefficient.
g) Write the normal equations.
h) Compute the regression parameters.
i) Interpret the meaning of the vertical axis (Y) intercept (β_0).
j) Interpret the meaning of the slope parameter.
k) Predict the weight, if the student's height is 172 centimetres.

7.

Demonstrate that if the slope of a linear regression equation is equal to zero, the coefficient of determination is also zero.

8.

Suppose that you have been asked to analyse the association between price (P) and quantity demanded for special Hungarian salami (Q). You collect data on price and demand from different supermarkets, and make a regression analysis. The price is measured in HUF/kg and the quantity demanded is given in kilograms. The result of your estimation for the regression model for the two variables is as follows:

$$Q = 12\,000 - 1.4 \cdot P.$$

a) Does the sign of the slope in the estimated equation match your expectations?
b) Illustrate the data on a scatter diagram.
c) Explain the economic meaning of the estimated regression coefficients.
d) Explain the economic meaning of the intercepts.
e) List two or three independent variables that may influence the dependent variable.

9.

The table shows the annual public expenditure on tertiary educational institutions per student (PUBEXP) and the gross domestic product per capita (current prices and current PPPs) (GDP) in 2013.

OECD	PUBEXP	GDP	OECD	PUBEXP	GDP
Australia	7 740	44 706	**Japan**	6 855	36 225
Austria	15 794	45 133	**Korea**	3 684	33 089
Belgium	13 808	41 595	**Mexico**	5 129	16 891
Chile	2 866	21 888	**Netherlands**	13 209	46 749
Czech Republic	6 753	28 963	**New Zealand**	7 570	34 989
Denmark	14 047	43 797	**Norway**	19 873	65 635
Estonia	7 068	26 160	**Poland**	6 544	23 616
Finland	17 168	40 017	**Portugal**	5 883	27 651
France	12 479	37 617	**Slovak Republic**	6 824	26 586
Germany	14 140	43 282	**Slovenia**	8 434	28 675
Hungary	6 275	23 507	**Spain**	8 685	32 546
Iceland	9 775	41 987	**Sweden**	20 167	44 586
Ireland	9 994	46 858	**Turkey**	6 935	18 599
Israel	6 892	32 713	**United Kingdom**	14 209	38 743
Italy	7 264	34 781	**United States**	10 134	52 592

Source: OECD (2017a,b)

Public expenditure on tertiary educational institutions per student is given in equivalent USD converted using PPPs for GDP, GDP data are given in US dollars.

a) Illustrate the data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram. Is there a positive or negative association between the two variables?
c) Compute the linear correlation coefficient.
d) Evaluate the result of the correlation coefficient.
e) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.
f) Give the interpretation of the slope of the estimated regression equation.
g) Compute SST, SSR and SSE.
h) Examine the goodness of fit.
i) What percentage of the variation in total public expenditure per student on tertiary education can be explained by the level of GDP?
j) Give other independent variables which may affect the dependent variable.

10.

Assume that your lecturers have collected data on the performance of students. These data are summarized in the Table below. The required minimum level of the econometrics test was 100 points, and the maximum was 200 points.

Students' performance		
Observation	Test points	Time spent on econometrics studies (minutes)
1	32	66
2	58	144
3	52	144
4	74	192
5	82	216
6	80	294
7	94	300
8	130	552
9	155	576
10	165	684
11	175	714
12	200	720
13	200	960
14	150	343.2
15	200	746.4

a) Illustrate the data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram.
c) Compute the linear correlation coefficient.
d) Evaluate the result of the correlation coefficient. Is there a strong or weak association between the two variables?
e) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.
f) Give the interpretation of the slope of the estimated regression equation.
g) Compute SST, SSR and SSE.
h) What is the interpretation of SSR?
i) Examine the goodness of fit.
j) What percentage of the variation in test points is explained by the time spent on econometrics studies?
k) Give other independent variables which may affect the dependent variable.

11.

Collect data on the unemployment rate and the inflation rate for different countries for a freely chosen year.

a) Is there any association between the two variables? Analyse the association between the two variables. Evaluate the result.
b) Give some variables that may influence the unemployment rate.
c) Give some variables that may influence the inflation rate.

12.

Suppose that you have been asked to analyse the success of a given festival. You collect 10-year time-series data on the revenues of the festival and expenditures spent on the performers at the festivals. Revenues and expenditures are measured in EUR. The result of your estimation for the regression model for the two variables is as follows:

$$Y = 30 + 2.1 \cdot X,$$

where Y is the revenues of the festival, and X the expenditures spent on the performers at the festivals.

a) Does the sign of the slope in the estimated equation match your expectations?
b) Explain the economic meaning of the estimated regression coefficients.
c) List two or three independent variables that may influence the dependent variable.

13.

The table shows the annual private expenditure on educational institutions per student (PRIVBEXP) (primary to tertiary) and the household net saving rate as a percentage of household net disposable income in 2013.

Country	Net Saving Rates (%)	Private Expenditure on Educational Institutions (USD)	Country	Net Saving Rates (%)	Private Expenditure on Educational Institutions (USD)
Australia	-0.2	6 199	**Korea**	5.0	3 298
Austria	-0.0	7 990	**Norway**	5.6	9 882
Belgium	0.7	10 890	**Poland**	5.7	2 531
Czech Republic	1.3	2 917	**Portugal**	7.3	2 081
Estonia	2.9	6 194	**Slovenia**	7.6	4 877
Finland	3.9	9 420	**Spain**	9.1	3 396
France	3.9	5 439	**Sweden**	9.7	10 528
Hungary	3.9	4 157	**United Kingdom**	9.7	10 489
Italy	4.9	2 413	**United States**	15.1	3 253

Source: OECD (2017a,b)

a) Illustrate the data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram.
c) Compute the linear correlation coefficient.
d) Evaluate the result of the correlation coefficient.
e) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.

f) Give the interpretation of the slope of the estimated regression equation.
g) Compute SST, SSR and SSE.
h) Examine the goodness of fit.
i) What percentage of the variation in total private expenditure per student on education can be explained by the amount of the household net saving rates?
j) Give other independent variables which may affect the dependent variable.

14.

The table below shows the employment rate (%) and the gross domestic product per capita (current prices and current PPPs) (GDP) in 2013.

Country	Employment rate (%)	GDP (USD)	Country	Employment rate (%)	GDP (USD)
Australia	72	44 706	Korea	64.4	33 089
Austria	71.4	45 133	Luxembourg	65.7	93 234
Belgium	61.8	41 595	Mexico	60.8	16 891
Canada	72.4	43 038	Netherlands	73.6	46 749
Chile	62.3	21 888	New Zealand	72.8	34 989
Czech Republic	67.7	28 963	Norway	75.4	65 635
Denmark	72.6	43 797	Poland	60	23 616
Estonia	68.5	26 160	Portugal	60.6	27 651
Finland	68.9	40 017	Slovak Republic	59.9	26 586
France	64.1	37 617	Slovenia	63.6	28 675
Germany	73.5	43 282	Spain	54.8	32 546
Greece	48.8	25 523	Sweden	74.4	44 586
Hungary	58.1	23 507	Switzerland	78.4	56 897
Iceland	81.1	41 987	Turkey	49.5	18 599
Ireland	60.5	46 858	United Kingdom	70.5	38 743
Israel	67.1	32 713	United States	67.4	52 592
Italy	55.5	34 781	Euro area	63.5	37 606
Japan	71.8	36 225			

Source: OECD (2017a,c)

a) Illustrate the data on a scatter diagram.
b) Compute the linear correlation coefficient.
c) Evaluate the result of the correlation coefficient.
d) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.

e) Give the interpretation of the slope of the estimated regression equation.
f) Compute SST, SSR and SSE.
g) Examine the goodness of fit.
h) Give other independent variables which may affect the dependent variable.

15.

Collect data on the unemployment rate and the GDP for different countries for a freely chosen year.

a) Illustrate the data on a scatter diagram.
b) Analyse the association between the two variables. Evaluate the result.
c) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.
d) Compute SST, SSR and SSE.
e) Show the relationship between SST, SSR and SSE.
f) Compute the coefficient of determination. Provide an interpretation of the result.
g) Give other independent variables which may affect the dependent variable. Give the direction of the association between the dependent and independent variable.

3. The Classical Model

In the previous chapter we studied simple linear regression analysis. We estimated the parameters of the regression equation and tried to analyse the goodness of fit. As we have discussed in the previous section, the theoretical regression equation is:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i, \quad (3.1)$$

where Y is the dependent variable, X is the independent variable, and (u) is the error term. The estimated regression equation is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot X_i.$$

If the estimation is completed, it is necessary to evaluate the results. We would like to know:

- How can we describe the goodness of fit of the data?
- Does the ordinary least square method provide the best estimation for our equation?
- Are all important variables included in the regression equation?
- Is the chosen function type appropriate for the data?

Some assumptions must be fulfilled in order for the ordinary least squares method to be applied as the best method for estimation. In this section, we point to each assumption.

3.1. The model assumptions

In the first step, we list the assumptions and try to interpret each of them.

The Classical Model Assumptions
I. The regression model is linear, and is correctly specified.
"II. The error term has a zero population mean.
III. All explanatory variables are uncorrelated with the error term.
IV. Observations of the error term are uncorrelated with each other (no serial correlation).
V. The error term has a constant variance (no heteroscedasticity).
VI. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity).
VII. The error term is normally distributed (this assumption is optional but usually is invoked)."

Studenmund (2014)

I. The regression model is linear, is correctly specified, and has an additive error term

The regression model is ***linear in parameters*** and the model has an additive error term. In this case the regression model can be written as follows:

$$Y = \beta_0 + \beta_1 \cdot X + u, \quad (3.1)$$

where Y is the dependent variable and X is the independent variable; β_0 and β_1 are the intercept and slope parameters in equation (3.1), respectively, and u is the error term.

Of course, not all economic relationships can be characterized by a linear function, so we should also use other (nonlinear) function specifications.
We will return later to the discussion of the nonlinearity problem in Chapter 6. We would like to note in advance that the nonlinear function may be transformed by using different techniques than the linear function (for example logarithmic transformation). This means that the OLS method can be applied to the transformed function, so the problem of nonlinearity can be manageable in most cases.

II. The error term has a zero population mean / the error term has zero expectation

The error term is a random variable with an expected value of zero (a mean of zero):

$$E(u) = 0$$

The error terms can be negative and positive in that they tend to offset each other.
The theoretical regression equation is:

$$Y = \beta_0 + \beta_1 \cdot X + u$$

If there is an intercept in the regression equation, in this case the assumption is satisfied, because the intercept picks up any constant tendency in Y that the independent variable does not explain.
Assume that the expected value of the error term is other than zero:

$$E(u) = c, \tag{3.2}$$

where $c \neq 0$.
Assume that u^* is equal to the difference between u and c:

$$u^* = u - c,$$
$$u = u^* + c.$$

Substitute u into the equation (3.1):

$$Y = \beta_0 + \beta_1 \cdot X + u^* + c,$$
$$Y = \beta_0 + c + \beta_1 \cdot X + u^*,$$

Define

$$\beta_0{}^* = \beta_0 + c,$$
$$Y = \beta_0{}^* + \beta_1 \cdot X + u^*. \tag{3.3}$$

In model (3.3) the error term has changed. The error term in model (3.3) satisfies the assumption II:

$$E(u^*) = E(u) - E(c),$$

On the basis of equation (3.2):

$$E(u) = c,$$

and

$$E(c) = c,$$
$$E(u^*) = E(u) - E(c) = c - c = 0.$$

III. All explanatory variables are uncorrelated with the error term

If the assumption is not met, the values of the explanatory variables would probably depend on the values of the error term. This means that the change in the error term may result in a change in the independent variables. However, we would like to analyse the variation in Y

that is explained by the change in the independent variables (X). If the independent variables and the error term are correlated, the Ordinary Least Squares method incorrectly describes the variation caused by X, since the X impact shown may also come from the error term.
In chapter 2.2, we mentioned that the error term may represent the omitted independent variables which may explain the behaviour of the dependent variable. If our model has an omitted variable, and the omitted variable and the one independent variable are correlated, this means that the independent variable and the error are also correlated. In this case the assumption is not valid.

There is some variation in the regressor (independent variable) in the sample

If the independent variable is constant, the values of the independent variable would be equal to their mean for all observations. In this case we cannot determine the regression parameters, since:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

the numerator and the denominator are equal to zero. If we cannot calculate the regression parameter of $\hat{\beta}_1$, we are not able to calculate the regression parameter of $\hat{\beta}_0$, because

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}.$$

IV. Observations of the error term are uncorrelated with each other (no serial correlation)

The values of the error term are independent. This means that the value of the error term for a value of the independent variable (X) is not related to the value of the error for other values of the independent variable (X). The Ordinary least Squares estimation provides an inefficient estimate if the assumption is not valid. In the analysis of time series data the error term correlation problem may occur. There is autocorrelation if the error terms in two different time periods are correlated. The autocorrelation can be detected using by different tests. Later, we will study some autocorrelation tests.

V. The error term has a constant variance (no heteroscedasticity)

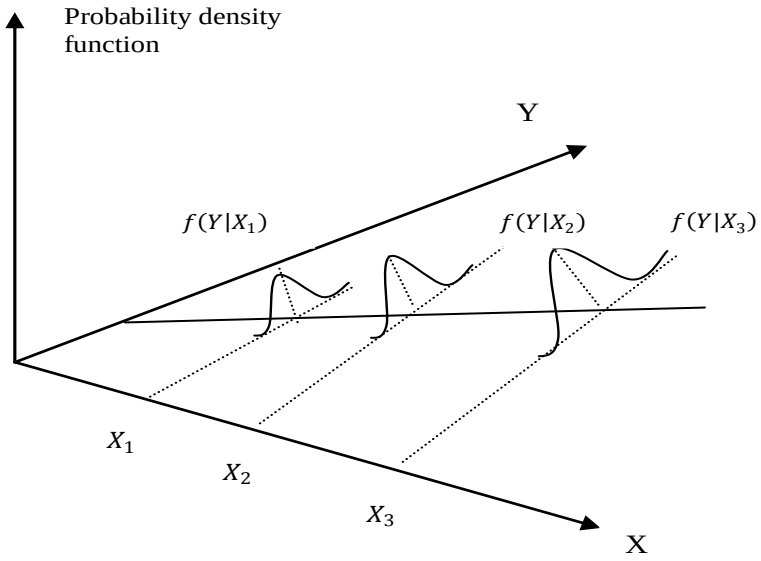
The variance of the error term should be constant in each observation:

$$\mathrm{Var}(u_i) = \sigma^2 \text{ for all i.}$$

This means that the variance of the error term has the same variance for any values of the independent variables. It is said that the error term is homoscedastic if the variance of the error term does not depend on the independent variable. In other cases, if the assumption is violated the variance of the error term depends on X, and the model shows heteroscedasticity (Figure 3.1):

$$\mathrm{Var}\big((u_i|X)\big) \neq \sigma^2.$$

Figure 3.1 The simple regression model under heterokedasticity



On the basis of assumption II, $\mathrm{E(u)} = \mathrm{c} = 0$, the variance of the error term is

$$\mathrm{E}[(\mathrm{u_i} - \mathrm{c})^2] = E(u_i^2) = \sigma^2 \text{ for all i.}$$

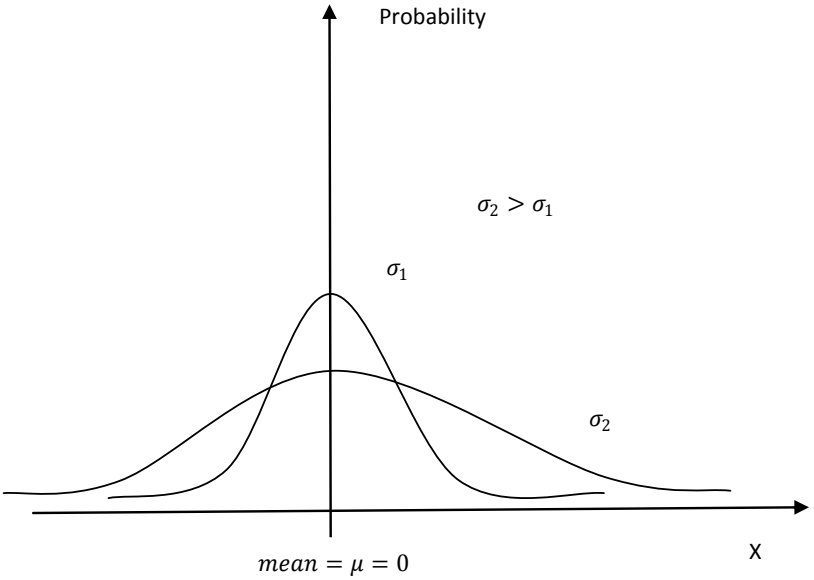
There are different tests for heteroscedasticity.

VI. The error term is normally distributed (this assumption is optional but is usually invoked)

We assume that the distribution of the error term is a normal, bell-shaped curve. However, this assumption is not required for an OLS estimation, although later we will see that it will be important in hypothesis testing.

The main characteristics of normal distribution

- Normal distribution can be characterised by the mean and the standard deviation. The two parameters determine the place of the curve along the horizontal axis and the shape of the normal curve.
- The distribution is the standard normal probability distribution if the random variable has normal distribution with zero mean and a standard deviation of one.
- The highest point of the bell-shaped curve can be found at the mean.
- The area under the normal curve represents the probabilities of the normal random variable. The amount of the total area under the curve is equal to one.
- The bell-shaped curve is symmetric.
- The mean is equal to the median and the mode of the distribution.
- The standard deviation determines the shape of the curve, how flat and how wide the normal curve is. If the standard deviation is greater, the curve is wider and flatter because the variability of data is greater.
- The tails of the curve go to infinity in both directions and never touch the horizontal axis.
- If the mean changes and the standard deviation remains the same, the normal curve moves along the horizontal axis.



VII. No explanatory variable is a perfect linear function of any other explanatory variable(s) (no perfect multicollinearity).

According to assumption VI, There are no perfect linear relationships among the independent variables, and the independent variables cannot be constant**.** ***Perfect collinearity*** among independent variables means that an independent variable is a perfect linear combination of the other independent variables.

In the model we make a distinction between dependent and independent variables. The dependent variable depends on the independent variables. What happens if the independent variables are correlated to some degree with one another in a multiple regression model? Let us assume that there are two independent variables in the following regression model:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + u$$

We suppose that there is a linear association between the two independent variables, so perfect collinearity can be shown. If there are two independent variables (X_1, and X_2), in the case of perfect collinearity the relationship between the two variables can be written as follows:

$$X_1 = \alpha \cdot X_2 + \omega.$$

There is a very strong linear correlation between the two independent variables. In this case it is very difficult to recognize the effect of the individual independent variable on the dependent variable:

$$Y = \beta_0 + \beta_1 \cdot (\alpha \cdot X_2 + \omega) + \beta_2 \cdot X_2 + u = \beta_0 + \beta_1 \cdot \omega + (\beta_1 \cdot \alpha + \beta_2) \cdot X_2 + u,$$

or

$$X_2 = \frac{1}{\alpha} \cdot X_1 - \frac{\omega}{\alpha}$$

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot \left(\frac{1}{\alpha} \cdot X_1 - \frac{\omega}{\alpha}\right) + u = \beta_0 - \beta_2 \cdot \frac{\omega}{\alpha} + \left(\beta_1 + \frac{\beta_2}{\alpha}\right) \cdot X_1 + u.$$

Actually, ***multicollinearity*** represents a correlation among the independent variables if there are several independent variables in the regression model. However, if the model includes more than two independent variables, we should check the possibility of multicollinearity. We have to apply one of the different methods that we will study in Chapter 7.

> *"Statisticians have developed several tests for determining whether multicollinearity is high enough to cause problems. According to the rule of thumb test, multicollinearity is a potential problem if the absolute value of the sample correlation coefficient exceeds 0.7 for any two of the independent variables."*
>
> *Anderson – Sweeney – Williams (2007)*

3.2. The Gauss-Markov theorem

The Gauss-Markov theorem says that the OLS estimators are the minimum variance linear unbiased estimators of β under the assumptions of the model. Therefore on the basis of the theorem, we can say that the OLS estimators are efficient if the assumptions I-VI are fulfilled (see Chapter 3.1). Efficient OLS estimation is also referred to as BLUE estimation. BLUE means that the estimation is the best, linear, and unbiased estimator of the regression coefficients (BLUE = the best, linear, unbiased estimator).

An estimator is called the best, linear, unbiased estimator (BLUE) if it is a linear function of the data and has minimum variance among linear unbiased estimators.

The best means that the ordinary least squares estimator of the regression parameters is the minimum variance estimator. An efficient estimator is the unbiased estimator with the smallest variance. On the basis of the classical model assumptions I-VII, we can state that the OLS method is the best unbiased estimator of all linear and nonlinear estimators. The OLS coefficient estimators have the following properties:

- The estimators are unbiased. "This means that the OLS estimates of the coefficients are centred around the true population values of the parameters being estimated" (Studenmund, 2014).
- The estimators are minimum variance estimators. "The distribution of the coefficient estimates around the true parameter values is as tightly or narrowly distributed as is possible for an unbiased distribution. No other unbiased estimator has a lower variance for each estimated coefficient than OLS" (Studenmund, 2014).
- The estimators are consistent. "As the sample size approaches infinity, the estimates converge to the true population parameters. Put differently, as the sample size gets larger, the variance gets smaller, and each estimate approaches the true value of the coefficient being estimated" (Studenmund, 2014).
- They are normally distributed. "Thus various statistical tests based on the normal distribution may indeed be applied to these estimates" (Studenmund, 2014).

3.3. Terms and Questions

autocorrelation,
BLUE estimation,
collinearity,
efficient estimator,
Gauss-Markov theorem,
heteroscedasticity,
homoscedasticity,
multicollinearity,
normal probability distribution,
perfect collinearity,
standard normal probability distribution.

Problems

Theoretical questions

1. What is the difference between homoscedasticity and heteroscedasticity?

2. What is perfect collinearity?

3. How can we characterize normal probability distribution?

4. Give the definition of collinearity.

5. What is the difference between correlation and autocorrelation?

6. What does BLUE estimation mean?

7. How can we interpret the Gauss-Markov theorem?

8. Give the definition of multicollinearity.

9. Explain what the difference is between normal probability distribution and standard normal probability distribution.

10. Give the definition of heteroscedasticity.

11. What does the efficient estimator mean?

Exercises

1.

Consider the following function:

$$Studpo = \beta_0 + \beta_1 \cdot time + u,$$

where $Studpo$ is the students' test results in econometrics, and $time$ is the Time spent on econometrics studies. u is:

$$u = time^{1/2} \cdot e,$$

where e is a random variable, its expected value is zero $E(e) = 0$, and its variance is $Var(e) = \sigma_e^2$.

a) Explain the economic meaning of the regression coefficients of β_1.
b) Prove that the error term has an expected value of zero.
c) Prove that homoscedasticity is not satisfied in $Var(u|time) = \sigma_e^2 \cdot time$.

2.

Consider the following function:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i,$$

a) Show the OLS is BLUE (linear function of Y, and unbiased) for the coefficient of β_1.
b) Show that the error term expected value is zero for all i.

4. Properties of the regression parameters

In the first part of this chapter we give the standard error of the estimation and analyse the variances of the regression parameters. We describe the properties of the sampling distribution of the estimated regression parameters. In the second part of this chapter we carry out a t-test and determine the confidence intervals of the regression parameters.

4.1. Standard error of the estimation

In Chapter 2, we analysed the decomposition of the total sum of squares into the explained sum of squares and the residual sum of squares. The residual shows the difference between the actual value of the dependent variable and the estimated value of the dependent variable:

$$e_i = Y_i - \hat{Y}_i.$$

The residual sum of squares (SSE) measures the variability of the actual value of Y around the estimated regression line ($\hat{Y}$):

$$\text{SSE} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2.$$

The equation of the theoretical regression line can be written as follows:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + u_i$$

The error term shows the difference between the actual value of the dependent variable and the value of the dependent variable on the theoretical regression line:

$$u_i = Y_i - (\beta_0 + \beta_1 \cdot X_i).$$

The variance of the error term is indicated by σ^2:

$$\sigma^2 = E[u_i - E(u_i)]^2,$$

on the basis of the assumption II (in Chapter 3), the expected value of the error term is zero:

$$E(u_i) = 0,$$

and we get that

$$\sigma^2 = E[u_i - E(u_i)]^2 = E(u_i^2).$$

On the basis of the definition of the variance, we may write that the estimation of the variance of the error term is the average of the squared errors:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} u_i}{n}. \tag{4.1}$$

However, we are not able to observe the values of the error term; as a result we cannot use the formula (4.1). We replace the unobservable error term with the residuals (e_i) of the regression estimation:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}. \tag{4.2}$$

We make an additional correction because the formula (4.2) is a biased estimation of the variance of the error term (σ^2). We subtract two from the value of n, since statisticians have shown that the residual sum of squares has $(n-2)$ degrees of freedom, since the number of regression parameters is two (β_0, β_1):

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}.$$

We have to compute two parameters in order to determine the residual sum of squares; this means that the degree of freedom of SSE is equal to two.

The given $\hat{\sigma}^2$ is an unbiased estimation of the variance of the error term (σ^2). The ***mean square error (MSE)*** is equal to the residual sum of squares divided by its degrees of freedom:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}.$$

The standard error of the estimation:

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}.$$

The estimated standard deviation around the regression line is equal to the square root of the mean square error (MSE).

4.2. Estimation of the standard deviation of the regression parameters

Let assume that we would like to analyse the relationship between the price and floor area of flats in Debrecen in 2017. We have a sample of twenty flats, and we estimate the regression parameters of our model. Suppose that we make a new regression analysis of a new random sample of twenty flats. In this case our newly estimated parameters are unlikely to remain the same, we get a different regression equation for the same problem. The parameters have normal distributions, since the linear combination of the normal distribution results in a normal distribution, and on the basis of assumption VI the error term has a normal distribution (Dougherty, 2011:116,118):

$$\hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} (a_i \cdot u_i),$$

where

$$a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})},$$

and

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} (c_i \cdot u_i),$$

where

$$c_i = \frac{1}{n} - a_i \cdot \bar{X}.$$

The estimations of the regression parameters vary from sample to sample. The variance of the regression parameters is equal to the average of the squared difference between the values of the random variable β_1 and its mean:

$$var(\hat{\beta}_1) = (E\left[\hat{\beta}_1 - E(\hat{\beta}_1)\right]^2,$$
$$E(\hat{\beta}_1) = \beta_1.$$

The variance of the regression coefficient β_1:

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

The estimated standard deviation of the parameter $\hat{\beta}_1$:

$$\sigma_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{SSE}{(n-2)\cdot\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

The variance of the parameter $\hat{\beta}_1$ is given by the following equation:

$$\sigma^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (4.3)$$

Actually, we can analyse the sampling precision of the estimator by examining the variance. The smaller the variance of the parameter, the better the sampling precision of the estimator. If we compare two estimators, we should choose the estimator with the smaller sampling variance. The variance of the regression coefficient $\hat{\beta}_0$ is given by:

$$\sigma^2_{\hat{\beta}_0} = \hat{\sigma}^2 \cdot \frac{\sum_{i=1}^{n} X_i^2}{n\cdot\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (4.4)$$

We can see from formula (4.3) that the greater the variability of the values of the independent variable about their sample mean ($\sum_{i=1}^{n}(X_i - \bar{X})^2$), the smaller the variance of the estimated regression parameter $\hat{\beta}_1$. The total variation in X ($SST_X = \sum_{i=1}^{n}(X_i - \bar{X})^2$) is greater if the number of the observations is greater and/or there are more deviations of X_i around the sample mean ($\bar{X}$). The variance of the estimated regression parameter $\hat{\beta}_1$ is proportional to the variance of the estimated standard deviation, and inversely proportional to the number of observations in the sample.

- The more sample values of the independent variable are spread out around their sample mean, the greater the total sum of squares of X.
- This means that the variance of the estimated coefficient is smaller. If the sample size n is larger, the variances of the estimated coefficients are smaller.
- If the variance of the estimated error term is greater, the precision of the estimations of the parameters are less favourable. The variance of the error term can be seen in the variance formula of the estimated parameters. If the variance of the error term is greater, the variance of the estimated parameters is greater, too. This means that the estimation is less precise; the uncertainty of the regression model is greater, too.

The term $\sum_{i=1}^{n} X_i^2$ can be found in the formula (4.4). If the squared distance of the data is greater from the axis $x = 0$, the estimation of the intercept parameter is less precise. This means that the uncertainty of the estimation of its value is greater.

Example 4.1

Let us determine the standard error of the estimation and the variance of the regression parameters in example 2.4. The standard error of the estimation is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SSE}{n-2}.$$

In example 2.5, we computed the total sum of squares, the residual and the explained sum of squares:

$$\mathrm{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 987.5124$$

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = 972.9288,$$

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = 14.5836.$$

Example 2.4	Y_i Million HUF	X_i Square metres	$Y_i - \bar{Y}$	$X_i - \bar{X}$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$	$(\hat{Y}_i - \bar{Y})$	$(Y_i - \hat{Y}_i)$
Sum	**461.95**	**1327.00**	**0**	**0**	**2811.17**		
Sum of squares	**11657.403**	**96169**	**987.512375**	**8122.55**		**972.9288**	**14.5836**

Let us substitute the amount of the residual sum of squares into SSE:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{14.5836}{18} = 0.8102.$$

$$\hat{\sigma} = 0.90011.$$

The estimated standard deviation of the parameter $\hat{\beta}_1$:

$$\sigma_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{0.90011}{\sqrt{8122.55}} = 0.009987,$$

and the estimated standard deviation of the parameter $\hat{\beta}_0$:

$$\sigma_{\hat{\beta}_0} = \hat{\sigma} \cdot \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{\mathrm{n} \cdot \sum_{i=1}^{n}(X_i - \bar{X})^2}} = 0.90011 \cdot \sqrt{\frac{96169}{20 \cdot 8122.55}} = 0.69255.$$

4.3. Hypothesis testing, test of significance

If we create a model, we would like to test the validity of our model, or we may test the validity of existing theories by using estimates based on a random sample. In fact, we can test the given hypothesis only on our sample, so we can examine whether the given sample corroborates the hypothesis. This means that if we confirm the hypothesis, our conclusion may state that the sample reinforces the hypothesis.

Firstly we should establish the hypothesis that we would like to test. We have to give a null hypothesis (H_0) and an alternative hypothesis (H_a). The null hypothesis (H_0) represents the statement that the researcher does not expect. The alternative hypothesis is a statement that the researcher expects. The null hypothesis and the alternative hypothesis are mutually exclusive, and collectively exhaustive.

Example 4.2

For example, if we expect that the regression parameter β_1 is greater than 1, the null hypothesis states that the parameter β_1 is equal to or less than 1:

$$H_0: \beta_1 \leq 1,$$

and the alternative hypothesis is:

$$H_a: \beta_1 > 1.$$

We can see that the H_0 null hypothesis is absolutely the opposite of the alternative hypothesis. "Classical hypothesis testing requires that the null hypothesis contain the equal sign in some form (whether it be $=$, $\leq$, or $\geq$). This requirement means that researchers are forced to put the value they expect in the null hypothesis if their expectation includes an equal sign" (Studenmund, 2014). We can make a distinction between a one-sided and a two-sided hypothesis test. In the case of the one-sided test (or one-tailed test), the alternative hypothesis has values on only one side of the null hypothesis. In example 4.2 the given statements can be seen as a one-sided hypothesis. We apply a two-sided test (or two-tailed test), if the alternative hypothesis has values on both sides of the null hypothesis.
An example of a one-sided test is the following:

$$H_0: \beta_1 \geq 0,$$

and

$$H_a: \beta_1 < 0.$$

As we can see, the alternative hypothesis ($H_a: \beta_1 < 0$) has values only one side, actually the opposite side of the range of the null hypothesis.
An example of a two-sided test is as follows:

$$H_0: \beta_1 = 1,$$

and

$$H_a: \beta_1 \neq -1.$$

As we can see, the alternative hypothesis ($H_a: \beta_1 \neq -1$) has values on two sides of the range of the null hypothesis.

Example 4.3
We may expect that the regression parameter β_1 is different from zero. In this case the null and the alternative hypothesis are as follows:

$$H_0: \beta_1 = 0,$$

and

$$H_a: \beta_1 \neq 0.$$

4.3.1. t-Test

Simple linear regression model
The main goal of the t-test is to test a hypothesis about the slope of the regression function. On the basis of the simple linear regression model, there is a linear association between the dependent variable (Y) and the independent variable (X):

$$Y = \beta_0 + \beta_1 \cdot X + u.$$

In the case of a linear association, the regression coefficient β_1 is not equal to zero. The t-test makes possible the separate testing of the regression parameters. It is especially important if our model contains more independent variables (multiple regression model), and what we would like to decide about the independent variables is whether we should keep a given variable in our model or leave it out.
The t-test for hypothesis testing is often used by researchers. The t-test can be applied when the stochastic error term is normally distributed, and when the variance of that distribution must be estimated.

Let us examine the regression parameters separated testing. The null hypothesis states that the β_1 regression parameter is equal to zero, and the alternative hypothesis states the opposite of the null hypothesis:

$$H_0: \beta_1 = 0,$$
$$H_a: \beta_1 \neq 0.$$

If the null hypothesis cannot be rejected, the regression parameter β_1 is equal to zero, so the independent variable (X) has no effect on the dependent variable. This means that the change in the independent variable (X) does not result in a change in the dependent variable (Y). There is no linear relationship between the dependent and independent variables. In the opposite case, if the null hypothesis is rejected, we can state that $\beta_1 \neq 0$, and there is a statistically significant relationship between the dependent and independent variables.

The t-statistic (or t-ratio) follows a *t* distribution with *(n-2)* degrees of freedom (df) (*n* is the number of observations):

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}, \qquad (4.5)$$

where $\sigma_{\hat{\beta}_1}$ is the estimated standard error of $\hat{\beta}_1, \mathrm{df} = \mathrm{n} - 2$.

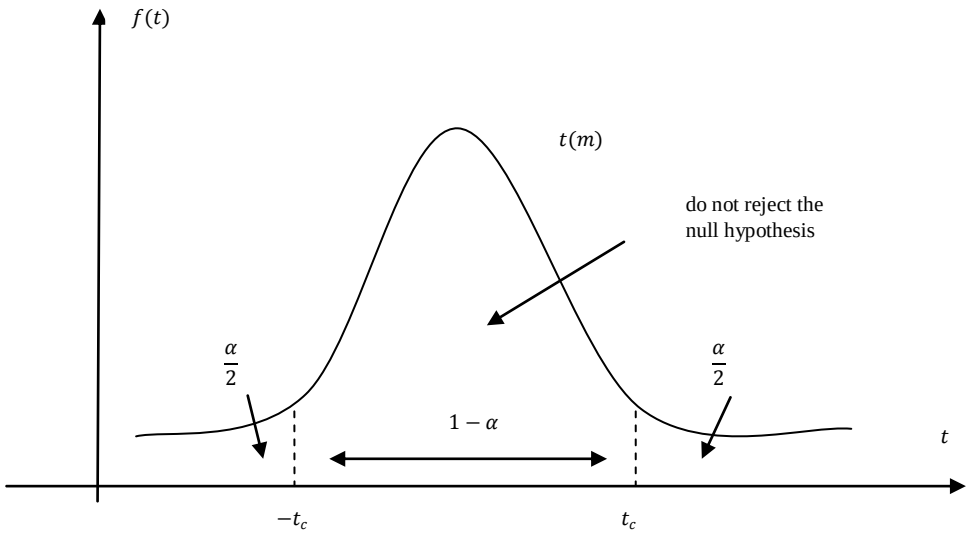
The proof of the t-distribution in (4.5) is not simple, and it does not represent any significant additional knowledge for students, so the proof will not be discussed in this chapter.

If the null hypothesis is not rejected ($\beta_1 = 0$), formula (4.5) is:

$$t = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}.$$

If we would like to decide about the null hypothesis, we have to compare the calculated t-value and the critical t-value (Figure 4.1).

Figure 4.1 *t-distribution* and the critical values



We can make a distinction between Type I and Type II errors.

The level of significance (α) of a test represents the probability that we reject the null hypothesis when it is true. This case is called a Type I error.

Type I error:
We reject the null hypothesis, but it is true.
Type II error:
We do not reject the null hypothesis, but it is false.

The critical t-value is determined from the t-table; the value separates the rejection region from the acceptance region. If the calculated (or empirical) t-value is less in absolute value than its theoretical value, the null hypothesis cannot be rejected. In this case the rejection rule can be written as follows:
$H_0: \beta_1 = 0$ is rejected if:

$$|t| > t_c,$$

Testing the null hypothesis ($H_0: \beta_1 = 0$) against the alternative hypothesis ($H_a: \beta_1 \neq 0$), in the case of a two-tail test we reject the null hypothesis ($H_0: \beta_1 = 0$) and accept the alternative hypothesis ($H_a: \beta_1 \neq 0$) if

$$t \leq t_{(\frac{\alpha}{2}, n-2)}$$

or

$$t \geq t_{(1-\frac{\alpha}{2}, n-2)}.$$

Example 4.4

In example 2.4 we estimated the regression parameters and provided the regression equation. Let us implement the t-test on example 2.4.
Let us remember that data on the price and floor area of 20 flats in Debrecen are given,

Data of flats		
Number	Price	Floor area
	Million HUF	Square metres
1	16.2	48
2	19.4	55
3	24.5	71
4	28.2	82
5	35.2	100
6	29.1	85
7	24.6	70
8	25.8	73
9	26.1	74
10	22.2	66
11	14.5	35
12	17.6	53
13	24.8	73
14	14.1	39
15	23.75	67
16	16.8	49
17	18.4	51
18	21.5	61
19	16.2	53
20	43	122

and the estimated regression equation is:

$$\hat{Y} = 0.134149 + 0.34609 \cdot X.$$

The first step is that we set the null hypothesis and give the alternative hypothesis:

$$H_0: \beta_1 = 0,$$
$$H_a: \beta_1 \neq 0.$$

The null hypothesis states that the β_1 regression parameter is equal to zero. In the second step we specify the test statistic and its distribution with the assumption of the null hypothesis:

$$t = \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}}.$$

We determine the value of α and the rejection region:

$$\alpha = 0.05,$$
$$df = 20 - 2 = 18,$$

the critical value of t is:

$$t_{(1-\frac{\alpha}{2},n-2)} = t_{0.975,18} = 2.101.$$

In the next step we calculate the value of the t-test statistic:

$$t = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_j}} = \frac{0.34609}{0.009987} = 34.654.$$

since $t = 34.654 > 2.101 = t_c$, we reject the null hypothesis (that $\beta_1 = 0$) and accept the alternative hypothesis that $\beta_1 \neq 0$. In this case we state that there is a statistically significant relationship between the floor area of flats and the prices of flats. We reject the null hypothesis that there is no relationship between floor area of flats and price. If the calculated value of *t* is greater than -2.101 and less than 2.101, we do not reject the null hypothesis (Figure 4.1):

$$-2.101 < t < 2.101.$$

Computer software such as the data analysis in Excel calculates the t-values for null hypothesis such that the regression coefficient is zero, and determines the standard error of the estimation, and the standard error of the parameters. See the Excel calculation in Appendix 4.1.

For the sake of completeness, let us give the t-test for testing the hypothesis that the intercept of the regression function is zero:

$$H_0: \beta_0 = 0,$$
$$H_a: \beta_0 \neq 0.$$

Let us give the calculated value of *t*:

$$t = \frac{\hat{\beta}_0}{\sigma_{\hat{\beta}_0}} = \frac{0.134149}{0.69255} = 0.1937.$$

The critical t-value is:

$$\alpha = 0.05,$$
$$df = 20 - 2 = 18,$$
$$t_c = t_{(1-\frac{\alpha}{2},n-2)} = t_{0.975,18} = 2.101.$$

Since $t = 0.1937 < 2.101 = t_c$, we do not reject the null hypothesis (that $\beta_0 = 0$) and do not accept the alternative hypothesis that $\beta_0 \neq 0$. In this case, there is no statistically

significant evidence against the null hypothesis, that the intercept of the estimated regression equation is not different from zero.
Note:

- If the t-value of a regression parameter is greater than the t-value of other parameters, this does not mean that the parameter with the greater t-value is more important, and its explanatory power in changes in the dependent variable is stronger than the other parameter's explanatory power.

To select the critical t-value (or theoretical t-value) from its table, we need the level of significance and the degrees of freedom, which is equal to the number of observations minus the number of parameters estimated (including the constant parameter) or $(n-k-1)$ (n is the number of observations, k is the number of estimated parameters without the constant parameter).

The level of significance

The level of significance means the probability that the calculated t-value is greater than the critical t-value, if the null hypothesis were valid. In other words, the level of significance is the probability of rejecting the null hypothesis when it is true. Let assume that we reject the null hypothesis at the 15% level of significance; in this case 15% percent is the probability that the null hypothesis was indeed correct.

The decision rule of the t-test can be applied:

- for a one-sided hypothesis around zero,

$$H_0: \beta_k \leq 0,$$
$$H_a: \beta_k > 0.$$

$$H_0: \beta_k \geq 0,$$
$$H_a: \beta_k < 0.$$

- for a two-sided hypothesis around zero,

$$H_0: \beta_0 = 0,$$
$$H_a: \beta_0 \neq 0.$$

- for a one sided hypothesis, where the hypothesized value is different from zero,

$$H_0: \beta_k \leq C,$$
$$H_a: \beta_k > C.$$

$$H_0: \beta_k \geq C,$$
$$H_a: \beta_k < C.$$

- for a two-sided hypothesis where the hypothesized value is different from zero.

$$H_0: \beta_0 = \mathrm{C},$$
$$H_a: \beta_0 \neq \mathrm{C}.$$

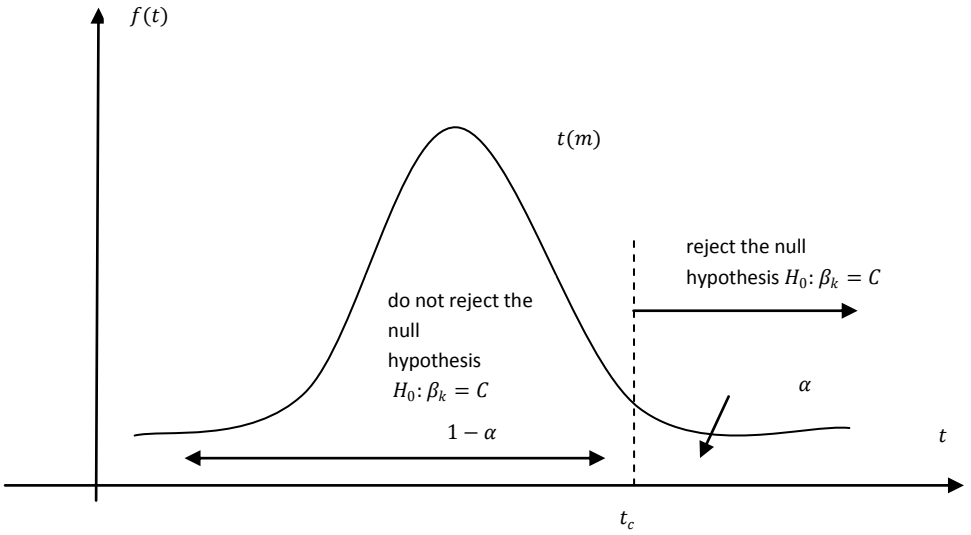
The examination of one-sided tests

1. Let us assume that the null hypothesis and the alternative hypothesis are:

$$H_0: \beta_k = C,$$
$$H_a: \beta_k > C.$$

We reject the null hypothesis if the calculated t-value is greater than the critical value for the given level of significance (α).

Figure 4.2 One-sided test of $H_0: \beta_k = C$ against $H_a: \beta_k > C$



The decision rule:

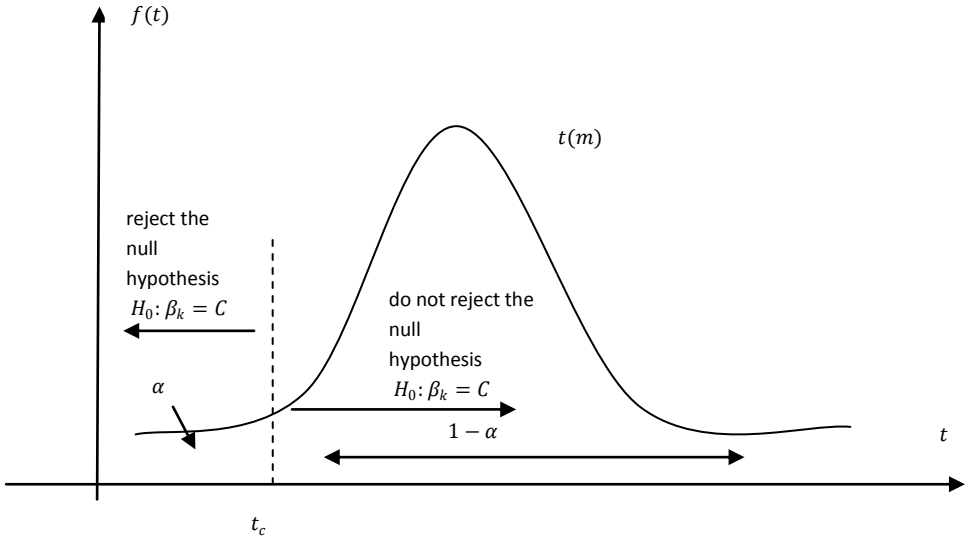
Reject the null hypothesis and accept the alternative hypothesis when $t \geq t_{(1-\alpha,n-2)}$ (Figure 4.2). The t statistic has a t-distribution, and its values fall in the non-rejection region with a probability of $(1-\alpha)$ if the null hypothesis is valid. If $t < t_{(1-\alpha,n-2)}$ is true, there is no statistically significant evidence against the null hypothesis (Figure 4.2).

2. Let us assume that the null hypothesis and the alternative hypothesis are:

$$H_0: \beta_k = C,$$
$$H_a: \beta_k < C.$$

We reject the null hypothesis if the calculated t-value is lower than the critical value for the given level of significance (α).

Figure 4.3 One-sided test of $H_0: \beta_k = C$ against $H_a: \beta_k < C$



The decision rule:
Reject the null hypothesis and accept the alternative hypothesis when $t \leq t_{(\alpha,n-2)}$ (Figure 4.3). The t statistic has a t-distribution, and its values fall in the non-rejection region with a probability of $(1-\alpha)$ if the null hypothesis is valid. If $t > t_{(\alpha,n-2)}$ is true, there is no statistically significant evidence against the null hypothesis (Figure 4.3).

4.3.2. Confidence interval

Confidence intervals are the ranges of values in which the parameters are likely to be located. If we estimate the slope of the linear regression model, we make a point estimate of the parameter. However, we can provide an interval that is likely to contain the parameter under the assumptions of the linear regression model. We provide an interval estimation if we give the confidence interval, that is the range of likely values for the parameter. Let assume that we estimate the 90 percent confidence interval for the slope of the regression model, and we have the possibility to take repeated samples. The 90 percent confidence interval means that the confidence interval would contain the true value in 90 out of 100 of the samples.

The interval estimation provides an interval that contains the given parameter with a given probability. The interval can be given by its two endpoints. The probability that the calculated t-value can be found between $-t_c \leq t \leq t_c$ is:

$$P(-t_c \leq t \leq t_c) = 1-\alpha, \tag{4.6}$$

since

$$P(t \leq -t_c) = P(t_c \leq t) = \frac{\alpha}{2},$$

where α is the probability (Figure 4.1).
Substitute the t-value (4.5) into equation (4.7) to get:

$$P(-t_c \leq \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \leq t_c) = 1-\alpha,$$

and

$$P(-\hat{\beta}_1 - t_c \cdot \sigma_{\hat{\beta}_1} \leq -\beta_1 \leq -\hat{\beta}_1 + t_c \cdot \sigma_{\hat{\beta}_1}) = 1-\alpha,$$

$$P(\hat{\beta}_1 - t_c \cdot \sigma_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_c \cdot \sigma_{\hat{\beta}_1}) = 1-\alpha, \tag{4.7}$$

The endpoints of the interval are random, since they vary from sample to sample:

$$\hat{\beta}_1 - t_c \cdot \sigma_{\hat{\beta}_1},$$
$$\hat{\beta}_1 + t_c \cdot \sigma_{\hat{\beta}_1}.$$

These endpoints provide an interval estimator of the regression parameter β_1. Equation (4.7) shows the probability that the interval containing the parameter β_1 is equal to $(1-\alpha)$. Endpoints $\hat{\beta}_1 - t_c \cdot \sigma_{\hat{\beta}_1}$, $\hat{\beta}_1 + t_c \cdot \sigma_{\hat{\beta}_1}$ determines the $(1-\alpha)\cdot 100\%$ confidence interval of β_1.
The level of significance (α) of a test represents the probability that we reject the null hypothesis, even if it is true.

To calculate the confidence interval, we need the two-sided critical t-value and the standard error of the estimated parameter:

$$\hat{\beta}_j - t_c \cdot \sigma_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_c \cdot \sigma_{\hat{\beta}_j}.$$

If the standard error of the estimated parameter is greater, the confidence interval is wider. In this case the sampling variability of the least square estimator is greater. In this case the least square estimator is less reliable.

Example 4.5

Let us give the confidence interval for the two parameters in the examples 2.4 and 4.4. The confidence intervals of the slopes of the regression equation for the confidence level of 90% and the confidence level of 95% are:

Confidence level of 90%:

$$\hat{\beta}_1 - t_c \cdot \sigma_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_c \cdot \sigma_{\hat{\beta}_1}.$$

$$0.34609 - 1.74 \cdot 0.009987 \leq \beta_1 \leq 0.34609 + 1.74 \cdot 0.009987$$
$$0.3287 \leq \beta_1 \leq 0.36347$$

Confidence level of 95%:

$$0.34609 - 2.101 \cdot 0.009987 \leq \beta_1 \leq 0.34609 + 2.101 \cdot 0.009987$$
$$0.325 \leq \beta_1 \leq 0.36707$$

Does the interval contain the true value of β_1? Actually, we do not know.

The confidence interval for the constant parameter of the regression equation is:

$$\hat{\beta}_0 - t_c \cdot \sigma_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_c \cdot \sigma_{\hat{\beta}_0}.$$

Confidence level of 90%:

$$0.134149 - 1.74 \cdot 0.69255 \leq \beta_0 \leq 0.134149 + 1.74 \cdot 0.69255$$
$$-1.07089 \leq \beta_0 \leq 1.339186$$

Confidence level of 95%:

$$0.134149 - 2.101 \cdot 0.69255 \leq \beta_0 \leq 0.134149 + 2.101 \cdot 0.69255$$
$$-1.3209 \leq \beta_0 \leq 1.589$$

There is a relationship between the two-tail test of significance and the examination of the confidence interval. If the hypothesized value of the parameter falls within the confidence interval of 95%, we will not reject the null hypothesis at the 5 percent level of significance in the (two tail) t-test. As we can see in example 4.5, the confidence interval of the constant parameter β_0 includes its hypothesized value (zero), and in example 4.4 we have not rejected the null hypothesis. However, the opposite can be experienced in the case of the β_1 parameter. The confidence interval does not include zero, and we have rejected the null hypothesis.

4.3.3. The p-value

The *p*-value provides the minimum significance level at which the null hypothesis would be rejected. Let us assume that we make a two-sided test of our model at the 5% significance level, According to the *t* statistic the calculated value of *t*: $t = 1.622$, and the critical value is equal to 1.725. In this case, we do not reject the null hypothesis, since the calculated value of t is less than its critical value. We can make one additional test at the 10% significance level. Instead of the new test, we can give the probability value (p-value), the lowest level of significance at which we reject the null hypothesis. If the significance level is 5% and the p-value is less than or equal to 5%, we can reject the null hypothesis.

The p-value decision rule:
Reject the null hypothesis H_0 if the p-value is less than the level of significance

$$p\ value \leq \alpha,$$

and if the regression parameter has the sign implied by H_a. Researchers generally use the p-value to evaluate overall or individual significance, since the result of the p-values are provided by statistical (econometrics) software packages. The results of the p-values calculated by computer show the p-values of each regression parameter, which is used to test the individual significance, and the F significance represents the p-value that can be used to test for overall significance. In section 5.4.1, we will define and give an example of an F statistic.

Example 4.6:
The result of the price and floor area calculated by a simple regression estimation can be found in Appendix 4.1.
The p value for the regression parameter β_1 is very low, which means that we can reject the null hypothesis in the t-statistic, since the value of p is less than the significance level ($\alpha = 0.05$). However, in the case of the intercept parameter the null hypothesis is true, as the previous calculation demonstrated. The p-value for β_0 is equal to 0.848578, i.e. it is greater than the significance level ($\alpha = 0.05$).

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.134149	0.692552	0.193703	0.848578	-1.32085	1.589147	-1.32085	1.589147
Square metre	0.346094	0.009987	34.6533	6.24E-18	0.325112	0.367077	0.325112	0.367077

4.4. Terms and Questions

alternative hypothesis,
calculated t-value,
confidence interval,
critical t-value,
degrees of freedom,
hypothesis testing,
interval estimation,
level of significance,
null hypothesis,
one-sided hypothesis test,
point estimation,
p-value,
standard error of the estimation,
test of significance,
t-statistic,
t-test,
t-test decision rule,
test of significance,
type I error,
type II error,
t-value,
two-sided hypothesis test,
variance of the error term.

Problems

Theoretical questions

1. What is the difference between a calculated t-value and a critical t-value?

2. What is the one-sided hypothesis test? Give an example of a one-sided test.

3. Give the definition of the confidence interval.

4. What is the decision rule for the t-test?

5. What does level of confidence mean?

6. Give the definition of the degrees of freedom.

7. Explain what the difference is between a null hypothesis and an alternative hypothesis.

8. How should we determine the calculated t-value?

9. What does type I error mean?

10. What are the main steps of a test of significance?

11. What is the difference between a one-sided hypothesis test and a two-sided hypothesis test?

12. Give an example of a null hypothesis and its alternative hypothesis.

13. What is the difference between a point estimation of the regression parameter β_1 and an interval estimation of the regression parameter β_1?

Exercises

1.

You have the following information about a sample survey:

X_i	Y_i
0	118
3	130
6	142
12	164
15	178
20	200
28	232

a) Illustrate these data on a scatter diagram.
b) Approximate the relationship between the two variables by using the scatter diagram.
c) Find the mean of data *X*.

d) Find the mean of data *Y*.

e) Compute the linear correlation coefficient. Evaluate the result of the correlation coefficient.

f) Compute the regression parameters. Give the estimated regression equation.

g) Predict the value of Y when $X = 25$, and determine the value of the residual.

h) Compute the mean square error by using

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-2}.$$

i) Compute the standard error of the estimation by using

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

j) Compute the estimated standard deviation of the parameter $\hat{\beta}_1$ by using:

$$\sigma_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

2.

Consider the following linear model:

$$Y = \beta_0 + \beta_1 \cdot X + u,$$

where u is the error term, and we assume that the assumptions of the regression models I-VI (Chapter 3.1) are valid. Let us assume that the value of the intercept (β_0) is zero.

a) Show that the estimated regression parameter is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sum_{i=1}^{n} X_i^2}$$

b) Show that the estimated regression parameter is a biased estimator of β_1 if the constant parameter is not zero.

c) Show that the estimated regression parameter is an unbiased estimator of β_1 if the constant parameter is not zero and $\bar{X} = 0$.

3.

You have the following information about a sample survey:

X_i	Y_i
5	240
10	310
50	802
68	1025
125	1748
140	1930
180	2428

a) Illustrate these data on a scatter diagram.

b) Compute the linear correlation coefficient. Evaluate the result of the correlation coefficient.
c) Compute the regression parameters. Give the estimated regression equation.
d) Predict the value of Y when $X = 150$, and determine the value of the residual.
e) Compute the mean square error.
f) Compute the standard error of the estimation.
g) Compute the estimated standard deviation of the parameter $\hat{\beta}_1$.
h) Compute the estimated standard deviation of the parameter $\hat{\beta}_0$.
i) Test the null hypothesis that the coefficient β_1 is zero against the alternative hypothesis that it is not, at the 5% level of significance.
j) Create a 95% confidence interval for the coefficient β_1. Give the interpretation of your result.

4.

Consider the following simple regression equation in which the independent variable (*NKM*) is the total number of kilometres driven of a given type of car. The *NKM* is given in thousands of kilometres. The dependent variable represents the price of the car (*PCAR*) in hundreds of EUR. The estimated regression equation based on the data consisting of 30 observations is:

$$\begin{aligned} PCAR &= \hat{\beta}_0 - 11.2 \cdot NKM, \\ standard\ errors&: \ (128), \ (2.8), \\ t\ statistics&: \ (12.8), \\ n &= 30. \end{aligned}$$

a) Interpret the coefficient of NKM.
b) Determine the regression parameter $\hat{\beta}_0$.
c) Create a 95% confidence interval for the coefficient of NKM. Give the interpretation of your result.
d) Test the null hypothesis that the coefficient of NKM is zero against the alternative hypothesis that it is not, at the 5% level of significance.
e) Compare the result of the t-test to the confidence interval. Give the interpretation of the relationship between the result of the hypothesis test and the confidence interval.

5.

Return to exercises 4.1 and 4.4 in which we have analysed the relationship between the price and the area of flats. Use the results of the exercises.

a) Interpret the regression coefficient $\hat{\beta}_1$.
b) Create a 99% confidence interval for the coefficient $\hat{\beta}_1$. Give the interpretation of your result.
c) Test the null hypothesis that the coefficient $\hat{\beta}_1$ is zero against the alternative hypothesis that it is not, at the 1% level of significance.

d) Compare the result of the t-test to the confidence interval. Give the interpretation of the relationship between the result of the hypothesis test and the confidence interval.

e) Test the hypothesis that the coefficient $\hat{\beta}_1$ is zero against the alternative hypothesis that it is positive, at the 5% level of significance. Give your conclusion.

6.

You have the following information about a sample survey:

X_i	Y_i
0	8000
50	7900
100	7600
150	7400
200	7100
250	6800
300	6600

a) Illustrate these data on a scatter diagram.

b) On the basis of your expectations, what will be the sign of the slope parameter?

c) Compute the linear correlation coefficient. Evaluate the result of the correlation coefficient.

d) Compute the regression parameters. Give the estimated regression equation.

e) Predict the value of Y when $X = 200$, and determine the value of the residual.

f) Compute the mean square error.

g) Compute the standard error of the estimation.

h) Compute the estimated standard deviation of the parameter $\hat{\beta}_1$.

i) Compute the estimated standard deviation of the parameter $\hat{\beta}_0$.

j) Test the null hypothesis that the coefficient β_1 is zero against the alternative hypothesis that it is not, at the 5% level of significance.

k) Create a 95% confidence interval for the coefficient β_1. Give the interpretation of your result.

l) Create a 95% confidence interval for the coefficient β_0. Give the interpretation of your result.

7.

Suppose that you would like to estimate the impact of Political Stability and Absence of Violence/Terrorism on GDP. According to the World Bank definition, the indicator of Political stability is:

"*Political Stability and Absence of Violence/Terrorism measures perceptions of the likelihood of political instability and/or politically-motivated violence, including terrorism" (World Bank, 2017a). "[An] estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e. ranging from approximately -2.5 to 2.*5" (World Bank,

2017b). A greater value of the political stability indicator represents greater stability in a country.

Observation	Political Stability and Absence of Violence/Terrorism	GDP per capita, PPP (constant 2011 $)	Observation	Political Stability and Absence of Violence/Terrorism	GDP per capita, PPP (constant 2011 $)
1	0.902	43631.24	19	1.408	93899.66
2	1.186	44048.43	20	-0.874	16490.35
3	0.603	41825.81	21	0.930	46353.85
4	1.242	42983.1	22	1.491	35158.64
5	0.960	30380.59	23	1.148	63649.51
6	0.892	45483.76	24	0.874	25322.54
7	0.621	27345.25	25	0.874	26548.62
8	1.039	38993.67	26	0.957	28254.26
9	0.273	37775	27	0.919	29097.34
10	0.715	43787.82	28	0.288	32218.54
11	-0.227	24094.79	29	0.965	45488.29
12	0.732	24831.35	30	1.311	56517.45
13	1.267	42704.42	31	-1.276	19460.48
14	0.927	61378.36	32	0.557	38509.21
15	-1.118	31970.69	33	0.699	52704.2
16	0.344	34219.76	34	-0.561	13572.19
17	-1.223	34386.57	35	-0.921	5733.452
18	0.452	23080.36			

a) Illustrate the data on a scatter diagram.
b) Compute the linear correlation coefficient.
c) Evaluate the result of the correlation coefficient.
d) Use the least squares method to estimate the regression parameters. Give the estimated regression equation.
e) Give the interpretation of the slope of the estimated regression equation.
f) Compute SST, SSR and SSE.
g) Examine the goodness of fit.
h) Give other independent variables which may affect the dependent variable.
i) Calculate the standard error of the regression coefficient.
j) Create a 95% confidence interval for the coefficient $\hat{\beta}_1$. Give the interpretation of your result.
k) Test the null hypothesis that the coefficient $\hat{\beta}_1$ is zero against the alternative hypothesis that it is not, at the 1% level of significance.
l) Compare the result of the t-test to the confidence interval. Give the interpretation of the relationship between the result of the hypothesis test and the confidence interval.

m) Test the hypothesis that the coefficient $\hat{\beta}_1$ is zero against the alternative hypothesis that it is positive, at the 5% level of significance. Give your conclusion.

8.

Consider the following simple regression equation in which the dependent variable (*ICECR*) is the household's demand for ice cream. The *ICECR* is given in kilograms. The independent variable represents the price of the product (*PICECR*), in EUR/kilograms. The estimated regression equation based on data consisting of 100 observations is:

$$ICECR = 280 - 12.5 \cdot PICECR,$$
$$standard\ errors: \ (112), \ (4.2),$$
$$n = 100.$$

a) Interpret the coefficient of the independent variable - $PICECR$.
b) Create a 95% confidence interval for the coefficient of $PICECR$. Give the interpretation of your result.
c) Test the null hypothesis that the coefficient of $PICECR$ is zero against the alternative hypothesis that it is not, at the 5% level of significance.
d) Compare the result of the t-test to the confidence interval. Give the interpretation of the relationship between the result of the hypothesis test and the confidence interval.
e) Test the hypothesis that the coefficient of $PICECR$ is zero against the alternative hypothesis that it is positive, at the 5% level of significance. Give your conclusion.
f) Give other independent variables which may affect the dependent variable.

Appendix 4.1

We illustrate the main steps to calculate the main results for Example 2.4 by using Excel's Regression tool.
An estate agent collected data on the price and floor area of 20 flats in Debrecen in September 2017. Remember we analysed the relationship between the price and floor area of 20 flats.

Data of flats		
Number	Price	Floor area
	Million HUF	Square metres
1	16.2	48
2	19.4	55
3	24.5	71
4	28.2	82
5	35.2	100
6	29.1	85
7	24.6	70
8	25.8	73
9	26.1	74
10	22.2	66
11	14.5	35
12	17.6	53
13	24.8	73
14	14.1	39
15	23.75	67
16	16.8	49
17	18.4	51
18	21.5	61
19	16.2	53
20	43	122

Step 1:
Enter the data
Enter the labels Number, Price and Floor area into the cell A1:C1 of the first worksheet. We would like to identify each of the observations, therefore enter the numbers of observations 1-20 into cell A2:A21, and the rest of the data into cell B2:C21.
Step 2:
Select the Tools menu and the Data Analysis possibility.
Step 3:
Choose regression from the list of possibilities of the Data Analysis, and fill the regression dialog box.

The result of the regression shows us the regression statistics under the summary output title. The regression statistics include the standard error, the number of observations, and the R^2 (Table 4A.1). The results of the Anova (analysis of variance examination) include the total, residual and explained sum of squares, degrees of freedom, and F test.

Below the Anova, we find the results of the linear regressions such as the intercept and slope of the estimated regression function (Table 4A.1). The table includes the standard error of the parameters, their t-values and p-values, and the confidence intervals (Table 4A.1).

Table A4.1 Results of the regression, using the Excel regression tool

Table 4A.1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,992589
R Square	0,985232
Adjusted R Square	0,984412
Standard Error	0,900111
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	972,9288	972,9288	1200,851	6,24E-18
Residual	18	14,58359	0,810199		
Total	19	987,5124			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	0,134149	0,692552	0,193703	0,848578	-1,32085	1,589147	-1,32085	1,589147
Square metre	0,346094	0,009987	34,6533	6,24E-18	0,325112	0,367077	0,325112	0,367077

LIST OF FIGURES

LIST OF TABLES

References

Anderson. D. R. – Sweeney. D. J. – Williams. T. A. (2007): Essentials of Modern Business Statistics with Microsoft® Excel. 3e. Thomson Higher Education. Student Edition: ISBN 0-324-31276-8.

Besanko. D. – Braeutigam. R. (2011): Microeconomics. International Students Version. John Wiley & Sons. Inc. Fourth Edition. ISBN: 978-0-470-64606-9.

Dougherty. C. (2011): Introduction to econometrics. Fourth Edition. Oxford University Press. ISBN: 978-0-19-956708-9.

Eurostat (2017a): Students in tertiary education - as % of 20-24 years old in the population. Eurostat database: Students in tertiary education by age groups - as % of corresponding age population (educ_uoe_enrt07). http://ec.europa.eu/eurostat/data/database. Download time. 24.07.2017. 17:06.

Hill. R. E. – W. E. Griffiths – G. E. Lin (2011): Principles of econometrics. Fourth Edition. John Wiley & Sons Inc. ISBN 978-0-470-62673-3.

Hungarian Central Statistical office (2017a): Per capita gross domestic product (GDP). Tables (STADAT) - Time series of annual data - National accounts. GDP. https://www.ksh.hu/stadat_annual_3_1. Download time. 24.07.2017. 14:03.

Hungarian Central Statistical office (2017b): Tertiary Education. Tables (STADAT) - Time series of annual data – Education. http://www.ksh.hu/docs/eng/xstadat/xstadat_annual/i_zoi007a.html. Download time. 24.07.2017. 14:32.

Leamer. E. E. (1983): Let's take the con out of econometrics. The American Economic Review. Vol. 73. No. 1. pp. 31-43.

OECD (2017a): National Accounts at a Glance. 2015. OECD Publishing. Paris. http://dx.doi.org/10.1787/na_glance-2015-en. ISBN: 9789264246799. Download time: 30.07.2017. 14:14.

OECD (2017b): Education at a Glance. 2016. OECD indicators. OECD Publishing. Paris. http://dx.doi.org/10.187/eag-2016-en. Download time: 30.07.2017. 15:04.

OECD (2017c): Labour: Labour market statistics. Main Economic Indicators (database). http://dx.doi.org/10.1787/data-00046-en. Download time: 30.07.2017. 18:11.

Studenmund. A. H. (2014): Using Econometrics. A Practical Guide. Sixth Edition. Pearson Education Limited. ISBN: 978-1-292-02127-0.

World Bank (2017a): Political Stability and Absence of Violence/Terrorism http://info.worldbank.org/governance/wgi/pdf/pv.pdf. Download time: 08.08.2017. 19.08.

World Bank (2017b): Political Stability and Absence of Violence/Terrorism. http://databank.worldbank.org/data/Views/Metadata/MetadataWidget.aspx?Name=Political%20Stability%20and%20Absence%20of%20Violence/Terrorism:%20Estimate&Code=PV.EST&Type=S&ReqType=Metadata&ddlSelectedValue=DZA&ReportID=30316&ReportType=Table. Download time: 08.08.2017. 20.21.