

DEALING WITH ECONOMETRICS

Real World Cases with Cross-Sectional Data

Jordi Ripollés

Inmaculada Martínez-Zarzoso

Maite Alguacil

Dealing with Econometrics

Dealing with Econometrics:

*Real World Cases with
Cross-Sectional Data*

By

Jordi Ripollés,
Inmaculada Martínez-Zarzoso
and Maite Alguacil

**Cambridge
Scholars
Publishing**



Dealing with Econometrics: Real World Cases with Cross-Sectional Data

By Jordi Ripollés, Inmaculada Martínez-Zarzoso
and Maite Alguacil

This book first published 2022

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2022 by Jordi Ripollés, Inmaculada Martínez-Zarzoso
and Maite Alguacil

All rights for this book reserved. No part of this book may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
electronic, mechanical, photocopying, recording or otherwise, without
the prior permission of the copyright owner.

ISBN (10): 1-5275-8500-X
ISBN (13): 978-1-5275-8500-3

The book covers the basic statistical tools needed to analyse cross-sectional data in order to identify, quantify and evaluate possible socio-economic relationships. It includes theoretical summaries as well as practical examples and exercises, some of which are solved using Excel or the Gretl software package. The exercises are based mostly on real-world data from Europe and Spain. Topics include the basic methodologies, principles and practices of cross-section econometrics, considering simple and multiple regression analysis, statistical inference, and the use of qualitative information in regression analysis. Essentially, the book is a practical manual for the Foundations of Econometrics courses commonly taught in Business Administration, Finance and Accounting, and Economics degree programmes in Europe.

January 2022

TABLE OF CONTENTS

List of abbreviations	ix
List of figures	x
List of tables	xi
1. An introduction to statistical analysis with Gretl and the simple regression model.....	1
1.1. Introduction	1
1.2. What is econometrics?.....	3
1.3. Describing the data sample.....	6
1.4. The simple regression model.....	13
Problem set 1A (with solutions)	27
Problem set 1B	34
Multiple-choice questions (Topic 1).....	36
2. The multiple regression model	39
Problem set 2A (with solutions)	47
Problem set 2B	61
Multiple-choice questions (Topic 2).....	65
3. Statistical inference in regression models.....	71
3.1. Simple hypothesis testing	72
3.2. Test of a linear combination of parameters.....	75
3.3. Multiple hypothesis testing.....	77
Problem set 3A (with solutions)	82
Problem set 3B	90
Multiple-choice questions (Topic 3).....	96
4. Other topics related to regression models.....	100
4.1. Rescaling of variables.....	100
4.2. Interactions between explanatory variables	106
4.3. Goodness-of-fit and model selection	106
Problem set 4A (with solutions)	109
Problem set 4B	115
Multiple-choice questions (Topic 4).....	119

5. Including dummy variables in regression analysis	122
5.1. Introduction	122
5.2. Interactions in regression analysis with dummy variables.....	126
Problem set 5A (with solutions)	129
Problem set 5B	133
Multiple-choice questions (Topic 5).....	137
6. Discrete choice models	141
6.1. Introduction	141
6.2. Statistical inference.....	148
6.3. Marginal effects.....	149
6.4. Odds ratio	150
6.5. Goodness-of-fit.....	151
6.6. Model selection and the Akaike Information Criterion	152
Problem set 6A (with solutions)	153
Problem set 6B	161
Multiple-choice questions (Topic 6).....	164
Solutions to the Multiple Choice Questions	167
References	171
Appendix A. Critical values (percentiles) for statistical distributions	173

LIST OF ABBREVIATIONS

CA	Coefficient Asymmetry
CK	Coefficient Kurtosis
DCM	Discrete Choice Models
DF	Degrees of Freedom
KE	Kurtosis Excess
LPM	Linear Probability Model
MLR	Multiple Linear Regression
MM	Method of Moments
NR	Non-Restricted Model
OLS	Ordinary Least Squares
PRF	Population Regression Function
R	Restricted Model
SRF	Sample Regression Function
SSR	Sum of Squares of the Residuals
TSS	Total Sum of Squares
ESS	Explained Sum of Squares

LIST OF FIGURES

Figure 1. Frequency distribution of the variable precio.....	6
Figure 2. Scatter plot of price (Y-axis) against size (X-axis) for all sampled apartments	12
Figure 3. Regression function.....	14
Figure 4. Observed values (y_i) versus estimated values (\hat{y}_i).....	16
Figure 5. Goodness-of fit in a simple regression analysis based on two different samples	18
Figure 6. Scatter plot prices-rooms.....	22
Figure 7. Level-level demand function.....	26
Figure 8. Log-log demand function	26
Figure 9. Scatter plot of life expectancy vs GDP per capita	27
Figure 10. Estimated consumption	30
Figure 11. Estimated marginal propensity to consume.....	30
Figure 12. Pairwise scatter plots.....	45
Figure 13. Frequency distribution for the variable internet	49
Figure 14. Scatter plot of distance from refinery/storage (distref) vs number of nearby rivals (rivals)	70
Figure 15. Regression model with a dummy variable	122
Figure 16. Regression model with multiple categories.....	125
Figure 17. Regression model with a dummy variable	126
Figure 18. Binary-choice model	142
Figure 19. Multiple-choice model	142
Figure 20. Linear Probability Model	144
Figure 21. Logistic function and cumulative normal function.....	147
Figure 22. Final Year Project submission.....	161

LIST OF TABLES

Table 1. Summary of Gretl commands.....	2
Table 2. Univariate descriptive statistics	11
Table 3. Matrix of correlations	13
Table 4. Functional forms in regression analysis	20
Table 5. Prices and number of rooms occupied.....	22
Table 6. OLS procedure step by step in Excel.....	23
Table 7. Numerical properties of SRF	24
Table 8. Calculating the coefficient of determination	25
Table 9. Consumption and disposable income for a sample of countries	28
Table 10. Estimated results with Data_Barcelona_cars_autocasion.gdt...	46
Table 11. Random sample	57
Table 12. Sample data in Lilliput	62
Table 13. Descriptive statistics for GDP per capita in “Data_convergence.gdt”	87
Table 14. Dataset of second-hand vehicles for sale.....	95
Table 15. Rescaling dependent or independent variables in a level-level model.....	102
Table 16. Rescaling dependent or independent variables in a level-log model.....	103
Table 17. Rescaling dependent or independent variables in a log-level model.....	104
Table 18. Rescaling dependent or independent variables in a log-log model.....	105
Table 19. Linear regression results for the determinants of the natural log of CO2 emissions, using a rescaled independent variable.....	108
Table 20. Linear regression results for the determinants of CO2 emissions	108
Table 21. Linear regression results for the determinants of the CO2 emissions, using a rescaled dependent variable.....	109
Table 22. OLS Regression results	109
Table 23. Regression results for exports and Covid-19 incidence.....	112
Table 24. Table of the Standard Normal Cumulative Distribution Function.....	146
Table 25. Possibilities to summarise marginal effects.....	149
Table 26. Logistic model regression results. Attending or no training course in Math.	159

CHAPTER 1

AN INTRODUCTION TO STATISTICAL ANALYSIS WITH GRETL AND THE SIMPLE REGRESSION MODEL

1.1. Introduction

Dealing with Econometrics: Real World Cases with Cross-Sectional Data is intended as a basic manual for the **Foundations of Econometrics** module commonly taught in degree courses in Business Administration, Finance and Accounting, Economics, and in joint honours programmes in Business Administration and Law. Most of the data samples used in the book are available here:

<https://drive.google.com/drive/folders/1yOnbfr19isWkqTKBL2HBmY0shhQadWd0>

The module typically includes laboratory sessions taught in computer rooms. In these sessions, students are introduced to the regression analysis of economic variables through exercises and problems to be solved using Microsoft Excel and the Gretl (Gnu Regression, Econometrics and Time-series Library) statistical package. Developed by Allin Cottrell of the University of Wake Forest, Gretl can be used to perform statistical analyses and estimations of econometric models. Gretl not only has an intuitive graphical user interface that makes carrying out a wide range of quantitative analyses relatively simple, but also contains a number of sample data sets taken from various econometrics manuals (Wooldridge, 2020; Stock and Watson, 2012; Verbeek, 2008; Ramanathan, 2002; among others).

Gretl is open-source software and can be downloaded from <http://gretl.sourceforge.net/>.

The Gretl commands used most frequently in the laboratory sessions are summarised in the following table:

Table 1. Summary of Gretl commands

Description	Path
Load sample data	<i>File / Open data / Sample file...</i>
Import external files from other formats, including CSV (.csv), ASCII (.txt), Excel (.xls, .xlsx) and Stata (.dta)	<i>File / Open data / User file...</i>
Inform the program what kind of data set we are going to be using: cross-sectional, time series or panel data	<i>Data / Dataset structure...</i>
Show descriptive statistics for a random variable (mean, median, minimum, maximum, standard deviation, coefficient of variation, coefficient of asymmetry, and coefficient of excess kurtosis)	<i>Right-click on the variable name / Summary statistics...</i>
Show the frequency distribution of a variable	<i>Right-click on the variable name / Frequency distribution...</i>
Show the matrix of correlations between two or more variables	<i>Selecting two or more variables (while pressing Ctrl) / Right-click on the variable name / Correlation matrix</i>
Show the scatter plot or X-Y plot	<i>View / Graph specified vars / X-Y scatter plot</i>
Estimate a model using ordinary least squares	<i>Model / Ordinary Least Squares</i>

For more information, a Gretl User's Guide can be found in the toolbar Help menu.

1.2. What is econometrics?

Econometrics is an area within economics that combines mathematics and statistics to study economic theories from an empirical perspective, with a view to verifying and quantifying them. According to Frisch (1933), econometrics should not be taken as synonymous with the application of mathematics and statistics to economics: “it is the unification of all three that is powerful. And it is this unification that constitutes econometrics” (*Econometrica* 1, pp. 1-2).

Why a separate discipline? Econometrics is based on economic models, which are crucial for interpreting the statistical results obtained. Moreover, the particular nature of the data, obtained outside of controlled experiments (i.e. the researcher collects data by passively observing the real world), makes this discipline more than just the application of mathematics and statistical methods.



What is econometrics for? Econometrics is widely used nowadays in economics and finance. The main applications of econometric tools include:

- The application of statistical methods *to test hypotheses* in economics and finance, e.g. the theoretical nexus between inflation and trade openness.¹
- The use of quantitative data and econometric models *to predict future economic trends*, e.g. the expected growth of public debt in Spain over the next few years.
- Econometrics can be used *to evaluate the implementation of certain economic policies*, e.g. the cost in jobs of an increase in the national minimum wage in the United Kingdom.
- *To estimate causal relationships*, e.g. the causal link between risk and return in equity investments.

¹ Romer, D. (1993). Openness and Inflation: Theory and Evidence. *The Quarterly Journal of Economics*, 108(4), 869-903.

How do econometricians proceed in their analysis of an economic problem?

The main steps in econometric analysis are as follows:

1. Statement of the research question or hypotheses

E.g. What are the main factors behind changes in labour productivity?

2. Specification of the economic model

Human capital theory states that workers can increase their productive capacity and thus their earnings through greater education and skills training:² $wages = f(\text{education}, \text{experience}, \text{skills})$.

3. Specification of the econometric model

The econometric model allows us to move from theoretical reflection (economic model) to its empirical counterpart. To do this we must specify the mathematical form of the function, $f(\cdot)$. How are the explained variable and the explanatory variables related?

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + \beta_3 skills + u$$

Parameters of the model

Error term

This captures the effect on *wage* of variables other than those included in the model (*education*, *experience* and *skills*). THE ERROR TERM IS CRUCIAL IN ECONOMETRIC ANALYSIS

4. Obtaining the data

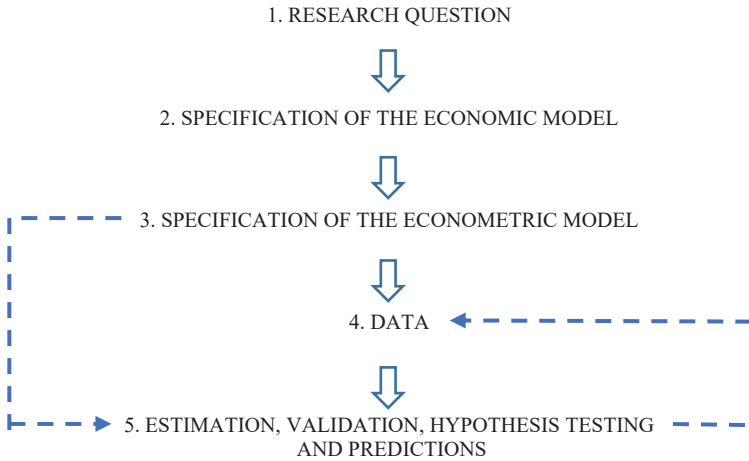
The data used in econometrics are not experimental data. They are collected by passively observing the real world.

5. Estimation, validation, hypothesis testing and prediction.

Once we have the data, our next task is to estimate the parameters of the econometric model. Then we validate our estimation by evaluating the results both from an economic point of view (Are the estimates, the signs and the magnitudes reasonable from the point of view of economic theory?)

² Becker, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

and from a statistical point of view (statistical tests on the significance of the parameters and the goodness-of-fit).



1.3. Describing the data sample

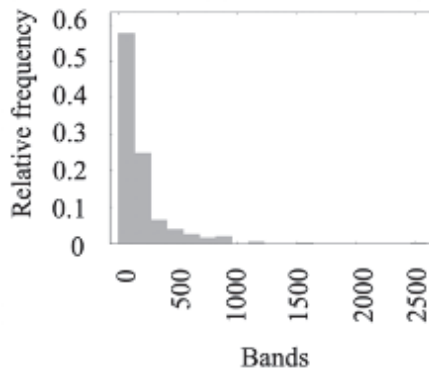
The examples and exercises in this book will be based on the analysis of cross-sectional data sets of samples of information on individuals, countries, firms and other entities collected at a given point in time. One of the basic assumptions is that the data have been collected by extracting a random sample from the underlying (unobserved) population.

In what follows we shall learn how to conduct a basic descriptive analysis of a data sample, using the file “Data_Valencia_pisos.gdt”, which contains information about a random sample of 387 apartments for sale in Valencia, taken from the Nestoria property search website (www.nestoria.es) on 15 April 2018. Specifically, we have cross-sectional data on the apartment selling price in thousands of euros (*precio*), the size of the apartment in square metres (*m2*) and the number of bedrooms (*dormitorios*).

1.3.1. Univariate descriptive analysis

The first stage in the data exploration commonly consists of a univariate descriptive analysis of the main variables of interest. For this purpose, we first obtain the **frequency distribution** of one of our variables (e.g. *precio*). The frequency distribution allows us to group the variable of interest in exclusive frequency bands of equal thickness and determine the number of observations in each band.

Figure 1. Frequency distribution of the variable *precio*



Number of bands = 19, mean = 195,642, std. dev. = 234,993

Band	Midpoint	Freq.	Rel.	Accum.
< 136.78	68.392	223	57.62%	57.62%
136.78 - 273.57	205.18	96	24.81%	82.43%
273.57 - 410.35	341.96	25	6.46%	88.89%
410.35 - 547.13	478.74	15	3.88%	92.76%
547.13 - 683.92	615.53	10	2.58%	95.35%
683.92 - 820.70	752.31	6	1.55%	96.90%
820.70 - 957.48	889.09	8	2.07%	98.97%
957.48 - 1094.3	1025.9	0	0.00%	98.97%
1094.3 - 1231.1	1162.7	2	0.52%	99.48%
1231.1 - 1367.8	1299.4	0	0.00%	99.48%
1367.8 - 1504.6	1436.2	0	0.00%	99.48%
1504.6 - 1641.4	1573.0	1	0.26%	99.74%
1641.4 - 1778.2	1709.8	0	0.00%	99.74%
1778.2 - 1915.0	1846.6	0	0.00%	99.74%
1915.0 - 2051.8	1983.4	0	0.00%	99.74%
2051.8 - 2188.5	2120.1	0	0.00%	99.74%
2188.5 - 2325.3	2256.9	0	0.00%	99.74%
2325.3 - 2462.1	2393.7	0	0.00%	99.74%
>= 2462.1	2530.5	1	0.26%	100.0

Note: Prepared by the authors based on Gretl output: Right-click on the variable precio / Frequency distribution. Data source:

Data_Valencia_pisos.gdt

Figure 1 shows the frequency distribution of the variable *precio*. As can be seen, the data have been grouped by default into 19 bands, 19 being the number closest to \sqrt{n} , where n is the number of apartments (387). In Gretl, the midpoints of the first and last bands usually correspond to the minimum and maximum values of the sample. As shown in the resulting frequency distribution table:

- Almost 60% of the apartments in the sample have a selling price lower than 136,780 euros.
- Almost 25% of the apartments in the sample have a selling price higher than or equal to 136,780 euros, but strictly lower than 273,570 euros.
- Only one apartment has a price higher than or equal to 2,462,100 euros.

The properties of the frequency distribution of a single variable can be formally described with **measures of location, dispersion, and shape**.

First, measures of location supply information about the central tendency of the data. They usually include the mean and the median. While the mean is appropriate for symmetric distributions without extreme values (outliers), the median is more useful for skewed data with outliers.

- The (arithmetic) mean, also known as the average, is the sum of the observations divided by the number of observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- The median is the middle value of the data set (from smallest to largest value):

$$\begin{aligned} Me(x_i) &= x_{(n+1)/2} && \text{if } n \text{ is odd,} \\ Me(x_i) &= (x_{(n/2)} + x_{(n/2)+1})/2 && \text{if } n \text{ is even.} \end{aligned}$$

Second, measures of dispersion inform us about the degree of homogeneity or heterogeneity of the data distribution.

- The range is the difference between the extreme observations of the sample:

$$Range(x_i) = Maximum(x_i) - Minimum(x_i)$$

- The standard deviation quantifies how much the individuals of a sample differ from the sampled mean value:

$$\widehat{sd}(x_i) = s_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

The more concentrated the data points are around \bar{x} , the closer $Range(x_i)$ and s_x will be to zero. In these cases, measures of position will be more representative of the set of observations.

However, the range and the standard deviation depend on the units of measurement of the variable being analysed, making it difficult to compare the representativeness of measures of position in two data sets expressed in different units. As a solution, one could calculate the coefficient of variation (CV), which is a standardized measure of the dispersion of a frequency distribution. It is expressed as the ratio of the standard deviation to the (absolute value) mean:

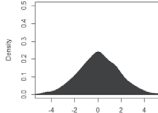
$$CV_x = \frac{s_x}{|\bar{x}|} \text{ if } \bar{x} \neq 0$$

Third, measures of shape describe the distribution of the data set in terms of skewness and kurtosis.

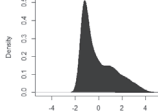
- The **coefficient of asymmetry (CA)** summarises the degree of asymmetry (or skewness) of the distribution:

$$CA = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \right)^{3/2}}$$

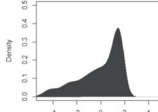
If $CA = 0 \rightarrow \bar{x} = Me(x)$ Data distributed symmetrically around the sample mean (\bar{x}).



If $CA > 0 \rightarrow \bar{x} > Me(x)$ The right tail of the distribution is longer.



If $CA < 0 \rightarrow \bar{x} < Me(x)$ The left tail of the distribution is longer.



- The **kurtosis excess (KE)** measures how deeply the tails of a distribution differ from the tails of a Gaussian (normal) distribution. Kurtosis excess can also be understood as a measure

of the width of a distribution, compared to a Gaussian distribution with the same mean and variance:

$$KE = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \right)^4} - 3$$

If $KE = 0$	Mesokurtic (or normal) distribution
If $KE > 0$	Leptokurtic distribution, with heavier tails than a Gaussian distribution
If $KE < 0$	Platykurtic distribution, with shorter tails than a Gaussian distribution. ³

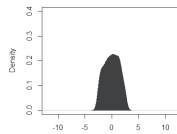
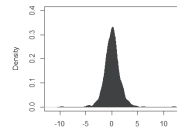
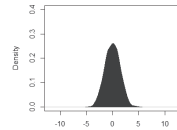


Table 1 presents a summary of the univariate descriptive statistics for the apartments' price and size. As can be seen, the sampled apartments in Valencia have an average selling price of 195,642 euros, with a standard deviation of 234,993 euros. In relative terms, this standard deviation is 120% of the average, indicating a relatively high dispersion of prices. In this case, therefore, the mean price and the median would be poor measures of the central tendency of the entire sample of prices, presumably due to the presence of extreme prices.⁴ The difference between the most expensive apartment and the cheapest one, i.e. the range, is 2,462,100 euros. Additionally, the distribution of prices has a positive asymmetry (where the mean is greater than the median), indicating that the more (less) frequent prices are below (above) the average. Finally, the distribution of prices is

³ The examples of distributions have been obtained from simulated random observations using the R package "*PearsonDS*".

⁴ This contrasts with the size variable ($m2$), whose standard deviation is 67.6% of its average, indicating that the size of the apartments in the sample is relatively homogeneous.

leptokurtic, with heavier tails than a Gaussian distribution. This last characteristic is consistent with the presence of extreme prices, lying far away from the sample average.

Table 2. Univariate descriptive statistics

	Price (precio)	Size (m2)
Mean	€195,642	118.59 m ²
Median	€123,000	100 m ²
Range	€2,462,100	881 m ²
Standard deviation	€234,993	80.119 m ²
CV	1.201 > 1	0.676 < 1
CA	4.207 > 0	6.116 > 0
KE	27.583 > 3	53.296 > 3

Note: Authors' elaboration based on Gretl output: Right-click on the variable precio / Summary statistics.

Data source: Data_Valencia_pisos.gdt.

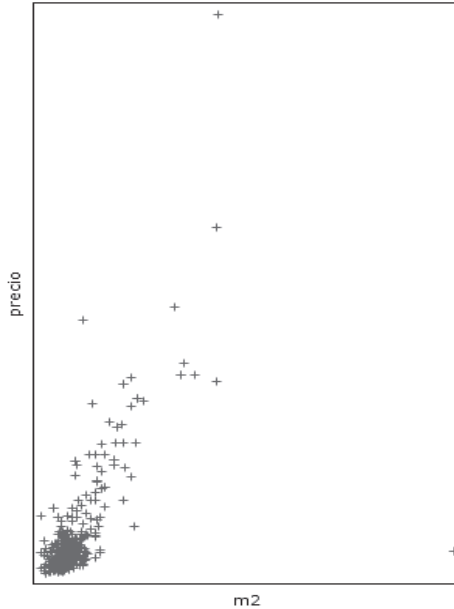
1.3.2. Multivariate descriptive analysis

So far, we have been considering univariate analysis. From now on we will explore the potential relationship between two or more variables.

To visually explore the relationship between two continuous variables, we can use a **scatter plot** (also known as an X-Y graph). A scatter plot uses Cartesian coordinates to display each pair of observations for two variables x_i and y_i for a set of data $i = 1, 2, \dots, n$. If the variables are correlated, the resulting cloud of points will form a line or curve. The stronger the correlation, the tighter the points will hug the line.

Figure 2 shows the scatter plot of apartment size ($m2$) and price ($precio$). As reasonably expected, the larger (smaller) apartments are mostly the more expensive (cheaper) ones. The cloud of points has a positive slope and a shape that closely approximates a straight line, indicating a strong positive linear relationship between the two variables.

Figure 2. Scatter plot of price (Y-axis) against size (X-axis) for all sampled apartments



Note: Authors' elaboration based on Gretl output: View / Multiple graphs / X-Y scatter plot.

Data source: Valencia_pisos.gdt

The correlation (or linear relationship) between a pair of quantitative variables x_i and y_i in sample data $i = 1, 2, \dots, n$ can be formally measured using **Pearson coefficient of correlation**:

$$r_{xy} = \frac{\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \cdot \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}}} \quad \text{where } -1 \leq r_{xy} \leq 1$$

- If $r_{xy} = 1$ → Perfect positive linear relationship between x_i and y_i .
- If $r_{xy} = 0$ → No linear relationship between x_i and y_i .
- If $r_{xy} = -1$ → Perfect inverse linear relationship between x_i and y_i .

Table 3 shows a matrix of simple correlations for each pair of variables. We see a relatively strong positive linear association between apartment size and price ($r_{m2,precio} = 0.5534 > 0.5$), i.e. as apartment size increases, apartment price also increases in a constant proportion.

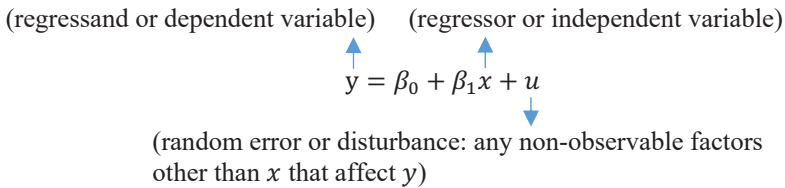
Table 3. Matrix of correlations

<i>m2</i>	<i>dormitorios</i>	<i>precio</i>	
1.0000	0.6608	0.5534	<i>m2</i>
	1.0000	0.4476	<i>dormitorios</i>
		1.0000	<i>precio</i>

Note: Authors' elaboration based on Gretl output: Select variables of interest / Right-click / Correlation matrix. Data source: Valencia_pisos.gdt.

1.4. The simple regression model

The relationship $y = f(x)$ can be studied through a simple linear **econometric model**:



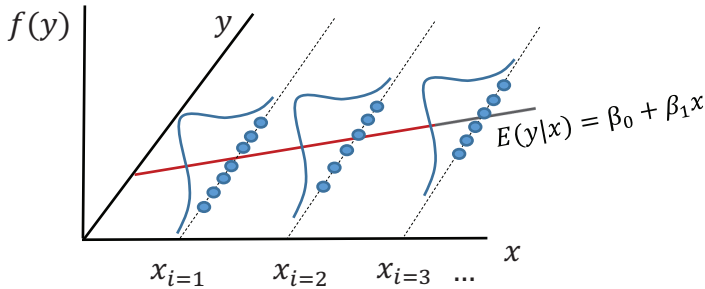
On the one hand, the constant parameter β_0 indicates the value taken by y when $x = 0$. On the other, the slope parameter β_1 provides information about how much y changes given one unit change in x , when other factors that may influence y are relegated to the disturbance term.⁵ For this latter to be the case, it must be possible to assume that the average value of u is independent of the value of x for all i (*zero conditional mean assumption*):

$$\beta_1 = \frac{\Delta y}{\Delta x} \Big|_{\Delta u=0} \leftarrow \boxed{\text{If } E(u|x)=E(u)=0}$$

⁵ The model is linear in the β parameters. That is to say, β_1 shows the change in y associated with a one-unit change in x , regardless of the level of x .

Consequently, the **population regression function (PRF)** provides a *linear relationship* between the mean of y , $E(y)$ and the different values of x presented by the individuals in a population: $E(y|x) = \beta_0 + \beta_1x$.

Figure 3. Regression function



The aim of the regression analysis is to assess the relationship between y and x by estimating the (fixed but unknown) population parameters β_0 and β_1 from a set of observations in a sample. With this purpose in mind, the following steps are taken:

1. We draw a random sample of the population, $\{(y_i, x_i): i = 1, 2, \dots, n\}$
2. We specify a model that is linear in the β parameters for each observation i in the sample: $y_i = \beta_0 + \beta_1x_i + u_i$
3. We estimate the **sample regression function (SRF)** for the model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_i$
4. As **estimation method**, we use the Method of Moments (MM) or Ordinary Least Squares (OLS). In our study framework, both methods yield the same result.⁶

⁶ There are other estimation methods, such as maximum likelihood estimation (MLE), which consists in selecting the values of the parameters that maximise the probability of obtaining the sample observations. In linear models, MLE, which has the desirable asymptotic properties under more general conditions, also coincides with OLS estimation for large samples.

Method of Moments (MM) involves finding estimates of the population parameters β_0 and β_1 that meet the following two restrictions:⁷

Population moment conditions:

- (1) $E(u) = E(y - \beta_0 - \beta_1 x) = 0$
- (2) $Cov(x, u) = E(xu) = E(x(y - \beta_0 - \beta_1 x)) = 0$

Sample versions of the moment conditions:

- (1) $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
- (2) $\frac{1}{n} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

Solving the last system of equations, we obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, which define the SRF: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$$(1) \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \rightarrow \quad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

$$(2) \frac{1}{n} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \cdot n;$$

$$\sum_{i=1}^n x_i (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i);$$

$$\sum_{i=1}^n x_i (y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x})$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \frac{\widehat{Cov}(x_i, y_i)}{\widehat{Var}(x_i)}$$

where

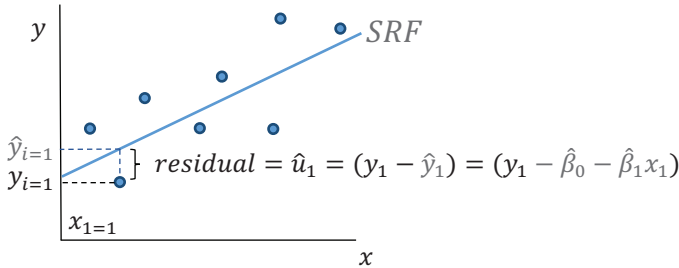
- s_{xy} is the sample covariance between x and y, defined as $s_{xy} = \widehat{Cov}(x_i, y_i) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

⁷ $Cov(x, u) = E[(x - E[x])(u - E[u])] = E(xu) - E(x)E(u) - E(x)E(u) + E(x)E(u) = E(xu)$

- s_x^2 is the sample variance of x . That is, $s_x^2 = \widehat{Var}(x_i) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$

The resulting SRF, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, provides the estimated value of the dependent variable y for each observable value of x . The residual for each observation $i = 1, 2, \dots, n$ is the difference between the actual value y_i and its own estimated value \hat{y}_i based on $x = x_i$. Therefore, each observation i of variable y_i may be expressed as the sum of its predicted value according to the SRF (\hat{y}_i) and its residual (\hat{u}_i): $y_i = \hat{y}_i + \hat{u}_i$. Figure 4 summarises the graphical representation of the SRF with respect to the observed sample data on variables y_i and x_i .

Figure 4. Observed values (y_i) versus estimated values (\hat{y}_i)



Note: The representation of observed values in an X-Y scatter plot is usually called a point cloud.

We arrive at the same result using the OLS method, which consists in obtaining estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimising the sum of squared residuals:

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (\hat{u}_i)^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

where the first-order conditions are, respectively, the sample analogue of population moments (1) and (2):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\frac{\partial \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0 \rightarrow -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0 \rightarrow -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Given the procedure described above, the OLS estimates and their related statistics present the following **numerical properties**:

1. (1) $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \sum_{i=1}^n \hat{u}_i = 0$
2. (2) $\frac{1}{n} x_i \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \rightarrow \sum_{i=1}^n x_i \hat{u}_i = 0$
3. $\bar{\hat{y}} = \bar{y}$ because $\bar{y} = \bar{\hat{y}} + \bar{\hat{u}}$, where $\bar{\hat{u}} = 0$ from (1) and $\bar{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$
4. The SRF is at point (\bar{x}, \bar{y})
5. From (1) and (2), $\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$.

Goodness-of-fit: Once we have estimated the SRF from a cross-sectional data set, it is useful to measure how well the regressor explains the sample variation in the dependent variable. To do that, we can use the coefficient of determination (R^2), which also tells us how well our SRF fits the observed point cloud of the sample. In other words, R^2 measures the quality of the linear approximation.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}; \quad 0 \leq R^2 \leq 1 \quad (\text{the higher the } R^2, \text{ the better the goodness-of-fit})$$

where TSS is the total sum of squares, $\sum_{i=1}^n (y_i - \bar{y})^2$,

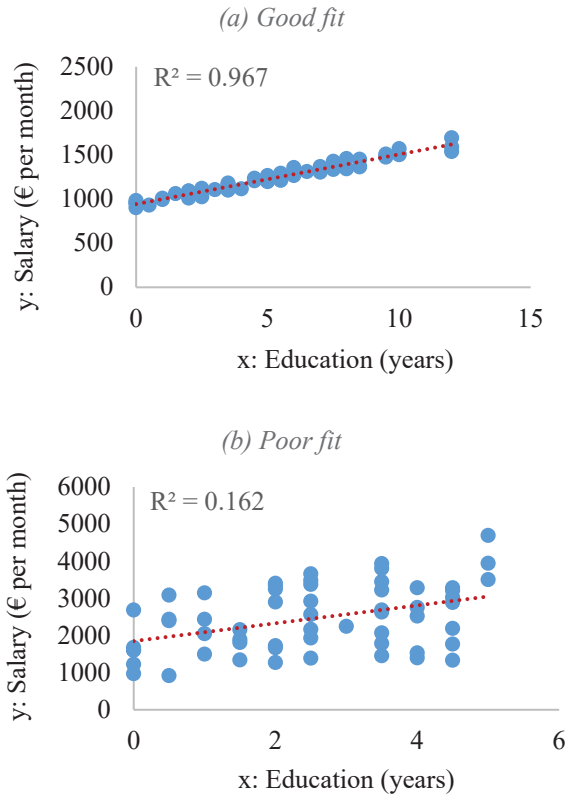
ESS is the explained sum of squares, $\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$,

SSR is the sum of squared residuals, $\sum_{i=1}^n \hat{u}_i^2$,

TSS = ESS + SSR.

In Figure 5 we present two SRFs, based on different simulated data sets of 60 individuals, of monthly salary (y) on years of education (x). In the first case (a), the data points lie close to the SRF and $R^2 = 0.967 (>0.50)$, which suggests that the OLS SRF provides a good fit to the data. In particular, we can conclude that 96.7% of the sample variation in salary is explained by education. In contrast, in the second case (b) the data points are relatively far away from the SRF and $R^2 = 0.162 (<0.50)$, suggesting that the OLS SRF provides a poor fit to the data. In this second case, 16.2% of the sample variation in salary is explained by education.

Figure 5. Goodness-of fit in a simple regression analysis based on two different samples



Note: Authors' elaboration based on Excel output: *Insert / Scatter (X, Y) Chart*

If R^2 is low, then the SRF will have a poor capacity to predict suitable estimated values of the dependent variable, \hat{y}_i , using observed values of $x = x_i$. Nevertheless, if the zero conditional mean assumption is fulfilled, the OLS SRF will still be able to properly estimate the *ceteris paribus* linkage between dependent and explanatory variable, regardless of the size of R^2 .

Finally, let us now illustrate one special case for regression analysis, which is based on a regression line that passes through the origin $(y, x) = (0, 0)$. This specification, commonly known as *regression through the origin*, is only appropriate if $E(y|x = 0) = 0$. Otherwise, the estimated coefficient β_1

will be biased. In regression through the origin, the OLS procedure for the model $y_i = 0 + \beta_1 x_i + u_i$ is given by:

$$\min_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^n (\tilde{u}_i)^2 = \min_{\tilde{\beta}_0, \tilde{\beta}_1} \sum_{i=1}^n (y_i - 0 - \tilde{\beta}_1 x_i)^2$$

$$\frac{\partial \sum_{i=1}^n (y_i - 0 - \tilde{\beta}_1 x_i)^2}{\partial \tilde{\beta}_1} = 0 \rightarrow -2 \sum_{i=1}^n x_i (y_i - 0 - \tilde{\beta}_1 x_i) = 0;$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

In this case, the TSS can be different from (ESS + SSR) since, according to the OLS procedure, $\sum_{i=1}^n \tilde{u}_i$ does not have to be equal to zero.

The following table shows the different functional forms we can work with, which are linear in the parameters, together with their main characteristics.

Table 4. Functional forms in regression analysis

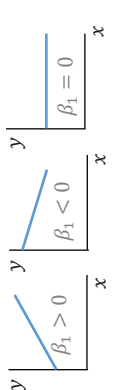

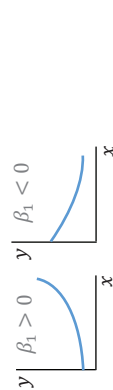
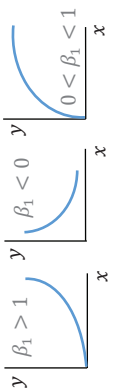

Type	Econometric model	Slope	Interpretation of slope	Regression function
Level-level	$y = \beta_0 + \beta_1 x + u$	$\frac{\Delta y}{\Delta x} = \beta_1$	For each unit increase in x , y changes by β_1 units, regardless of the level of x .	
Level-log	$e^y = e^{\beta_0} x^{\beta_1} e^u$ $y = \beta_0 + \beta_1 \log(x) + u$	$\frac{\Delta y}{\Delta \log(x)} = \frac{\Delta y}{\Delta x/x} = \beta_1$; $\frac{\Delta y}{\Delta x} = \frac{\beta_1}{100}$	A 1% increase in x causes y to change by $(\beta_1/100)$ units.	
Log-level or semi-elasticity	$y = e^{(\beta_0 + \beta_1 x + u)}$ $\log(y) = \beta_0 + \beta_1 x + u$	$\frac{\Delta \log(y)}{\Delta x} = \frac{\Delta y/y}{\Delta x} = \beta_1$; $\frac{\% \Delta y}{\Delta x} = \beta_1 \cdot 100$	For each unit increase in x , y changes by $(\beta_1 \cdot 100)\%$.	
Log-log or elasticity	$y = e^{\beta_0} x^{\beta_1} e^u$ $\log(Y) = \beta_0 + \beta_1 \log(x) + u$	$\frac{\Delta \log(y)}{\Delta \log(x)} = \frac{\Delta y/y}{\Delta x/x} = \beta_1$; $\frac{\% \Delta y}{\% \Delta x} = \beta_1$	A 1% increase in x causes y to change by $\beta_1\%$, i.e., β_1 is the elasticity of y with respect to x .	

Table 4 (continued)

<p>Quadratic⁸</p>	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$\frac{\Delta y}{\Delta x} = \beta_1 + 2\beta_2 x$	<p>For each unit increase in x, y changes by $(\beta_1 + 2\beta_2 x)$ units, i.e. the marginal effect of x on y depends on the starting level of x.</p>	
------------------------------	---	--	---	---

Note: *log* refers to the natural logarithm.

Properties and approximations used:

$$\log(x^y) = y \cdot \log(x)$$

$$\log(x^\alpha \cdot y^\beta) = \alpha \cdot \log(x) + \beta \cdot \log(y)$$

$$\log(e^\alpha) = \alpha \cdot \log(e) = \alpha$$

$$\Delta \log(y) = \log(y_1) - \log(y_0) = \log\left(\frac{y_1}{y_0}\right) = \log\left(\frac{\Delta y + y_0}{y_0}\right) = \log\left(\frac{\Delta y}{y_0} + 1\right) \approx \frac{\Delta y}{y} \quad (\text{when } \frac{\Delta y}{y} \text{ is close to zero})$$

⁸ In this case, the x value that maximises or minimises the SRF $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ is given by $\frac{\partial \hat{y}}{\partial x} = 0 \rightarrow \hat{\beta}_1 + 2\hat{\beta}_2 x = 0$; $x = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$.

To illustrate simple regression analysis, let us now use an example. The following table shows, for a group of hotels in a town, the price per night of a room and the average number of rooms occupied per day.

Table 5. Prices and number of rooms occupied

i	Price (euros/night)	Number of rooms occupied
1	35	150
2	100	20
3	90	50
4	115	10
5	70	100
6	60	130
7	50	180
8	80	100

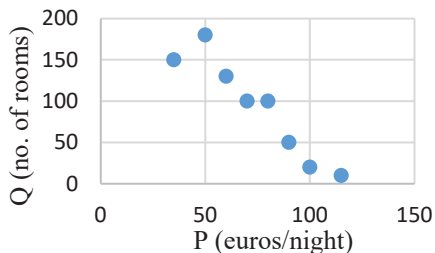
Imagine that we want to explain the relationship between hotel demand and price using the model:

$$Q_i = \beta_0 + \beta_1 P_i + u_i \quad \text{with } i = 1, 2, \dots, 8 \text{ hotels}$$

where Q represents the number of rooms occupied and P is the room price per night.

a) With the information provided, we first draw and interpret a scatter plot (X - Y plot) that shows the relationship between the two variables (response and explanatory).

Figure 6. Scatter plot prices-rooms



There is an inverse relationship between price and number of rooms occupied. The hotels with higher (lower) prices have lower (higher) room occupancy

Note: Authors' elaboration based on Excel output: *Insert / Scatter (X, Y) Chart*

b) Now we use the OLS procedure to estimate the SRF of the proposed model and interpret the estimated values of the constant and the slope.

Table 6. OLS procedure step by step in Excel

	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
hotel (id)	P_i	Q_i	$(P_i - \bar{P})$	$(Q_i - \bar{Q})$	$(P_i - \bar{P})(Q_i - \bar{Q})$	$(P_i - \bar{P})^2$
1	35	150	-40	57.5	-2300	1600
2	100	20	25	-72.5	-1812.5	625
3	90	50	15	-42.5	-637.5	225
4	115	10	40	-82.5	-3300	1600
5	70	100	-5	7.5	-37.5	25
6	60	130	-15	37.5	-562.5	225
7	50	180	-25	87.5	-2187.5	625
8	80	100	5	7.5	37.5	25
	\bar{P}	\bar{Q}			$\sum_{i=1}^n (P_i - \bar{P})(Q_i - \bar{Q})$	$\sum_{i=1}^n (P_i - \bar{P})^2$
	=75	=93			= -10800	= 4950

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (P_i - \bar{P})(Q_i - \bar{Q})}{\sum_{i=1}^n (P_i - \bar{P})^2} = \frac{-10800}{4950} = -2.1818$$

$$\hat{\beta}_0 = \bar{Q} - \hat{\beta}_1 \bar{P} = 93 - (-2.1818)75 = 256.1364$$

Therefore, the SRF is given by

$$\hat{Q} = 256.1364 - 2.1818 P$$

whose estimated coefficients can be interpreted as follows:

- If the price per night were zero, the estimated number of rooms occupied would be 256.
- When the price increases by 1 euro/night, the number of rooms occupied is predicted to decrease by 2.1818 units, holding other factors constant (if $E[u|P] = E[u] = 0$).

c) Additionally, considering the SRF, we can now evaluate its numerical properties:

Table 7. Numerical properties of SRF

hotel (id)	x_i	y_i	Adjusted values		Residuals		
			\hat{y}_i	\hat{Q}_i	$\hat{u}_i = (y_i - \hat{y}_i)$	$x_i \hat{u}_i$	$\hat{y}_i \hat{u}_i$
1	35	150	$\hat{Q}_1 = 256.1364 + (-2.1818) \cdot (35) = 179.773$		-29.773	-1042.06	-352.382
2	100	20	$\hat{Q}_2 = 256.1364 + (-2.1818) \cdot (100) = 37.955$		-17.955	-1795.5	-681.482
3	90	50	$\hat{Q}_3 = 256.1364 + (-2.1818) \cdot (90) = 59.773$		-9.773	-879.57	-584.162
4	115	10	$\hat{Q}_4 = 256.1364 + (-2.1818) \cdot (115) = 5.227$		4.773	548.895	24.948
5	70	100	$\hat{Q}_5 = 256.1364 + (-2.1818) \cdot (70) = 103.409$		-3.409	-238.63	-352.521
6	60	80	$\hat{Q}_6 = 256.1364 + (-2.1818) \cdot (60) = 125.227$		4.773	286.38	597.708
7	50	200	$\hat{Q}_7 = 256.1364 + (-2.1818) \cdot (50) = 147.045$		32.955	1647.75	4845.868
8	80	100	$\hat{Q}_8 = 256.1364 + (-2.1818) \cdot (80) = 81.591$		18.409	1472.72	1502.009
	$\bar{P} = 75$	$\bar{Q} = 93$		$\bar{Q} = 93$	$\sum_{i=1}^n \hat{u}_i = 0$	$\sum_{i=1}^n x_i \hat{u}_i = 0$	$\sum_{i=1}^n \hat{y}_i \hat{u}_i = 0$

d) Concerning the goodness-of-fit, the coefficient of determination (R^2) allows us to evaluate what proportion of the sample variation in demand is explained by price. It is calculated as follows:

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

where $SSR = \sum_{i=1}^n \hat{u}_i^2$ and $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$

Table 8. Calculating the coefficient of determination

	x_i	y_i	Squared residuals	$(y_i - \bar{y})^2$
hotel (id)	P_i	Q_i	\hat{u}_i^2	$(Q_i - \bar{Q})^2$
1	35	150	$(-29.773)^2 = 886.432$	3306.25
2	100	20	$(-17.955)^2 = 322.366$	5256.25
3	90	50	$(-9.773)^2 = 95.506$	1806.25
4	115	10	$(4.773)^2 = 22.779$	6806.25
5	70	100	$(-3.409)^2 = 11.622$	56.25
6	60	80	$(4.773)^2 = 22.779$	1406.25
7	50	200	$(32.955)^2 = 1086.002$	7656.25
8	80	100	$(18.409)^2 = 338.895$	56.25
			$\sum_{i=1}^n \hat{u}_i^2$ = 2,786.364	$\sum_{i=1}^n (y_i - \bar{y})^2$ = 26,350

Then, $R^2 = 1 - \frac{2786.364}{26,350} = 0.8943$

That is, 89.43% of the sample variation in demand is explained by price. In this case, the OLS SRF fits the observed data relatively well.

e) Based on the SRF obtained, we predict the average number of rooms that would be occupied if the price were set at 75 euros per night.

$$\hat{Q} = 256.1364 - 2.1818 (75) = 92.5 \text{ rooms occupied}$$

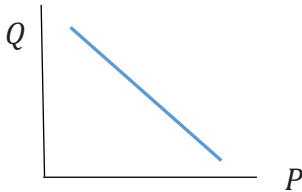
f) From the regression function obtained, we calculate the price elasticity of the demand for a price of 75 euros per night.⁹

⁹ Remember that the price elasticity of demand is defined as $\epsilon_p^d = \frac{\Delta Q}{\Delta P} \cdot \frac{P}{Q}$ where $\frac{\Delta Q}{\Delta P} = \beta_1$ in the regression.

$$\hat{\epsilon}_p^d = \frac{\Delta \hat{Q}}{\Delta P} \cdot \frac{P}{\hat{Q}} = -2.1818 \cdot \frac{75}{92.5} = -1.769$$

For a price of 75 euros/night, an increase of 1% in the price is estimated to reduce the number of rooms occupied by 1.769%. Figure 7 shows the shape of the estimated demand function.

Figure 7. Level-level demand function

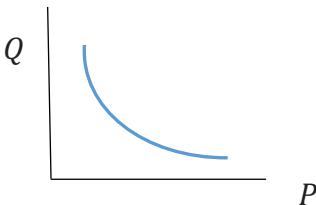


g) Specify a model that can be used to directly obtain the (constant) price elasticity of the demand and explain which parameter in that model would be the elasticity.

Log-log model: $\log(Q_i) = \beta_0 + \beta_1 \log(P_i) + u$

In this case, the parameter β_1 would directly indicate the price elasticity of (constant) demand. The model would represent an isoelastic demand function; i.e. with constant elasticity irrespective of price level, as shown in Figure 8.

Figure 8. Log-log demand function



Problem set 1A (with solutions)

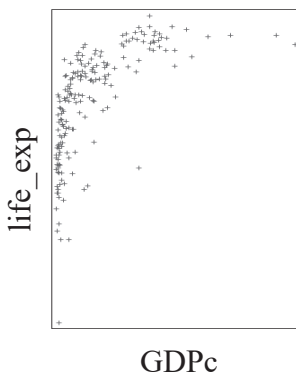
EXERCISE 1A.1 The file “Data_Gapminder_2010.gdt” contains information on various macroeconomic variables, including gross domestic product expressed in dollars per capita (GDPc), and life expectancy in years (life_exp), for different countries in 2010, taken from Gapminder (free material from www.gapminder.org).

- a) Draw the scatter plot (X-Y plot) with GDP on the horizontal axis and life expectancy on the vertical axis. What kind of relationship would you say there is between GDP and life expectancy?
- b) Based on the above plot, now propose an econometric model that better describes the relationship between the two variables. Use OLS to estimate the model.
- c) Interpret the estimated parameters of the SRF based on the proposed model.

SOLUTION 1A.1:

- a) In Gretl: View / Multiple graphs / X-Y scatters...

Figure 9. Scatter plot of life expectancy vs GDP per capita



The figure shows that the countries with high (low) levels of economic development are the ones with high (low) life expectancy. However, the relationship is not monotonically increasing, since successive increases in GDP are associated with smaller increases in life expectancy.

b) Given the scatter plot, a log-log specification could be suitable to model the exhibited linkage across life expectancy and GDP:

$$\log(\widehat{life_exp}) = \beta_0 + \beta_1 \log(GDPc) + u$$

where $\widehat{life_exp}$ represents life expectancy in years and $GDPc$ is gross domestic product, expressed in dollars per capita. Finally, u is the error term, capturing other unobserved factors that could affect life expectancy.

In Gretl (Model / Ordinary least Squares), we can obtain the following SRF using the OLS estimator:

$$\log(\widehat{life_exp}) = 3.462 + 0.087 \log(GDPc)$$

$$n = 189 \quad R^2 = 0.558$$

c) On the one hand, according to the estimated parameter $\hat{\beta}_1 = 0.087$, a 1% increase in GDP per capita is associated with a 0.087% increase in life expectancy, holding other factors constant (if a zero conditional mean can reasonably be assumed). On the other hand, the estimated constant $\hat{\beta}_0$ provides information on $\log(\widehat{life_exp})$ when $\log(GDPc) = 0$. That is, $\widehat{life_exp} = e^{3.462} = 31.88$ years when $GDPc = 1$ \$ per capita.

EXERCISE 1A.2 The file “Data_cons_inc.xlsx” contains information from Eurostat (European Statistical Office, reference: non-financial transactions, “nasq_10_nf_tr”) on consumption (*cons*) and disposable income (*inc*) in 2016, both expressed in millions of euros, for 15 European countries.

Table 9. Consumption and disposable income for a sample of countries

<i>i</i> Country	<i>cons</i> (<i>y</i>)	<i>inc</i> (<i>x</i>)
1. Germany	1674394	1970801
2. Austria	186225	213596
3. Belgium	216574	241024
4. Denmark	131609	139498
5. Spain	644719	700113
6. Finland	119005	127195
7. France	1232883	1425435
8. Greece	121737	114009
9. Ireland	90847	94739
10. Italy	1022411	1137017

11. Luxembourg	16037	20071
12. Netherlands	310692	337048
13. Portugal	121335	128768
14. United Kingdom	1577330	1626064
15. Sweden	205911	235318

- a) Use the random sample of size $n = 15$ to estimate the following model by OLS:

$$cons_i = \beta_0 + \beta_1 inc_i + u_i \quad \text{where } i = 1, 2, \dots, 15.$$

Interpret the estimated constant and slope parameters. Based on the estimated equation, by how much will consumption change if disposable income increases by one million euros?

- b) Based on the estimates obtained, calculate the predicted consumption when disposable income is 50 billion euros.
- c) Based on the estimated results, represent graphically and comment on the behaviour of the following measures in relation to disposable income:

- Estimated consumption, \widehat{cons}
- Estimated marginal propensity to consume, $PMgC = \frac{\partial \widehat{cons}}{\partial inc}$

SOLUTION 1A.2:

- a) In Gretl (Model, Ordinary Least Squares), we obtain the following SRF, using the OLS estimator with the 15 observations:

$$\widehat{cons} = 8065.9 + 0.887 inc$$

$$n = 15 \quad R^2 = 0.995$$

where $cons$ and inc refer to consumption and income. On the one hand, according to the estimated slope parameter $\hat{\beta}_1$, each additional million euros of disposable income causes a change of 0.887 million euros in consumption, holding other unobserved factors constant, if the zero conditional mean assumption holds. On the other hand, according to the estimated constant parameter $\hat{\beta}_0$, the average level of consumption is 8065.9 million euros when disposable income is 0.

b) When disposable income is 50 billion euros (i.e., $inc = 50,000$), then the predicted consumption is $\widehat{cons} = 8065.9 + 0.887(50000) = 52415.9$ million euros.

c) As can be seen, Figure 10 shows the estimated consumption for different values of income, according to the estimated SRF. Additionally, Figure 11 shows the estimated slope parameter, which represents the marginal propensity to consume. In this case, using a level-level model, consumption changes by 0.887 million euros for each additional million euros of income, regardless of the initial level of income.

Figure 10. Estimated consumption

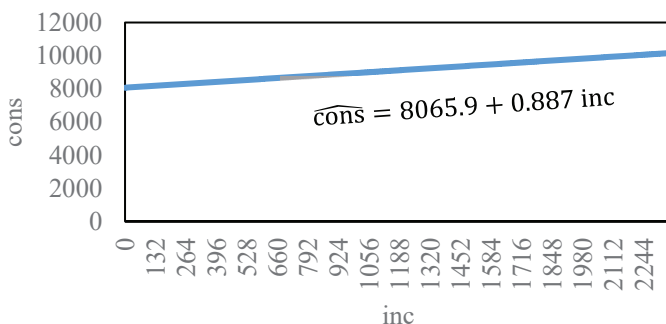
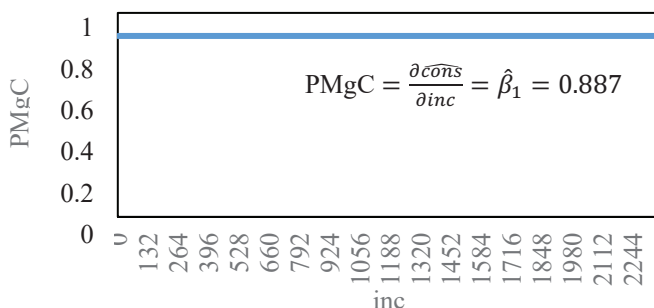


Figure 11. Estimated marginal propensity to consume



EXERCISE 1A.3 The file “Data_salarios2014ESP.gdt” contains information for 2014 about the wages (*salbase*, expressed in euros/month) and years of service (*antig*) of a sample of workers living and working in Spain. This information has been extracted from the Spanish National Institute of Statistics (INE) Wage Structure Survey.

a) Calculate the mean and standard deviation of both the wages and the length of service of the workers in the sample.

b) What proportion of the individuals in the sample have less than one year’s service ($antig < 1$)? What is the greatest length of service in the sample?

c) Estimate the following regression model: $salbase = \beta_0 + \beta_1 antig + u$ and show the results of the SRF. Based on the estimated equation, interpret the constant term and the slope.

d) According to the model estimated in c), what proportion of the sample variation in wages is explained by length of service?

e) Specify and estimate a regression model that predicts the percentage change in wages for each additional year of service. Interpret the constant term. What is the estimated percentage increase in wages when length of service increases by 15 years?

f) Specify and estimate a regression model that allows you to directly obtain the elasticity of wages with respect to length of service. Interpret the constant term. What would be the estimated percentage increase in wages if length of service were doubled?

SOLUTION 1A.3:

a) In Gretl: select the variable of interest / right click / Summary statistics

$$\overline{salbase} = \frac{\sum_{i=1}^n salbase_i}{n} = 1285.9 \text{ euros/month}$$

$$S_{salbase} = \sqrt{\frac{\sum_{i=1}^n (salbase_i - \overline{salbase})^2}{n-1}} = 766.83 \text{ euros/month}$$

$$\overline{antig} = \frac{\sum_{i=1}^n antig_i}{n} = 9.966 \text{ years}$$

$$S_{antig} = \sqrt{\frac{\sum_{i=1}^n (antig_i - \overline{antig})^2}{n-1}} = 9.689 \text{ years}$$

b) In Gretl: select the variable of interest / right-click / Frequency distribution / Minimum value, left bin: 0 and bin width: 1

Frequency distribution for antig, obs 1-208675
number of bins = 60, mean = 9.96613, sd = 9.68928

Interval	Midpt	Frequency	Rel.	Cum.
< 1.0000	0.50000	28518	13.67%	13.67% ****
1.0000 - 2.0000	1.5000	16211	7.77%	21.43% **
2.0000 - 3.0000	2.5000	11121	5.33%	26.76% *
3.0000 - 4.0000	3.5000	10039	4.81%	31.57% *
...				

Therefore, 13.67% of individuals in the sample have less than one year's service ($antig < 1$).c) In Gretl (Model / Ordinary least squares), we obtain the following SRF for a level-level model using the OLS estimator:

$$\widehat{salbase} = 1124.73 + 16.170 \text{ antig}$$

$$n = 208,675 \quad R^2 = 0.042$$

As can be seen, according to the estimated constant parameter, the average wage is 1124.73 euros/month when years of service are 0. Regarding the

estimated slope parameters, the results suggest that the average level of wages increases by 16.170 euros/month for each additional year of service.

d) According to the coefficient of determination, only 4.2% of the sample variability of wages is explained by the individuals' years of service.

e) In this case, a log-level model should be estimated: $\log(\widehat{salbase}) = \beta_0 + \beta_1 \log(antig) + u$, where $\beta_1 \cdot 100 = \% \frac{\Delta \widehat{salbase}}{\Delta \log(antig)}$. In Gretl we need to calculate the logarithm of the dependent variable (Add / Logs of selected variables) and then estimate the log-level model (Model / Ordinary least squares) using the OLS estimator. The resulting SRF is given by:

$$\log(\widehat{salbase}) = 6.883 + 0.011 \log(antig) \\ n = 208,631 \quad R^2 = 0.033$$

where wages change by 1.1% for each additional year of service. Therefore, the estimated percentage increase in wages when length of service increases by 15 years is 16.5% (= 1.1% · 15).

f) If we are interested in directly obtaining the elasticity of wages with respect to length of service, we need to estimate a log-log model, as follows: $\log(\widehat{salbase}) = \beta_0 + \beta_1 \log(antig) + u$, where $\beta_1 = \frac{\% \Delta \widehat{salbase}}{\% \Delta \log(antig)}$. Again, in Gretl (Model / Ordinary least squares) we can obtain the logarithms of the dependent and explanatory variables (Add / Logs of selected variables) and obtain the corresponding SRF using the OLS estimator (Model / Ordinary least squares):

$$\log(\widehat{salbase}) = 6.821 + 0.104 \log(antig) \\ n = 180,114 \quad R^2 = 0.030$$

where the results suggest that an increase of 1% in years of service causes an increase of 0.104% in average wages. Therefore, the estimated percentage increase in wages if length of service were doubled (i.e. a 100% increase in *antig*) would be 10.4% (= 0.104% · 100).

Problem set 1B

EXERCISE 1B.1 Search online for a cross-sectional data set for two economic and business variables that you think may be related.¹⁰ Using that data set, perform the following tasks:

a) Using Excel, save the two variables in columns, naming and sorting them, together with an index variable $i = 1, 2, \dots, N$ to represent the cross-sectional dimension (e.g. individuals, countries, households, companies, etc.). Do not forget to state the source for the data and the meaning of each variable and its unit of measurement.

b) Use economic theory or logical reasoning to explain which is the dependent variable, Y , and which the explanatory variable, X .

c) Import the Excel file into Gretl and graph and interpret the frequency distribution and descriptive statistics of the dependent variable that is to be explained (see, for example, the solution provided in section 1.3.1.).

d) In Excel or Gretl, draw and interpret a scatter plot (X-Y plot) that shows the relationship between the two selected variables. Additionally, calculate and interpret the correlation coefficient between Y and X (see, for example, the solution provided in section 1.3.2.).

e) Based on the above plot, now propose an econometric model that better describes the relationship between the two variables. You can take as a reference the information provided in Table 4. Use OLS to estimate the model. *Show the calculations performed in Excel in detail and explain your reasoning.*

f) What proportion of the sample variation in the dependent variable is explained by the regressor? Use Excel to calculate your answer and show the calculations in detail.

¹⁰ Possible data sources: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (<https://ec.europa.eu/eurostat/data/database>), OECD (<https://stats.oecd.org/>), World Bank (<https://data.worldbank.org/>), UNCTADSTAT (<https://unctadstat.unctad.org/>), FAOSTAT (<http://www.fao.org/faostat/en/#data>), others (<https://www.economicsnetwork.ac.uk/links/sources>, <https://db.nomics.world/>).

EXERCISE 1B.2 Using a random sample of 1573 Spanish individuals from the 2012 European Social Survey (www.europeansocialsurvey.org), the following estimated model relating individual wellbeing to age was obtained:

$$\log(\text{happy}) = 2.255 - 0.037 \log(\text{age})$$
$$n = 1573 \text{ and } R^2 = 0.002$$

where *happy* is a variable that captures the score from 1 to 11 the respondents gave to the question: How happy are you? and *age* refers to the respondents' age in years at the time of the survey.

- a) Interpret the estimated values of the constant and the coefficient associated with $\log(\text{age})$.
- b) State which other variables could influence individuals' happiness and explain whether and how any of them could be related to age. If so, could we rely on the results of the simple regression model provided in the problem statement? Why?

EXERCISE 1B.3 The data set "Data_RD_scoreboard.gdt" (source: EU R&D Scoreboard: <http://iri.jrc.ec.europa.eu/scoreboard16.html>) contains information on the 2500 companies worldwide that invested most in research and development (R&D) in the reporting period. For the abovementioned sample of companies, the variables *rd* and *sales* represent R&D expenditure and sales, both expressed in millions of euros, for the year 2015.

- a) Using the data set described above, estimate the following model specifications. In each case, write out the estimated equations in the usual form and interpret their results.

$$rd = \beta_0 + \beta_1 \text{sales} + u$$

$$\log(rd) = \beta_0 + \beta_1 \text{sales} + u$$

$$rd = \beta_0 + \beta_1 \log(\text{sales}) + u$$

$$\log(rd) = \beta_0 + \beta_1 \log(\text{sales}) + u$$

- b) How would the results in point a) change if R&D expenditure were expressed in euros instead of millions of euros?

Multiple-choice questions (Topic 1)

1.1. After estimating by OLS a simple regression model for a random sample of the dependent (y) and explanatory (x) variables, an estimated constant $\hat{\beta}_0 = 5$ and an estimated slope $\hat{\beta}_1 = 3$ have been obtained. If we know that the mean value of x is equal to 2, what is the mean value of y ?

- (a) The mean value of y is 11 units.
- (b) The mean value of y is 13 units.
- (c) The mean value of y is 8 units.
- (d) None of the above is correct.

1.2. In a simple level-level linear regression, the quantity that indicates how much y changes for each unit change in the variable x is called:

- (a) Determination coefficient.
- (b) Slope of the regression line.
- (c) Correlation coefficient.
- (d) None of the above.

1.3. The slope parameter β_1 in a simple regression model has a *ceteris paribus* interpretation (that is, keeping other factors constant) when:

- (a) The coefficient of determination exceeds 0.50.
- (b) The expected value of the error term u does not depend on the explanatory variable.
- (c) The correlation between x and y is high.
- (d) None of the above.

1.4. A simple regression model, such as $y_i = \beta_0 + \beta_1 x_i + u_i$, has been estimated by OLS for a random sample of size $n=5$. The following information is available:

$$\widehat{Cov}(x_i, y_i) = s_{xy} = 2$$

$$\widehat{Var}(x_i) = s_x^2 = 0.8$$

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{20}{5} = 4$$

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{15}{5} = 3$$

What is the OLS estimate of the parameter $\hat{\beta}_0$?

- (a) -3.5
- (b) 4
- (c) 9.5
- (d) None of the above.

1.5. After obtaining a SRF $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ from a random sample for the dependent (y) and explanatory (x) variables, we know that the sample variation in the fitted values is 0.60 and the sum of squared residuals is 0.20. With the information available, indicate what fraction of the sample variation in y would be explained by x :

- (a) 0.80
- (b) 0.75
- (c) 0.40
- (d) None of the above.

1.6. In a simple regression model, $y = \beta_0 + \beta_1 x + u$, the coefficient of determination is defined as:

- (a) The change in y resulting from a one-unit change in x .
- (b) The correlation between the predicted y and the observed y in the sample.
- (c) The difference between the total sum of squares and the sum of squared residuals, divided by the total sum of squares.
- (d) None of the above.

1.7. In the simple regression model, the SRF $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ has certain numerical properties. Based on those properties, which of the following equalities is true?

- (a) $\sum_{i=1}^n x_i \hat{y}_i = 0$
- (b) $\sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
- (c) $\sum_{i=1}^n \hat{u}_i = 1$
- (d) $\sum_{i=1}^n \hat{y}_i y_i = 0$

1.8. In the simple regression model, the SRF $\tilde{y} = \tilde{\beta}_1 x$ has certain numerical properties. Based on them, which of the following equalities is true?

- (a) TSS = ESS + SSR
- (b) $\sum_{i=1}^n x_i (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i) = 0$
- (c) $\sum_{i=1}^n \tilde{y}_i y_i = 0$
- (d) None of the above.

1.9. Considering the simple regression model $y_i = \beta_0 + \beta_1 x_i + u_i$, which of the following statements is FALSE?

- (a) The total sum of squares is equal to the explained sum of squares plus the sum of squares of the residuals.
- (b) The mean of the fitted values of y is equal to the mean of the variable to be explained y .
- (c) There are always the same number of points above and below the regression line.
- (d) The point (\bar{x}, \bar{y}) is always on the regression line.

1.10. The coefficients of the regression line obtained by OLS are determined by minimising:

- (a) The sum of squared residuals.
- (b) The sum of squares of the x-coordinates.
- (c) The total sum of squares.
- (d) The explained sum of squares.

1.11. Consider the following information on a paired data set:

$$n = 10$$

$$\sum_{i=1}^n x_i = 15$$

$$\sum_{i=1}^n y_i = 25$$

$$\sum_{i=1}^n x_i y_i = 100 \quad \sum_{i=1}^n x_i^2 = 50$$

What is the sample covariance between x and y ?

[In other words, calculate s_{xy}]

- (a) 6.9
- (b) 2
- (c) 0.266
- (d) None of the above.

1.12. Which of the following models can be estimated by OLS?

- (a) $y = \beta_0 + \beta_1 x_1 + u$
- (b) $\log(y) = \beta_0 + \beta_1 \log(x) + u$
- (c) $e^y = e^{\beta_0} x^{\beta_1} e^u$
- (d) All of the above.

CHAPTER 2

THE MULTIPLE REGRESSION MODEL

Consider now a multiple linear regression (MLR) model with k regressors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + u$$

where the zero conditional mean assumption $E(u|x_1, x_2, \dots, x_k) = E(u) = 0$ ensures that the slope parameter estimates have a *ceteris paribus* interpretation. In any case, unlike the simple linear model, the parameter β_j ($j = 1, 2, \dots, k$) of the MLR model measures the **partial effect** of y on k , holding constant all the other regressors included in the model other than x_j :

$$\beta_1 = \frac{\Delta y}{\Delta x_1} \Big|_{\substack{\Delta x_2=0, \dots, \Delta x_k=0 \\ \Delta u=0 \leftarrow \text{If } E(u|x_1, x_1, \dots, x_k)=E(u)=0}}$$

Again, based on a sample of data for y_i and x_{ij} , we can obtain the estimated values of the parameters $\beta_0, \beta_1, \dots, \beta_k$ using the OLS method:

$$\min \sum_{i=1}^n (\hat{u}_i)^2 = \min_{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2$$

where the first-order conditions are:

$$\frac{\partial \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2}{\partial \hat{\beta}_0} = 0 \rightarrow -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik} x_i) = 0$$

$$\frac{\partial \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2}{\partial \hat{\beta}_1} = 0 \rightarrow -2 \sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

...

$$\frac{\partial \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik})^2}{\partial \hat{\beta}_k} = 0 \rightarrow -2 \sum_{i=1}^n x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) = 0$$

Using a shorthand notation, this optimisation problem can be also expressed as follows:

Vector notation

$$\min_{\hat{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}_i' \hat{\beta})^2$$

$$\frac{\partial \sum_{i=1}^n (y_i - \mathbf{x}_i' \hat{\beta})^2}{\partial \hat{\beta}} = 0 \rightarrow -2 \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}_i' \hat{\beta}) = 0 \rightarrow$$

$$\hat{\beta} = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i=1}^n \mathbf{x}_i y_i$$

where $\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$ and $\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$.

Matrix notation

$$\min_{\hat{\beta}} (y - \mathbf{X} \hat{\beta})' (y - \mathbf{X} \hat{\beta})$$

$$\frac{\partial (y - \mathbf{X} \hat{\beta})' (y - \mathbf{X} \hat{\beta})}{\partial \hat{\beta}} = 0 \rightarrow -2(\mathbf{X}'y - \mathbf{X}'\mathbf{X}\hat{\beta}) = 0 \rightarrow$$

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'y$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_N \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \cdots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix},$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}, \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Therefore, a unique solution for $\hat{\beta}$ in the system of first-order conditions can be obtained if $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ is invertible. Then, the resulting SRF is given by:

$$\hat{y}_i = \mathbf{x}_i' \hat{\beta} \rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik}.$$

Regarding the goodness-of-fit of a multiple regression function, here too we can calculate and interpret the coefficient of determination: $R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$; $0 \leq R^2 \leq 1$. However, R^2 has a drawback because it increases when additional regressors are included in the model (given that SSR decreases), regardless of its real explanatory power. For this reason, R^2 should not be used to decide whether one or more regressors should be included in the model. The criterion for that purpose is statistical inference (see Topic 3).

Gauss-Markov assumptions of the MLR model with cross-sectional data:

MLR1. Linearity in the β parameters (the β are only raised to power 1). In the population, the linkage between the dependent variable and the regressors is linear in parameters.

MLR2. The data for the dependent variable and regressors is a **random sample** drawn from population.

MLR3. Zero conditional mean, $E(u|x_1, x_2, \dots, x_k) = E(u) = 0$. That is, independence between the regressors and all other unobservable factors that may explain the dependent variable (which are contained in the error term, u) is required.

MLR4. No perfect collinearity between regressors: (a) no explanatory variable is constant for all i , and (b) the explanatory variables are not perfectly correlated. If MLR4 is not satisfied, then $(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')$ is not invertible and then OLS estimator cannot be uniquely defined from the first-order conditions.

MLR5. Homoskedasticity:

$$var(u|x_1, x_2, \dots, x_k) = \sigma^2 \rightarrow var(y|x_1, x_2, \dots, x_k) = \sigma^2$$

Compliance with MLR1–4 guarantees the statistical property of unbiasedness of the OLS estimator ($E[\hat{\beta}_j] = \beta_j$), while MLR5 is added to guarantee the statistical property of (relative) efficiency of the OLS estimator (min $var(\hat{\beta}_j)$ among all the unbiased linear estimators).

Under **unbiasedness** of the OLS estimator, one can expect that in repeated sampling the estimated parameters would be on average equal to the true parameter β .

Unbiasedness in a simple regression: $y_i = \beta_0 + \beta_1 x_i + u_i$

$$\begin{aligned} E[\hat{\beta}_1] &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\ &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_{i1} + u_i - \beta_0 - \beta_1 \bar{x}_1 - \bar{u})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\ &= \beta_1 + E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] = \beta_1 \end{aligned}$$

Unbiasedness in a multiple regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i \rightarrow y_i = y_i + \mathbf{x}'_i \boldsymbol{\beta} + u_i$$

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i] = [(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_{i=1}^n \mathbf{x}_i (\mathbf{x}'_i \boldsymbol{\beta} + u_i)] \\ &= \boldsymbol{\beta} + E[(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_{i=1}^n \mathbf{x}_i u_i] = \boldsymbol{\beta} \end{aligned}$$

In both cases, the last step here is possible if $\sum_{i=1}^n (x_i - \bar{x})(u_i - \bar{u}) = 0$ (simple case) or $(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_{i=1}^n \mathbf{x}_i u_i = 0$ (multiple case), which occurs when the zero conditional mean assumption (MLR3) is satisfied, together with MLR1, MLR2 and MLR4.

What happens when we omit a regressor from the model? To answer this question, we evaluate the unbiasedness of the OLS estimator when a regressor is omitted. To do this, we compare a hypothetical true model (which satisfies the first four MLR assumptions) with a model with one omitted regressor:

a. True model, satisfying MLR1 – MLR4: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$

b. Estimated model by OLS, omitting the regressor x_{2i} : $\check{y}_i = \check{\beta}_0 + \check{\beta}_1 x_{1i}$

$$\begin{aligned} E[\check{\beta}_1] &= E\left[\frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}\right] \\ &= E\left[\frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + u_i - \beta_0 - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \bar{u})}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}\right] \\ &= \beta_1 + \beta_2 E\left[\frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}\right] \end{aligned}$$

Therefore, the OLS estimator will be unbiased, $E[\check{\beta}_1] = \beta_1$ when

- The omitted regressor is irrelevant ($\beta_2 = 0$) and/or
- There is no correlation between the omitted regressor (x_{i2}) and the included one (x_{i1}). That is, $Corr(x_{i1}, x_{i2}) = 0$

In contrast, the inclusion of one or more irrelevant variables in the model has no effect on the unbiasedness of the OLS estimator, provided that MLR1 – 4 are satisfied.

Under assumptions MLR1 – 4, conditional on the sample values of the regressors, one can obtain the sampling variance of the OLS estimator as follows:

The sampling variance for the simple case: $y_i = \beta_0 + \beta_1 x_{i1} + u_i$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} = \beta_1 + \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(u_i - \bar{u})}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} \\ var(\hat{\beta}_1 | x_{i1}) &= E[\hat{\beta}_1 - E(\hat{\beta}_1)]^2 \\ &= E\left[\beta_1 + \frac{\sum_{i=1}^n (x_{i1} - \bar{x}_1)(u_i - \bar{u})}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} - \beta_1\right]^2 \\ &= \frac{E(u_i - \bar{u})^2}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} = \frac{\sigma^2}{TSS_1} \end{aligned}$$

The sampling variance for the multiple case:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + u_i \rightarrow y_i = y_i + \mathbf{x}'_i \boldsymbol{\beta} + u_i \\ var(\hat{\boldsymbol{\beta}} | \mathbf{x}'_i) &= E[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]^2 = E[\boldsymbol{\beta} + (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \sum_{i=1}^n \mathbf{x}_i u_i - \boldsymbol{\beta}]^2 = \\ &= \sigma^2 (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}'_i)^{-1} \end{aligned}$$

or, alternatively,

$$\text{var}(\hat{\beta}_j | x_i) = \frac{\sigma^2}{TSS_j(1 - R_j^2)}$$

where the (relative) efficiency of the OLS estimator for the slope parameter is higher:

- The lower the error variance, $\sigma^2 = E[u - E(u)]^2$,
- The higher the total sample variation of the regressor x_{ij} , $TSS_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$
- The lower the degree of collinearity of all the independent variables, which is measured by the R_j^2 obtained by regressing x_{ij} on the other regressors in the model.

Let us now illustrate the multiple regression analysis with a real example. In order to evaluate the market value of one horsepower in second-hand vehicles, a linear regression model is used that describes the price of each used vehicle (*price*) as a function of its power, expressed in horsepower (*power_hp*), age (*age*) and mileage, expressed in kilometres (*km*). We use a cross-sectional data set containing information on 50 second-hand vehicles offered for sale in Barcelona, collected from the website **Error! Hyperlink reference not valid.** on 10 July 2021.

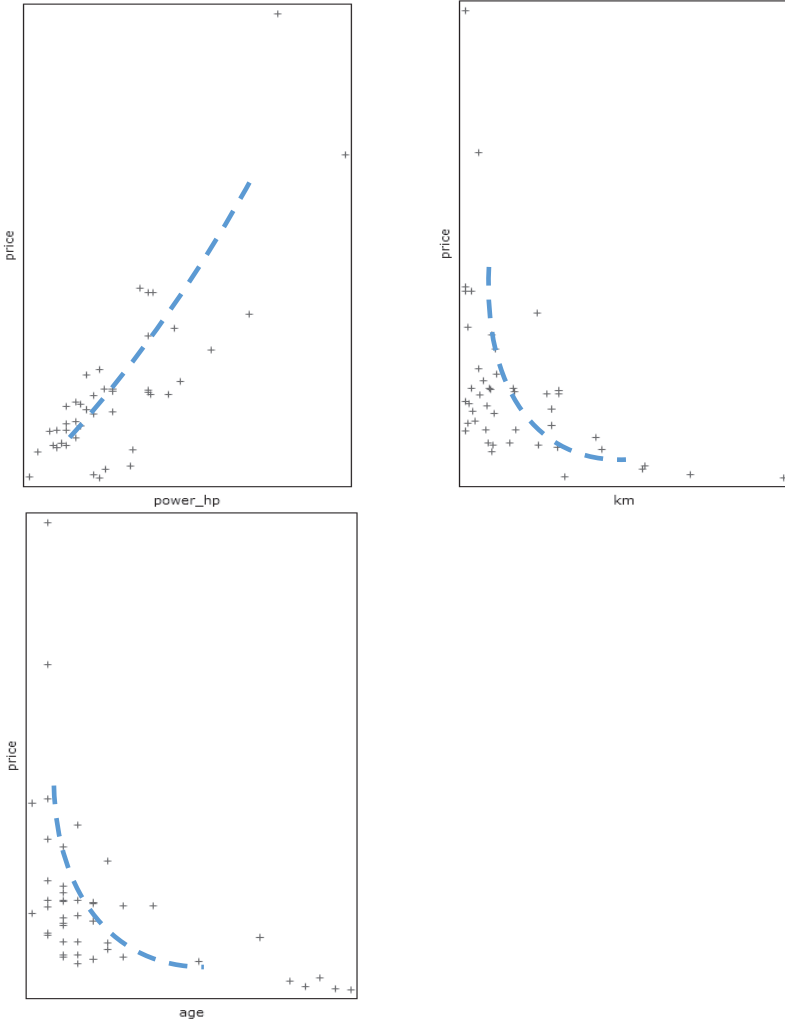
First of all, we need to identify the appropriate regression specification to evaluate the linkages of interest. To do so, we must ideally use economic theory, combined with a graphical analysis based on the data sample. In this case, we graphically evaluate the pairwise linkages *price-power_hp*, *price-km*, and *price-age* in order to deduce which functional form would be the most suitable for the regression model. For this purpose, we can focus on the shapes of the observed point clouds resulting from the sampled pairwise relationships between the dependent variable and each independent variable. As can be seen in Figure 9, the *price-power_hp* linkage could be summarised with a positive slope linear function; that is, price increases in a relatively constant proportion as power increases. However, the *price-age* and *price-km* linkages could be summarised with an inverse nonlinear function. We therefore specify the following econometric model:

$$\text{price} = \beta_0 + \beta_1 \text{power_hp} + \beta_2 \log(\text{km}) + \beta_3 \log(\text{age}) + u$$

where β_1 represents our parameter of interest, giving information about the

average change in the second-hand vehicle price in response to a one unit increase in horsepower, holding vehicle mileage and age constant.

Figure 12. Pairwise scatter plots



Note: Authors' elaboration based on Gretl output: View / Multiple graphs / X-Y scatter plot. Data source: Data_Barcelona_cars_autocasion.gdt

Next, we proceed to estimate the previously specified model by OLS. Table 10 shows the estimation results, obtained using Gretl. The estimated slope parameter can be interpreted as follows:

- $\hat{\beta}_1 = 215,385$. The average price increase by 215,385 euros for each additional horsepower, holding vehicle mileage and age constant.
- $\hat{\beta}_2 = -2,273.06$. The average price decreases by 22.731 euros for each additional 1% increase in mileage, holding horsepower and age constant.
- $\hat{\beta}_3 = -4,912.85$. The average price decreases by 49.128 euros for each additional 1% increase in vehicle age, holding horsepower and mileage constant.

Table 10. Estimated results with Data_Barcelona_cars_autocasion.gdt

Model 1: OLS, using observations 1-50 (n = 47)

Missing or incomplete observations dropped: 3

Dependent variable: price

	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	18964.100	6654.430	2.850	0.007	***
power_hp	215.385	20.051	10.74	0.000	***
l_km	-2273.060	635.909	-3.575	0.001	***
l_age	-4912.850	1687.660	-2.911	0.006	***
Mean dependent var	22401.060	S.D. dependent var	20356.770		
Sum squared resid	3.41e+09	S.E. of regression	8903.470		
R-squared	0.821	Adjusted R-squared	0.809		
F(3, 43)	65.823	P-value(F)	4.10e-16		
Log-likelihood	-492.027	Akaike criterion	992.054		
Schwarz criterion	999.455	Hannan-Quinn	994.839		

Note: Authors' elaboration based on Gretl output: Model / Ordinary Least Squares / Dependent variable: price, Regressors: power_hp, l_km and l_age

From the Gretl output, we can also obtain the coefficient of determination, R^2 . As can be seen, our SRF has a relatively high goodness-of-fit. More specifically, horsepower, mileage and age together explain 82.112% of the sample variation in vehicle price. Therefore, the specified regression would be suitable for predicting individual values for the dependent variable using the observed independent variables. In other words, the estimated regression model could be used to appraise vehicles based on their characteristics. For

example, according to our estimated regression function, a five-year-old car with a 100 hp engine and 15,000 km on the clock should be worth 10,738,37 euros:

$$p\hat{r}ice = 18964.1 + 215.385 \text{ power_hp} - 2273.06 \log(km) - 4912.85 \log(age)$$

$$p\hat{r}ice = 18964.1 + 215.385 (100) - 2273.06 \log(15000) - 4912.85 \log(5) = 10738,37 \text{ euros}$$

Problem set 2A (with solutions)

EXERCISE 2A.1 Which of the following models satisfy the assumption of linearity in the parameters and could therefore be estimated by the least ordinary squares (OLS) method:

- a) $y = \beta_0 + \beta_1 x + u$
- b) $\log(y) = \beta_0 + \beta_1 x + u$
- c) $y = \beta_0 + \sqrt{\beta_1} x + u$
- d) $y = e^{\beta_0} x^{\beta_1} e^u$
- e) $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$
- f) $y = \beta_1 + \beta_2 \left(\frac{1}{x}\right) + u$
- g) $y = \beta_1 + \log(x_1) + u$

SOLUTION 2A.1:

The following models (directly or indirectly) satisfy the assumption of linearity in the parameters and so could be estimated by the OLS method: a, b, d, e, f and g.

a) $y = \beta_0 + \beta_1 x + u$. Linear in variables and parameters.

b) $\log(y) = \beta_0 + \beta_1 x + u$. Not linear in variables, but linear in parameters.

c) $y = \beta_0 + \sqrt{\beta_1} x + u$. Linear in variables, but not linear in parameters.

d) $y = e^{\beta_0} x^{\beta_1} e^u$. Not linear in parameters and not linear in variables, but we could apply the OLS method after linearising the function in the following way:

$$y = e^{\beta_0} x^{\beta_1} e^u \rightarrow \log(y) = \beta_0 \log(e) + \beta_1 \log(x) + u \cdot \log(e) \\ \rightarrow \log(y) = \beta_0 + \beta_1 \log(x) + u$$

e) $y = \beta_1 + \beta_2 x + \beta_3 x^2 + u$. Not linear in variables, but linear in parameters.

f) $y = \beta_1 + \beta_2 \left(\frac{1}{x}\right) + u$. Not linear in variables, but linear in parameters.

g) $y = \beta_1 + \log(x_1) + u$. Not linear in variables, but linear in parameters.

EXERCISE 2A.2 Consider the MLR model presented below.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Assuming the zero conditional mean assumption is satisfied:

- What is the expected change in y if x_1 increases by 5 units and x_2 is held fixed?
- What is the expected change in y if x_2 decreases by 3 units and x_1 is held fixed?
- What is the expected change in y if x_1 increases by 5 units and x_2 decreases by 3 units?

SOLUTION 2A.2:

a) Given that $\beta_1 = \frac{\Delta y}{\Delta x_1} \Big|_{\substack{\Delta x_2=0 \\ \Delta u=0}} \rightarrow \Delta y = \beta_1 \Delta x_1 \rightarrow \Delta y = \beta_1 5$.

b) Given that $\beta_2 = \frac{\Delta y}{\Delta x_2} \Big|_{\substack{\Delta x_1=0 \\ \Delta u=0}} \rightarrow \Delta y = \beta_2 \Delta x_2 \rightarrow \Delta y = \beta_2 (-3)$.

c) Based on the previous considerations, $\Delta y = \beta_1 5 - \beta_2 3$.

EXERCISE 2A.3 The data set “Data_Internet2017_WB” contains information for 129 countries in 2017, taken from the World Bank, on the percentage of individuals using internet (*internet*), GDP per capita in thousand dollars (*gdpdollar_capita_thousand*) and the number of children out of school as a percentage of primary school age children (*children*).

- a) Present and interpret the descriptive statistics for the percentage of individuals using internet.
- b) Using the data set, estimate by OLS the following econometric model and write out the results in equation form.

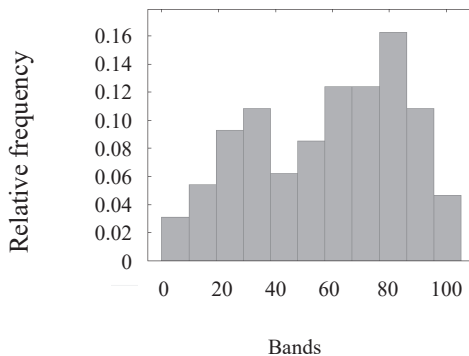
$$internet = \beta_0 + \beta_1 gdpdollar_capita_thousand + \beta_2 children + u$$

- c) What is the estimated increase in internet usage in a country for every additional thousand dollars of income, keeping the percentage number of children out of school constant?
- d) How much of the sample variation in the percentage of individuals using internet is explained by GDP and the number of children out of school?

SOLUTION 2A.3:

a)

Figure 13. Frequency distribution for the variable *internet*



Number of bands = 11, mean = 58.179, std.dev. = 26.405

Bands	Midpoints	Freq.	Rel.	Accum.
< 9.5594	4.7797	4	3.10%	3.10% *
9.5594 - 19.119	14.339	7	5.43%	8.53% *
19.119 - 28.678	23.899	12	9.30%	17.83% ***
28.678 - 38.238	33.458	14	10.85%	28.68% ***
38.238 - 47.797	43.018	8	6.20%	34.88% **
47.797 - 57.357	52.577	11	8.53%	43.41% ***
57.357 - 66.916	62.136	16	12.40%	55.81% ****
66.916 - 76.476	71.696	16	12.40%	68.22% ****
76.476 - 86.035	81.255	21	16.28%	84.50% *****
86.035 - 95.594	90.815	14	10.85%	95.35% ***
>= 95.594	100.37	6	4.65%	100.00% *

Note: Authors' elaboration based on Gretl output: Right-click on the variable *internet* / Frequency distribution. Data source: "Data_Internet2017_WB.gdt".

Figure 1 displays the frequency distribution of the variable *internet*. In this case, the data set has been grouped into 11 bands $\approx \sqrt{n} = \sqrt{129}$. As can be seen, 16.28% of sampled countries (i.e. 21 out of 129 observations) have a percentage of individuals using internet equal to or higher than 76.476% and strictly lower than 86.035%. Only the 3.10% of the sampled countries (i.e. 4 out of 129 observations) have a percentage of individuals using internet lower than 9.559%. Below we present the main properties of the distribution frequency of our variable of interest:

Measures of location

- $\overline{internet} = 58.179\%$
- $Me(internet_i) = 63.186\%$

Measures of dispersion:

- $Range (internet_i) = Max(internet_i) - Min(internet_i) = 98.225 - 2.6607 = 95.564\%$
- $S_{internet} = \sqrt{\frac{\sum_{i=1}^n (internet_i - \overline{internet})^2}{n-1}} = 26.405\%$
- $CV_{internet} = \frac{S_{internet}}{\overline{internet}} = 0.454 < 1$

Measures of shape:

- $$CA(\text{internet}_i) = \frac{\frac{1}{n} \sum_{i=1}^n (\text{internet}_i - \overline{\text{internet}})^3}{\left(\sqrt{\frac{\sum_{i=1}^n (\text{internet}_i - \overline{\text{internet}})^2}{n}} \right)^3} = -0.354 < 0$$
- $$KE(\text{internet}_i) = \frac{\frac{1}{n} \sum_{i=1}^n (\text{internet}_i - \overline{\text{internet}})^4}{\left(\sqrt{\frac{\sum_{i=1}^n (\text{internet}_i - \overline{\text{internet}})^2}{n}} \right)^4} - 3 = -1.075 < 0$$

The average percentage of internet users in the sampled countries is 58.179%, with a standard deviation of 26.405%. According to the coefficient of variation (CV = 0.454), this standard deviation is 45.4% of the average, indicating that the sample is relatively homogenous across countries. The distribution of the percentage of internet users has a negative asymmetry, with a mean lower than the median. Lastly, according to the measure of kurtosis excess, the distribution is platykurtic, with shorter tails than a Gaussian distribution.

b) Using the OLS estimator, we obtain the following SRF:

$$\widehat{\text{internet}} = 52.970 + 0.0007 \text{ gdpdollar}_{\text{capita}_{\text{thousand}}} - 1.246 \text{ children}$$

$$n = 129 \quad R^2 = 0.620$$

In Gretl: Model / Ordinary Least Squares.

c) Holding constant the number of children out of school, the estimated variation in internet usage for every additional thousand dollars of income is 0.0007 percentage points.

d) According to the coefficient of determination, 62% of the sample variation in the percentage of internet users (dependent variable) is explained by GDP and the number of children out of school.

EXERCISE 2A.4 The data set “Data_Palma_Mallorca_alquileres.gdt” contains information taken from the Nestoria property search website (<https://www.nestoria.es/>) on 27 August 2018 for a sample of rental apartments in Palma de Mallorca. Use the data set to estimate the econometric model specified below:

$$precio = \beta_0 + \beta_1 m2 + \beta_2 dormitorios + \beta_3 dist_centro + u$$

where *precio* is the rental price expressed in euros per month, *m2* is the indoor floor area expressed in square metres, *dormitorios* is the number of bedrooms, *dist_centro* is the distance from the city centre (expressed in kilometres) and *u* is the error term.

a) Write out the estimated equation in the usual form. What percentage of the change in rental price is explained by floor area, number of bedrooms and distance from the city centre.

b) What would be the estimated change in an apartment’s rental price with one additional bedroom, keeping floor area and distance from the city centre fixed? Interpret the result. Does it make sense?

c) What would be the estimated change in an apartment’s rental price with one additional bedroom that adds approximately 10 square metres to the apartment’s floor area, keeping distance from the city centre fixed? Compare the result with your answer in section b).

d) Based on the SRF, obtain the predicted rental price for an apartment with a floor area of 110 m² and two bedrooms, located 2 km from the city centre.

e) Assume the rental price of the apartment described in section d) turns out to be 1500 euros per month. Calculate the residual for this apartment. Assuming the estimated model is true, do you think the rent for the apartment is expensive?

f) Estimate the following econometric model and interpret all the estimated parameters:

$$\log(precio) = \beta_0 + \beta_1 \log(m2) + \beta_2 \log(dormitorios) + \beta_3 dist_centro + u$$

g) Using the SRF based on the model in section f), obtain the predicted value of *precio* when $m2 = 110$, *dormitorios* = 2 and *dist_centro* = 2. Is this one-off prediction better or worse than the one obtained in section (e)? Why?

SOLUTION 2A.4:

a) The results have been obtained by using Gretl: Model / Ordinary Least Square:

Model 1: OLS, using observations 1-350					
Dependent variable: precio					
	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	687.136	71.821	9.567	0.000	***
m2	7.448	0.627	11.88	0.000	***
dormitorios	-46.316	29.860	-1.551	0.122	
dist_centro	8.999	24.024	0.375	0.708	
Mean dependent var	1397.469	S.D. dependent var	580.948		
Sum squared resid	72870083	S.E. of regression	458.920		
R-squared	0.381	Adjusted R-squared	0.376		
F(3, 346)	71.092	P-value(F)	7.69e-36		
Log-likelihood	-2639.723	Akaike criterion	5287.446		
Schwarz criterion	5302.878	Hannan-Quinn	5293.589		

Using OLS method, we obtain the following SRF:

$$\widehat{precio} = 687.136 + 7.448m2 - 46.316dormitorios + 8.999dist_centro$$

$$\begin{matrix} (71.802) & (0.627) & (29.860) & (24.024) \\ n = 350 & R^2 = 0.381 & SCE = 72.870.083 \end{matrix}$$

According to coefficient of determination (R^2), 38.1% of the sample variation in apartment rental price is explained by floor area, number of bedrooms and distance from the city centre.

b) According to our estimates, apartment rental prices decrease by 46.316 euros per month for each additional bedroom, holding floor area and distance from the city centre constant. The estimated effect is reasonable, taking into consideration that, keeping apartment size constant, adding a bedroom entails subdividing an existing space into two rooms.

c) To calculate the change in an apartment's rental price with one additional bedroom that adds 10 square metres to the apartment's total floor area, keeping distance from the city centre constant, we proceed as follows:

$$\Delta \widehat{p\grave{r}e\grave{c}i\grave{o}} = 7.448(10) - 46.316(1) = 28.164 \text{ monthly euros.}$$

d) The predicted rental price for an apartment located 2 kilometres from the city centre, with a floor area of 110 square metres and two bedrooms, is given by $\widehat{p\grave{r}e\grave{c}i\grave{o}} = 687.136 + 7.448(110) - 46.316(2) + 8.999(2) = 1431.782$ monthly euros.

e) $\hat{u} = 1500 - 1431.782 = 68.218$ monthly euros. Therefore, the rental price is higher than the estimated price, based on the characteristics of the apartment (floor area, number of rooms and distance from the city centre).

f) In Gretl: Model / Ordinary Least Square:

Model 2: OLS, using observations 1-350					
Dependent variable: l_precio					
	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	4.829	0.220	21.94	0.000	***
l_m2	0.522	0.054	9.753	0.000	***
l_dormitorios	-0.055	0.047	-1.175	0.241	
dist_centro	-0.009	0.016	-0.5453	0.586	
Mean dependent var	7.173	S.D. dependent var	0.360		
Sum squared resid	30.621	S.E. of regression	0.297		
R-squared	0.322	Adjusted R-squared	0.316		
F(3, 346)	54.734	P-value(F)	5.62e-29		
Log-likelihood	-70.283	Akaike criterion	148.566		
Schwarz criterion	163.998	Hannan-Quinn	154.709		

Therefore, the SRF is given by the following expression:

$$\log(\widehat{p\grave{r}e\grave{c}i\grave{o}}) = 4.829 + 0.522\log(m2) - 0.055\log(dormitorios) - 0.009\text{dist_centro}$$

$$\begin{matrix} (0.220) & (0.054) & (0.047) & (0.016) \\ n = 350 & R^2 = 0.322 & \bar{R}^2 = 0.316 \end{matrix}$$

Estimated constant $\hat{\beta}_0 = 4.829$. When $\log(m2) = 0$ (i.e. one square metre), $\log(dormitorios) = 0$ (i.e. one bedroom) and $dist_centre = 0$ (i.e. 0 kilometres), then $\log(\widehat{precio}) = 4.829$. Therefore, in this case, the rental price for the apartment is given by the corresponding antilogarithm, $\widehat{precio} = e^{4.829} = 125.086$ euros/month.

Estimated slope $\hat{\beta}_1 = 0.522$. A 1% increase in the number of square metres leads to a price increase of 0.522%, holding the number of bedrooms and the distance from the centre constant.

Estimated slope $\hat{\beta}_2 = 0.055$. A 1% increase in the number of bedrooms leads to a price decrease of 0.055%, holding the number of square metres and the distance from the centre constant.

Estimated slope $\hat{\beta}_3 = -0.009$. Each additional kilometre of distance from the city centre leads to a 0.9% decrease in price, holding the number of square metres and the number of bedrooms constant.

g) $\log(\widehat{precio}) = 4.829 + 0.522\log(110) - 0.055\log(2) - 0.009 \cdot 2 = 7.227$. Therefore, $\widehat{precio} = e^{7.227} = 1375.438$ euros/month.

EXERCISE 2A.5 The following model explains the selling price (in euros) of a washing machine ($plav$) on the second-hand market in terms of the number of washes the machine has done ($usage$) and its age (age):

$$plav = \beta_0 + \beta_1 usage + \beta_2 age + u$$

Assuming this model satisfies the Gauss-Markov assumptions, what is the likely bias we would obtain from a simple linear regression of $plav$ on $usage$.

SOLUTION 2A.5

Multiple regression function: $\widehat{plav} = \hat{\beta}_0 + \hat{\beta}_1 usage + \hat{\beta}_2 age$

Simple regression function: $\widetilde{plav} = \check{\beta}_0 + \check{\beta}_1 usage$

To evaluate the possible bias in the OLS estimator arising from the omission of an explanatory variable in the regression analysis, we can use the following expression, which relates the expected values of the estimated slope parameters from simple regressions in different random samples from a population, $E[\check{\beta}_1]$, and the true slope parameter in the population, β_1 :

$$E[\check{\beta}_1] = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (usage_i - \overline{usage})(age_i - \overline{age})}{\sum_{i=1}^n (usage_i - \overline{usage})^2}$$

which implies that bias in $\check{\beta}_1$ is given by $\beta_2 \frac{\sum_{i=1}^n (usage_i - \overline{usage})(age_i - \overline{age})}{\sum_{i=1}^n (usage_i - \overline{usage})^2}$, where:

- β_2 represents the true partial effect of the omitted explanatory variable, that is, the true effect of age on a washing machine's selling price, holding the number of washes constant. It is reasonable to expect that the selling price will decrease with each additional year of age. Therefore, $\beta_2 < 0$.
- $\frac{\sum_{i=1}^n (usage_i - \overline{usage})(age_i - \overline{age})}{\sum_{i=1}^n (usage_i - \overline{usage})^2}$ is the estimated slope parameter of regressing age_i on $usage_i$ by OLS, whose sign will depend on the correlation between these two variables. In this case, it is reasonable to expect that older (newer) machines will have higher (lower) usage. Therefore, $\frac{\sum_{i=1}^n (usage_i - \overline{usage})(age_i - \overline{age})}{\sum_{i=1}^n (usage_i - \overline{usage})^2} > 0$.

Therefore, if we estimate by OLS the simple regression function $\overline{plav} = \check{\beta}_0 + \check{\beta}_1 usage$ on all the potential random samples from a population, omitting the explanatory variable age_i , the estimated parameter $\check{\beta}_1$ will tend to be upward biased with respect the true population parameter β_1 . That is, $E[\check{\beta}_1] > \beta_1$.

EXERCISE 2A.6 Considering the information presented in the table below, imagine that we are interested in estimating the multiple regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$:

Table 11. Random sample

y_i	x_{i1}	x_{i2}
15	-3	9
14	-2	8
7	1	3
11	0	5
20	2	7
3	3	2

Under this framework, answer the following questions:

- Obtain the vector of explanatory variables, \mathbf{x}_i .
- Calculate $\sum_{i=1}^n \mathbf{x}_i y_i$.
- Calculate $\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ and $(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')^{-1}$.
- Calculate the vector of OLS estimated coefficients $\hat{\boldsymbol{\beta}}$.
- Write the SRF, according to the obtained results.
- Obtain the estimated values for the dependent variable.
- Obtain the vector of residuals.
- Calculate the sum of squared residuals (SSR).

SOLUTION 2A.6

$$\text{a) } \mathbf{x}_i = \mathbf{X}' = \begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & 1 & 0 & 2 & 3 \\ 9 & 8 & 3 & 5 & 7 & 2 \end{pmatrix}$$

b)

$$\sum_{i=1}^n \mathbf{x}_i y_i = \mathbf{X}' \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & 1 & 0 & 2 & 3 \\ 9 & 8 & 3 & 5 & 7 & 2 \end{pmatrix} \begin{pmatrix} 15 \\ 14 \\ 7 \\ 11 \\ 20 \\ 3 \end{pmatrix} = \begin{pmatrix} 70 \\ -17 \\ 469 \end{pmatrix}$$

c)

$$\begin{aligned} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' &= \mathbf{X}' \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ -3 & -2 & 1 & 0 & 2 & 3 \\ 9 & 8 & 3 & 5 & 7 & 2 \end{pmatrix} \begin{pmatrix} 1 & -3 & 9 \\ 1 & -2 & 8 \\ 1 & 1 & 3 \\ 1 & 0 & 5 \\ 1 & 2 & 7 \\ 1 & 3 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 6 & 1 & 34 \\ 1 & 27 & -20 \\ 34 & -20 & 232 \end{pmatrix} \end{aligned}$$

$$\left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} = (\mathbf{X}' \mathbf{X})^{-1} = \begin{pmatrix} 2.464 & -0.383 & -0.394 \\ -0.383 & 0.099 & 0.065 \\ -0.394 & 0.065 & 0.068 \end{pmatrix}$$

d)

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \sum_{i=1}^n \mathbf{x}_i y_i \rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 2.464 & -0.383 & -0.394 \\ -0.383 & 0.099 & 0.065 \\ -0.394 & 0.065 & 0.068 \end{pmatrix} \begin{pmatrix} 70 \\ -17 \\ 469 \end{pmatrix} = \begin{pmatrix} -5.856 \\ 1.838 \\ 3.038 \end{pmatrix}$$

e) The SRF is given by the equation $\hat{y}_i = -5.856 + 1.838 x_{i1} + 3.038 x_{i2}$

f)

$$\hat{y}_1 = -5.856 + 1.838 (-3) + 3.037 (9) = 15.972$$

$$\hat{y}_2 = -5.856 + 1.838 (-2) + 3.037 (8) = 14.772$$

$$\hat{y}_3 = -5.856 + 1.838 (1) + 3.037 (3) = 5.096$$

$$\hat{y}_4 = -5.856 + 1.838 (0) + 3.037 (5) = 9.334$$

$$\hat{y}_5 = -5.856 + 1.838 (2) + 3.037 (7) = 19.086$$

$$\hat{y}_6 = -5.856 + 1.838 (3) + 3.037 (2) = 5.734$$

g)

$$\hat{u}_1 = -0.972$$

$$\hat{u}_2 = -0.772$$

$$\hat{u}_3 = 1.904$$

$$\hat{u}_4 = 1.666$$

$$\hat{u}_5 = 0.914$$

$$\hat{u}_6 = -2.734$$

h)

$$SSR = \sum_{i=1}^n \hat{u}_i^2 = \hat{\mathbf{u}}' \hat{\mathbf{u}}$$

$$\hat{\mathbf{u}}' \hat{\mathbf{u}} =$$

$$(-0.972 \quad -0.772 \quad 1.904 \quad 1.666 \quad 0.914 \quad -2.734) \begin{pmatrix} -0.972 \\ -0.772 \\ 1.904 \\ 1.666 \\ 0.914 \\ -2.734 \end{pmatrix}$$

$$SSR = -0.972^2 + (-0.772)^2 + 1.904^2 + 1.666^2 + 0.914^2 + (-2.734)^2 = 16.25$$

Problem set 2B

EXERCISE 2B.1 The following model is often used to explain people's wages:

$$wage = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 exper + u$$

where *wage* is the wage in euros/month, *age* is the person's age in years, *educ* is the person's total years of training, and *exper* is the person's years of job experience.

We have a data set with information on wages, age and years of education (academic training) for a sample of individuals. Unfortunately, we do not have information on job experience. As an alternative, we have used a measure of potential experience, defined as $exper = age - educ - 3$ (people generally start school at age three). Explain why in this case the parameters of the proposed model could not be estimated.

EXERCISE 2B.2 Are the following statements true or false? Justify your answer with a short sentence:

- In order to estimate an econometric model using the least ordinary squares method, the model has to be linear in the regressors.
- The fact that variables x and y are correlated means that if we know the values of x , we can predict the average value of y .
- If the explanatory variable (x) is constant, it is impossible to estimate the effect it has on the dependent variable (y).
- The variance of the OLS estimator is positively related to the number of observations, so that decreasing the sample size can improve the efficiency of our estimation.
- Including an irrelevant regressor can cause bias in OLS estimators if the estimator is correlated with the other regressors included in the model.

EXERCISE 2B.3 The following data on sales of five different brands of mobile phone in Lilliput are available:

Table 12. Sample data in Lilliput

Brand	<i>SL</i>	<i>RP</i>	<i>AP</i>
Elephone	10	8	5.5
Nikita	8	12	8.5
Saoni	7	13	9.0
Plophon	6	24	12.5
Pepaphone	13	9	6.5

where *SL* is annual sales, expressed in gold coins, *RP* is an index of relative prices and *AP* is annual expenditure on advertising and promotions, also expressed in gold coins.

Based on the above information:

- Use OLS to estimate the coefficients of the following model: $SL_i = \beta_0 + \beta_1 RP_i + u_i$.
- Obtain the coefficient of determination (R^2) of this regression and interpret the calculated value.
- Obtain the correlation coefficient between *RP* and *AP*. Would it be a good idea to add advertising and promotion as an additional explanatory variable in our regression to improve the goodness-of-fit? Give reasons for your answer.

EXERCISE 2B.4 The following model describes the price of diesel fuel in euros per litre (p_goa) set by a city's petrol stations based on the number of nearby¹¹ petrol stations operating under rival brands ($rivals$) and the distance in km from the nearest refinery-storage facility ($distref$):

$$p_goa = \beta_0 + \beta_1 rivals + \beta_2 \log(distref) + u$$

- a) What are the expected signs of β_1 and β_2 ?
- b) Using a cross-sectional data set for 597 Valencian petrol stations ($i = 1, 2, \dots, 597$), downloaded on 15 January 2017 from a Ministry of Energy, Tourism and Digital Agenda website (<http://geoportalgasolineras.es/>), the following table of results has been obtained after estimating an OLS model:

Model 1: OLS, using observations 1-597

Dependent variable: p_goa

	<i>Coefficient</i>	<i>Standard Error</i>	<i>t-statistic</i>	<i>p-value</i>	
<i>const</i>	1.075	0.025	42.48	0.000	***
<i>rivals</i>	-0.006	0.002	-3.855	0.000	***
<i>l_distref</i>	0.013	0.006	2.193	0.029	**
Mean of the dep. var.	1.125	S.D. of the dep. var.		0.044	
Sum of sq. residuals	1.097	S.D. of the regression		0.043	
<i>R</i> -squared	0.038	Adjusted <i>R</i> -squared		0.035	
<i>F</i> (2, 594)	11.712	<i>p</i> -value (of <i>F</i>)		0.000	
Log-likelihood	1033.280	Akaike criterion		-2060.560	
Schwarz criterion	-2047.384	Hannan-Quinn criterion		-2055.430	

Write out the estimated equation in the usual form and interpret the estimated parameters associated with the constant and the explanatory variables $rivals$ and $\log(distref)$.

- c) What proportion of the total variation in p_goa is explained by $rivals$ and $\log(distref)$? Justify your answer.

¹¹ Nearness has been defined by drawing a radius of 500 metres around each filling station.

d) Holding $\log(\text{distref})$ fixed, by how much would *rivals* have to increase for the price of fuel to be reduced by 0.05 euros/litre?

e) Assuming the initial econometric model satisfies the Gauss-Markov assumptions and given that the areas furthest from the refinery-storage facility are mostly the ones with the lowest density of petrol stations, what is the likely bias we would obtain from a simple linear regression of p_goa on *rivals*? Why?

Multiple-choice questions (Topic 2)

2.1. Assume that an econometric model meets the Gauss-Markov assumptions. In this case, the OLS estimate of the slope parameter will be more precise:

- (a) if the sample size decreases.
- (b) by reducing the number of explanatory variables.
- (c) if the explanatory variables show a greater correlation between them.
- (d) the greater the sample variation in the regressors.

2.2. In the context of regression analysis, select the correct answer:

- (a) Multicollinearity can bias the OLS estimates.
- (b) It is possible to use the coefficient of determination (R^2) to compare models with a different number of explanatory variables.
- (c) The variance of OLS estimates of the slopes tends to increase with the correlation among regressors.
- (d) Homoskedasticity implies that the mean of the errors is constant for all possible values of the explanatory variables.

2.3. In the context of regression analysis, check the correct answer:

- (a) OLS estimates are biased in the presence of perfect collinearity, since in this case there are exact linear relationships between the explanatory variables and the error term.
- (b) OLS estimates are always biased when a relevant variable is omitted in the regression model.
- (c) It is possible to use the coefficient of determination (R^2) to compare models with a different number of explanatory variables.
- (d) Adding an irrelevant variable to a model can lead to an increase in the variance of the OLS estimates.

2.4. Please indicate which of the following statements is correct:

- (a) In a multiple regression model, the explanatory variables cannot be correlated with each other.
- (b) The MLR model, by allowing the inclusion of more than one explanatory variable, guarantees fulfilment of the null conditional mean assumption.
- (c) Unlike in simple regression, in multiple regression the coefficient of determination cannot be used as a measure of goodness-of-fit.

(d) The multiple regression model allows us to evaluate the effects that changes in an independent variable have on the dependent variable, holding the rest of the explanatory variables constant.

2.5. The assumption of homoskedasticity:

- (a) Guarantees the unbiasedness of the OLS estimator.
- (b) Implies that the variance of the explained variable, conditional on the explanatory variables, depends on at least one of the explanatory variables.
- (c) Is necessary to guarantee that, among all the linear and unbiased estimators, the OLS estimators have the smallest variance.
- (d) Implies that the variance of the error term, conditional on the explanatory variables, changes with at least one of the explanatory variables.

2.6. The error term in the model $y_i = \beta_0 + \beta_1 x_i + u_i$ is said to be homoskedastic when:

- (a) The variance of the error term depends on x_i .
- (b) The explanatory variable x_i is constant.
- (c) The mathematical expectation of the error term is independent of x_i ; that is, $E[u|x] = 0$.
- (d) None of the above.

2.7. In a multiple regression model, the null conditional mean assumption (MLR3) can fail when:

- (a) An irrelevant explanatory variable is included in the regression model.
- (b) The variance of the error term depends on the values taken by the explanatory variables.
- (c) The functional relationship between the explained and the explanatory variables is not properly specified.
- (d) An explanatory variable is an exact linear combination of one or more other explanatory variables.

2.8. Consider an econometric model aimed at explaining the salary, in hundreds of dollars, of a set of NBA basketball players (*wage*) based on their age (*age*), years of professional experience (*exper*) and the average number of points scored during the season (*points*):

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{age} + \beta_2 \log(\text{exper}) + \beta_3 \text{points} + u$$

Using a cross-sectional data sample for 269 players, the following results have been obtained:

Model 1: OLS, using observations 1-269
 Dependent variable: l_wage

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	5.943	0.533	11.15	0.000	***
<i>age</i>	-0.015	0.023	-0.630	0.530	
<i>l_exper</i>	0.45	0.105	4.039	0.000	***
<i>points</i>	0.080	0.0070	11.49	0.000	***
Mean dependent var	6.952	S.D. dependent var		0.881	
Sum squared resid	108.215	S.E. of regression		0.639	
<i>R</i> -squared	0.480	Adjusted <i>R</i> -squared		0.474	
<i>F</i> (3, 265)	81.626	<i>p</i> -value(<i>F</i>)		1.99e-37	
Log-likelihood	-259.220	Akaike criterion		526.440	
Schwarz criterion	540.819	Hannan-Quinn		532.215	

where $l_exper = \log(exper)$

Based on the SRF obtained, select the correct answer:

- (a) If we choose two players, A and B, with the same age and experience but with a different average number of points scored, the model predicts that, for each point of difference scored, the salary increases by 8.0332%.
- (b) If we choose two players A and B with the same age and experience but with a different average number of points scored, the model predicts that, for each point of difference scored, the salary increases by 0.080332%.
- (c) Holding points scored and age constant, it is estimated that salary increases by 42.4685% for each additional 1% of experience.
- (d) Holding points scored and age constant, it is estimated that salary increases by 0.424685% for each additional year of experience.

2.9. Consider a regression model to explain the price in euros per litre of fuel (*price*) set by a city's petrol stations based on the number of nearby petrol stations operating under rival brands (*rivals*), the number of nearby petrol stations operating under the same brand (*samebrand*) and the distance in km from the nearest refinery or storage facility (*distref*):

$$price = \beta_0 + \beta_1 rivals + \beta_2 samebrand + \beta_3 \log(distref) + u$$

Using a sample of cross-sectional data for 597 petrol stations in Valencia ($i = 1, 2, \dots, 597$) as of 15 January 2017, the following table of results has been obtained using Gretl:

Model 1: OLS, using observations 1-597 ($n = 583$)
 Missing or incomplete observations dropped: 14
 Dependent variable: *price*

	<i>Coefficient</i>	<i>Standard. error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	1.208	0.026	46.82	0.000	***
<i>rivals</i>	-0.004	0.002	-2.740	0.006	***
<i>samebrand</i>	0.032	0.006	5.237	0.000	***
<i>l_distref</i>	0.009	0.006	1.523	0.128	
Mean dependent var	1.247	S.D. dependent var		0.045	
Sum squared resid	1.093	S.E. of regression		0.043	
<i>R</i> -squared	0.070	Adjusted <i>R</i> -squared		0.065	
<i>F</i> (3, 579)	14.531	<i>p</i> -value(<i>F</i>)		3.89e-09	
Log-likelihood	1003.226	Akaike criterion		-1998.452	
Schwarz criterion	-1980.979	Hannan-Quinn		-1991.641	

Based on these results, which of the following statements is correct?

- (a) Given an increase of 1% in the distance from the refinery/storage facility, it is estimated that the average price of fuel increases by 0.00893199 euros / litre, keeping the number of nearby rivals and the number of nearby same-brand petrol stations constant.
- (b) Given an increase of 1% in the distance from the refinery/storage facility, it is estimated that the average price of fuel increases by 0.0000893199 euros / litre, keeping the number of nearby rivals and the number of nearby same-brand petrol stations constant.
- (c) If we choose two petrol stations, A and B, each located at the same distance from the refinery / storage facility and each having the same number of nearby rivals, the model predicts that for each nearby same-brand petrol station the price will increase by 3%.
- (d) According to the determination coefficient, only 0.07% of the sample price variation is explained by nearby rivals, nearby same-brand petrol stations and distance from the refinery / storage facility.

2.10. Using a sample of cross-sectional data for 597 petrol stations ($i = 1, 2, \dots, 597$) in a city as of 15 January 2017, the table below presents the results of an estimated model that explains the price in euros per litre (*price*) set by petrol stations based on the number of nearby rivals (petrol stations operating under different brands (*rivals*) and the distance in km from the closest refinery/storage facility (*distref*):

Model 1: OLS, using observations 1-597 ($n = 583$)

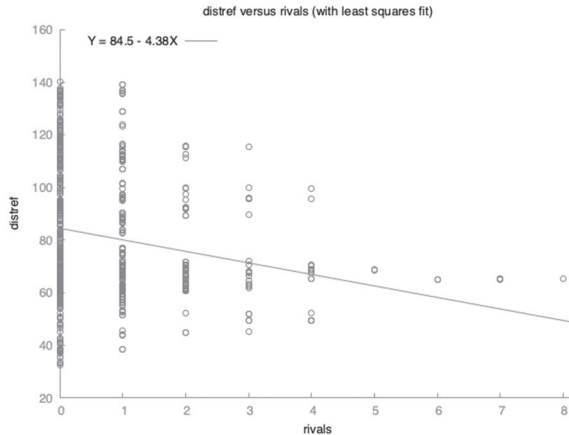
Missing or incomplete observations dropped: 14

Dependent variable: *price*

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	1.23679	0.00654304	189.0	<0.0001	***
<i>rivals</i>	-0.00408824	0.00155320	-2.632	0.0087	***
<i>samebrand</i>	0.0324026	0.00619066	5.234	<0.0001	***
<i>distref</i>	0.000127104	0.00007269	1.749	0.0809	*
Mean dependent var	1.246954	S.D. dependent var	0.044933		
Sum squared resid	1.091378	S.E. of regression	0.043416		
<i>R</i> -squared	0.071198	Adjusted <i>R</i> -squared	0.066386		
<i>F</i> (3, 579)	14.79461	<i>p</i> -value(<i>F</i>)	2.71e-09		
Log-likelihood	1003.596	Akaike criterion	-1999.193		
Schwarz criterion	-1981.720	Hannan-Quinn	-1992.382		

We know that the estimated model satisfies the Gauss-Markov assumptions, and all the explanatory variables are relevant. In addition, we have the following information on how the distance from the refinery/storage facility (*distref*) and the number of close rivals (*rivals*) are related in the sample:

Figure 14. Scatter plot of distance from refinery/storage (*distref*) vs number of nearby rivals (*rivals*)



With the available information and knowing that petrol stations tend to set higher prices the further they are from the refinery (since they incur higher transport costs), indicate the probable bias we would obtain from a simple linear regression of *price* on *rivals*, omitting *distref*.

- The bias will be zero.
- Upward bias.
- Downward bias.
- If the OLS estimator based on the multiple regression model is unbiased, when we omit a relevant variable the estimator will remain unbiased.

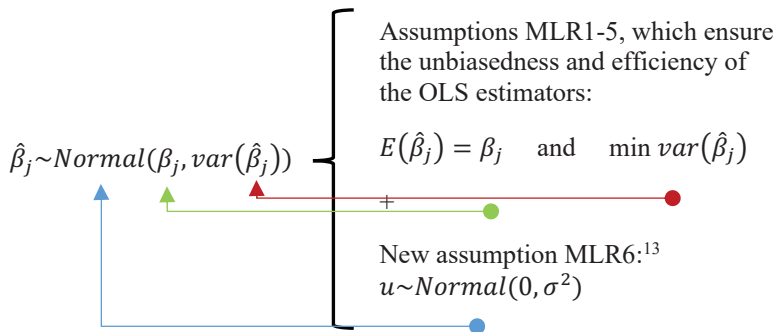
2.11. The omission of an explanatory variable from a model generates a bias in the OLS estimation of the slope of the regression line when:

- The omitted explanatory variable is correlated with the explanatory variable included in the model.
- The omitted explanatory variable is not correlated with the explanatory variable included in the model.
- The omitted explanatory variable is correlated with both the dependent variable and the explanatory variable included in the model.
- None of the above.

CHAPTER 3

STATISTICAL INFERENCE IN REGRESSION MODELS

We want to test hypotheses about a population using the SRF. To do that, besides knowing $var(\hat{\beta}_j)$ and $E[\hat{\beta}_j]$ of the OLS estimators, we need to know their sample distribution¹², which depends on the distribution of u_i .



¹² The sample distribution of the OLS estimator is the frequency distribution of the values $\hat{\beta}_j$ obtained from estimating a model for all possible random samples of a population. Once we know that sample distribution, we will be able to obtain, from a single sample, the probability that our estimate approximates the population parameter.

¹³ The reasoning is that u aggregates a large number of different unobservable factors. According to the central limit theorem, the distribution of the sum of a set of independent equally distributed random variables tends to be Gaussian as n increases.

We can perform simple hypothesis tests using the standardised OLS estimator, which is also known as the t -statistic:

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} \quad \text{where} \quad se(\hat{\beta}_j) = \sqrt{\frac{\hat{\sigma}^2}{TSS_j (1 - R_j^2)}}$$

$\hat{\sigma}^2$ is an unbiased estimation of $var(u_i|x_i)$

$$\frac{\sum_{i=1}^n \hat{u}_i^2}{n - k - 1}$$

The resulting statistic $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$ has a Student's t -distribution, which depends on sample size, n , and the number of model parameters other than the constant, k . Its magnitude indicates how many standard errors (s.e.) the point estimate $\hat{\beta}_j$ is from the hypothesised value $\beta_j = a_j$. Sampling error is taken into account through the s.e.

3.1. Simple hypothesis testing

To infer the relationship between the independent and dependent variables we use simple hypothesis testing. In what follows we describe the steps to be taken in the process.

1. There are three potential null (H_0) and alternative (H_1) hypotheses about the population parameters that can be used for inference:

a) Two-tailed test:

$H_0: \beta_j = a_j$ The effect of x_j on y is equal to a_j , once the effect of the other x 's has been controlled for.

$H_1: \beta_j \neq a_j$ The effect of x_j on y and is not equal to a_j , once the effect of the other x 's has been controlled for.

b) Right-tailed test:

$H_0: \beta_j \leq a_j$ The effect of x_j on y is equal or lower than a_j , once the effect of the other x 's has been controlled for.

$H_1: \beta_j > a_j$ The effect of x_j on y is greater than a_j , once the effect of the other x 's has been controlled for.

c) Left-tailed test:

- $H_0: \beta_j \geq a_j$ The effect of x_j on y is equal or higher than a_j , once the effect of the other x 's has been controlled for.
- $H_1: \beta_j < a_j$ The effect of x_j on y is less than a_j , once the effect of the other x 's has been controlled for.

The researcher has to decide between a), b) and c) depending on the empirical model and variables used.

2. The next step is to construct the t -statistic for $\hat{\beta}_j$: $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - a_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$

When $\hat{\beta}_j$ is “sufficiently” different from the hypothesised value a_j , taking the sampling error, $se(\hat{\beta}_j)$, into account, we will reject H_0 . What do we mean by “sufficiently”?

3. Once the t -statistic is constructed, we have to choose a significance level (α), the probability of committing a Type I error (rejecting H_0 when it is in fact true) we are willing to assume in the test. Generally, $\alpha = 0.1, 0.05$ or 0.01 . Since the 2000s given the increasing abundance of big datasets, some econometricians have proposed the use of even smaller significance levels, such as $\alpha = 0.05, 0.01$ or 0.001 .

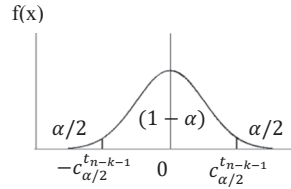
4. In order to decide whether to reject the null, $\hat{\beta}_j$ is “sufficiently” different from a_j , considering $se(\hat{\beta}_j)$, when $t_{\hat{\beta}_j}$ is more extreme than the critical value (c), which defines the $(1 - \alpha)$ percentile in the distribution t with $n - k - 1$ degrees of freedom.¹⁴

¹⁴ In Appendix A we present the critical values for a Student's t distribution, at different significance levels and with different degrees of freedom.

5. The decision rule corresponding to each hypothesis is

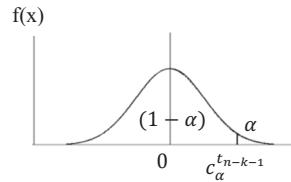
a) Two-tailed test:

We reject H_0 when $|t_{\hat{\beta}_j}| > c_{\alpha/2}^{t_{n-k-1}}$



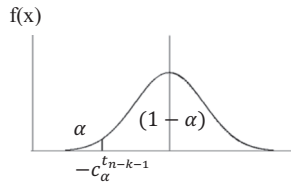
b) Right-tailed test:

We reject H_0 when $t_{\hat{\beta}_j} > c_{\alpha}^{t_{n-k-1}}$



c) Left-tailed test:

We reject H_0 when $t_{\hat{\beta}_j} < -c_{\alpha}^{t_{n-k-1}}$



6. We conclude by indicating whether the null hypothesis has been rejected or not at the corresponding level α at which the test has been performed.

Another possibility, instead of choosing a significance level, is to use a p -value. Considering the $t_{\hat{\beta}_j}$ statistic obtained, the p -value is the smallest significance level at which H_0 would be rejected. When the p -value $\leq \alpha \rightarrow$ we reject H_0 at a significance level α .

Alternatively, confidence intervals could be constructed for the population parameters and hypothesis testing could proceed by observing whether the population value is within the constructed intervals or whether it lies outside of the given lower and upper bounds.

Confidence intervals: Under assumptions MLR1–6, we can construct a confidence interval (CI) for the population parameter β_j :

$$\Pr\left(\hat{\beta}_j - c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j) \leq \beta_j \leq \hat{\beta}_j + c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)\right) = (1 - \alpha)$$

The lower bound, $\hat{\beta}_j - c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)$, and the upper bound, $\hat{\beta}_j + c_{\frac{\alpha}{2}}^{t_{n-k-1}} se(\hat{\beta}_j)$, contain the population value β_j in $100 \times (1 - \alpha)\%$ of all possible random samples. The CI thus contains all the values for which $H_0: \beta_j = a_j$ could not be rejected at one α (versus $H_1: \beta_j \neq a_j$). It is useful for performing *two-sided tests*.

3.2. Test of a linear combination of parameters

Given the model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$, we want to perform the following test:

- $H_0: \beta_1 - \beta_2 = a_j$ The difference in the effects of x_1 and x_2 on y is equal to a_j , once the effect of x_3 has been controlled for.
- $H_1: \beta_1 - \beta_2 \neq a_j$ The difference in the effects of x_1 and x_2 on y is not equal to a_j , once the effect of x_3 has been controlled for.

In cases like this, we cannot follow the same procedure as in a simple hypothesis test, since the outputs of the programs usually used in introductory econometrics courses do not provide all the information needed to construct the t -statistic for a linear combination of parameters:

$$t_{\hat{\beta}_1 + \hat{\beta}_2} = \frac{\hat{\beta}_1 - \hat{\beta}_2 - a_j}{se(\hat{\beta}_1 - \hat{\beta}_2)} \sim t_{n-k-1}$$

where

$$se(\hat{\beta}_1 - \hat{\beta}_2) = \sqrt{\widehat{Var}(\hat{\beta}_1 - \hat{\beta}_2)} = \sqrt{\widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_2) - 2\widehat{Cov}(\hat{\beta}_1, \hat{\beta}_2)}$$

To solve this problem, we therefore recommend redefining the model and proceeding as follows:

1. We redefine the linear combination of parameters, $\beta_1 - \beta_2 = \delta_1$, and reformulate the test accordingly:

$H_0: \delta_1 = a_j$ The difference in the effects of x_1 and x_2 on y (δ_1) is equal to a_j , once the effect of x_3 has been controlled for.

$H_1: \delta_1 \neq a_j$ The difference in the effects of x_1 and x_2 on y (δ_1) is not equal to a_j , once the effect of x_3 has been controlled for.

2. Given that $\beta_1 - \beta_2 = \delta_1$, we substitute one of the original parameters in the model. For example, we substitute $\beta_1 = \delta_1 + \beta_2$ and rearrange the expression:

$$\begin{aligned} y &= \beta_0 + (\delta_1 + \beta_2)x_1 + \beta_2x_2 + \beta_3x_3 + u \\ y &= \beta_0 + \delta_1x_1 + \beta_2(x_1 + x_2) + \beta_3x_3 + u \end{aligned}$$

3. We estimate the SRF of the reparameterised model presented in point 2 above. by OLS:

$$\hat{y} = \hat{\beta}_0 + \hat{\delta}_1x_1 + \hat{\beta}_2(x_1 + x_2) + \hat{\beta}_3x_3$$

$$\begin{array}{cccc} se(\hat{\beta}_0) & se(\hat{\delta}_1) & se(\hat{\beta}_2) & se(\hat{\beta}_3) \\ n & & R^2 & \end{array}$$

4. With the SRF of the reparameterised model, we now construct the t statistic for $\hat{\delta}_1$, which will allow us to test the proposed linear combination of parameters, given that $\beta_1 - \beta_2 = \delta_1$:

$$t_{\hat{\delta}_1} = \frac{\hat{\delta}_1 - a_j}{se(\hat{\delta}_1)} \sim t_{n-k-1}$$

5. We proceed as in the case of simple hypothesis testing and follow the abovementioned steps 3-6.

3.3. Multiple hypothesis testing

With the following MLR model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$, we want to test multiple hypotheses about the parameters, such as:

$H_0: \beta_1 = 0, \beta_2 = 0.$ x_1 and x_2 have no jointly significant effect on y , once the effect of x_3 has been controlled for.

$H_1: H_0$ is not true. x_1 and x_2 have a jointly significant effect on y , once the effect of x_3 has been controlled for.

Unrestricted model (NR): $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$

Model restricted (R) by H_0 : $y = \beta_0 + \beta_3 x_3 + \epsilon$
(model true only if H_0 is true)

If H_0 is not true, switching from the NR model to the R model will make the fit of the regression worse: $SSR_{nr} < SSR_r$.¹⁵ In this case, therefore, the inference is based on the rate of variation in the SSRs on switching from an NR model to an R model, adjusted for their respective degrees of freedom (df):¹⁶

$$F = \frac{(SSR_R - SSR_{NR})/q}{SSR_{NR}/(n - k - 1)} \sim F_{q, n-k-1}$$

Under H_0 , the F statistic has a Snedecor's F distribution with q and $n - k - 1$ df.¹⁷ Why? Under assumptions MLR1–6:

¹⁵ $SSR_{nr} = \sum_i^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \hat{u}_i^2$ and $SSR_r = \sum_i^n (y_i - \tilde{y}_i)^2 = \sum_{i=1}^n \tilde{\epsilon}_i^2$.

¹⁶ Difference of degrees of freedom between R and NR models = $(n - (k - q) - 1 - n + k + 1) = q$.

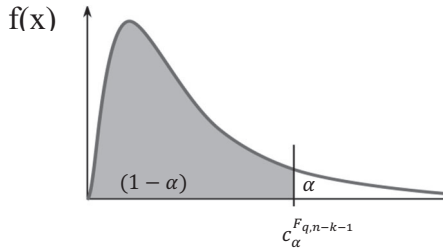
¹⁷ This F statistic can likewise be used, by comparing the unrestricted model and the restricted model, to test linear constraints such as those indicated in the previous section resulting from a linear combination of parameters. For example, assuming the null hypothesis $H_0: \beta_1 - \beta_2 = a_j$ is true, the restricted version of model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$ would be $y = \beta_0 + (a_j + \beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \rightarrow (y - a_j) = \beta_0 + x_1 + \beta_2(x_2 + x_1) + \beta_3 x_3 + \epsilon$.

$$F = \frac{(\sum_{i=1}^n \tilde{\epsilon}^2 - \sum_{i=1}^n \hat{u}_i^2)/q}{\sum_{i=1}^n \hat{u}_i^2 / (n - k - 1)} \sim \frac{\mathcal{X}_q^2/q}{\frac{\mathcal{X}_{n-k-1}^2}{n-k-1}} \sim F_{q, n-k-1}$$

where \mathcal{X}_g^2 are independent chi-square distributions with g degrees of freedom.

When is F “sufficiently” large to reject H_0 ?

- We choose the significance level α as indicated above
- We reject H_0 when the F statistic is more extreme than the critical value (c), which marks the $(1 - \alpha)$ percentile of an F distribution with q and $(n - k - 1)$ degrees of freedom in the numerator and denominator, respectively.¹⁸



Note that, where the unrestricted and restricted models have the same dependent variable, the F statistic can also be expressed in terms of the coefficients of determination, R^2 , of each model:

$$R^2 = 1 - \frac{SSR}{TSS}; \quad SSR = (1 - R^2)TSS$$

$$F = \frac{(SSR_R - SSR_{NR})/q}{SSR_{NR}/(n - k - 1)} = \frac{(R_{NR}^2 - R_R^2)/q}{(1 - R_{NR}^2)/(n - k - 1)}$$

¹⁸ In Appendix A, we present the critical values for the Snedecor’s F distribution, according to the significance level and the degrees of freedom of the numerator and denominator of the statistic.

Note: Certain restrictions (e.g. $H_0: \beta_1 = 1, \beta_2 = 0$) may alter the dependent variable of the restricted model, thus making it impossible to use this latter expression.

Summarising, hypothesis testing can be done either using the t-test or by building confidence intervals. Analytically, we can use the previously described procedure following the abovementioned six steps for simple hypothesis testing or construct confidence intervals. In both cases it is always useful to use a graphical representation of the rejection and non-rejection regions, which helps to visually identify how precise or imprecise are the estimated parameters are.

To illustrate statistical inference in regression analysis, let us now use an example. In a study of ice cream consumption in a set of regions, $i = 1, 2, \dots, 200$, the following estimated model is obtained by OLS:

$$\hat{D}_i = 300,28 - 0,74 P_i + 8,04 T_i$$

$$(78,31) \quad (0,05) \quad (2,98)$$

$$\sum_{i=1}^n (D_i - \bar{D})^2 = 83021 \quad \sum_{i=1}^n (\hat{D}_i - \bar{\hat{D}})^2 = 80814$$

where:

D_i is the number of ice creams consumed per capita in each region,

P_i is the price of an ice cream (in euros) in each region,

T_i denotes the average temperature (in degrees Celsius) in each region.

The standard errors of each estimated parameter is are presented in parentheses.

1. Using the regressions results, we are asked to construct a confidence interval for the parameter estimated for average temperature (at the 10 percent significance level) and determine whether the variable is statistically significant. The answer should include (i) the interpretation of the estimated parameter, (ii) the null and alternative hypotheses, (iii) the calculated interval, and (iv) the decision and the rejection rule.

(i) Interpretation: An increase of 1° C is associated with an additional 8 ice creams consumed per person.

(ii) $H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

(iii) The lower bound is given by: $8.04 - \frac{c_{\frac{10}{2}}^{t_{200-2-1}}}{2} 2.98$, and the upper bound by: $8.04 + \frac{c_{\frac{10}{2}}^{t_{200-2-1}}}{2} 2.98$. Searching in the corresponding statistical tables for the Student's t with degrees of freedom (d.f.)=197, we find the critical value (c)=1.64. Therefore, the resulting interval is given by: [3.15, 12.93].

(iv) We reject the null hypothesis, given that the value zero is not within the 90% confidence interval.

2. Next, we show how to present and perform a joint significance test of the estimated model. We have to include (i) the null and alternative hypotheses, (ii) the calculated statistic needed for the test and (iii) the decision and the rejection rule.

$H_0: \beta_1 = 0, \beta_2 = 0.$ x_1 and x_2 have no jointly significant effect on y .

$H_1: H_0$ is not true. x_1 and x_2 have a jointly significant effect on y .

Considering that $R^2 = \frac{\sum_{i=1}^n (\hat{D}_i - \bar{\bar{D}})^2}{\sum_{i=1}^n (D_i - \bar{D})^2} = 80814/83021 = 0.9733$; then we can obtain the F statistic as follows: $F = \frac{(R^2)/q}{(1-R^2)/(n-k-1)} = (0.9733/2)/((1-0.9733)/197) = 3606.79$. Therefore, we can reject the null hypothesis at the 5% level of significance, since the F statistic is more extreme than the critical value (c), which marks the $(1 - \alpha)$ percentile of an F distribution with q and $(n - k - 1)$ degrees of freedom in the numerator and denominator, respectively. In this case, $F = 3606,79 > c_{\alpha=0.05}^{F_{2,200-2-1}} = 3.07$.

3. Next, the researcher would like to test whether the effect of a 1 euro increase in price of has the same effect on ice cream consumption as a 1°C decrease in temperature. (i) Indicate what would be the null and alternative hypotheses and (ii) how to proceed to perform the test.

(i)

$H_0: \beta_1 + \beta_2 = 0$ The difference in the effects of x_1 and x_2 on y is equal to 0,

$H_1: \beta_1 + \beta_2 \neq 0$ The difference in the effects of x_1 and x_2 on y is not equal to 0.

(ii) We redefine the linear combination of parameters, $\beta_1 + \beta_2 = \delta_1$, and reformulate the test accordingly:

$H_0: \delta_1 = 0$ The difference in the effects of x_1 and x_2 on y (δ_1) is equal to 0,
 $H_1: \delta_1 \neq 0$ The difference in the effects of x_1 and x_2 on y (δ_1) is not equal to 0.

Given that $\beta_1 + \beta_2 = \delta_1$, we substitute one of the original parameters in the model. For example, we substitute $\beta_1 = \delta_1 - \beta_2$ and rearrange the expression:

$$y = \beta_0 + (\delta_1 - \beta_2)x_1 + \beta_2x_2 + u$$

$$y = \beta_0 + \delta_1x_1 + \beta_2(x_2 - x_1) + u$$

We estimate the reparameterised model presented by OLS as indicated above and with the SRF of the reparameterised model construct the t statistic for $\hat{\delta}_1$, which will allow us to test the proposed linear combination of parameters, given that $\beta_1 + \beta_2 = \delta_1$:

$$t_{\hat{\delta}_1} = \frac{\hat{\delta}_1 - 0}{se(\hat{\delta}_1)} \sim t_{n-k-1}$$

We proceed as in the case of simple hypothesis testing.

4. Knowing that warmer regions have higher prices, would the simple regression of D_i on P_i (excluding T_i) produce an estimator of D_i with an upward or downward bias? Why?

Original specification: $D_i = \beta_0 + \beta_1P_i + \beta_2T_i + u_i$

Subspecified model: $D_i = \beta_0 + \beta_1P_i + \epsilon_i$, where the corresponding SRF would be $\tilde{D}_i = \tilde{\beta}_0 + \tilde{\beta}_1P_i$.

$\beta_2 = \frac{\partial D_i}{\partial T_i} > 0$ (given that the hotter the region, the higher the demand)

Assuming that higher temperatures can put upward pressure on demand and price: $\text{Corr}(T, P) > 0$. Therefore,

the single regression model will produce an upward bias in the estimated coefficient of P_i . A different correlation between T and P could be argued, but if the explanation is adequate, the argument will be correct.

Problem set 3A (with solutions)

EXERCISE 3A.1. Consider the following model: $y_i = \beta_0 + \beta_1 x_{i1} + u_i$, where y_i is the mark obtained by a group of students $i = 1, 2, \dots, n$ in the course Foundations of Econometrics course and x_{i1} is the overall average mark obtained by the same group for the bachelor's degree (excluding Foundations of Econometrics). The results obtained using a sample of 100 students from the file "Data_marks.xlsx" are as follows:

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = 467.558$$

$$TSS_1 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 = 41.594$$

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 145.345 \quad \bar{x} = 6.417$$

$$SSR = \sum_{i=1}^n \hat{u}_i^2 = 322.213 \quad \bar{y} = 5.310$$

$$\sum_i^n (x_{i1} - \bar{x}_1)(y_i - \bar{y}) = \sum_i^n (x_{i1} - \bar{x}_1)y_i = 77.752$$

a) Using the available information, obtain the parameters of the proposed model by OLS. Interpret the estimated slope coefficient.

b) Does the overall average mark influence the mark obtained in Foundations of Econometrics? Specify and perform the test at a 5% significance level. (Note: you will need to calculate the standard errors (s.e.) associated with the estimated parameters using the available information).

c) Consider a different model in which the hours each student has spent studying for the econometrics exam in the subject (x_{i2}) and the number of times they have previously taken the exam (x_{i3}) have been included as additional regressors. Using the sample of 100 students, the following information is known:

$$\hat{y}_i = -1.597 + 0.382x_{i1} + 0.042x_{i2} + 0.142x_{i3}$$

$$\begin{matrix} (0.698) & (&) & (&) & (&) \\ n = 100 & & & R^2 = 0.9206 \end{matrix}$$

$$TSS_1 = \sum_i^n (x_{i1} - \bar{x}_1)^2 = 41.594 \qquad R_1^2 = 0.2598$$

$$TSS_2 = \sum_i^n (x_{i2} - \bar{x}_2)^2 = 211,091.508 \qquad R_2^2 = 0.2486$$

$$TSS_3 = \sum_i^n (x_{i3} - \bar{x}_3)^2 = 21.310 \qquad R_3^2 = 0.0269$$

$$SSR = \sum_{i=1}^n \hat{u}_i^2 = 37.117$$

Using the available information, obtain the missing s.e., then specify and perform an individual significance test (at the 5% level of significance) on the partial effect of study hours (x_{i2}) on the econometrics mark (y_i).

SOLUTION 3A.1:

a) Using OLS, the SRF is $\hat{y}_i = -6.685 + 1.869 x_i$.

$$\hat{\beta}_1 = \frac{\sum_i^n (x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum_i^n (x_{i1} - \bar{x}_1)^2} = \frac{77.752}{41.594} = 1.869 \qquad \text{and}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 5.310 - 1.869 \cdot 6.417 = -6.685$$

According to the estimates' constant $\hat{\beta}_1$, a one-point increase in the average mark for the bachelor's degree is estimated to generate an increase of 1.869 points in the mark obtained in the Foundations of Econometrics course.

b) The standard error associated with the estimated slope parameter $\hat{\beta}_1$ is calculated by using the following expression:

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{TSS_1(1-R_1^2)}} = \sqrt{\frac{\frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}}{TSS_1(1-R_1^2)}} = \sqrt{\frac{322.213}{41.594(1-0)}} = 0.281$$

We can proceed specifying and performing the hypothesis test as follows:

$$H_0: \beta_1 = 0$$

(The overall average mark does not influence the Econometrics' mark)

$$H_1: \beta_1 \neq 0$$

(The overall averaged mark influences on the Econometrics' mark)

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{1.869}{0.281} = 6.648$$

$$c_{5\%}^{t_{n-k-1}} = c_{5\%}^{t_{100-1-1}} = 1.99$$

Given that $|t_{\hat{\beta}_1}| > \left| c_{\frac{5\%}{2}}^{t_{n-k-1}} \right|$, we can reject the null hypothesis ($H_0: \beta_1 = 0$) at the 5% level of significance. Therefore, we can conclude that the overall average mark in the bachelor's degree significantly influences the Econometrics' mark.

c)

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}}{TSS_1(1-R_1^2)}} = \sqrt{\frac{\frac{37.117}{100-1-1}}{41.594(1-0.2598)}} = 0.111$$

$$s.e.(\hat{\beta}_2) = \sqrt{\frac{\frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}}{TSS_2(1-R_2^2)}} = \sqrt{\frac{\frac{37.117}{100-1-1}}{211,091.508(1-0.2486)}} = 0.002$$

$$s.e.(\hat{\beta}_3) = \sqrt{\frac{\frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1}}{TSS_3(1-R_3^2)}} = \sqrt{\frac{\frac{37.117}{100-1-1}}{21.310(1-0.0269)}} = 0.135$$

We can specify and perform the hypothesis test as follows:

$H_0: \beta_2 = 0$ (The number of study hours does not influence on the Econometrics' mark)

$H_1: \beta_2 \neq 0$ (The number of study hours influences on the Econometrics' mark)

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{0.042}{0.002} = 21$$

$$c_{\frac{5\%}{2}}^{t_{n-k-1}} = c_{\frac{5\%}{2}}^{t_{100-2-1}} = 1.99$$

Given that $|t_{\hat{\beta}_2}| > \left| c_{\frac{5\%}{2}}^{t_{n-k-1}} \right|$, we can reject the null hypothesis ($H_0: \beta_2 = 0$) at the 5% level of significance. Therefore, we can conclude that, after controlling for the influence of the overall averaged mark in the bachelor's degree and the number of times students have previously taken the exam, the number of study hours significantly influences on the Econometrics' mark.

EXERCISE 3A.2 Consider the following multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$$

where y_i represents investment in R&D (expressed in million euros), x_{i1} is the number of employees, and x_{i2} is sales (in million euros), for a set of companies $i = 1, 2, \dots, n$.

Using a sample of 25 companies, the following results have been obtained from a regression analysis using OLS. With the available information, calculate the missing values [A] and [B]. Is x_{i2} statistically significant at the 5% level? Please show the calculations and justify your answers.

	Dependent variable: y		
	Estimated coeff.	s.e.	
constant	23.015	7.318	
x_1	0.397	0.292	
x_2	1.859	[A]	
Sample size (n)	25		
R^2	[B]		
Adjusted R^2	0.295		

$SSR = \sum_{i=1}^n \hat{u}_i^2 = 3334$	
$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 1822$	$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k-1} = \frac{3334}{25-2-1} = 152$
$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = 5156$	
$TSS_1 = \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 = 1787$	$R_1^2 = 0.0036$
$TSS_2 = \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2 = 423$	$R_2^2 = 0.0036$

SOLUTION 3A.2:

$$[A] = s. e. (\hat{\beta}_2) = \sqrt{\frac{\hat{\sigma}^2}{TSS_2(1-R_2^2)}} = \sqrt{\frac{\frac{\sum_{t=1}^n \hat{u}_t^2}{n-k-1}}{TSS_2(1-R_2^2)}} = \sqrt{\frac{152}{423(1-0.0036)}} = 0.601$$

$$[B] = R^2 = \frac{ESS}{TSS} = \frac{1822}{5156} = 0.353$$

$$H_0: \beta_2 = 0$$

(the number of employees does not affect the investment in R&D)

$$H_0: \beta_2 \neq 0$$

(the number of employees affects the investment in R&D)

$$t_{\hat{\beta}_2} = \frac{\hat{\beta}_2 - 0}{s.e.(\hat{\beta}_2)} = \frac{1.859}{0.601} = 3.093$$

$$c_{\frac{5\%}{2}}^{t_{n-k-1}} = c_{\frac{10\%}{2}}^{t_{25-1-1}} = 2.07$$

Given that $|t_{\hat{\beta}_2}| > \left| c_{\frac{5\%}{2}}^{t_{n-k-1}} \right|$, we can reject the null hypothesis ($H_0: \beta_2 = 0$) at the 5% level of significance.

EXERCISE 3A.3 The following model has been proposed to assess whether a group of countries have converged or diverged in per capita income. In the latter case, the differences in per capita income would have increased. The period of analysis is from 1980 to 2016.

$$meangr_i = \beta_0 + \beta_1 \ln(GDP_cap_{i1980}) + u_i$$

where:

- $meangr_i$ is the average over time of the annual per capita GDP growth rates (in US dollars) in each country i from 1980 to 2016, defined as $\frac{1}{T-1} \sum_{t=1980}^T (\ln(GDP_cap_{it}) - \ln(GDP_cap_{it-1}))$
- $\ln(GDP_cap_{i1980})$ is the logarithm of the initial level of per capita GDP of each country i

When the poorer countries grow at higher rates than the richer countries, in the long run all the countries gradually tend to the same level of per capita income. This tendency to converge therefore manifests itself when there is

an inverse relationship between the average over time of the annual per capita GDP growth rates and the initial level of GDP, $\beta_1 < 0$.

The data set “Data_convergence.gdt” (source: World Bank) contains information on the per capita GDP of 141 countries from 1980 to 2016, expressed in constant dollars. These data have been used to obtain the average over time of the countries’ annual per capita GDP growth rates between 1980 and 2016 ($meangr_i$) and the corresponding level of initial GDP per capita in 1980, expressed in logarithms ($\ln(GDP_cap_{i1980})$). Using that data set, answer the following questions:

- a) Compare the descriptive statistics of per capita GDP for 1980 and 2016.
- b) Use OLS to estimate the equation of the model proposed in the problem statement. Report the results in the usual form and interpret the estimated parameter associated with $\ln(GDP_cap_{i1980})$.
- c) Test the null hypothesis $H_0: \beta_1 \geq 0$ against the alternative of convergence $H_1: \beta_1 < 0$. Perform the test at a 5% significance level.

SOLUTION 3A.3:

- a) In Gretl: Summary statistics.

Table 13. Descriptive statistics for GDP per capita in “Data_convergence.gdt”

	gdp_cap1980	gdp_cap2016
Mean	10496	17054
Median	3302.7	6681.3
Minimum	190.48	219.21
Maximum	113.68	191590
Std. Dev.	17217	25262
C. V.	1.6404	1.481
Coefficient of assymmetry	3.39	3.244
Kurtosis excess	15.074	15.979

b) In Gretl: Model / Ordinary least squares:

$$\widehat{meangr}_i = 0.031 - 0.002 \ln(GDP_{cap_{i1980}})$$

$$(0.008) \quad (0.001)$$

$$n = 141 \quad R^2 = 0.032 \quad SSR = 0.040$$

According to the SRF, a 1% increase in GDP per capita in 1980 gives rise to a 0.2% decrease in the average over time of annual growth rates between 1980 and 2016.

c)

$$H_0: \beta_1 \geq 0$$

$$H_1: \beta_1 < 0$$

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{-0.002}{0.001} = -2$$

$$-c_{5\%}^{t_{n-k-1}} = -c_{5\%}^{t_{141-1-1}} = -1.64$$

Given that $t_{\hat{\beta}_1} < -c_{5\%}^{t_{141-1-1}}$, we can reject the null hypothesis $H_0: \beta_1 \geq 0$ vs $H_1: \beta_1 < 0$ at the 5% level of significance.

EXERCISE 3A.4 The following model aims to explain the number of children people have (*ceb*) using four explanatory variables: monthly earnings in US dollars (*sal*), age (*age*), years of education or training (*ed*) and years of being a smoker (*sm*):

$$ceb = \beta_0 + \beta_1 \log(sal) + \beta_2(age) + \beta_3 \log(ed) + \beta_3 sm + u$$

a) Based on the specified model, state the null hypothesis that years of being a smoker has no effect on number of children, after controlling for earnings, age and years of education. State the alternative that years of being a smoker has a negative effect on the number of children.

b) Using a sample of 462 individuals living in Colombia in 2009 taken from the Latin American Migration Project, the following regression output was obtained by applying OLS:

$$\widehat{ceb} = -5.506 + 0.005 \log(sal) + 2.335 \log(age) - 0.459 \log(ed) - 0.005 sm$$

$$\begin{matrix} (0.982) & (0.067) & (0.236) & (0.127) \\ & & & (0.004) \end{matrix}$$

$$n = 462 \quad R^2 = 0.249$$

What is the estimated difference in number of children between a person who has never smoked and a person who has smoked for 60 years, when holding the variables earnings, age and years of education constant?

c) Carry out the test proposed in section (a) at the 10% significance level.

d) Based on the results, would you include *sm* in the final model explaining number of children? Give reasons.

SOLUTION 3A.4:

a)

$$H_0: \beta_3 \geq 0$$

$$H_1: \beta_3 < 0$$

b)

$$\frac{\widehat{ceb}_A - \widehat{ceb}_B}{\widehat{ceb}_A - \widehat{ceb}_B} = \frac{-5.506 + 0.005 \log(sal_A) + 2.335 \log(age_A) - 0.459 \log(ed_A) - 0.005 sm_A - [-5.506 + 0.005 \log(sal_B) + 2.335 \log(age_B) - 0.459 \log(ed_B) - 0.005 sm_B]}{-0.005(sm_A - sm_B)} = \frac{-0.005(0 - 60)}{-0.005(0 - 60)} = 0.3 \text{ years}$$

c)

$$t_{\widehat{\beta}_2} = \frac{\widehat{\beta}_3 - 0}{s.e.(\widehat{\beta}_3)} = \frac{-0.005}{0.004} = 1.25 \quad c_{10\%}^{t_{n-k-1}} = c_{10\%}^{t_{462-4-1}} = 1.28$$

Given that $t_{\widehat{\beta}_2} < c_{10\%}^{t_{n-k-1}}$, we cannot reject the null hypothesis ($H_0: \beta_2 \geq 0$) at the 10% level of significance.

d) The partial effect of *sm* on *ceb* is not statistically significant at the 10% level of significance, so we conclude that this regressor is irrelevant. The

inclusion of an irrelevant regressor does not violate any regression assumption, as long as the regressor in question is not perfectly collinear with the other regressors in the model.

Problem set 3B

EXERCISE 3B.1 An entrepreneur wants to study the relationship between the production (Y) of a group of subsidiaries she owns and the factors of production used in the production process: capital (K) and labour (L).

a) Based on the Cobb-Douglas production function $Y = AK^{\beta_1}L^{\beta_2}$, specify an econometric model that is linear in its parameters and hence can be estimated by OLS.

b) The data set “Data_production.xlsx” contains information on the number of goods produced in 2018 for a set of 100 subsidiaries, in which the number of workers and the capital goods (equipment and machinery) each subsidiary has used in the year’s production are known. Using the available information, estimate the parameters of the proposed model by OLS.

c) Test the null hypothesis that there are constant or increasing returns to scale, that is, $H_0: \beta_1 + \beta_2 \geq 1$, against the alternative that there are decreasing returns to scale, $H_1: \beta_1 + \beta_2 < 1$. Explain the significance of the conclusion reached.

EXERCISE 3B.2. Consider a model that explains the mark obtained in Econometrics in a Spanish university (*nota_econometria*) with the following independent variables: the average mark at the university (*nota_media*), the number of times the exam was taken (*convocatoria*) and the number of tutoring sessions each student has attended (*tutorias*). Below is the SRF obtained by OLS for a sample of 187 students (data set: Data_marks2):

$$\widehat{\text{nota_econometrica}} = -0.711 + 0.988 \text{ nota_media} \\ (1.222) \quad (0.180)$$

$$-0.226 \text{ convocatoria} + 0.481 \text{ tutorias} \\ (0.256) \quad (0.151)$$

$$n = 187 \quad R^2 = 0.207$$

- a) Calculate a 95% confidence interval for the parameter associated with the number of tutoring sessions.
- b) Using the estimated confidence interval, test whether the number of tutoring sessions attended influences the mark obtained in Econometrics at a confidence level of 95%.
- c) Test whether $\beta_{tutorias} = 0.5$ at a confidence level of 95%.

EXERCISE 3B.3. Consider the following equation, which relates the rental price of homes to the number of rooms (*dormitorios*), the number of bathrooms (*banos*), the size of the home in square metres (*m2*), the distance from the city centre (*dist_centro*), and the number of Airbnb homes nearby (*n_airbnb*):¹⁹

$$\log(\text{precio}) = \beta_0 + \beta_1 \text{dormitorios} + \beta_2 \text{banos} + \beta_3 m2 + \beta_4 \text{dist_centro} + \beta_5 n_airbnb + u$$

The following SRF was obtained using the data set “Data_Palma_Mallorca_alquileres.gdt”, which contains information taken from the Nestoria property search website (<https://www.nestoria.com/>) and Inside Airbnb (<http://insideairbnb.com/>) for a sample of rental homes in Palma de Mallorca on 27 August 2019:

$$\begin{aligned} \log(\widehat{\text{precio}}) = & 6.489 - 0.015 \text{ dormitorios} + 0.175 \text{ banos} + 0.003 m2 \\ & (0.078) \quad (0.019) \qquad \qquad (0.031) \qquad \qquad (0.0004) \\ & + 0.030 \text{ dist_centro} + 0.001 n_airbnb \\ & (0.023) \qquad \qquad \qquad (0.0006) \end{aligned}$$

$$n = 348 \qquad SCE = 25.763 \qquad R^2 = 0.424$$

- a) Specify and carry out a test to determine whether, *ceteris paribus*, the number of nearby Airbnb homes is responsible for the increase in rental prices.
- b) Now specify and carry out a test to determine whether a home’s internal characteristics (*dormitorios*, *banos* and *m2*) are jointly significant. For this purpose, we know that $\sum_i^n (\log(\text{precio}_i) - \hat{\beta}_0 - \hat{\beta}_4 \text{dist_centro}_i - \hat{\beta}_5 n_airbnb_i)^2 = 45.008$.

¹⁹ Nearby is defined as within a radius of 500 metres around each home.

EXERCISE 3B.4. The following model describes the price of diesel fuel in euros per litre (p_goa) set by a city's petrol stations based on the number of nearby petrol stations operating under rival brands (*rivals*), the number of nearby petrol stations operating under the same brand (*samebrand*) and the distance in km from the nearest refinery-storage facility (*distref*):²⁰

$$p_goa = \beta_0 + \beta_1 rivals + \beta_2 samebrand + \beta_3 \log(distref) + u$$

Using a cross-sectional data set for 597 petrol stations in Valencia ($i = 1, 2, \dots, 597$), downloaded on 15 February 2017 from the Ministry of Energy, Tourism and Digital Agenda website (<http://geoportalgasolineras.es/>), the following table of results was obtained:

Model 1: OLS, using observations 1-597					
Dependent variable: p_goa					
	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
const	1.075	0.025	43.46	0.000	***
<i>rivals</i>	-0.005	0.001	-3.381	0.001	***
<i>samebrand</i>	0.032	0.006	5.349	0.000	***
$\log(distref)$	0.012	0.0060	2.094	0.037	**
Mean of the dep. var.	1.125	S.D. of the dep. var.	0.044		
Sum of sq. residuals	1.046	S.D. of the regression	0.042		
<i>R</i> -squared	0.082	Adjusted <i>R</i> -squared	0.078		
$F(3, 593)$	17.709	<i>p</i> -value (of F)	5.08e-11		
Log-likelihood	1047.347	Akaike criterion	-2086.694		
Schwarz criterion	-2069.126	Hannan-Quinn criterion	-2079.853		

- Interpret the estimated coefficient associated with $\log(distref)$.
- Test the hypothesis that the price for diesel fuel set by the petrol stations does not change with distance from the refinery-storage facility, against the alternative that it increases. Carry out the test at the 1%, 5% and 10% significance levels.

²⁰ Nearness has been defined by drawing a radius of 500 metres around each petrol station.

c) A report by the National Commission on Markets and Competition (CNMC) states that the price of diesel fuel increases by 0.05 euros/litre for each additional nearby same-brand petrol station. Based on the results of the estimated model, construct a 95% confidence interval for the price increase per additional nearby same-brand petrol station and use it to test the statement in the CNMC report.

d) State the null hypothesis that the effect on the fuel price of the addition of one more nearby rival petrol station is offset by the effect of the addition of one nearby same-brand petrol station. Explain why the results given in the problem statement cannot be used to test the proposed hypothesis. Specify a model that directly provides the t -statistic for testing this hypothesis and explain how you would carry out the test.

e) Specify and carry out a joint significance test of the previous regression at a 1% significance level.

EXERCISE 3B.5. Using a sample of houses for rent in Madrid, the following models have been estimated:

Dependent variable: price		

	Model A	Model B

Constant	684.2*** (88.778)	77.65 (252.157)
size	14.31*** (1.039)	14.49*** (1.040)
size_sq	-0.00856*** (0.001)	-0.00837*** (0.001)
baths	399.7*** (54.038)	412.4*** (54.220)
distc	-156.9*** (24.868)	-152.3*** (24.901)

l_rooms	-569.3*** (79.934)	-588.8*** (80.210)
$D_elevator$		574.0** (224.013)
$D_terrace$		37.35 (95.504)

N	1993	1993
R ²	0.389	0.391
Adjusted R ²	0.387	0.389

s.e. between parentheses.

The dependent variable in the two models, price, is the monthly rental price in euros. Regarding the explanatory variables, *size* is the size of the house in square metres; *size_sq* is the square of size; *baths* is the number of bathrooms; *distc* is the distance in kilometres of each home from the city center; *l_rooms* is the logarithm of the number of rooms in each house; *D_elevator* is a dummy variable that takes the value 1 if the building has an elevator and 0 otherwise; and *D_terrace* is a dummy variable that takes the value 1 if the house has a terrace, 0 otherwise. With the information provided, answer the following questions:

- According to the regression output of model A, interpret the parameter estimated by OLS associated with the explanatory variable *l_rooms*.
- From the results of model A, calculate a 95% confidence interval for the parameter β_{baths} and use it to test whether an additional bathroom raises the rental price by 300 euros. Your answer should include (i) null and alternative hypotheses, (ii) the calculated interval, and (iii) the decision and the rejection rule.
- Using model B, calculate the estimated difference in rental prices between (I) homes that have both an elevator and a terrace, and (II) homes that do not have an elevator or a terrace, *ceteris paribus*. Is this difference statistically significant at 5%? Your answer should include (i) the estimated difference in rental prices between profiles I and II, (ii) the null and alternative hypothesis, (iii) the calculated statistic needed for the test, and (iv) the decision and the rejection rule.

EXERCISE 3B.6. Imagine that we are interested in estimating the following model with a data sample of second-hand vehicles for sale:

$$price = \beta_0 + \beta_1 age + \beta_2 km + u$$

where *price* is the sale price expressed in euros, *age* is the age in years, and *km* is the mileage in kilometres of each vehicle $i=1, 2, \dots, n$.

Considering the sampled data presented in the following table, answer the questions a) to c) below, showing and explaining the calculations in each case.

Table 14. Dataset of second-hand vehicles for sale

<i>i</i>	<i>price</i>	<i>age</i>	<i>km</i>
1	17900	2	9000
2	6500	10	121000
3	21500	2	24943
4	11490	7	35000
5	18800	4	135500
6	15400	3	15257
7	22300	4	38200
8	13900	1	28000
9	16495	3	2907
10	21000	5	94000

Source: Random drawn from <https://www.autocasion.com/> (Spain), extracted the 25 of August 2021.

- Use OLS to obtain the estimated parameters.
- Are the variables *age* and *km* individually significant?
- Test the null hypothesis $H_0: \beta_1 = \beta_2$ versus the alternative hypothesis $H_1: \beta_1 \neq \beta_2$.

Multiple-choice questions (Topic 3)

3.1. The Phillips curve theory postulates an inverse relationship between inflation (π) and the unemployment rate (ur):

$$\pi = \beta_0 + \beta_1 ur + u$$

Which of the following would be the most appropriate null and alternative hypotheses in this case to carry out a significance test of β_1 ?

- a) $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0.$
- b) $H_0: \beta_1 \leq 0, H_1: \beta_1 > 0.$
- c) $H_0: \beta_1 \geq 0, H_1: \beta_1 < 0.$
- d) Either of the above.

3.2. Consider a model in which the dependent variable is CO2 emissions per capita ($CO2c$) and the explanatory variables are income per capita ($GDPc$) and income per capita squared ($GDPc^2$):

$$CO2c = \beta_0 + \beta_1 GDPc + \beta_3 GDPc^2 + u$$

Which of the following is the most appropriate way to test for the existence of an Environmental Kuznets Curve (i.e., an inverse U-shaped relationship between emissions and income per capita)?

- a) $H_0^a: \beta_1 \leq 0$ and $H_1^a: \beta_1 > 0$; $H_0^b: \beta_1 \geq 0$ and $H_1^b: \beta_1 < 0$
- b) $H_0: \beta_1 = \beta_2 = 0, H_1: H_0$ is not valid.
- c) $H_0: \beta_1 \leq 0, H_1: \beta_1 > 0.$
- d) All the above are incorrect.

3.3. To tackle the health and economic crisis, a few months after the outbreak of the Covid-19 pandemic in 2020, many governments implemented fiscal expansionary policies. According to economic theory, these measures should have a positive effect on income. Imagine that we are interested in testing whether a 1 percentage point increase in government expenditure increases GDP per capita proportionally. Using data on quarterly gross domestic product (GDP) and fiscal expending ($gexpending$), which of the following is the correct way to do this, using the following log-log specification:

$$\log(GDP) = \beta_0 + \beta_1 \log(gexpending) + u$$

- a) $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$.
- b) $H_0: \beta_1 \leq 0, H_1: \beta_1 > 0$.
- c) $H_0: \beta_1 = 0, H_1: \beta_1 < 0$.
- d) All the above options are inappropriate.

3.4. Which of the following answers are correct?

In a simple linear regression model, if the explanatory variable is divided by a constant λ ,

- a) The coefficient of determination is multiplied by the same constant λ .
- b) The coefficient of the explanatory variable and its standard error estimated by OLS are both divided by the same constant λ .
- c) The coefficient of the explanatory variable and its standard error estimated by OLS are both multiplied by the same constant λ .
- d) All the above answers are incorrect.

3.5. A researcher has estimated a model to evaluate the determinants of innovation using sales and public investment and would like to test whether the variables are jointly significant. Which of the following options are correct?

- a) An F test can be used to test whether the two variables are jointly significant only if we know the R^2 .
- b) A Student's t -test can be performed to determine the statistical significance of each of the two variables and if in both cases the null is rejected, the two variables are jointly significant.
- c) An F test can be used to test whether the two variables are jointly significant only if we know the sum of squares of the residuals.
- d) None of the above are correct.

3.6. In a simple regression model, imagine that you are interested in testing a hypothesis at the 1% significance level for the slope parameter. Which of the following options would be correct?

- a) Construct a 99% confidence interval for the parameter in question and see whether the hypothesised value is contained within the interval. If it is not, you reject the null hypothesis.
- b) Use a Student's t -test for the corresponding parameter and hypothesised value and compare the resulting t -statistic with the critical value "c" in the statistical t table using the corresponding degrees of freedom and the given significance level.
- c) An F -test can be used of joint significant of the regression, given that there is only one regressor.
- d) All the above answers are correct.

3.7. The estimated slope of a regression model of sales on innovation expenditure, estimated by OLS and obtained for a sample of 23 individual firms, is $\hat{\beta} = 0.7$, with the corresponding standard error being (0.102).

The following confidence interval for the β parameter has been obtained at a confidence level of 95%: $(\hat{\beta} - 2.086 \text{ s.e.}(\hat{\beta}), \hat{\beta} + 2.086 \text{ s.e.}(\hat{\beta}))$. Which of the following statements are correct?

- a) The hypothesis $\beta = 1$ will be rejected at the 5% significance level since $\beta = 1$ is outside the 95% confidence interval for β .
- b) The hypothesis $\beta = 1$ will be rejected at the 10% significance level since $\beta = 1$ is outside the 95% confidence interval for β .
- c) The hypothesis $\beta = 1$ will be rejected at the 1% significance level since $\beta = 1$ is outside the 95% confidence interval for β .
- d) None of the above are correct.

3.8. In 2020 many governments took mobility restriction measures aimed at reducing the spread of Covid-19 and the number of deaths. A researcher wants to estimate whether and to what extent the number of new cases will decrease for each additional day of lockdown. Which of the following statements are correct?

- a) An OLS model can be estimated using the number of Covid-19 cases as dependent variable and the number of days of lockdown as independent variable. Then the slope of the regression can be used to test the null hypothesis that the corresponding β is equal to zero versus the alternative that it is different from zero.
- b) An OLS model can be estimated using the number of Covid-19 cases as dependent variable and the number of days of lockdown as independent variable, while also controlling for other relevant factors. Then the slope of the regression can be used to test the null hypothesis that the corresponding β is equal to zero versus the alternative that it is different from zero.
- c) After estimating the OLS model indicated in a), a confidence interval can be constructed for the slope parameter; and if that interval contains zero, it means the lockdown has no significant effect on the number of Covid-19 cases.
- d) All the above answers are correct.

3.9. In a model aimed at evaluating the determinants of home rental prices in big cities in Spain, which include the size of the home in square metres, the size of the garden in square metres, the number of rooms and the distance to the city centre, what is the correct way to test whether a one square metre increase in home size has the same effect on prices as a one square metre increase in garden size:

- a) $H_0: \beta_1 + \beta_2 = 0$, $H_1: H_0$ is not true.
- b) $H_0: \beta_1 - \beta_2 = 0$, $H_1: H_0$ is not true.
- c) $H_0: -\beta_1 + \beta_2 = 0$, $H_1: H_0$ is not true.
- d) None of the above is correct.

3.10. A researcher would like to test the effect of an increase in income per capita on inequality, measured using the GINI coefficient. The researcher knows that, according to economic theory, a higher GINI is related to a lower income per capita for lower income levels. What is the correct way to specify the model and test whether income per capita affects inequality.

- a) Taking logs of income per capita and specifying a log-level model, test the null hypothesis that β , the coefficient on the GINI, is equal to zero versus the alternative that it is different from zero.
- b) Taking logs of income per capita and specifying a log-level model, test the null hypothesis that β , the coefficient on the GINI, is higher than or equal to zero versus the alternative that it is lower than zero.
- c) Taking logs of income per capita and specifying a log-level model, test the null hypothesis that β , the coefficient on the GINI, is lower than or equal to zero versus the alternative that it is higher than zero.
- d) All the above answers are correct.
- b) Taking logs of income per capita and specifying a log-level model, test the null hypothesis that β , the coefficient on the GINI, is higher or equal to zero versus the alternative that it is lower than zero.
- c) Taking logs of income per capita and specifying a log-level model, test the null hypothesis that β , the coefficient on the GINI, is lower or equal to zero versus the alternative that it is higher than zero.
- d) All the above answers are correct.

CHAPTER 4

OTHER TOPICS RELATED TO REGRESSION MODELS

4.1. Rescaling of variables

In what follows, we will evaluate the changes an SRF undergoes when variables are rescaled.²¹ To do so, we will consider three different cases:

A. Estimated level-level model:

$$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$$
$$se(\hat{\beta}_0) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$$

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

B. Estimated level-log model:

$$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$$
$$se(\hat{\beta}_0) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$$

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

C. Estimated log-level model:

$$\log(\hat{y}_A) = \hat{\beta}_0 + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$$
$$se(\hat{\beta}_0) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$$

$$SSR = \sum_{i=1}^n (\log(y_i) - \log(\hat{y}_i))^2, \quad R^2 = \frac{\sum_{i=1}^n (\log(\hat{y}_i) - \log(\bar{y}))^2}{\sum_{i=1}^n (\log(y_i) - \log(\bar{y}))^2}$$

²¹ Note: we use c to denote a constant other than 0.

D. Estimated log-log model:

$$\log(\hat{y}_A) = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)} + \frac{\hat{\beta}_1 \log(x_{1A})}{se(\hat{\beta}_1)} + \frac{\hat{\beta}_2 \log(x_{2A})}{se(\hat{\beta}_2)}$$

$$SSR = \sum_{i=1}^n (\log(y_i) - \log(\hat{y}_i))^2, \quad R^2 = \frac{\sum_{i=1}^n (\log(\hat{y}_i) - \log(\bar{y}))^2}{\sum_{i=1}^n (\log(y_i) - \log(\bar{y}))^2}$$

Table 15. Rescaling dependent or independent variables in a level-level model

A. Estimated level-level model		
Rescaling	Sample regression function	Sum of squared residuals and coefficient of determination
$y_A/c = y_B$	$\hat{y}_B = \hat{\beta}_0/c + \hat{\beta}_1/c x_{1A} + \hat{\beta}_2/c x_{2A}$ $se(\hat{\beta}_0)/c \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)/c$	$\frac{SSR}{c^2} \quad R^2$
$y_A \cdot c = y_B$	$\hat{y}_B = \hat{\beta}_0c + \hat{\beta}_1c x_{1A} + \hat{\beta}_2c x_{2A}$ $se(\hat{\beta}_0)c \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)c$	$SSR \cdot c^2 \quad R^2$
$x_{1A} \cdot c = x_{1B}$	$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1/c x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)$	$SSR \quad R^2$
$x_{1A}/c = x_{1B}$	$\hat{y}_A = \hat{\beta}_0 + \hat{\beta}_1c x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)$	$SSR \quad R^2$

Table 16. Rescaling dependent or independent variables in a level-log model

B. Estimated level-log model		Sum of squared residuals and coefficient of determination
Rescaling	Sample regression function	
$y_A/c = y_B$	$\hat{y}_B = \hat{\beta}_0/c + \hat{\beta}_1/c \log(x_{1A}) + \hat{\beta}_2/c \log(x_{2A})$ $se(\hat{\beta}_0)/c \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)/c$	$\frac{SSR}{c^2}$ R^2
$y_A \cdot c = y_B$	$\hat{y}_B = \hat{\beta}_0c + \hat{\beta}_1c \log(x_{1A}) + \hat{\beta}_2c \log(x_{2A})$ $se(\hat{\beta}_0)c \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)c$	$SSR \cdot c^2$ R^2
$\log(x_{1A} \cdot c) = \log(x_{1B})$	$\hat{y}_A = [\hat{\beta}_0 + \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 + \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$\log(x_{1A}/c) = \log(x_{1B})$	$\hat{y}_A = [\hat{\beta}_0 - \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 - \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2

Table 17. Rescaling dependent or independent variables in a log-level model

C. Estimated log-level model		
Rescaling	Sample regression function	Sum of squared residuals and coefficient of determination
$\log(Y_A/c) = \log(Y_B)$	$\log(\hat{Y}_B) = [\hat{\beta}_0 - \log(c)] + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 - \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$\log(Y_A \cdot c) = \log(Y_B)$	$\log(\hat{Y}_B) = [\hat{\beta}_0 + \log(c)] + \hat{\beta}_1 x_{1A} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0 + \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$x_{1A} \cdot c = x_{1B}$	$\log(\hat{Y}_A) = \hat{\beta}_0 + \hat{\beta}_1/c \quad x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)/c \quad se(\hat{\beta}_2)$	SSR R^2
$x_{1A}/c = x_{1B}$	$\log(\hat{Y}_A) = \hat{\beta}_0 + \hat{\beta}_1 c \quad x_{1B} + \hat{\beta}_2 x_{2A}$ $se(\hat{\beta}_0) \quad se(\hat{\beta}_1)c \quad se(\hat{\beta}_2)$	SSR R^2

Table 18. Rescaling dependent or independent variables in a log-log model

D. Estimated log-log model		
Rescaling	Sample regression function	Sum of squared residuals and coefficient of determination
$\log(y_A/c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 - \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 - \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$\log(y_A \cdot c) = \log(y_B)$	$\log(\hat{y}_B) = [\hat{\beta}_0 + \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 + \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$\log(x_{1A} \cdot c) = \log(x_{1B})$	$\log(\hat{y}_A) = [\hat{\beta}_0 + \hat{\beta}_1 \log(c)] + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 + \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2
$\log(x_{1A}/c) = \log(x_{1B})$	$\log(\hat{y}_A) = \hat{\beta}_0 - \hat{\beta}_1 \log(c) + \hat{\beta}_1 \log(x_{1A}) + \hat{\beta}_2 \log(x_{2A})$ $se(\hat{\beta}_0 - \hat{\beta}_1 \log(c)) \quad se(\hat{\beta}_1) \quad se(\hat{\beta}_2)$	SSR R^2

4.2. Interactions between explanatory variables

We can **interact explanatory variables** in the regression model in order to modulate the marginal effects. For example, when we interact two different explanatory variables, we allow the marginal effect of one variable (x_1) on the dependent variable (y) to depend on the value of another variable (x_2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u, \quad \frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

4.3. Goodness-of-fit and model selection

We mentioned earlier that adding regressors to a model, whether or not they are relevant, decreases the SSR and thus increases R^2 . For that reason, the R^2 should not be used to decide whether to add one or more regressors to a model.

Solution:

- Statistical inference (t -test or F -test, as the case may be) is particularly useful for **selecting models with or without a constant** and, in general, for comparing **nested models**. See Section 3. Statistical inference in regression models.
- Comparing the adjusted coefficient of determination (\bar{R}^2) is useful for choosing between different **non-nested models** provided you have the same dependent variable (y) and the same sample size (n).

$$\bar{R}^2 = 1 - \frac{\frac{SSR}{n - k - 1}}{\frac{TSS}{n - 1}}$$

\bar{R}^2 is especially useful when we want to decide between alternative regressors to capture a particular aspect that may explain the dependent variable, or between regressors that represent different functional forms.

To illustrate the consequences of rescaling certain variables in the regression analysis, let us now use an example. Specifically, imagine that we are interested in studying the determinants of CO_2 emissions in gigatons per capita ($\ln CO_2 pc$) for a cross-section of countries $i=1, 2, \dots, 116$ in 2019.

With this purpose in mind, the following explanatory variables are used: the natural logarithm of GDP per capita ($\ln GDPpc$) and its squared term ($\ln GDPpc^2$), the natural log of exports, and an index of air quality. The OLS estimated results are presented below:

Table 19. Linear regression results for the determinants of the natural log of CO2 emissions

<i>ln CO2pc</i>	Coef.	Std. Err.	t-stat
<i>ln GDPpc</i>	2.391	0.448	5.34
<i>ln GDPpc²</i>	-0.075	0.027	-2.75
<i>ln Exports</i>	0.031	0.015	2.13
<i>Air_quality</i>	-0.022	0.005	-4.12
<i>constant</i>	-13.963	1.970	-7.09
<i>R²</i>	0.833		
<i>Number of observations</i>	116		
<i>F(4, 111)</i>	144.18		
<i>Prob>F</i>	0.000		

Note: The GDP in thousand US dollars per inhabitant has been collected from the World Development Indicators. Exports are in current US dollars and have been obtained from UN Comtrade. *Air_quality* is an index that varies between 0 and 100, with higher values indicating stricter environmental air quality legislation, and is obtained from <https://epi.yale.edu>.

As explained above, when the units of measurement change, part of the outcome of the regression results will vary. In this case, we illustrate below the results obtained when the variable *Exports* is rescaled to 1000 US dollars, that is, we divide the original variable by 1000 and take the natural log:

Table 19. Linear regression results for the determinants of the natural log of CO2 emissions, using a rescaled independent variable

<i>ln CO2pc</i>	Coef.	Std. Err.	t-stat
<i>ln GDPpc</i>	2.391	0.448	5.34
<i>ln GDPpc²</i>	-0.075	0.027	-2.75
<i>ln Exports</i>	0.031	0.015	2.13
<i>Air_quality</i>	-0.022	0.005	-4.12
constant	-13.746	1.956	-7.03
<i>R²</i>	0.833		
Number of observations	116		
<i>F(4, 111)</i>	144.18		
<i>Prob>F</i>	0.000		

where the new constant coefficient is given by: $(\hat{\beta}_0 - \hat{\beta}_3 \log(c)) = (\hat{\beta}_0 - \hat{\beta}_3 \log(1000)) = -13.963 - 0.031 * 6.9077 = -13.746$, and the corresponding standard error is calculated as: $se(\hat{\beta}_0 - \hat{\beta}_3 \log(c)) = 1.970 - 0.031 * 6.9077 = 1.956$.

Next, we consider the same baseline model but now the dependent variable is expressed in levels. The regression results are given by:

Table 20. Linear regression results for the determinants of CO2 emissions

CO2pc	Coef.	Std. Err.	t-stat
<i>ln GDPpc</i>	-12.356	5.159	-2.39
<i>ln GDPpc²</i>	0.984	0.352	2.8
<i>ln Exports</i>	0.082	0.135	0.61
<i>Air_quality</i>	-0.195	0.070	-2.8
<i>constant</i>	42.358	21.431	1.98
<i>R²</i>	0.511		
<i>Number of observations</i>	116		
<i>F(4, 111)</i>	27.64		
<i>Prob>F</i>	0.00		

Consider how the results will be modified if the dependent variable is rescaled to be expressed in tonnes instead of gigatons per capita, that is, it is multiplied by 1000. As can be seen below, all the coefficients and s.e. are multiplied by 1000, whereas the t-stats and R^2 are unchanged.

Table 21. Linear regression results for the determinants of the CO2 emissions, using a rescaled dependent variable

CO2pc	Coef.	Std. Err.	t-stat
<i>ln GDPpc</i>	-12355.700	5159.060	-2.39
<i>ln GDPpc</i> ²	984.059	351.547	2.8
<i>ln Exports</i>	81.818	135.072	0.61
<i>air_quality</i>	-195.306	69.761	-2.8
constant	42358.410	21431.200	1.98
<i>R</i> ²	0.511		
Number of observations	116		
<i>F</i> (4, 111)	27.64		
Prob> <i>F</i>	0.00		

Problem set 4A (with solutions)

EXERCISE 4A.1. A researcher aims at investigating the determinants of income inequality for 34 OECD countries. The dependent variable is the GINI coefficient in percent and the regressors are participation in global value chains in percent, GDP per capita (in thousand US dollars) and the share of FDI on GDP. The model is estimated with the GINI and the GDP per capita in levels in columns (1) and (2), and with the GDP in natural logs in columns (3) and (4). In addition, the squared term of the GDP per capita enters the model in column (2) and the squared of the natural log of the GDP per capita is added in column (4). Using the regression results presented in Table 23:

Table 22. OLS Regression results

Dependent variable:	(1)	(2)	(3)	(4)
	GINI	GINI	GINI	GINI
<u>Explanatory variables:</u>				
<i>GVC_participation</i>	-0.515*** (0.162)	-0.542*** (0.102)	-0.485*** (0.133)	-0.455*** (0.096)
<i>ln_FDI_IN</i>	-0.780 (0.772)	-1.156** (0.471)	-0.376 (0.723)	-0.712* (0.393)
<i>GDP_pc</i>	-0.105 (0.084)	-0.742*** (0.116)		
<i>GDP_pc_sq</i>		0.007*** (0.001)		
<i>ln_GDP_pc</i>			-6.782***	-65.201***

			(2.281)	(7.562)
<i>ln_GDP_pc_sq</i>				8.556*** (1.113)
<i>Constant</i>	60.814*** (8.158)	74.790*** (4.336)	78.632*** (9.282)	175.928*** (13.062)
<i>Observations</i>	34	34	34	34
<i>R-squared</i>	0.487	0.767	0.616	0.789
<i>R-squared adjusted</i>	0.436	0.735	0.578	0.760

Note: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

- Indicate whether you will consider more appropriate to introduce the GDP variable in levels or in natural logs in the model and reason why. Compare results and test shown in columns (1) and (2) with those in columns (3) and (4).
- In models (2) and (4) a non-linearity is introduced in the model. Calculate the partial effects of GDP per capita on inequality in models (2) and (4) and interpret the results. Please consider that the average income per capita in the sample is 35000 US dollars.
- What is the turning point for GDP per capita in models estimated in columns (2) and (4)? That is, what is the level of GDP per capita for which the effect of this variable on the GINI changes from negative to positive. Please consider the average income per capita given in b).
- Is FDI a relevant regressor in model (4), please interpret the estimated coefficient.

SOLUTION 4A.1:

a) The models with GDP per capita in logs (3) and (4) show a higher adjusted R^2 than the corresponding models in levels, (1) and (2). For instance, compare 0.616 with 0.487 and 0.789 with 0.767. In addition, GDP per capita is not statistically significant when introduced in levels in model (1). Moreover, GDP is a variable in monetary units that can take very high values and it is convenient to take natural logs to smooth the series. For all these reasons, the specifications with GDP per capita in natural logs are preferred.

b) In order to calculate the partial effects, we have to consider the partial derivative of the GINI with respect to GDP per capita. The partial effect equals the coefficient of GDP per capita plus the coefficient of its square term multiplied by two and by the average value in the sample, that is:

Model (2): Partial effect = $-0.742 + 2 * 35 * 0.007 = -0.259$

Model (4): Partial effect= $-65.201 + \ln(35) * 8.556 = -4.36$ to be interpreted as in a level-log model.

c) The turning points are calculated considering that the relationship between GNI and GDP per capita presents a U-shape. Therefore, the minimum is calculated by taking the first derivative, then equating the resulting expression to zero, and obtaining the corresponding value for the GDP per capita, that is:

In model (2): GDP per capita= $0.074 / (2 * 0.007) = 195$ thousand US dollars.

In model (4): \ln of GDP per capita= $65.201 / (2 * 8.556) = 10.71 \rightarrow \exp(10.71) = 45$ thousand US dollars.

d) FDI is only significant at the 10 percent level in model (4), which is the preferred model according to the value of the adjusted R^2 , therefore its inclusion depends on the decision of the researcher.

EXERCISE 4.A.2. A regression analysis has been used to identify the effect of the Covid-19 pandemic on trade for a global sample of countries that comprises 97 exporters and 169 importers. The dependent variable is the natural log of bilateral exports from country i to country j and the explanatory variables are the natural log of GDP in the exporter ($\ln y_i$) and the importer ($\ln y_j$) countries, the average number of covid cases per 10000 in the exporter countries weighted by the distance to each trading partner ($wcovcai$), a similar variable for the importer countries ($wcovcaj$) and several proxies for trade costs, namely, whether countries belong to the same regional trade agreement, the natural log of distance between countries ($\ln dij$), and three dummy variables for: common language, common border and the existence of a past or present colonial relationship. A pooled OLS regression using data for 2009 and 2020 provides the results given in the following table:

Table 23. Regression results for exports and Covid-19 incidence

Dependent variable:	(1)	(2)	(3)
	Ln Exports	Ln Exports	Ln Exports
<u>Explanatory variables:</u>			
<i>lnyi</i>	0.776*** (0.005)	0.797*** (0.004)	0.797*** (0.004)
<i>lnyj</i>	0.643*** (0.004)	0.646*** (0.004)	0.646*** (0.004)
<i>lnwcovcai</i>	-0.843*** (0.032)		
<i>lnwcovcaj</i>	-1.957*** (0.046)		
<i>wcovidcai</i>		-2.414*** (0.099)	-2.261*** (0.107)
<i>wcovidcaj</i>		-1.319*** (0.047)	-1.343*** (0.049)
<i>RTAwcovi</i>			-0.436*** (0.119)
<i>RTAwcovj</i>			0.181*** (0.058)
<i>Indij</i>	-0.924*** (0.008)	-0.917*** (0.008)	-0.917*** (0.008)
<i>Regional Trade Agreement dummy</i>	1.261*** (0.015)	1.307*** (0.015)	1.316*** (0.016)
<i>Common Language dummy</i>	0.204*** (0.020)	0.197*** (0.020)	0.198*** (0.020)
<i>Contiguity dummy</i>	1.464*** (0.040)	1.462*** (0.040)	1.461*** (0.040)
<i>Former Colony dummy</i>	1.482*** (0.037)	1.496*** (0.037)	1.496*** (0.037)
<i>Constant</i>	-14.901*** (0.179)	-15.567*** (0.177)	-15.583*** (0.177)
<i>Observations</i>	188,840	188,840	188,840
<i>R-squared</i>	0.491	0.486	0.486
<i>R-squared adjusted</i>	0.491	0.486	0.486

Note: Robust standard errors in parentheses***. p<0.01, ** p<0.05, * p<0.1.

- a) Interpret the coefficient of the estimated parameters in column (1) for the covid incidence variables. Notice that they are entered in natural logs. How will the estimated coefficients and the standard errors of these variables change if the covid cases are expressed as raw number of cases, instead of as number of covid cases per 10000.
- b) In column (2) the covid variables are entered in levels, how does the interpretation change? Indicate how the coefficient will be affected if the covid cases are expressed as raw number of cases.
- c) In column (3) the covid variables are interacted with the regional trade agreement dummy (RTA) variable, calculate the partial effect of an increase in the number of covid cases in the exporter country for countries that belong to the same regional trade agreement.
- d) How will the coefficient of the variable *wcovidcai* will change in column (2) if the number of cases is measured in number of cases per million of inhabitants. Indicate whether the standard errors will change and whether the constant parameter of the model will be the same.

SOLUTION 4A.2:

- a) Since both dependent and independent variables are in natural logs, the coefficients can be interpreted as elasticities: For each 1 percent increase in the average number of cases in the exporter countries, exports are expected to decrease by 0,84 percent; whereas for each 1 percent increase in the number of cases in the importer countries, exports are expected to decrease by 1,95 percent, that is, more than proportionally. The coefficients will not change since the model is a log-log model, as can be seen in the theory section in case D (Table 18).
- b) When the covid variables are entered in levels the interpretation corresponds to a log-level model, therefore, an increase in 1 case per ten thousand inhabitants in the exporter country, will results in a decrease of exports by $(2,41/100)=0.02$ percent. Similarly, for the importer country the decrease will be 0.013 percent. For easier interpretation we can refer to an increase in 100 cases (per each ten thousand inhabitants) and say that for each additional 100 cases (per each ten thousand inhabitants) in the exporter country exports will decrease by around 2 percent.

c) According to the theory section, the marginal effect will be calculated as,

$$\ln exports = \beta_0 + \beta_1 \ln y_i + \dots + \beta_3 w_{covidca_i} + \dots + \beta_5 RTA_{ij} + \beta_6 w_{covidca_i} * RTA_{ij} + u,$$

$$\frac{\Delta \ln export}{\Delta w_{covidca_i}} = \beta_3 + \beta_6 RTA_{ij} = -2.26 - 0.436 = -2.69$$

d) The variable $w_{covidca_i}$ will change in column (2) if the variable number of cases is measured as number of cases per million of inhabitants, since the results have been obtained using $w_{covidca_i}/10000$, we will have to consider that now the variable equals $(w_{covidca_i}/10000)/100$, therefore, given that the model has a log-level form for this variable, the coefficient, according to Table 17 in the theory section, would have to be multiplied by 100, resulting in $-2.414 * 100 = -241$. The corresponding standard error will also be multiplied by 100.

Problem set 4B

EXERCISE 4B.1 The following model describes the relationship between the level of pollution in different countries and their income (environmental Kuznets curve), where $CO2c$ is the level of pollution in terms of CO2 emissions, measured in metric tons per capita, and $GDPc$ is the gross domestic product (GDP), expressed in millions of PPP dollars per capita.

$$CO2c = \beta_0 + \beta_1 GDPc + \beta_2 GDPc^2 + u$$

From the information available on the World Bank website <https://data.worldbank.org>, cross-sectional data on the variables of interest for 2014 were downloaded for a sample of 182 countries ($i = 1, 2, \dots, 182$). The table below shows the main results of the estimation:

Model 1: OLS, using observations 1-182
Dependent variable: $CO2c$

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
const	-0.417	0.341	-1.222	0.223	
$GDPc$	351.913	36.938	9.527	0.000	***
$GDPc_sq$	-3161.380	696.8	-4.537	0.000	***
Mean of the dep. var.	3.738	S.D. of the dep. var.		3.793	
Sum of sq. residuals	968.739	S.D. of the regression		2.326	
R-squared	0.628	Adjusted R-squared		0.624	
$F(2, 179)$	151.041	p -value (of F)		3.73e-39	
Log-likelihood	-410.398	Akaike criterion		826.796	
Schwarz criterion	836.408	Hannan-Quinn criterion		830.692	

a) Calculate and interpret the marginal effect of income on the level of pollution. Also, according to the estimated model, graph the relationship between pollution and income, quantifying the point of origin of the SRF, its slope and possible turning point.

b) Based on the estimated results, should the model contain the quadratic term as a regressor? Give reasons for your answer.

c) The per capita GDP of Norway in 2014 was \$0.066 million and that of Spain, \$0.034 million. Based on the results obtained, could economic development policies in these two countries harm the environment? Give reasons for your answer.

d) Knowing that 1 dollar = 0.89 euros, write the estimated equation (including the standard errors and R -squared) we would obtain if we expressed GDP in millions of euros per capita, instead of millions of dollars per capita.

e) Below are three equations estimated using the data set described above. Which of the proposed models would be preferable? Give reasons for your answer.

$$\widehat{CO_2c} = -0.417 + 351.913 \text{ GDPc} - 3161.38 \text{ GDPc}^2$$

$$\begin{array}{ccc} (0.341) & (36.938) & (696.8) \\ R^2 = 0.628 & \bar{R}^2 = 0.624 & n = 182 \end{array}$$

$$\widehat{CO_2c} = 0.620 + 192.628 \text{ GDPc}$$

$$\begin{array}{ccc} (0.267) & (192.628) \\ R^2 = 0.585 & \bar{R}^2 = 0.583 & n = 182 \end{array}$$

$$\widehat{CO_2c} = 15.818 + 2.610 \log(\text{GDPc})$$

$$\begin{array}{ccc} (0.799) & (0.168) \\ R^2 = 0.573 & \bar{R}^2 = 0.570 & n = 182 \end{array}$$

EXERCISE 4B.2 Search online for a *cross-sectional* data set for two economic and business variables that you think may have a quadratic relationship.²² Using that data set, perform the following tasks:

a) Using EXCEL, save the two variables in columns, naming and sorting them, together with an index variable $i = 1, 2, \dots, N$ to represent the cross-sectional dimension (e.g. individuals, countries, households, companies, etc.). Do not forget to state the source for the data and the meaning of each variable and its units of measurement.

b) Use economic theory or logical reasoning to explain which is the dependent variable and which the explanatory variable(s).

c) Using EXCEL or Gretl, draw and interpret a scatter plot (X-Y plot) that shows the relationship between the variables.

d) Using Gretl, estimate the following model by OLS: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$ and write out the resulting SRF in the usual form. Based on the results obtained, explain the estimated marginal effect of x_1 on y . Calculate and interpret the turning point in the relationship between y and x_1 .

e) Should the model contain the quadratic term as a regressor? Give reasons for your answer.

f) Now rescale the regressor x as you choose (e.g. from years to months, from euros to hundreds of euros, etc.). Express the SRF, after rescaling, and explain how the estimated coefficients, the standard errors, the t -statistics, the coefficient of determination and the SSR will change.

²² Possible data sources: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (<https://ec.europa.eu/eurostat/data/database>), OECD (<https://stats.oecd.org/>), World Bank (<https://data.worldbank.org/>), UNCTADSTAT (<https://unctadstat.unctad.org/>), FAOSTAT (<http://www.fao.org/faostat/en/#data>), others (<https://www.economicsnetwork.ac.uk/links/sources>, <https://db.nomics.world/>).

EXERCISE 4B.3. Consider the following SRFs:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 \log(x_1) + \hat{\beta}_2 \log(x_2)$$

$$\log(\hat{y}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- a) Explain what would happen to the OLS estimators, s.e. and R^2 based on the models considered if we multiplied all the values of the dependent variable, y , by 100.
- b) Assume, instead of (a), that all values of x_1 are multiplied by 100. Explain how the OLS estimators, s.e. and R^2 would be affected in this case.

EXERCISE 4B.4. Consider the following econometric model:

$$\log(\text{precio}) = \beta_0 + \beta_1 \log(\text{dorm}) + \beta_2 \log(m2) + \beta_3 \log(\text{dorm}) \cdot \log(m2) + \beta_4 \log(\text{banos}) + u$$

where *precio* is the selling price of homes expressed in euros, *dorm* the number of bedrooms, *m2* the size of the home expressed in square metres and *banos* the number of bathrooms. Based on this information, answer the following questions:

- a) According to this model, what is the marginal effect of the number of bedrooms on price?
- b) Using the data set with information on 1300 homes for sale in Castellón on 11 February 2019 (source: Nestoria), the following SRF was estimated:

$$\log(\widehat{\text{precio}}) = 9.344 - 2.877 \log(\text{dorm}) + 0.463 \log(m2)$$

(0.565) (0.390) (0.133)

$$+ 0.576 \log(\text{dorm}) \cdot \log(m2) + 0.608 \log(\text{banos})$$

(0.089) (0.042)

$$n = 1213 \quad R^2 = 0.590$$

Interpret the estimated coefficient associated with the interaction term. For this purpose, assume that the size is 150 square metres. What if it were 300 square metres?

- c) Test whether the effect of the number of bedrooms on price depends on the size of the home or not.

Multiple-choice questions (Topic 4)

4.1. Which of the following answers is correct?

In a SRF derived from a simple log-log model, if the dependent variable is divided by a constant λ ,

- a) The coefficient of the explanatory variables, estimated by OLS, and its standard errors are both divided by the same constant λ .
- b) The coefficient of the explanatory variables, estimated by OLS, and its standard errors are both multiplied by the same constant λ .
- c) Only the coefficient of the constant, estimated by OLS, and its standard error vary in the new regression results.
- d) All the above answers are incorrect.

4.2. Which of the following answers is correct?

In a sample regression derived from a level-level model, if the dependent variable is divided by a constant λ ,

- a) The coefficient of the explanatory variables, estimated by OLS, and its standard errors are both divided by the same constant λ .
- b) The coefficient of the explanatory variables, estimated by OLS, and its standard errors are both multiplied by the same constant λ .
- c) Only the coefficient of the constant, estimated by OLS, and its standard error vary in the new regression results.
- d) All the above answers are incorrect.

4.3. According to the following model:

$$p_i = \beta_0 + \beta_1 size_i + \beta_2 rooms_i + \beta_3 size_i \cdot rooms_i + u_i$$

where

p_i is the selling price for a set of homes $i = 1, 2, 3, \dots, n$.

$size_i$ represents the floor area of the homes, measured in square metres

$rooms_i$ is the number of rooms in each home.

$size_i \cdot rooms_i$ denotes the interaction term between $size_i$ and $rooms_i$.

With the information provided, which of the following statements are correct?

- The response of prices to a marginal effect of size is given by the parameter β_1
- The response of prices to a marginal effect of size is given by $(\beta_1 + \beta_3 \text{rooms}_i)$.
- The response of prices to a marginal effect of size is given by $(\beta_1 + \beta_2)$.
- All the above answers are incorrect.

4.4. Consider the regression results of the following two estimated models, obtained from a sample of 100 households, where *sav* refers to annual household savings in dollars, *inc* is the annual household income in dollars, and *sq_inc* is the square term of the annual household income in dollars

Estimated model 1	Estimated model 2
$\widehat{sav} = 124.842 + 0.147 \text{ inc}$ <p style="text-align: center;">(655.393) (0.058)</p>	$\widehat{sav} = -295.643 + 0.22 \text{ inc}$ <p style="text-align: center;">(1249.67) (0.193)</p>
$n = 100 \quad R^2 = 0.062 \quad \bar{R}^2 = 0.053$	$-0.00000234875 \text{ inc}^2$ <p style="text-align: center;">(0.00000593)</p>
	$n = 100 \quad R^2 = 0.064 \quad \bar{R}^2 = 0.044$

Considering the information provided, which model would be preferable?

- Model 1 is preferable because the adjusted coefficient of determination, \bar{R}^2 , is higher.
- Model 1 is preferable because the coefficient of determination, R^2 , is higher.
- Model 1 is preferable because *sq_inc* is not statistically significant in Model 2.
- The preferred model is always the one with the smallest number of independent variables.

4.5. Consider the regression results of the following two estimated models, obtained from a sample of 50 second-hand cars, where *price* is the selling price in euros on the second-hand market, *age* is the car's age in years, *mileage* is the number of kilometres travelled by car, and *power_hp* is the engine horsepower.

Model 1: OLS, using observations 1-50
Dependent variable: *price*

	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	-4345.420	4041.80	-1.075	0.288	
<i>age</i>	-310.029	476.173	-0.651	0.518	
<i>km</i>	-0.106	0.050	-2.124	0.039	**
<i>power_hp</i>	232.199	22.437	10.350	0.000	***
Mean dependent var	23434.300	S.D. dependent var	20413.730		
Sum squared residuals	4.79e+09	S.E. of regression	10206.670		
<i>R</i> -squared	0.765	Adjusted <i>R</i> -squared	0.750		
<i>F</i> (3, 46)	50.002	<i>p</i> -value(<i>F</i>)	1.61e-14		
Log-likelihood	-530.402	Akaike criterion	1068.804		
Schwarz criterion	1076.453	Hannan-Quinn	1071.717		

Model 2: OLS, using observations 1-50
Dependent variable: *price*

	<i>Coefficient</i>	<i>Std. Error</i>	<i>t-ratio</i>	<i>p-value</i>	
const	31435.900	3471.810	9.055	0.000	***
<i>age</i>	-1613.230	462.620	-3.487	0.001	***
Mean dependent var	23434.300	S.D. dependent var	20413.730		
Sum squared resid	1.63e+10	S.E. of regression	18423.220		
<i>R</i> -squared	0.202	Adjusted <i>R</i> -squared	0.186		
<i>F</i> (1, 48)	12.160	<i>p</i> -value(<i>F</i>)	0.001		
Log-likelihood	-560.995	Akaike criterion	1125.989		
Schwarz criterion	1129.814	Hannan-Quinn	1127.446		

Considering the information provided, which model would be preferable?

- Model 1 is preferable because the adjusted coefficient of determination, \bar{R}^2 , is considerably higher.
- Model 1 is preferable because the coefficient of determination, R^2 , is considerably higher.
- Model 1 is preferable because the independent variables *km* and *power_hp* are jointly statistically significant.
- The preferred model is always the one with the smallest number of independent variables.

CHAPTER 5

INCLUDING DUMMY VARIABLES IN REGRESSION ANALYSIS

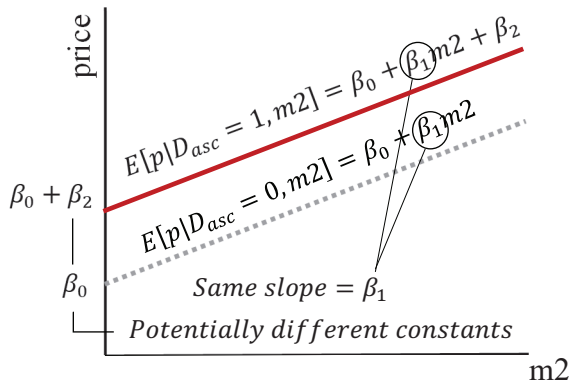
5.1. Introduction

We can include qualitative information in regression analysis by using **dummy variables**. A dummy variable is a binary variable that is equal to one when the cross-sectional unit i belongs to a particular category and zero otherwise. For example, consider the following econometric model, which is intended to explain the price of apartments in terms of their size and whether or not there is a lift in the building:

$$p = \beta_0 + \beta_1 m2 + \beta_2 D_{asc} + u$$

where p is the price of the apartment, $m2$ is the floor area in square metres and D_{asc} is a dummy variable that is equal to one if the building has a lift and 0 if it does not. Example:

Figure 15. Regression model with a dummy variable



Then, β_2 captures the lift premium, i.e. the average price difference between apartments with and without a lift, for a given apartment size.

$$\beta_2 = E[p|D_{asc} = 1, m2] - E[p|D_{asc} = 0, m2]$$

Its statistical significance can be tested with the t -statistic for $\hat{\beta}_2$:

- If $\beta_2 = 0$, having a lift is not important in determining the price of an apartment, *ceteris paribus*.
- If $\beta_2 > 0$, apartments with a lift are more expensive than those without a lift, *ceteris paribus* (the case depicted in figure).
- If $\beta_2 < 0$, apartments with a lift are cheaper than those without a lift, *ceteris paribus*.

Avoid the dummy variable trap: If the model contains a constant, including a dummy variable for each possible category will give rise to a problem of perfect collinearity. That is why in the previous example we include only one dummy variable to model two possible categories (with a lift and without a lift).

What if we have more than two categories? If we want to assess possible price differences between apartments in different parts of a city (e.g., centre, north, south, east and west), we should omit one category so as to avoid perfect collinearity. In this case, a valid model could be as follows:

$$p = \beta_0 + \beta_1 m2 + \beta_2 D_c + \beta_3 D_n + \beta_4 D_s + \beta_5 D_e + u$$

where

D_c is equal to one if the apartment is in the city centre and zero otherwise

D_n is equal to one if the apartment is in the north of the city and zero otherwise

D_s is equal to one if the apartment is in the south of the city and zero otherwise

D_e is equal to one if the apartment is in the east of the city and zero otherwise

We do not include a dummy variable for the “west of the city” category, which will represent the base group. In this case, therefore, the coefficients associated with each dummy variable $\beta_2, \beta_3, \beta_4, \beta_5$ must be interpreted relative to the omitted category. For example, β_3 would indicate the average price difference between apartments in the north of the city and apartments in the west (omitted base category):

$$\beta_3 = E \left[p \begin{array}{l} D_C = 0 \\ D_N = 1 \\ D_S = 0 \\ D_E = 0 \end{array}, m2 \right] - E \left[p \begin{array}{l} D_C = 0 \\ D_N = 0 \\ D_S = 0 \\ D_E = 0 \end{array}, m2 \right]$$

The t -statistic for $\hat{\beta}_3$ could be used to test whether the average price difference between apartments in the north and apartments in the west is statistically significant at the usual levels.

But what if we want to know the average price difference between apartments in the city centre and apartments in the north for a given size of apartment? Neither of the two compared categories (centre and north) is the base category (west). We can still measure the difference, however, by comparing the regression functions for the two categories:

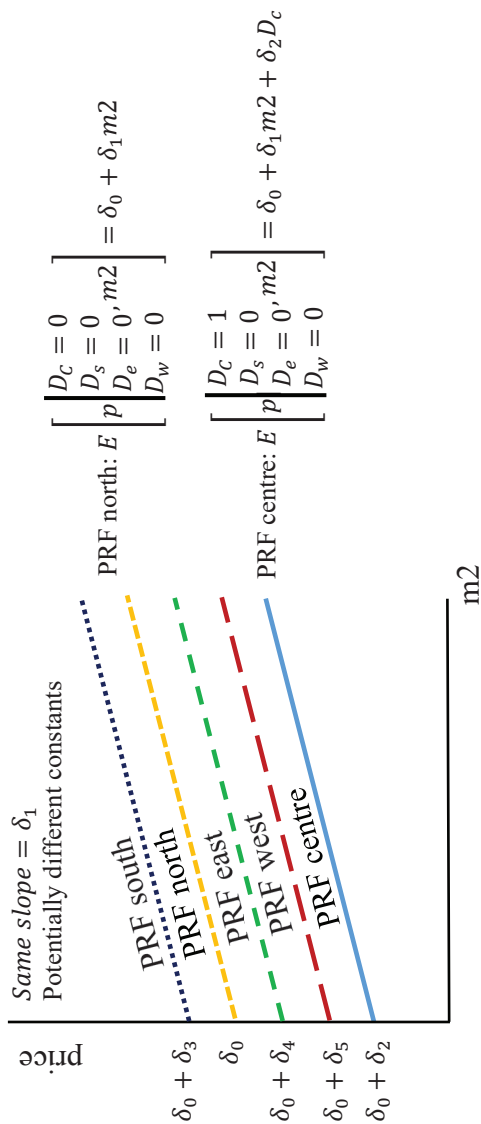
$$E \left[p \begin{array}{l} D_C = 1 \\ D_N = 0 \\ D_S = 0 \\ D_E = 0 \end{array}, m2 \right] - E \left[p \begin{array}{l} D_C = 0 \\ D_N = 1 \\ D_S = 0 \\ D_E = 0 \end{array}, m2 \right] = \beta_2(1) - \beta_3(1)$$

The simplest way to test whether that difference is statistically significant, however, would be to reformulate the model and omit the dummy variable for the category in respect of which we want to make the comparison:

$$p = \delta_0 + \delta_1 m2 + \delta_2 D_C + \delta_3 D_S + \delta_4 D_E + \delta_5 D_W + u$$

In this last case, D_W is a dummy variable equal to one if the apartment is in the west of the city and zero otherwise, and the parameter δ_2 will directly tell us the price difference between apartments in the city centre and apartments in the north (now our base category, omitted in the reformulated model), for a given apartment size. Also, the t -statistic for $\hat{\delta}_2$ could be used to carry out the appropriate significance test. For example, in this case, regardless of the apartment size, if prices in the city centre were lower than in the north ($\delta_2 < 0$), we would have the following:

Including dummy variables in regression analysis
 Figure 16. Regression model with multiple categories



5.2. Interactions in regression analysis with dummy variables

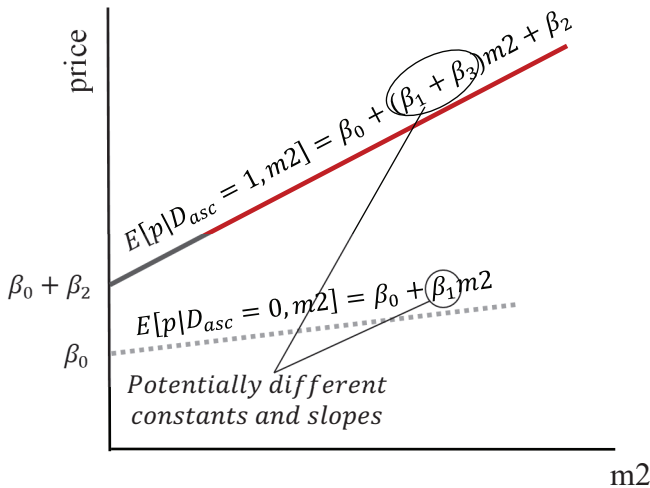
The **interaction between a continuous variable and a dummy variable** allows the partial effect of the continuous variable to depend on whether the individual belongs to a particular category. Continuing with the first example of the apartments, we now interact the continuous variable $m2$ with the dummy variable D_{asc} ($=1$ if the apartment has a lift, 0 otherwise):

$$p = \beta_0 + \beta_1 m2 + \beta_2 D_{asc} + \beta_3 (m2 \cdot D_{asc}) + u$$

We thus allow the partial effect of floor size on price to differ depending on whether the apartment has a lift or not: $\frac{\Delta p}{\Delta m2} = \beta_1 + \beta_3 D_{asc}$.

- If there is a lift, $\beta_1 + \beta_3$ represents the partial effect of size on price
- If there is not a lift, β_1 represents the partial effect of size on price

Figure 17. Regression model with a dummy variable



β_2 captures the average price difference between apartments with and without a lift, for a given apartment size.

$$\beta_2 = E[p|D_{asc} = 1, m2] - E[p|D_{asc} = 0, m2]$$

Its statistical significance can be tested with the t -statistic for $\hat{\beta}_2$

β_3 captures the difference in the partial effect of size on price between apartments with and without a lift. Its statistical significance can be tested with the t -statistic for $\hat{\beta}_3$:

- If $\beta_3 = 0 \rightarrow$ The partial effect does not depend on having a lift.
- If $\beta_2 > 0 \rightarrow$ The partial effect is greater when there is a lift (the case depicted in the graph).
- If $\beta_2 < 0 \rightarrow$ The partial effect is smaller when there is a lift.

Interaction between dummy variables can be used to create different categories based on the values of the dummy variables and assess the differences between those categories. For example, consider the following model:

$$p = \beta_0 + \beta_1 D_{asc} + \beta_2 D_{terr} + \beta_3 (D_{asc} \cdot D_{terr}) + \gamma_1 m2 + u$$

where D_{asc} is a dummy variable that is equal to one if the apartment has a lift and zero otherwise, D_{terr} is a dummy variable that is equal to one if the apartment has a terrace and zero otherwise, and $m2$ represents the apartment's floor area.

For a given apartment size, the model will tell us the average price difference between the following categories:

- Between apartments with a lift and a terrace (A) and apartments without a lift or terrace (B):

$$\begin{aligned} E[p|D_{asc} = 1, D_{terr} = 1 | m2] - E[p|D_{asc} = 0, D_{terr} = 0 | m2] \\ = \beta_1 + \beta_2 + \beta_3 \end{aligned}$$

- Between apartments without a lift but with a terrace (A) and apartments without a lift or terrace (B):

$$E[p|D_{asc} = 0, D_{terr} = 1 | m2] - E[p|D_{asc} = 0, D_{terr} = 0 | m2] = \beta_2$$

- Between apartments with a lift but without a terrace (A) and apartments without a lift or terrace (B):

$$\begin{aligned} E[p|D_{asc} = 1, D_{terr} = 0 | m2] - E[p|D_{asc} = 0, D_{terr} = 1 | m2] \\ = \beta_1 - \beta_2 \end{aligned}$$

The statistical relevance of these differences can be tested using a t test or F test, as appropriate.

Different SRFs according to categories: Finally, we can allow both the intercept and the partial effect of all the regressors (β_j) to differ between categories. To do that, we must interact all the continuous explanatory variables of the model with a dummy variable that represents the fact of belonging to the categories of interest. Consider, for example, the following specification:

$$p = \beta_0 + \beta_1 m2 + \beta_2 \text{rooms} + \gamma_0 D_{asc} + \gamma_1 (D_{asc} \cdot m2) + \gamma_2 (D_{asc} \cdot \text{rooms}) + u$$

where a dummy variable D_{asc} has been included that is also interacted with $m2$ and $rooms$. Then:

- γ_0 is the difference in the intercept between apartments with a lift and apartments without a lift,
- γ_1 is the difference in the marginal effect of size on price between apartments with a lift and apartments without a lift,
- γ_2 is the difference in the marginal effect of the number of bedrooms on price between apartments with a lift and apartments without a lift.

This type of specification is commonly used to assess structural changes in the model between categories. In our example, we could assess whether the price of the apartments follows the same model for apartments with a lift and apartments without a lift by specifying and performing the following joint significance test, through an F -test:²³

$H_0: \gamma_0 = 0, \gamma_1 = 0, \gamma_2 = 0$ (price follows the same model in both categories)

$H_1: H_0$ is not true (price does not follow the same model in both categories).

This test is known as the **Chow test**. In Gretl, once we have estimated a model by OLS (*Model / Ordinary Least Squares*), we can run the test from the results window by selecting *Tests / Chow Test*.

²³ Alternatively, the test could also be performed by estimating the same model with subsamples.

Problem set 5A (with solutions)

EXERCISE 5A.1. Consider the following gravity equation for bilateral trade flows. Exports from country i to country j are explained by the gross domestic product (y) of i and j and geographical and cultural variables, including the geographical distance (D) between the capitals of i and j and three dummy variables: common border (*contig*), which is equal to one if the countries share a border and zero otherwise; common language (*comlang_off*), which is equal to one if the countries have the same official language and zero otherwise; and colonial relationship (*col_to*), which is equal to one if the countries have a colonial relationship or had one in the past. The following model was estimated using the data file taken from the CEPII “gravity” data set (www.cepii.fr) for a sample of 17,088 export flows in 2006 (<http://www.cepii.fr>):

$$\ln X_{ij} = \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j + \beta_3 \ln D_{ij} + \beta_4 \text{Contig}_{ij} + \beta_5 \text{comlang_off}_{ij} + 6 \text{col_to}_{ij} + u_{ij}$$

The following table of results was obtained:

Model: OLS, using observations 1-30569 ($n = 17,088$)
Dependent variable: $\ln x$

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
const	-10.423	0.251	-41.51	0.000	***
$\ln Y_i$	1.247	0.009	146.2	0.000	***
$\ln Y_j$	0.927	0.008	112.3	0.000	***
$\ln D$	-1.369	0.025	-54.56	0.000	***
<i>contig</i>	1.162	0.124	9.344	0.000	***
<i>col_to</i>	0.224	0.193	1.155	0.248	
<i>comlang_off</i>	1.187	0.054	21.84	0.000	***
Mean of the dep. var.	1.200	S.D. of the dep. var.	4.119		
Sum of sq. residuals	103530.200	S.D. of the regression	2.462		
<i>R</i> -squared	0.643	Adjusted <i>R</i> -squared	0.643		
$F(6, 17081)$	5123.366	<i>p</i> -value (of F)	0.000		
Log-likelihood	-39638.730	Akaike criterion	79291.460		
Schwarz criterion	79345.680	Hannan-Quinn crit.	79309.33		

a) Interpret the coefficients of the three dummy variables. Do countries that share a border trade with one another more than pairs of countries that do not share a border? Quantify the effect.

b) The model was re-estimated adding an interaction between the common border (*contig*) and common language (*comlang_off*) variables. The estimated coefficients for *contig* and (*contig*comlang_off*) are, respectively (standard errors in parentheses):

$$\begin{array}{r} 1.513 \quad -0.892 \\ (0.155) \quad (0.236). \end{array}$$

Compute the partial effect of a common border on exports in the expanded model.

c) The model was re-estimated, adding an interaction between a continuous variable ($\ln D = \text{distance in natural logs}$) and the colonial relationship. The estimated coefficient for the interaction is not statistically significant. How could the result be interpreted?

SOLUTION 5A.1:

a) Countries that share a border trade with one another a 116% more than pairs of countries that do not share a border, given the economic development, distance, language, and colonial relationships.

Countries that share a common language trade with one another a 119% more than pairs of countries that do not share a language, given the economic development, distance, common borders, and colonial relationships.

Countries that have shared a colonial relationship trade with one another a 22% more than pairs of countries that do not have shared a colonial relationship, given the economic development, distance, common borders, and language.

b) According to the extended regression function, the estimated partial effect on exports of a common border is given by:

$$\frac{\Delta \ln \widehat{X}_{ij}}{\Delta \text{contig}} = 1.513 - 0.892 \cdot \text{comlang_off}$$

Then, the effect is 62% for countries with a common border, while it is 151% for countries that do not share a common border.

c) In this case, if the interaction term is not statistically significant, it implies that the partial effect of the distance between countries on the exports is independent from the colonial ties.

EXERCISE 5B.2. We have a data set that contains information on the average wage by Autonomous Community (*salario*, expressed in euros per year) and by gender (*mujer*) is a dummy variable that is equal to one for women and zero for men). The data set also contains two additional dummy variables: *sur* is equal to one for the autonomous communities of southern Spain, zero otherwise; and *islas* is equal to one for Spain's islands, zero otherwise. The data are for 2016 for a sample of workers living and working in Spain. This information has been extracted from the Wage Structure Survey published by INE (www.ine.es).

a) Using the data set described in the previous paragraph, interpret the results obtained after estimating the following SRF by OLS:

$$\log(\widehat{\text{salario}}) = 10.128 - 0.261 \text{ mujer}$$

$$(0.024) \quad (0.034)$$

$$n = 34 \quad R^2 = 0.653$$

b) With the same data, the model whose results are shown below was estimated. What is the wage gap if the model is estimated with the wage variable in levels? Also interpret the constant of the regression.

Model: OLS, using observations 1-34
Dependent variable: *salario*

	Coefficient	Standard error	t-statistic	p-value	
const	25159.700	563.300	44.66	0.000	***
<i>mujer</i>	-5808.580	796.627	-7.291	0.000	***
Mean of the dep. var.	22255.400	S.D. of the dep. var.	3731.119		
Sum of sq. residuals	1.73e+08	S.D. of the regression	2322.545		
R-squared	0.624	Adjusted R-squared	0.613		
F (1, 32)	53.165	p-value (of F)	0.000		

c) Another model was estimated including the variables *sur* and *islas*. In view of the results obtained in the following model, is the average wage significantly lower in the south than in the other regions?

Model: OLS, using observations 1-34

Dependent variable: $l_salario$

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
const	10.168	0.023	448.3	0.000	***
sur	-0.117	0.034	-3.407	0.002	***
mujer	-0.261	0.028	-9.178	0.000	***
islas	-0.109	0.045	-2.421	0.022	**
Mean of the dep. var.	9.997	S.D. of the dep. var.		0.164	
Sum of sq. residuals	0.206	S.D. of the regression		0.083	
R-squared	0.767	Adjusted R-squared		0.744	
F (3, 30)	32.991	p-value (of F)		0.000	

SOLUTION 5B.2.

a) According to the estimated slope parameter, women earn 26% less than men, on average.

b) According to the estimated slope parameter, women earn 5808.58 euros per year less than men. Regarding the estimated constant parameter, it suggests that the mean salary for men is 25159.7 euros per year.

c) The average wage in the south is 11.66% lower than the average salary in other regions, regardless of the gender and insularity. The t -statistic for the $\hat{\beta}_{islas}$ is -2.421, which is below the corresponding value of a Student's t distribution, considering two-tails, $34 - 3 - 1$ degrees of freedom and 5% level of significance ($c_{\frac{0.05}{2}}^{34-3-1} = 2.042$). Therefore, we can conclude that the average difference in salaries between south and other regions is statistically significant at the 5% level of significance.

Problem set 5B

EXERCISE 5B.1. Data on the annual sales and number of permanent employees of a sample of companies in Egypt in 2013 were obtained using data from the World Bank (World Bank Doing Business: <http://www.doingbusiness.org>). With this information, dummy variables were constructed to indicate whether the companies in the sample export or not, whether they have foreign capital and whether the top manager is a woman.

a) Interpret the results obtained in the following linear regression model, where the dependent variable is the logarithm of labour productivity (*llabpro*) and the explanatory variables are: the top manager's experience (*exper*), the company's age or years it has been operating (*age*) and two dummy variables that indicate whether the company exports (*exporter*) and whether it has foreign capital (*foreign*).

Model: OLS, using observations 1-2897 ($n = 2408$)

Dependent variable: *llabpro*

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
const	11.317	0.062	182.8	0.000	***
<i>age</i>	-0.013	0.002	-6.157	0.000	***
<i>exporter</i>	0.605	0.074	8.140	0.000	***
<i>exper</i>	0.006	0.003	2.087	0.037	**
<i>foreign</i>	0.192	0.102	1.875	0.061	*
Mean of the dep. var.	11.315	S.D. of the dep. var.		1.419	
Sum of sq. residuals	4632.163	S.D. of the regression		1.388	
<i>R</i> -squared	0.044	Adjusted <i>R</i> -squared		0.043	
<i>F</i> (4, 2403)	27.752	<i>p</i> -value (of <i>F</i>)		0.000	
Log-likelihood	-4204.494	Akaike criterion		8418.987	
Schwarz criterion	8447.920	Hannan-Quinn criterion		8429.511	

b) Dummy variables were also constructed from the company size variable (*size_cat*), which classifies companies in three categories (large=*cat_1*, medium=*cat_2*, small=*cat_3*) based on the number of employees. Interpret the results obtained after estimating the following extended model:

Model: OLS, using observations 1-2897 ($n = 2408$)

Dependent variable: *llabpro*

	<i>Coefficient</i>	<i>Standard error</i>	<i>t-statistic</i>	<i>p-value</i>	
<i>const</i>	11.218	0.066	169.6	0.000	***
<i>age</i>	-0.013	0.002	-6.231	0.000	***
<i>exporter</i>	0.526	0.079	6.686	0.000	***
<i>exper</i>	0.004	0.003	1.631	0.103	
<i>foreign</i>	0.168	0.103	1.638	0.102	
<i>Dsize_cat_1</i>	0.263	0.064	4.085	0.000	***
<i>Dsize_cat_2</i>	0.239	0.083	2.869	0.004	***
Mean of the dep. var.	11.315	S.D. of the dep. var.		1.419	
Sum of sq. residuals	4596.945	S.D. of the regression		1.384	
<i>R</i> -squared	0.051	Adjusted <i>R</i> -squared		0.049	
<i>F</i> (6, 2401)	21.693	<i>p</i> -value (of <i>F</i>)		0.000	
Log-likelihood	-4195.305	Akaike criterion		8404.610	
Schwarz criterion	8445.115	Hannan-Quinn criterion		8419.343	

c) Why were two dummy variables created rather than three? Interpret the coefficients of the variables *Dsize_cat_1* and *Dsize_cat_2*.

d) Why are *exper* and *foreign* no longer statistically significant in the above model?

EXERCISE 5B.2. Search online for a cross-sectional data set for three economic and business variables you think may be related.²⁴ Using that data set, perform the following tasks:

a) Using EXCEL, save the four variables in columns, naming and sorting them, together with an index variable $i = 1, 2, \dots, N$ to represent the cross-sectional dimension. Do not forget to state the source of the data and the meaning of each variable and its units of measurement. In addition, construct a fourth dummy variable that is equal to one or zero depending on whether the individuals i in the sample belong to a particular category (e.g. gender, race, nationality, continent, sector, developed countries, language, etc.)

b) Estimate by OLS and interpret the corresponding SRF based on a model of the type:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_0 D + u \quad (\text{b1}),$$

where Y is the dependent variable, X_1 and X_2 are the continuous explanatory variables and D is a dummy explanatory variable. Use economic theory or logical reasoning to explain which is the dependent variable and which the explanatory variable(s).

c) Based on the proposed model (b1), specify, explain and perform an overall significance test of the regression.

d) Based on the proposed model (b1), specify, explain and perform a Chow test.

e) Based on the variables available, specify and estimate a model to test whether or not the marginal effect of X_2 on Y depends on whether i belongs to one of the categories included in the dummy variable D .

e.1 Specify and perform that test.

²⁴ Possible data sources: Gapminder (www.gapminder.org), Goolzoom (www.goolzoom.es), Instituto Nacional de Estadística (www.ine.es), Eurostat (<https://ec.europa.eu/eurostat/data/database>), OECD (<https://stats.oecd.org/>), World Bank (<https://data.worldbank.org/>), UNCTADSTAT (<https://unctadstat.unctad.org/>), FAOSTAT (<http://www.fao.org/faostat/en/#data>), others (<https://www.economicsnetwork.ac.uk/links/sources>, <https://db.nomics.world/>).

e.2 Graph the relationship between Y and X_2 according to the estimated model.

EXERCISE 5B.3. In the following econometric model, we are interested in explaining the average mark obtained by a group of university students:

$$\text{Mark} = \beta_0 + \beta_1 H\text{hours} + \beta_2 H\text{Smark} + \beta_3 S\text{network} + \beta_4 \text{Female} + \beta_5 (S\text{network} \cdot \text{Female}) + u$$

where:

- *Mark*: the average mark obtained at the university (from 1 to 4 points),
- *Hours*: the average number of hours per week students spend preparing the subjects,
- *HSmark*: the average mark obtained in high school (from 1 to 4 points),
- *Snetworks*: the average number of hours per week students spend on social media,
- *Female*: dummy variable that takes the value 1 if the student is a woman and 0 otherwise.

According to this econometric model, what would be the expected effect of an additional hour per week spent on social media on the average university mark, *ceteris paribus*?

Multiple-choice questions (Topic 5)

5.1. Determine which of the following statements are correct, considering the following specification:

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 educ * female + u$$

where *wage* is the monthly salary in dollars, *educ* refers to the years of education, and *female* is a dummy variable that takes the value 1 for women and 0 otherwise.

- a) The average wage difference between women and men is given by the parameter β_2 .
- b) The wage variation resulting from each additional year of education is given by the parameter β_1 for women.
- c) The wage variation resulting from each additional year of education is given by the parameters $(\beta_1 + \beta_3)$ for men.
- d) All the above answers are incorrect.

5.2. The following model tries to explain the migratory balance, defined as the difference between immigrants and emigrants of the European Union member states (*MB*), in terms of the following explanatory variables:

- *EMPL*: Employment rate (% of employees with respect to the total population between 16 and 64 years-old)
- *AGGL*: Agglomeration rate (total immigration with respect to total population)
- *E_TERC*: Employment in the tertiary sector (% of total employment in the service sector with respect to total employment in the country in time *t*).
- *NEUMS*: New European Member State (takes the value of 1 if the country joined the EU in 2004 or after; and 1 if the country belongs to the EU before 2004)

All data are from 2019 and were obtained from Eurostat and the World Bank. In parenthesis we present the standard errors.

$$\widehat{MB}_i = -0.006 + 0.018 \text{ EMPL}_i + 0.0729 \text{ E_TERC}$$

$$\quad (0.009) \quad (0.012) \quad (0.089)$$

$$+ 0.483 \text{ AGGL} + 0.176 (\text{AGGL} \cdot \text{NEUMS})$$

$$\quad (0.092) \quad (0.079)$$

$$n = 27; R^2 = 0.882$$

Considering the above regression model, select the correct answer:

- a) The dummy variable NEUMS is capturing potentially different constant terms in the model depending on the country.
- b) Irrespective of the employment and the agglomeration rates, countries that joined the EU before 2004 have a lower migratory balance than the rest of the EU countries.
- c) In the New European Union Member States, the agglomeration effect has a higher impact on the migration balance than in the rest of EU countries, holding the employment rate and the employment in the tertiary sector constant.
- d) None of the above answers is correct.

5.3. From the model of the exercise 5.2.

$$MB = \beta_0 + \beta_1 \text{EMPL} + \beta_2 \text{E_TERC} + \beta_3 \text{AGGL}$$

$$+ \beta_4 (\text{AGGL} \cdot \text{NEUMS}) + u$$

we want to estimate the effect of the agglomeration effect on the migratory balance of the EU member states. To do that we must test the following null hypothesis,

- a) $\beta_3 = 0$
- b) $\beta_3 = 0; \beta_4 = 0$
- c) Both of the above answers are correct.
- d) $\beta_4 = 0$

5.4. The number of dummy variables to be included in a regression model with a constant is equal to,

- a) The number of categories minus one to avoid the dummy variable trap.
- b) The number of categories plus one to avoid the dummy variable trap.
- c) The number of categories to avoid the dummy variable trap.
- d) None of the above answers is correct.

5.5. Consider that we want to test if labor productivity (lp) and its relationship with the technological level of a firm ($tech$) is different for firms with an international expansion (part of its production is sold abroad via either exports or foreign investments) than those whose market is domestically oriented. To do that, we take a sample of n domestic firms. Let define $D_i = 1$ if firm i sells part of its production abroad, and $D_i = 0$ if it sells its production domestically. From the following models,

$$lp_i = \alpha_1 + \alpha_2 tech_i + \delta_1 D_i + u_{1i} \quad (1)$$

$$lp_i = \beta_1 + \beta_2 tech_i + \delta_2 D_i + \delta_3 (D_i \cdot tech_i) + u_{1i} \quad (2)$$

We can infer

- a) Irrespective of the technological level, the labor productivity is higher for firms selling abroad than for firms selling abroad, as far as $\delta_1 > 0$ from Model (1).
- b) An improvement in the technological level will enhance labor productivity more in firms selling abroad than in those domestically-market oriented, if $\delta_3 > 0$ from Model (2).
- c) The null hypothesis to test from Model (2) that the effect of the technological level on labor productivity does not depend on whether the firm sells abroad or not is $\delta_3 = 0$.
- d) All the above answers above are correct.

5.6. From Model (1) in the previous question, we can deduce the following,

- a) $E[lp_i | D_i = 1, tech_i] = \beta_1 + \delta_1$
- b) $E[lp_i | D_i = 1, tech_i] - E[lp_i | D_i = 0, tech_i] = \alpha_1 + \delta_1$
- c) $E[lp_i | D_i = 1, tech_i] - E[lp_i | D_i = 0, tech_i] = \delta_1$
- d) None of the above answer is correct.

5.7. If from Model 2 of question 5.5. we want to test the hypothesis that the productivity level is independent of whether the firm sells or not abroad, the null should be written as,

- a) $\delta_2 = 0; \delta_3 = 0$.
- b) $\delta_3 = 0$.
- c) Both a) and b) are correct.
- d) None of the above answers is correct.

CHAPTER 6

DISCRETE CHOICE MODELS

6.1. Introduction

The econometric models that we have seen in previous chapters studied the relationship between continuous and quantitative dependent variables – e.g. annual sales, labour productivity, exports, etc., and explanatory variables that were either quantitative, qualitative (or dummy) or a mixture of both (interaction terms). However, many economic phenomena of interest have a qualitative nature – e.g. political colour of government, mode of transportation, rejection or acceptance of a project or a loan, etc. In these cases, we may be interested in the factors behind the making-decision of individuals, firms, governments, etc. Models developed for this purpose, where the dependent variable is qualitative, are known as discrete choice models (DCM). Discrete choice modelling can help us understand why some people buy houses while others rent, or why some workers choose to go to work by car instead of using the train or the bus.

In DCM, the model may represent two or more exclusive options (binary-choice models and multiple-choice models, respectively). In our example, 40 students from a secondary school in Barcelona have the possibility of choosing a second foreign language as an optional subject. The results of this binary decision are depicted in Figure 18. These students must also opt for one of the four types of Baccalaureates (upper secondary education): Arts, Sciences, Humanities or Social Sciences. This latter choice would be represented by a multiple-choice model instead, as shown in Figure 19.

Figure 18. Binary-choice model

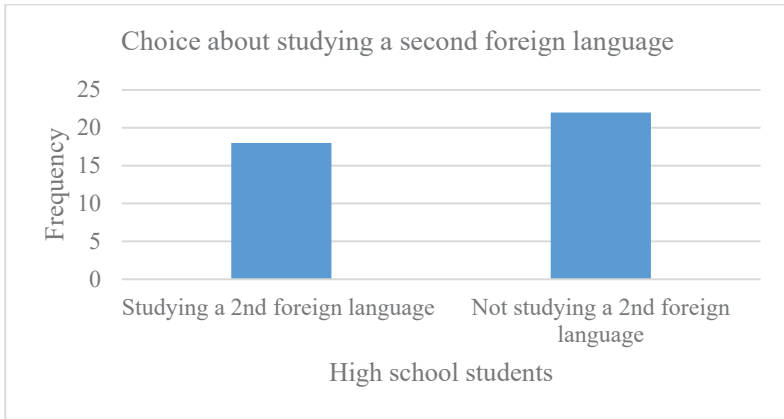
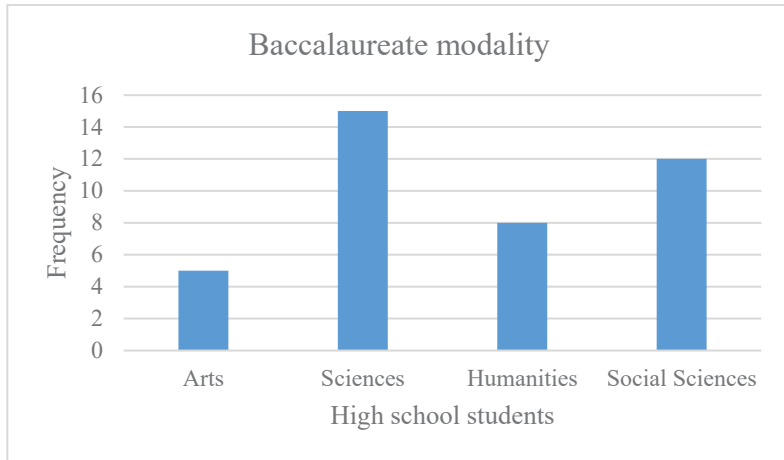


Figure 19. Multiple-choice model



Given that this is an introductory course, we will focus exclusively on the study of binary-choice models. In this case, the response variable has only two possible outcomes and is assigned a value of 1 for all observations in the data for which the event of interest has happened (“success”) and 0 for all other observations (“failure”).

$$y = \begin{cases} 1 & \text{If the event occurs (with probability } p) \\ 0 & \text{Otherwise (with probability } 1 - p) \end{cases}$$

Whenever the variable we want to model is qualitative, we can think in terms of probabilities. The objective of the regression is to estimate changes in the probability of “success”, p , when attributes vary. The regression model must include at least one explanatory variable, x , as the probability of success ($y = 1$) depends on the value of x . That is,

$$P(y = 1|x) = P(y = 1|x_1, x_2, \dots, x_k)$$

Where x_1, x_2, \dots, x_k denote the set of explanatory variables, which can be categorical or quantitative.

For example, if y represents whether or not to study a second foreign language, x may include individual factors, such as parents’ level of education, country of origin or average score in language-related subjects.

There are three approaches to estimating a binary-choice regression model:

- a) The linear probability model (LPM)
- b) The logit model
- c) The probit model

In the LPM, we assume that the probability of “success” can be represented similarly to the linear regression model. Hence, assuming $E(u_i|x) = 0$, we obtain the conditional probability of y_i being equal to one,

$$E(y_i|x_1, x_2, \dots, x_k) = P(y = 1|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

An advantage of the LPM is that it is easy to estimate. Under the Gauss-Markov assumptions of the MLR model, we can obtain unbiased estimations of the parameters of the model by OLS. We can also directly apply everything we know about hypothesis testing (t -tests, F -tests, confidence intervals, etc.) and goodness-of-fit.

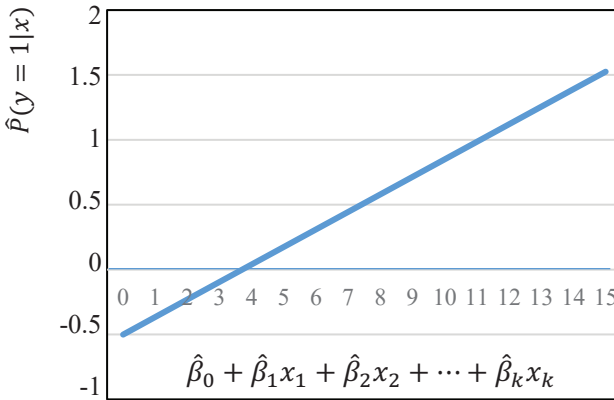
The interpretation of the β parameters in the LPM is straightforward, as they capture the change in the probability that the event of interest will occur when the associated explanatory variable changes by one unit, keeping all other covariates constant (partial effect):

$$\frac{\Delta P(y_i = 1|x_1, x_2, \dots, x_k)}{\Delta x_j} = \beta_j$$

However, the LPM has some drawbacks that should be taken into account. The two most important limitations are:

- i) The LPM measures the probability of an event occurring, given x , so its values must lie between 0 and 1. However, the fitted probabilities in this model can be less than zero or greater than one. This is a crucial problem with the LPM that we need to solve, because it does not make sense to have a probability below 0 or above 1.
- ii) The partial effect of covariates (when appearing in level form) is constant, which is not a realistic assumption. The probability of the event occurring when an explanatory variable increases by one unit, changes by the same amount no matter whether the value of $\beta_0 + \beta_1x_1 + \dots + \beta_kx_k$ is high or low.

Figure 20. Linear Probability Model



These limitations of the LPM are overcome using nonlinear binary response models. In general, these models can be represented as follows,

$$P(y = 1|x) = G(\beta_0 + \beta_1x_1 + \dots + \beta_kx_k) = G(z)$$

Where G is a nonlinear function taking values between zero and one $0 < G(z) < 1$ for all real numbers $z = \beta_0 + \beta_1x_1 + \dots + \beta_kx_k$.

The nonlinear functions for G most commonly suggested in the literature are the logistic cumulative distribution, the logit model, and the standard normal cumulative distribution, the probit model. Thus,

- In the logit model, $G(z) = \frac{\exp(z)}{1+\exp(z)} = \frac{1}{1+\exp(-z)} = \Lambda(z)$

- In the probit model,

$$G(z) = \Phi(z) = \int_{-\infty}^z \phi(v) dv = (2\pi)^{-1/2} \int_{-\infty}^z \exp\left(-\frac{z_i^2}{2}\right) dz_i$$

▲ *Standard normal $N(0,1)$ density function*

In both cases, the estimated probability, $P(y = 1|x)$ varies with x , but never exceeds the 0 – 1 interval.

- Example of a logit model:

Suppose we have only one regressor and $z = -1 + 2x$. We want to determine the probability that $y = 1$ when $x = 0.3$.

$$z = -1 + 0.6 = -0.4$$

$$P(y = 1|x = 0.3) = P(z \leq -0.4) = \Lambda(-0.4)$$

$$P(y = 1|x = 0.3) = \frac{1}{1 + e^{0.4}} = 0.401$$

- Example of a probit model

Suppose we have only one regressor and $z = -1 + 2x$. We want to determine the probability that $y = 1$ when $x = 0.3$.

$$z = -1 + 0.6 = -0.4$$

$$P(y = 1|x = 0.3) = P(z \leq -0.4) = \Phi(-0.4)$$

$$P(y = 1|x = 0.3) = 0.3446$$

We can obtain this result using the table of the standard normal CDF or using Excel: =DISTR.NORM.ESTAND(-0.4)

Table 24. Table of the Standard Normal Cumulative Distribution Function

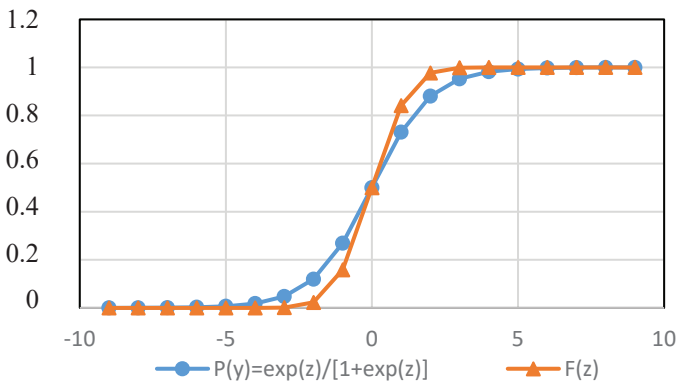
z	$\Phi(z)$
-3.4	0.000337
-3.3	0.000483
-3.2	0.000687
-3.1	0.000968
-3	0.00135
-2.9	0.001866
-2.8	0.002555
-2.7	0.003467
-2.6	0.004661
-2.5	0.00621
-2.4	0.008198
-2.3	0.010724
-2.2	0.013903
-2.1	0.017864
-2	0.02275
-1.9	0.028717
-1.8	0.03593
-1.7	0.044565
-1.6	0.054799
-1.5	0.066807
-1.4	0.080757
-1.3	0.0968
-1.2	0.11507
-1.1	0.135666
-1	0.158655
-0.9	0.18406
-0.8	0.211855
-0.7	0.241964
-0.6	0.274253
-0.5	0.308538
-0.4	0.344578
-0.3	0.382089
-0.2	0.42074
-0.1	0.460172
0	0.5
0.1	0.539828
0.2	0.57926
0.3	0.617911

We can easily extend the logit and probit regression models by including additional explanatory variables.

In the logit and probit regression models the relationship between the conditional probability and the covariates are nonlinear. The marginal effect of a change in the explanatory variables on $P(y = 1|x)$ will depend on the specific value of x . Specifically, it will be smaller for extreme values and greater for values in the central part of the distribution, as can be seen in the following figure.

In contrast to linear regression, logistic regression does not require the residuals of the model to be normally distributed or homoskedastic.

Figure 21. Logistic function and cumulative normal function



As the logistic and the cumulative normal are nonlinear functions, the coefficients (the β s) cannot be estimated by OLS. Instead, we should use nonlinear regression methods, such as maximum likelihood (ML) estimation. The ML method selects the value of the coefficients, β , that maximises the likelihood that our data fit the assumptions made about the distribution of the dependent variable. The intuition behind this estimation method is that the ML estimates are the values of the β s that best describe the full distribution of the data.

In most cases, calculating the estimated parameters using the ML method requires complex operations and iterations, although in most cases these are easily obtained using a package routine.

6.2. Statistical inference

- Simple hypothesis testing: We can construct t test and confidence intervals, as in the linear regression model. We calculate estimates of the model parameters and standard errors for the estimates. Confidence intervals are formed in the usual way, but we use standard normal z -values rather than critical values from t distribution.

- For exclusion restrictions: Rather than the F -statistic test, we use the likelihood ratio (LR). Computing the LR requires estimating both models: the unrestricted model and the restricted model (true under the null hypothesis). Like the F -statistic test, this test is based on the difference between the log-likelihood functions from the unrestricted model, l_{ur} , and the restricted model, l_r .

Steps to follow:

1. Specify the null and alternative hypotheses. E.g.:

$$H_0: \beta_1 = 0, \beta_2 = 0 \quad x_1 \text{ and } x_2 \text{ have no jointly significant effect on } y, \text{ once the effect of } x_3 \text{ has been controlled for}$$

$$H_1: H_0 \text{ is not true} \quad x_1 \text{ and } x_2 \text{ have a jointly significant effect on } y, \text{ once the effect of } x_3 \text{ has been controlled for}$$

2. Estimate the unrestricted (ur) model and keep the log-likelihood (l_{ur}):

$$\text{Unrestricted model: } P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

3. Estimate the restricted model (r) and keep the log-likelihood (l_r)

$$\text{Model restricted by } H_0: P(y = 1|x) = G(\beta_0 + \beta_1 x_1)$$

4. Compute the likelihood ratio, $LR = 2 * (l_{ur} - l_r)$

Under the H_0 , $LR \rightarrow \chi_q^2$, where q is the number of restrictions

5. Compute the LR p-value under the null: $P(\chi_q^2 > LR)$

6. Conclude

6.3. Marginal effects

The marginal effects of a continuous covariate on the conditional probability ($P(y = 1|x)$) is obtained from the partial derivative:

$$\frac{\partial P(y_i = 1|x_1, x_2, \dots, x_k)}{\partial x_j} = G'(\beta_0 + x_i' \beta) \cdot \beta_k = g(\beta_0 + x_i' \beta) \cdot \beta_k$$

Since both the logistic cumulative distribution and the standard normal cumulative distribution are positive functions, the sign of the coefficients in the logit and the probit regression models will indicate the direction of the marginal effect. However, unlike the LPM, marginal effects in logit and probit models are not constant but depend on the values of x for observation i .

General rule for obtaining marginal effects. Steps:

- 1.- Using the estimated regression equation, we obtain $P(Y = 1|x_0)$ at the original value of x (from which we want to measure the effect).
- 2.- We calculate the probability again at $x_j + \Delta x_j$.
- 3.- The difference between the two probabilities will be the effect of the change in x_j , on the probability of success.

Table 25. Possibilities to summarise marginal effects

<ul style="list-style-type: none"> • Marginal effect for the “mean type” 	Considering the average value of the explanatory variables (i.e., $x_i = \bar{x}_i$), we estimate the effect of one unit increase in one covariate	$G'(\beta_0 + \beta_1 \bar{x}_1 + \dots + \beta_K \bar{x}_K) \cdot \beta_k$
<ul style="list-style-type: none"> • Average marginal effect 	We calculate the average marginal effects across all i	$\frac{1}{N} \sum_{i=1}^N G'(\beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK}) \cdot \beta_k$
<ul style="list-style-type: none"> • Marginal effect for some specific type 	We estimate the marginal effect of a specific profile ($x_{ij} = x_{ij}^*$)	$G'(\beta_0 + \beta_1 x_{i1}^* + \dots + \beta_K x_{iK}^*) \cdot \beta_k$

6.4. Odds ratio

If p is the probability of success and $1 - p$ the probability of failure, we can define the odds as,

$$\frac{p}{1 - p} = \frac{\text{probability of success}}{\text{probability of failure}}$$

The odds tell us the number of times the event is more likely to occur than not to occur.

Example: Let the probability of success be 0.8, in which case the probability of failure is 0.2 (1-0.8). The odds are defined as the ratio of the probability of success to the probability of failure. In our example, the ratio is 4 (0.8/0.2), so the odds are 4 to 1. If the probability of success is 0.5, the odds are 1 to 1.

From the logit regression it is easy to define the odds, as

$$\frac{p}{1 - p} = \frac{P(y = 1|x)}{1 - P(y = 1|x)} = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK})$$

The odds ratio is the ratio of the odds of each of two different values of an explanatory variable.

In the case of a binary explanatory variable, D , the odds ratio is given by

$$\text{OR} = \text{odds ratio} = \frac{\text{odds when } D = 1}{\text{odds when } D = 0} = \exp^{\delta}$$

where δ is the coefficient on the binary regressor of the logit model.

The odds ratio for a change in a continuous explanatory variable, Δx_j , is given by

$$\text{OR} = \text{odds ratio} = \exp^{\beta_j \Delta x_j}$$

Note that when the odds ratio for a regressor is greater than 1, it describes a positive relationship between the regressor and the probability of success. Conversely, if the odds ratio is smaller than 1 implies a negative impact on the regressor and the probability of success.

We can also work with the natural logarithm of the odds. Doing this we model the log odds as a linear function of the coefficients.

$$\log\left(\frac{p}{1-p}\right) = z = \beta_0 + \beta_1 x_{i1} + \dots + \beta_K x_{iK}$$

Because log is a monotonic transformation, the greater the odds ratio, the greater the log of OR and vice versa.

6.5. Goodness-of-fit

Conventional R^2 is not very meaningful in probit or logit models. Many alternative measures have been proposed to evaluate the goodness-of-fit. The most widely used are:

- Count R^2 : $\frac{\# \text{ of correct predictions}}{\text{Total \# of observations}}$
- McFadden R^2 or Pseudo $R^2 = 1 - \frac{\log L}{\log L_0}$

where $\log L$ is the max. log-likelihood value of the model of interest, while $\log L_0$ represents the max. log-likelihood value of the model with only a constant term.

- Count R^2 and Pseudo R^2 increase as the fit of the model improves
- Both are bounded by 0 and 1
- If all the slope coefficients are zero, then $L = L_0 \rightarrow$ Pseudo $R^2 = 0$
- If $prob_i = y_i$ for all i , then $\log L = 0$ (the log of 1) and Pseudo $R^2 = 1$ (indicative of a “perfect fit”)

6.6. Model selection and the Akaike Information Criterion

Our goal now is to choose one among a set of logit or probit candidate models with the same response variable and observations, but different explanatory variables. For probit and logit models the model selection can be done through the Akaike Information Criterion (AIC).

AIC is a method that allows us to evaluate a model fit based on its log-likelihood and complexity (number of covariates). AIC is used to compare different possible models and determine which one is the best fit for the data based on the maximum likelihood estimate, L , and the number of parameters (explanatory variables). AIC is obtained through the following formula,

$$AIC = 2k - 2 \ln(L)$$

Lower AIC scores are better, so when comparing different models, we will select the model with the lowest AIC.

If a model is more than 2 AIC units lower than another, it is considered significantly better than that other one.

Problem set 6A (with solutions)

EXERCISE 6A.1 To analyse the role of macroeconomic stability and openness in the political colour of governments, we employ data for 36 democratic OECD and/or EU-member countries in 2018 from Comparative Political Data Set²⁵. With this data, we estimate a logit model using *gov_party_right* as dependent variable. This variable takes the value 1 when the government cabinet is dominated by right-wing and centre parties, and zero otherwise. To explain the political complexion of governments we use the following independent variables:

- *unempl*: Unemployment rate, percentage of civilian labour force.
- *openc*: Openness of the economy, measured as total trade (sum of imports and exports) as a percentage of GDP
- *debt_hist*: Gross general government debt as a percentage of GDP
- *sstran*: Social security transfer as a percentage of GDP

The following table presents the results of the logit regression:

Dependent variable: <i>gov_party_right</i> . Year: 2018						
Explanatory variables	Coef.	Std. Error	z	P> z	95% Conf. Interval	
<i>unemp</i>	-0.272	0.159	-1.71	0.088	-0.584	0.041
<i>openc</i>	-0.008	0.007	-1.26	0.209	-0.022	0.004
<i>debt_hist</i>	0.013	0.014	0.89	0.375	-0.015	0.043
<i>sstran</i>	-0.055	0.138	-0.41	0.689	-0.326	0.215
<i>Constant</i>	3229	1.979	1.63	0.103	-0.649	7.108
<i>Observations</i>	36					

²⁵ The Comparative Political Data Set 1960-2018 (CPDS) is a collection of political and institutional data, which have been assembled in the context of the research projects “Die Handlungsspielräume des Nationalstaates” and “Critical junctures. An international comparison” directed by Klaus Armingeon and founded by the Swiss National Science Foundation. This data set consists of annual data for 36 democratic OECD and/or EU-member countries. The CPDS is available on line: <https://www.cpds-data.org/index.php/data#CPDS>. Armingeon, Klaus, Virginia Wenger, Fiona Wiedemeier, Christian Isler, Laura Knöpfel, David Weisstanner and Sarah Engler. 2020. Comparative Political Data Set 1960-2018. Bern: Institute of Political Science, University of Berne.

From the above regression estimates, answer the following questions:

- a) Given the following values in the explanatory variables: $openc = 50$; $debt_hist = 100$; $sstran = 10$; $unemp = 10$, what is the estimated probability that a country has a government cabinet dominated by right-wing and center parties?
- b) Holding the explanatory variables fixed at the mean value ($openc = 116.836$; $debt_hist = 78.275$; $sstran = 12.941$; $unemp = 6.031$), what is the estimated difference in the probability of having a government cabinet dominated by right-wing and centre parties when the unemployment rate increases by 2 points.
- c) Fill the gaps in the following sentence: Holding other regressors constant, as the rate $openc$ increases by ten units, the log-odds of having a “centre-right government” (as opposed to a “left-wing government”) decreases by _____ times (a _____ % drop).

SOLUTIONS 6A.1:

a) According to the estimated logit model, the probability that a country has a government dominated by right-wing and center parties ($Y = 1$) is given by the following expression,

$$P(Y = 1|\mathbf{X}) = \Lambda(\hat{z}) = \Lambda(3.229 - 0.272 unemp - 0.008 openc + 0.013 debt_hist - 0.055sstran)$$

$$\Lambda(\hat{z}) = \frac{\exp(\hat{z})}{1 + \exp(\hat{z})} = \frac{1}{1 + \frac{1}{\exp(\hat{z})}}$$

For $unemp = 10$, $openc = 50$, $debt_hist = 100$, and $sstran = 10$, $z = 0.859$, and the probability that a country has a government dominated by right and centre parties is equal to: $P(Y = 1|\mathbf{X}) = 0.702$ (70.2%).

- b) At the mean value of the explanatory variables, the probability of having a centre-right government is equal to $P(Y = 1|\mathbf{X}) = \Lambda(0.9597) = 0.723$ (72.3%). Holding the values of $openc$, $debt_hist$ and $sstran$ fixed, an increase of 5 points in the unemployment rate reduces the expected probability that gov_party_right equals 1 to 0.602 (60.2%). Therefore, the expected probability of having a centre-right government drops 10 percent.
- c) 0.08 (8%).

EXERCISE 6A.2 We are interested in understanding the relationship between the probability of being unemployed and the years of education, marriage status and gender of individuals currently in work or actively searching for a job.

To study the factors behind this relationship, we consider the following variables:

- *unemp*: which is equal to one if the worker is unemployed and zero otherwise
- *yearseduc*: which represents the years of education
- *married*: which is equal to one if the worker is married and zero otherwise
- *woman*: which takes the value of 1 when the worker declares herself to be a woman, and zero otherwise

Estimating the probit model with a sample of 1256 Spanish citizens and using *unempl* as dependent variable provides the following results:

$$\hat{Y} = P(Y = 1) = \Phi(-0.461 - 0.101\text{yearseduc} + 0.347\text{married} + 0.367\text{woman})$$

$$\begin{array}{cccc} [0.105] & [0.010] & [0.084] & [0.091] \end{array}$$

Where Φ is the normal cumulative density function, and the values in brackets are the standard errors:

Based on the above results, answer the following questions:

- a) What is the probability that an unmarried man with 7 years of education will be unemployed?
- b) How does this probability change (compared to the previous question) if the person is a woman (*woman* = 1)?
- c) After considering years of education and gender, does marriage status have a significant impact on the probability of being unemployed?

SOLUTIONS 6A.2:

a) We estimate the probability that a single man with 7 years of education will be unemployed by substituting the specific values of the covariates in the probit regression. On doing so, we first obtain the value of z . Then we find the value of the Normal CDF at point z in the Standard Normal CDF table, or else enter the formula = DISTR.NORM.ESTAND(z) in Excel:

$$P(\text{unemp} = 1) = \Phi(-0.461 - 0.101 * 7 + 0.347 * 0 + 0.367 * 0) = \Phi(-1.168) = 0.121(12.1\%)$$

From the above results, we can say that for a person with the specified characteristics ($\text{yearseduc} = 12$; $\text{married} = 0$; $\text{woman} = 0$) the probability of being unemployed is 12.1%.

b) How does this probability change (compared to the previous question) if the person is a woman ($\text{woman} = 1$)?

According to the probit regression, the probability that a single woman with 12 years of education will be unemployed is 21.2%.

$$P(\text{unemp} = 1) = \Phi(-0.461 - 0.101 * 7 + 0.347 * 0 + 0.367 * 1) = \Phi(-0.801) = 0.212(21.2\%)$$

Comparing the answers to questions a) and b), we find that for a woman the probability of being unemployed is $0.212 - 0.121 = 0.091$ (9.1%) higher than for a man with the same years of education and marriage status.

c) We start by specifying the null and alternative hypotheses,

$$H_0: \beta_{\text{married}} = 0$$

$$H_1: \beta_{\text{married}} \neq 0$$

Then, to get the significant impact of marriage status we compute its z -score and compare it to a critical value for two-tailed p -value. We have the values of the estimated coefficients and the standard errors, so the z -score is equal to,

$$z = \frac{0.347 - 0}{0.084} = 4.131$$

For $\alpha = 0.05$, the critical value is 1.96. Given that our t -statistic is greater than the critical value, being married has a positive and significant impact on the probability of being unemployed.

EXERCISE 6A.3 In this problem, we investigate the effect that a family history of breast cancer has on the probability that a woman will purchase private insurance. The investigation uses a logit model, which shows the following results:

Explanatory variable	Definition	Estimate parameter	Standard error
<i>fam_cancer</i>	=1 if she has direct relatives that have suffered breast cancer, = 0 otherwise	0.35	0.10
<i>age</i>	Age in years	0.04	0.02
<i>years educ</i>	Years of education	0.03	0.01
C	Constant term	-0.05	0.02

- What is the odds ratio for getting a private insurance for every 10-year increase in age?
- What is the odds ratio for having a family history of cancer (*fam_cancer* = 1)

SOLUTIONS 6A.3:

a) The odds ratio for a change in a continuous variable, Δx_j , in a logit model is given by $\exp\beta_j\Delta x_j$.

For an increase of 10 years in age, the odds ratio is $OR = \exp(0.04 * 10) = 1.49$. Accordingly, we conclude that the odds of purchasing private insurance increases by 1.49 times (49%) for every 10 years increase in age.

b) What is the odds ratio for having a family history of cancer (*fam_cancer* = 1)

In this case, the odds ratio is given by $OR = \exp(0.35) = 1.42$. This can be interpreted that having a positive family history increase the odds for getting a private insurance 1.42 times (42% higher than in the case of not having a positive family history).

EXERCISE 6A.4 Read these statements carefully and say whether they are True or False. *Give reasons for your answer:*

- a) Like linear regression, logistic regression requires that the residual of the model be normally distributed.
- b) In a probit model, goodness-of-fit can be measured based on the percentage correctly predicted by the model.
- c) In logit and probit regression models, the direction of the effect will coincide with the sign of the parameters.
- d) An odds ratio between 0 and 1 describes a positive relationship between the regressor and the probability of success.

SOLUTIONS 6A.4:

- a) FALSE. In contrast to linear regression, logistic regression does not require the residuals of the model to be normally distributed.
- b) TRUE. In a probit model, goodness-of-fit can be measured based on the percentage correctly predicted by the model. A predicted value of success is defined if the predicted probability is more than 0.5 and zero otherwise.
- c) TRUE. Since both the logistic cumulative distribution and the standard normal cumulative distribution are positive functions, the sign of the coefficients in the logit and the probit regression models will indicate the direction of the marginal effect.
- d) FALSE. An odds ratio smaller than 1 implies a negative impact of the regressor on the probability of success. Conversely, if the odds ratio for a regressor is greater than 1, the relationship between the regressor and the probability of success is positive.

EXERCISE 6A.5 In order to predict whether second-year Economics students would be interested in a training course to reinforce their skills in maths applied to economics, we estimated several logistic models. In these models, we considered factors such as the overall average marks obtained in the first year of the degree programme (*firstyear_grade*), the average grade obtained in high school (*highschool_grade*), the score obtained in the university entrance exam (*score_univentrance*) and the number of credits already passed (*credits_passed*). The following table shows the results of this study (where the dependent variable is defined as $Y = 1$ if the student declared herself or himself interested in taking the mathematics course and zero otherwise).

Table 26. Logistic model regression results. Attending or no training course in Math.

	Model A	Model B
<i>firstyear_grade</i>	-0.003 (0.001)	-0.002 (0.001)
<i>credits_passed</i>	-0.004 (0.003)	-0.007 (0.004)
<i>highschool_grade</i>	0.0418 (0.009)	
<i>score_univentrance</i>	0.0056 (0.006)	
Number of observations	1054	1054
Log Likelihood	-401.09	-484.22

Based on the estimates from the above table,

- According to the AIC, which model would best fit the data sample?
- In model 1, can the average grade obtained in high school (*highschool_grade*) and the score in the university entrance exam (*score_univentrance*) be considered jointly significant?

SOLUTIONS 6A.5:

a) Lower AIC scores are better, so when comparing different models, we will select the model with the lower AIC.

$$\text{AIC in Model A: } AIC = 2k - 2 \ln(L) = 2(5) - 2(-401.09) = 812.18$$

$$\text{AIC in Model B: } AIC = 2k - 2 \ln(L) = 2(3) - 2(-484.22) = 974.44$$

In this case, we will choose Model A.

$$\text{b) } H_0: \beta_3 = 0, \beta_4 = 0$$

highschool_grade and *score_univentrance* have no jointly significant effect on Y , once the effect of *firstyear_grade* and *credits_passed* has been controlled for.

$H_1: H_0$ is not true

To test the above hypotheses, we first compute the likelihood ratio (LR), taking into account the likelihood ratio of the unrestricted and restricted models.

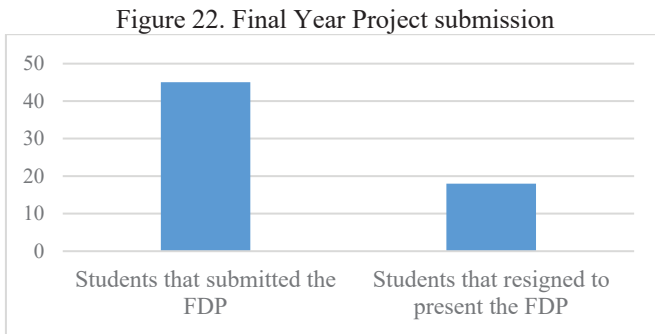
$$LR = 2 * (l_{ur} - l_r) = 2 * (-401.09 + 484.22) = 166.26$$

Critical value of the chi-square with 2 degrees of freedom at 5% significance level is equal to 5.99. Given that our LR statistic is greater than 5.99, we can reject the null at 5% significance level, concluding that the average grade obtained in high school (*highschool_grade*) and the score of the university entrance exam (*score_univentrance*) are jointly significant in the explanation of our dependent variable, once we take into account *firstyear_grade* and *credits_passed*.²⁶

²⁶ The critical values of the Chi-square distribution are presented in the Table A5, from Appendix A.

Problem set 6B

EXERCISE 6B.1 Some tutors in a Faculty of Economics want to analyse why so many final-year students in the Economics degree programme choose not to submit the Final Year Project (FYP) at the end of the course. They estimate a binary-choice model with information from the previous academic year, using the binary variable *submit*, which takes the value 1 for all students that submitted the FYP that year and 0 otherwise, as the dependent variable. According to our data, 63 students were enrolled in the Economics course that year, but only 45 submitted their Final Year Project at the end of the year. The following figure represents these data.



To study the factors behind this binary decision, they consider the following explanatory variables:

- Average grade of the student's academic record - on a scale of 1 to 4 (*avg_grade*),
- Whether or not the student is a repeater (*repeater* = 1 if the student is a repeater; 0 otherwise),
- The number of ECTS credits passed by the student (*n_passed*), and
- Whether the student is female, according to her ID card (*female* = 1) or not (*female* = 0).

The model has been estimated by OLS (LPM), using both logit and probit models. The outcomes are presented in the following table:

Dependent variable: *submit*

Explanatory variables	LPM	Logit model	Probit model
<i>n_passed</i>	0.268*** (0.0745)	0.268*** (0.0745)	0.159*** (0.0387)
<i>repeater</i>	-0.0244 (0.792)	-0.0244 (0.792)	-0.0650 (0.428)
<i>avg_grade</i>	0.042*** (0.005)	0.044*** (0.005)	0.045*** (0.003)
<i>female</i>	0.758 (0.803)	0.758 (0.803)	0.365 (0.442)
<i>Constant</i>	-62.05*** (17.18)	-62.05*** (17.18)	-36.77*** -8.967
<i>R</i> ²	0.4180	-	-
Pseudo <i>R</i> ²		0.3913	0.4004
log_likelihood	-	-22.944	-22.599
Observations	63	63	63

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

From the above regression estimates, answer the following questions:

a) According to the linear probability model outcomes, what is the expected effect of a one-point increase in *av_grade* on the probability of submitting the FYP, other factors being equal?

b) State the null hypothesis that, other things being equal, the variables *repeater* and *female* have no effect on the probability of submitting the FYP. If the log likelihood function obtained from the estimation of a logit model with only *n_passed* and *avg_grade* as explanatory variables is equal to -23.414, explain how to test the above null hypothesis through the likelihood ratio and show the results obtained.

c) According to the probit model estimates, what is the approximate change in the probability of submitting the FYP for a male, non-repeating student with an average score of 2.5, when his number of ECTS credits passed increases from 230 to 235?

d) From the above results, do you think the fact of being female has a significant impact on the probability of submitting the FYP, other things being equal? Give reasons for your answer.

EXERCISE 6B.2 Let Y be a dummy variable that takes the value one for all cross-sectional observations for which the event of interest happened and zero for all other observations. Explain how to use multiple regression to estimate the probability the event happens conditional on the values of x_1 and x_2 .

How would you estimate the marginal effects? Consider all possibilities remarking the limitations and the strengths.

EXERCISE 6B.3 Provide some examples in which the dependent variable has a quantitative meaning (with a binary outcome). State which explanatory variables you would include to explain these phenomena through a binary choice model.

EXERCISE 6B.4 Match each part of the following sentences with the correct answer:

In a logit or probit model, how much does $P(y = 1|x)$ change as we increase $z = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ (that is, how big are the marginal effects) when

- | | |
|-------------------------------------|-------------|
| 1. z is very high? | a. A little |
| 2. z is very low? | b. A lot |
| 3. z is neither big high nor low? | c. A little |

EXERCISE 6B.5 Describe step by step how we can use a binary-choice regression model to analyse whether the use of social media (such as WhatsApp, Facebook, Twitter or Instagram) has become more popular among people over the age of 65 since the outbreak of the COVID crisis, depending on their income level, years of education, whether they live alone or not, and whether they have direct relatives outside the city where they live. Show all options and mention all the limitations and the strengths.

Multiple-choice questions (Topic 6)

6.1. Which of the following statements are correct? In a logit model, the partial effects of the covariates...

- a) are constant, as in the probit model.
- b) are constant, as in the linear probability model.
- c) depend on x , as in the linear probability model.
- d) depend on x , as in the probit model.

6.2. In a binary-choice model $P(Y = 1|x) = G(\beta_0 + \beta_1x)$,

- a) We can interpret β as marginal effect of x on the probability that $Y = 1$.
- b) The slope parameter of the model, β , tells us the direction of the effect that an increase in x has on the probability that Y is equal to 1.
- c) β cannot be negative because a probability always takes a positive value between 0 and 1.
- d) β measures the probability that $Y = 1$.

6.3. In a linear probability model with the explanatory variables in levels, the slope coefficients

- a) have any interpretation, as the dependent variable is 0 or 1.
- b) give us the change in the dependent variable when the associated regressor changes by one unit, holding the rest of the covariates constant.
- c) represent the percentage change in the dependent variable.
- d) give us the direction of the marginal effect but do not have a direct interpretation in respect of the associated changes in the explanatory variable.

6.4. The logit model can be estimated by

- a) OLS.
- b) Maximum Likelihood.
- c) Both a) and b).
- d) None of the above answer is correct.

6.5. Which of the following options is NOT true about the goodness-of-fit in probit and logit models?

- a) The model with the highest Pseudo R^2 is preferred.
- b) Count R^2 and Pseudo R^2 are bounded by 0 and 1.
- c) Conventional R^2 is very meaningful in probit and logit models.
- d) The model with the lowest value of AIC is preferred.

6.6. In a probit or logit regression, $P(Y = 1|x) = G(\beta_0 + \beta_1 x)$. By changing the value of x from x_0 to x_1 we can get a new value of $P(Y = 1|x_1)$ in the range of

- a) $(0, +\infty)$
- b) $(-\infty, 1)$
- c) $(-\infty, +\infty)$
- d) $(0, 1)$

6.7. Which of the following statements is true?

- a) In logistic regression, the error terms are normally distributed.
- b) Linear regression errors must be normally distributed, as in probit regression.
- c) In both probit and logit regressions, the error values must be normally distributed.
- d) Errors must be normally distributed in linear regression but not in probit regression.

6.8. Suppose that in a probit or logit regression $P(Y = 1|x) = G(\beta_0 + \beta_1 x)$, the slope parameter β_1 is positive. Which of the following statements is correct?

- a) If x increases, the probability that $Y = 1$ increases for all initial values of x .
- b) If x increases, the probability that $Y = 1$ increases only for small values of x .
- c) If x increases, the probability that $Y = 1$ decreases for small values of x .

6.9. Which of the following statements is true?

- a) The probit model is used to predict continuous values of the dependent variable.
- b) In the logit model the sum of squares calculation is used to measure goodness-of-fit.
- c) AIC can be used to evaluate the performance of a logit model.
- d) Regression coefficients in probit models are estimated using the OLS method.

6.10. Which of the following statements is NOT true?

- a) The logit model explains a success-or-failure response variable in terms of at least one explanatory variable.
- b) If p is the proportion of successes, then the odds of a success are $p/(1-p)$, the ratio of the proportion of successes to the proportion of failures.
- c) The joint significance of regression in a probit model with more than one explanatory variable is tested using the F -statistic, like the joint significance test in linear regression.
- d) Statistical inference for logistic regression with one explanatory variable is like statistical inference for linear regression.

SOLUTIONS TO THE MULTIPLE-CHOICE QUESTIONS

Topic 1

1.1. (d) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$; $5 = \bar{y} - 3 \cdot 2$; $\bar{y} = 5 - 6 = -1$

1.2. (b)

1.3. (b)

1.4. (a) $\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = \frac{2}{0.8} = 2.5$; $\hat{\beta}_0 = 4 - 2.5 \cdot 3 = -3.5$

1.5. (b) $R^2 = \frac{ESS}{TSS} = \frac{0.60}{0.80} = 0.75$

1.6. (c)

1.7. (b)

1.8. (d)

1.9. (c)

1.10. (a)

1.11 (a) $x_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \sum_{i=1}^n y_i}{n-1} = \frac{100 - \left(\frac{15}{10}\right)25}{9} = 6.94$

1.12. (d)

Topic 2

2.1. (d)

2.2. (c)

2.3. (d)

2.4. (d)

2.5. (c)

2.6. (d)

2.7. (c)

2.8. (a)

2.9. (b)

2.10. (c)

2.11. (c)

Topic 3

3.1. (c)

3.2. (a)

3.3. (d)

3.4. (c)

3.5. (d)

3.6. (a)

3.7. (a) $1 \notin [0.487; 0.913]$

3.8. (b)

3.9. (b)

3.10. (b)

Topic 4

4.1. (d)

4.2. (a)

4.3. (b)

4.4. (c)

4.5. (c)

Topic 5

5.1. (d)

5.2. (c)

5.3. (b)

5.4. (a)

5.5. (d)

5.6. (c)

5.7. (a)

Topic 6

6.1. (d)

6.2. (b)

6.3. (b)

6.4. (b)

6.5. (c)

6.6. (d)

6.7. (d)

6.8. (a)

6.9. (c)

6.10. (c)

REFERENCES

Basic References

- Matilla García, M.; Pérez Pascual, P.; Sanz Carnero, B. (2017). *Econometría y Predicción*. McGraw Hill. UNED.
- Wooldridge, J.M. (2020). *Introductory Econometrics: A Modern Approach*. 7th Edition, CENGAGE Learning.

Supplementary References

- Baiocchi, G.; Distaso, W. (2003). GRET: Econometric software for the GNU generation. *Journal of Applied Econometrics*, 18(1), 105-110.
- Greene, W.H. (1999). *Análisis Económico*. 3rd edition, McGraw-Hill.
- Gujarati, D.N.; Porter, D.C. (2009). *Econometría*. 5th edition, McGraw-Hill.
- Ramanathan R. (2002). *Introductory Econometrics with Applications*. 5th edition, Harcourt College Publishers: Orlando, FL.
- Stock, J.H.; Watson, M.W. (2012). *Introduction to Econometrics*. 3rd Edition, Pearson.
- Verbeek, M. (2017). *A Guide to Modern Econometrics*. Wiley Custom.

Data sources

- Annual wage structure survey, Instituto Nacional de Estadística (INE), (<http://www.ine.es/>): Data_salarios2014ESP.gdt
- Autocasión (<https://www.autocasion.com/>): Data_Barcelona_cars_10july2021
- CEPII "gravity" (<http://www.cepii.fr>)
- EU R&D Scoreboard (<http://iri.jrc.ec.europa.eu/scoreboard16.html>): Data_RD_scoreboard.gdt
- European Social Survey (www.europeansocialsurvey.org)
- Eurostat (<https://ec.europa.eu/eurostat/data/database>): Data_cons_inc.xlsx
- Gapminder (www.gapminder.org): Data_Gapminder_2010.gdt
- Goolzoom (www.goolzoom.es)

- Hydrocarbons geoportal (<http://geoportalgasolineras.es>), Spanish Ministry of Energy, Tourism and Digital Agenda
- Inside Airbnb (<http://insideairbnb.com/>)
- Latin American Migration Project (LAMP), COL14 (2009) (<https://lamp.opr.princeton.edu/>)
- Nestoria (<https://www.nestoria.es/>): Data_Valencia_pisos.gdt (15 April 2018), Data_Palma_Mallorca_alquileres.gdt (27 August 2018)
- Simulated data: Data_marks.xls, Data_marks2.xls, Data_production.xlsx.
- World Bank (<https://data.worldbank.org/>): Data_convergence.gdt, Data_Internet2017_WB.gdt
- World Bank, Doing Business (<http://www.doingbusiness.org>)

Other sources of interest

- DB Nomics (<https://db.nomics.world/>)
- Food and Agriculture Organization of the United Nations, FAOSTAT (<http://www.fao.org/faostat/en/#data>),
- Organization for Economic Cooperation and Development, OECD (<https://stats.oecd.org/>)
- The Economics Network (<https://www.economicsnetwork.ac.uk/links/sources>)
- The Comparative Political Data Set 1960-2018, CPDS (<https://www.cpbs-data.org/index.php/data#CPDS>)
- United Nations Conference on Trade and Development, UNCTADSTAT (<https://unctadstat.unctad.org/>)

APPENDIX A.

CRITICAL VALUES (PERCENTILES) FOR STATISTICAL DISTRIBUTIONS

Table A1. Critical values for the Student's t distribution

	Level of significance (α)				
	0.1	0.05	0.01	0.025	0.005
One-tail	0.1	0.05	0.01	0.025	0.005
Two-tails	0.2	0.1	0.02	0.05	0.01
Degrees of freedom					
1	3.078	6.314	31.821	12.706	63.657
2	1.886	2.920	6.965	4.303	9.925
3	1.638	2.353	4.541	3.182	5.841
4	1.533	2.132	3.747	2.776	4.604
5	1.476	2.015	3.365	2.571	4.032
6	1.440	1.943	3.143	2.447	3.707
7	1.415	1.895	2.998	2.365	3.499
8	1.397	1.860	2.896	2.306	3.355
9	1.383	1.833	2.821	2.262	3.250
10	1.372	1.812	2.764	2.228	3.169
11	1.363	1.796	2.718	2.201	3.106
12	1.356	1.782	2.681	2.179	3.055
13	1.350	1.771	2.650	2.160	3.012
14	1.345	1.761	2.624	2.145	2.977
15	1.341	1.753	2.602	2.131	2.947
16	1.337	1.746	2.583	2.120	2.921
17	1.333	1.740	2.567	2.110	2.898
18	1.330	1.734	2.552	2.101	2.878
19	1.328	1.729	2.539	2.093	2.861
20	1.325	1.725	2.528	2.086	2.845
21	1.323	1.721	2.518	2.080	2.831
22	1.321	1.717	2.508	2.074	2.819
23	1.319	1.714	2.500	2.069	2.807
24	1.318	1.711	2.492	2.064	2.797
25	1.316	1.708	2.485	2.060	2.787

26	1.315	1.706	2.479	2.056	2.779
27	1.314	1.703	2.473	2.052	2.771
28	1.313	1.701	2.467	2.048	2.763
29	1.311	1.699	2.462	2.045	2.756
30	1.310	1.697	2.457	2.042	2.750
31	1.309	1.696	2.453	2.040	2.744
32	1.309	1.694	2.449	2.037	2.738
33	1.308	1.692	2.445	2.035	2.733
34	1.307	1.691	2.441	2.032	2.728
35	1.306	1.690	2.438	2.030	2.724
36	1.306	1.688	2.434	2.028	2.719
37	1.305	1.687	2.431	2.026	2.715
38	1.304	1.686	2.429	2.024	2.712
39	1.304	1.685	2.426	2.023	2.708
40	1.303	1.684	2.423	2.021	2.704
41	1.303	1.683	2.421	2.020	2.701
42	1.302	1.682	2.418	2.018	2.698
43	1.302	1.681	2.416	2.017	2.695
44	1.301	1.680	2.414	2.015	2.692
45	1.301	1.679	2.412	2.014	2.690
46	1.300	1.679	2.410	2.013	2.687
47	1.300	1.678	2.408	2.012	2.685
48	1.299	1.677	2.407	2.011	2.682
49	1.299	1.677	2.405	2.010	2.680
50	1.299	1.676	2.403	2.009	2.678
51	1.298	1.675	2.402	2.008	2.676
52	1.298	1.675	2.400	2.007	2.674
53	1.298	1.674	2.399	2.006	2.672
54	1.297	1.674	2.397	2.005	2.670
55	1.297	1.673	2.396	2.004	2.668
56	1.297	1.673	2.395	2.003	2.667
57	1.297	1.672	2.394	2.002	2.665
58	1.296	1.672	2.392	2.002	2.663
59	1.296	1.671	2.391	2.001	2.662
60	1.296	1.671	2.390	2.000	2.660
61	1.282	1.646	2.330	1.962	2.581
62	1.282	1.646	2.330	1.962	2.581
63	1.282	1.646	2.330	1.962	2.581
64	1.282	1.646	2.330	1.962	2.581
65	1.282	1.646	2.330	1.962	2.581
66	1.282	1.646	2.330	1.962	2.581

67	1.282	1.646	2.330	1.962	2.581
68	1.282	1.646	2.330	1.962	2.581
69	1.282	1.646	2.330	1.962	2.581
70	1.282	1.646	2.330	1.962	2.581
71	1.282	1.646	2.330	1.962	2.581
72	1.282	1.646	2.330	1.962	2.581
73	1.282	1.646	2.330	1.962	2.581
74	1.282	1.646	2.330	1.962	2.581
75	1.282	1.646	2.330	1.962	2.581
76	1.282	1.646	2.330	1.962	2.581
77	1.282	1.646	2.330	1.962	2.581
78	1.282	1.646	2.330	1.962	2.581
79	1.282	1.646	2.330	1.962	2.581
80	1.282	1.646	2.330	1.962	2.581
81	1.282	1.646	2.330	1.962	2.581
82	1.282	1.646	2.330	1.962	2.581
83	1.282	1.646	2.330	1.962	2.581
84	1.282	1.646	2.330	1.962	2.581
85	1.282	1.646	2.330	1.962	2.581
86	1.282	1.646	2.330	1.962	2.581
87	1.282	1.646	2.330	1.962	2.581
88	1.282	1.646	2.330	1.962	2.581
89	1.282	1.646	2.330	1.962	2.581
90	1.282	1.646	2.330	1.962	2.581
91	1.282	1.646	2.330	1.962	2.581
92	1.282	1.646	2.330	1.962	2.581
93	1.282	1.646	2.330	1.962	2.581
94	1.282	1.646	2.330	1.962	2.581
95	1.282	1.646	2.330	1.962	2.581
96	1.282	1.646	2.330	1.962	2.581
97	1.282	1.646	2.330	1.962	2.581
98	1.282	1.646	2.330	1.962	2.581
99	1.282	1.646	2.330	1.962	2.581
100	1.282	1.646	2.330	1.962	2.581
101	1.282	1.646	2.330	1.962	2.581
102	1.282	1.646	2.330	1.962	2.581
103	1.282	1.646	2.330	1.962	2.581
104	1.282	1.646	2.330	1.962	2.581
105	1.282	1.646	2.330	1.962	2.581
106	1.282	1.646	2.330	1.962	2.581
107	1.282	1.646	2.330	1.962	2.581

108	1.282	1.646	2.330	1.962	2.581
109	1.282	1.646	2.330	1.962	2.581
110	1.282	1.646	2.330	1.962	2.581
111	1.282	1.646	2.330	1.962	2.581
112	1.282	1.646	2.330	1.962	2.581
113	1.282	1.646	2.330	1.962	2.581
114	1.282	1.646	2.330	1.962	2.581
115	1.282	1.646	2.330	1.962	2.581
116	1.282	1.646	2.330	1.962	2.581
117	1.282	1.646	2.330	1.962	2.581
118	1.282	1.646	2.330	1.962	2.581
119	1.282	1.646	2.330	1.962	2.581
120	1.282	1.646	2.330	1.962	2.581
∞	1.282	1.646	2.330	1.962	2.581

The critical values have been obtained with EXCEL, using the function T.INV.2T(α ; degrees of freedom).

Table A2. Critical values (percentiles) for the Snedecor's F distribution at the 1% level of significance

Degrees of freedom in the denominator ($n - k - 1$)	Degrees of freedom in the numerator (q)									
	1	2	3	4	5	6	7	8	9	10
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508

19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979
31	7.530	5.362	4.484	3.993	3.675	3.449	3.281	3.149	3.043	2.955
32	7.499	5.336	4.459	3.969	3.652	3.427	3.258	3.127	3.021	2.934
33	7.471	5.312	4.437	3.948	3.630	3.406	3.238	3.106	3.000	2.913
34	7.444	5.289	4.416	3.927	3.611	3.386	3.218	3.087	2.981	2.894
35	7.419	5.268	4.396	3.908	3.592	3.368	3.200	3.069	2.963	2.876
36	7.396	5.248	4.377	3.890	3.574	3.351	3.183	3.052	2.946	2.859
37	7.373	5.229	4.360	3.873	3.558	3.334	3.167	3.036	2.930	2.843
38	7.353	5.211	4.343	3.858	3.542	3.319	3.152	3.021	2.915	2.828
39	7.333	5.194	4.327	3.843	3.528	3.305	3.137	3.006	2.901	2.814
40	7.314	5.179	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801
41	7.296	5.163	4.299	3.815	3.501	3.278	3.111	2.980	2.875	2.788
42	7.280	5.149	4.285	3.802	3.488	3.266	3.099	2.968	2.863	2.776

43	7.264	5.136	4.273	3.790	3.476	3.254	3.087	2.957	2.851	2.764
44	7.248	5.123	4.261	3.778	3.465	3.243	3.076	2.946	2.840	2.754
45	7.234	5.110	4.249	3.767	3.454	3.232	3.066	2.935	2.830	2.743
46	7.220	5.099	4.238	3.757	3.444	3.222	3.056	2.925	2.820	2.733
47	7.207	5.087	4.228	3.747	3.434	3.213	3.046	2.916	2.811	2.724
48	7.194	5.077	4.218	3.737	3.425	3.204	3.037	2.907	2.802	2.715
49	7.182	5.066	4.208	3.728	3.416	3.195	3.028	2.898	2.793	2.706
50	7.171	5.057	4.199	3.720	3.408	3.186	3.020	2.890	2.785	2.698
51	7.159	5.047	4.191	3.711	3.400	3.178	3.012	2.882	2.777	2.690
52	7.149	5.038	4.182	3.703	3.392	3.171	3.005	2.874	2.769	2.683
53	7.139	5.030	4.174	3.695	3.384	3.163	2.997	2.867	2.762	2.675
54	7.129	5.021	4.167	3.688	3.377	3.156	2.990	2.860	2.755	2.668
55	7.119	5.013	4.159	3.681	3.370	3.149	2.983	2.853	2.748	2.662
56	7.110	5.006	4.152	3.674	3.363	3.143	2.977	2.847	2.742	2.655
57	7.102	4.998	4.145	3.667	3.357	3.136	2.971	2.841	2.736	2.649
58	7.093	4.991	4.138	3.661	3.351	3.130	2.965	2.835	2.730	2.643
59	7.085	4.984	4.132	3.655	3.345	3.124	2.959	2.829	2.724	2.637
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632
61	7.070	4.971	4.120	3.643	3.333	3.113	2.948	2.818	2.713	2.626
62	7.062	4.965	4.114	3.638	3.328	3.108	2.942	2.813	2.708	2.621
63	7.055	4.959	4.109	3.632	3.323	3.103	2.937	2.808	2.703	2.616
64	7.048	4.953	4.103	3.627	3.318	3.098	2.932	2.803	2.698	2.611
65	7.042	4.947	4.098	3.622	3.313	3.093	2.928	2.798	2.693	2.607
66	7.035	4.942	4.093	3.618	3.308	3.088	2.923	2.793	2.689	2.602

67	7.029	4.937	4.088	3.613	3.304	3.084	2.919	2.789	2.684	2.598
68	7.023	4.932	4.083	3.608	3.299	3.080	2.914	2.785	2.680	2.593
69	7.017	4.927	4.079	3.604	3.295	3.075	2.910	2.781	2.676	2.589
70	7.011	4.922	4.074	3.600	3.291	3.071	2.906	2.777	2.672	2.585
71	7.006	4.917	4.070	3.596	3.287	3.067	2.902	2.773	2.668	2.581
72	7.001	4.913	4.066	3.591	3.283	3.063	2.898	2.769	2.664	2.578
73	6.995	4.908	4.062	3.588	3.279	3.060	2.895	2.765	2.660	2.574
74	6.990	4.904	4.058	3.584	3.275	3.056	2.891	2.762	2.657	2.570
75	6.985	4.900	4.054	3.580	3.272	3.052	2.887	2.758	2.653	2.567
76	6.981	4.896	4.050	3.577	3.268	3.049	2.884	2.755	2.650	2.563
77	6.976	4.892	4.047	3.573	3.265	3.046	2.881	2.751	2.647	2.560
78	6.971	4.888	4.043	3.570	3.261	3.042	2.877	2.748	2.644	2.557
79	6.967	4.884	4.040	3.566	3.258	3.039	2.874	2.745	2.640	2.554
80	6.963	4.881	4.036	3.563	3.255	3.036	2.871	2.742	2.637	2.551
81	6.959	4.877	4.033	3.560	3.252	3.033	2.868	2.739	2.634	2.548
82	6.954	4.874	4.030	3.557	3.249	3.030	2.865	2.736	2.632	2.545
83	6.950	4.870	4.027	3.554	3.246	3.027	2.863	2.733	2.629	2.542
84	6.947	4.867	4.024	3.551	3.243	3.025	2.860	2.731	2.626	2.539
85	6.943	4.864	4.021	3.548	3.240	3.022	2.857	2.728	2.623	2.537
86	6.939	4.861	4.018	3.545	3.238	3.019	2.854	2.725	2.621	2.534
87	6.935	4.858	4.015	3.543	3.235	3.017	2.852	2.723	2.618	2.532
88	6.932	4.855	4.012	3.540	3.233	3.014	2.849	2.720	2.616	2.529
89	6.928	4.852	4.010	3.538	3.230	3.012	2.847	2.718	2.613	2.527
90	6.925	4.849	4.007	3.535	3.228	3.009	2.845	2.715	2.611	2.524

91	6.922	4.846	4.004	3.533	3.225	3.007	2.842	2.713	2.609	2.522
92	6.919	4.844	4.002	3.530	3.223	3.004	2.840	2.711	2.606	2.520
93	6.915	4.841	3.999	3.528	3.221	3.002	2.838	2.709	2.604	2.518
94	6.912	4.838	3.997	3.525	3.218	3.000	2.835	2.706	2.602	2.515
95	6.909	4.836	3.995	3.523	3.216	2.998	2.833	2.704	2.600	2.513
96	6.906	4.833	3.992	3.521	3.214	2.996	2.831	2.702	2.598	2.511
97	6.904	4.831	3.990	3.519	3.212	2.994	2.829	2.700	2.596	2.509
98	6.901	4.829	3.988	3.517	3.210	2.992	2.827	2.698	2.594	2.507
99	6.898	4.826	3.986	3.515	3.208	2.990	2.825	2.696	2.592	2.505
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503
101	6.893	4.822	3.982	3.511	3.204	2.986	2.821	2.692	2.588	2.501
102	6.890	4.819	3.980	3.509	3.202	2.984	2.820	2.691	2.586	2.500
103	6.888	4.817	3.978	3.507	3.200	2.982	2.818	2.689	2.584	2.498
104	6.885	4.815	3.976	3.505	3.198	2.980	2.816	2.687	2.583	2.496
105	6.883	4.813	3.974	3.503	3.197	2.979	2.814	2.685	2.581	2.494
106	6.880	4.811	3.972	3.501	3.195	2.977	2.813	2.684	2.579	2.493
107	6.878	4.809	3.970	3.500	3.193	2.975	2.811	2.682	2.578	2.491
108	6.876	4.807	3.968	3.498	3.191	2.973	2.809	2.680	2.576	2.489
109	6.873	4.805	3.967	3.496	3.190	2.972	2.808	2.679	2.574	2.488
110	6.871	4.803	3.965	3.495	3.188	2.970	2.806	2.677	2.573	2.486
111	6.869	4.802	3.963	3.493	3.187	2.969	2.805	2.676	2.571	2.485
112	6.867	4.800	3.961	3.491	3.185	2.967	2.803	2.674	2.570	2.483
113	6.865	4.798	3.960	3.490	3.184	2.966	2.801	2.673	2.568	2.482
114	6.863	4.796	3.958	3.488	3.182	2.964	2.800	2.671	2.567	2.480

Appendix A

115	6.861	4.795	3.957	3.487	3.181	2.963	2.799	2.670	2.565	2.479
116	6.859	4.793	3.955	3.485	3.179	2.961	2.797	2.668	2.564	2.477
117	6.857	4.791	3.954	3.484	3.178	2.960	2.796	2.667	2.563	2.476
118	6.855	4.790	3.952	3.482	3.176	2.959	2.794	2.666	2.561	2.475
119	6.853	4.788	3.951	3.481	3.175	2.957	2.793	2.664	2.560	2.473
120	6.851	4.787	3.949	3.480	3.174	2.956	2.792	2.663	2.559	2.472
∞	6.635	4.606	3.782	3.320	3.018	2.802	2.640	2.512	2.408	2.321

The critical values have been obtained with EXCEL, using the function F.INV.RT(0.01; degrees of freedom in the numerator; degrees of freedom in the denominator).

Table A3. Critical values (percentiles) for the Snedecor's F distribution at the 5% level of significance

Degrees of freedom in the denominator ($n - k - 1$)	Degrees of freedom in the numerator (q)									
	1	2	3	4	5	6	7	8	9	10
1	161.448	199.500	215.707	224.583	230.162	233.986	236.768	238.883	240.543	241.882
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.786
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637
8	5.318	4.459	4.066	3.838	3.687	3.581	3.500	3.438	3.388	3.347
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378

20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323	2.255	2.199	2.153
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313	2.244	2.189	2.142
33	4.139	3.285	2.892	2.659	2.503	2.389	2.303	2.235	2.179	2.133
34	4.130	3.276	2.883	2.650	2.494	2.380	2.294	2.225	2.170	2.123
35	4.121	3.267	2.874	2.641	2.485	2.372	2.285	2.217	2.161	2.114
36	4.113	3.259	2.866	2.634	2.477	2.364	2.277	2.209	2.153	2.106
37	4.105	3.252	2.859	2.626	2.470	2.356	2.270	2.201	2.145	2.098
38	4.098	3.245	2.852	2.619	2.463	2.349	2.262	2.194	2.138	2.091
39	4.091	3.238	2.845	2.612	2.456	2.342	2.255	2.187	2.131	2.084
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
41	4.079	3.226	2.833	2.600	2.443	2.330	2.243	2.174	2.118	2.071
42	4.073	3.220	2.827	2.594	2.438	2.324	2.237	2.168	2.112	2.065
43	4.067	3.214	2.822	2.589	2.432	2.318	2.232	2.163	2.106	2.059

44	4.062	3.209	2.816	2.584	2.427	2.313	2.226	2.157	2.101	2.054
45	4.057	3.204	2.812	2.579	2.422	2.308	2.221	2.152	2.096	2.049
46	4.052	3.200	2.807	2.574	2.417	2.304	2.216	2.147	2.091	2.044
47	4.047	3.195	2.802	2.570	2.413	2.299	2.212	2.143	2.086	2.039
48	4.043	3.191	2.798	2.565	2.409	2.295	2.207	2.138	2.082	2.035
49	4.038	3.187	2.794	2.561	2.404	2.290	2.203	2.134	2.077	2.030
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
51	4.030	3.179	2.786	2.553	2.397	2.283	2.195	2.126	2.069	2.022
52	4.027	3.175	2.783	2.550	2.393	2.279	2.192	2.122	2.066	2.018
53	4.023	3.172	2.779	2.546	2.389	2.275	2.188	2.119	2.062	2.015
54	4.020	3.168	2.776	2.543	2.386	2.272	2.185	2.115	2.059	2.011
55	4.016	3.165	2.773	2.540	2.383	2.269	2.181	2.112	2.055	2.008
56	4.013	3.162	2.769	2.537	2.380	2.266	2.178	2.109	2.052	2.005
57	4.010	3.159	2.766	2.534	2.377	2.263	2.175	2.106	2.049	2.001
58	4.007	3.156	2.764	2.531	2.374	2.260	2.172	2.103	2.046	1.998
59	4.004	3.153	2.761	2.528	2.371	2.257	2.169	2.100	2.043	1.995
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
61	3.998	3.148	2.755	2.523	2.366	2.251	2.164	2.094	2.037	1.990
62	3.996	3.145	2.753	2.520	2.363	2.249	2.161	2.092	2.035	1.987
63	3.993	3.143	2.751	2.518	2.361	2.246	2.159	2.089	2.032	1.985
64	3.991	3.140	2.748	2.515	2.358	2.244	2.156	2.087	2.030	1.982
65	3.989	3.138	2.746	2.513	2.356	2.242	2.154	2.084	2.027	1.980
66	3.986	3.136	2.744	2.511	2.354	2.239	2.152	2.082	2.025	1.977
67	3.984	3.134	2.742	2.509	2.352	2.237	2.150	2.080	2.023	1.975

68	3.982	3.132	2.740	2.507	2.350	2.235	2.148	2.078	2.021	1.973
69	3.980	3.130	2.737	2.505	2.348	2.233	2.145	2.076	2.019	1.971
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
71	3.976	3.126	2.734	2.501	2.344	2.229	2.142	2.072	2.015	1.967
72	3.974	3.124	2.732	2.499	2.342	2.227	2.140	2.070	2.013	1.965
73	3.972	3.122	2.730	2.497	2.340	2.226	2.138	2.068	2.011	1.963
74	3.970	3.120	2.728	2.495	2.338	2.224	2.136	2.066	2.009	1.961
75	3.968	3.119	2.727	2.494	2.337	2.222	2.134	2.064	2.007	1.959
76	3.967	3.117	2.725	2.492	2.335	2.220	2.133	2.063	2.006	1.958
77	3.965	3.115	2.723	2.490	2.333	2.219	2.131	2.061	2.004	1.956
78	3.963	3.114	2.722	2.489	2.332	2.217	2.129	2.059	2.002	1.954
79	3.962	3.112	2.720	2.487	2.330	2.216	2.128	2.058	2.001	1.953
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
81	3.959	3.109	2.717	2.484	2.327	2.213	2.125	2.055	1.998	1.950
82	3.957	3.108	2.716	2.483	2.326	2.211	2.123	2.053	1.996	1.948
83	3.956	3.107	2.715	2.482	2.324	2.210	2.122	2.052	1.995	1.947
84	3.955	3.105	2.713	2.480	2.323	2.209	2.121	2.051	1.993	1.945
85	3.953	3.104	2.712	2.479	2.322	2.207	2.119	2.049	1.992	1.944
86	3.952	3.103	2.711	2.478	2.321	2.206	2.118	2.048	1.991	1.943
87	3.951	3.101	2.709	2.476	2.319	2.205	2.117	2.047	1.989	1.941
88	3.949	3.100	2.708	2.475	2.318	2.203	2.115	2.045	1.988	1.940
89	3.948	3.099	2.707	2.474	2.317	2.202	2.114	2.044	1.987	1.939
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
91	3.946	3.097	2.705	2.472	2.315	2.200	2.112	2.042	1.984	1.936

92	3.945	3.095	2.704	2.471	2.313	2.199	2.111	2.041	1.983	1.935
93	3.943	3.094	2.703	2.470	2.312	2.198	2.110	2.040	1.982	1.934
94	3.942	3.093	2.701	2.469	2.311	2.197	2.109	2.038	1.981	1.933
95	3.941	3.092	2.700	2.467	2.310	2.196	2.108	2.037	1.980	1.932
96	3.940	3.091	2.699	2.466	2.309	2.195	2.106	2.036	1.979	1.931
97	3.939	3.090	2.698	2.465	2.308	2.194	2.105	2.035	1.978	1.930
98	3.938	3.089	2.697	2.465	2.307	2.193	2.104	2.034	1.977	1.929
99	3.937	3.088	2.696	2.464	2.306	2.192	2.103	2.033	1.976	1.928
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
101	3.935	3.086	2.695	2.462	2.304	2.190	2.102	2.031	1.974	1.926
102	3.934	3.085	2.694	2.461	2.303	2.189	2.101	2.030	1.973	1.925
103	3.933	3.085	2.693	2.460	2.303	2.188	2.100	2.030	1.972	1.924
104	3.932	3.084	2.692	2.459	2.302	2.187	2.099	2.029	1.971	1.923
105	3.932	3.083	2.691	2.458	2.301	2.186	2.098	2.028	1.970	1.922
106	3.931	3.082	2.690	2.457	2.300	2.185	2.097	2.027	1.969	1.921
107	3.930	3.081	2.689	2.457	2.299	2.184	2.096	2.026	1.969	1.920
108	3.929	3.080	2.689	2.456	2.298	2.184	2.096	2.025	1.968	1.919
109	3.928	3.080	2.688	2.455	2.298	2.183	2.095	2.024	1.967	1.919
110	3.927	3.079	2.687	2.454	2.297	2.182	2.094	2.024	1.966	1.918
111	3.927	3.078	2.686	2.453	2.296	2.181	2.093	2.023	1.965	1.917
112	3.926	3.077	2.686	2.453	2.295	2.181	2.092	2.022	1.964	1.916
113	3.925	3.077	2.685	2.452	2.295	2.180	2.092	2.021	1.964	1.915
114	3.924	3.076	2.684	2.451	2.294	2.179	2.091	2.021	1.963	1.915

115	3.924	3.075	2.683	2.451	2.293	2.178	2.090	2.020	1.962	1.914
116	3.923	3.074	2.683	2.450	2.293	2.178	2.089	2.019	1.962	1.913
117	3.922	3.074	2.682	2.449	2.292	2.177	2.089	2.018	1.961	1.913
118	3.921	3.073	2.681	2.449	2.291	2.176	2.088	2.018	1.960	1.912
119	3.921	3.072	2.681	2.448	2.290	2.176	2.087	2.017	1.959	1.911
120	3.920	3.072	2.680	2.447	2.290	2.175	2.087	2.016	1.959	1.910
∞	3.842	2.996	2.605	2.372	2.214	2.099	2.010	1.939	1.880	1.831

The critical values have been obtained with EXCEL, using the function F.INV.RT(0.05; degrees of freedom in the numerator; degrees of freedom in the denominator).

Table A4. Critical values (percentiles) for the Snedecor's F distribution at the 10% level of significance

Degrees of freedom in the denominator ($n - k - 1$)	Degrees of freedom in the numerator (q)									
	1	2	3	4	5	6	7	8	9	10
1	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.858	60.195
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920

22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819
31	2.875	2.482	2.270	2.136	2.042	1.973	1.920	1.877	1.842	1.812
32	2.869	2.477	2.263	2.129	2.036	1.967	1.913	1.870	1.835	1.805
33	2.864	2.471	2.258	2.123	2.030	1.961	1.907	1.864	1.828	1.799
34	2.859	2.466	2.252	2.118	2.024	1.955	1.901	1.858	1.822	1.793
35	2.855	2.461	2.247	2.113	2.019	1.950	1.896	1.852	1.817	1.787
36	2.850	2.456	2.243	2.108	2.014	1.945	1.891	1.847	1.811	1.781
37	2.846	2.452	2.238	2.103	2.009	1.940	1.886	1.842	1.806	1.776
38	2.842	2.448	2.234	2.099	2.005	1.935	1.881	1.838	1.802	1.772
39	2.839	2.444	2.230	2.095	2.001	1.931	1.877	1.833	1.797	1.767
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763
41	2.832	2.437	2.222	2.087	1.993	1.923	1.869	1.825	1.789	1.759
42	2.829	2.434	2.219	2.084	1.989	1.919	1.865	1.821	1.785	1.755
43	2.826	2.430	2.216	2.080	1.986	1.916	1.861	1.817	1.781	1.751
44	2.823	2.427	2.213	2.077	1.983	1.913	1.858	1.814	1.778	1.747
45	2.820	2.425	2.210	2.074	1.980	1.909	1.855	1.811	1.774	1.744

46	2.818	2.422	2.207	2.071	1.977	1.906	1.852	1.808	1.771	1.741
47	2.815	2.419	2.204	2.068	1.974	1.903	1.849	1.805	1.768	1.738
48	2.813	2.417	2.202	2.066	1.971	1.901	1.846	1.802	1.765	1.735
49	2.811	2.414	2.199	2.063	1.968	1.898	1.843	1.799	1.763	1.732
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729
51	2.807	2.410	2.194	2.058	1.964	1.893	1.838	1.794	1.757	1.727
52	2.805	2.408	2.192	2.056	1.961	1.891	1.836	1.791	1.755	1.724
53	2.803	2.406	2.190	2.054	1.959	1.888	1.833	1.789	1.752	1.722
54	2.801	2.404	2.188	2.052	1.957	1.886	1.831	1.787	1.750	1.719
55	2.799	2.402	2.186	2.050	1.955	1.884	1.829	1.785	1.748	1.717
56	2.797	2.400	2.184	2.048	1.953	1.882	1.827	1.782	1.746	1.715
57	2.796	2.398	2.182	2.046	1.951	1.880	1.825	1.780	1.744	1.713
58	2.794	2.396	2.181	2.044	1.949	1.878	1.823	1.779	1.742	1.711
59	2.793	2.395	2.179	2.043	1.947	1.876	1.821	1.777	1.740	1.709
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707
61	2.790	2.392	2.176	2.039	1.944	1.873	1.818	1.773	1.736	1.705
62	2.788	2.390	2.174	2.038	1.942	1.871	1.816	1.771	1.735	1.703
63	2.787	2.389	2.173	2.036	1.941	1.870	1.814	1.770	1.733	1.702
64	2.786	2.387	2.171	2.035	1.939	1.868	1.813	1.768	1.731	1.700
65	2.784	2.386	2.170	2.033	1.938	1.867	1.811	1.767	1.730	1.699
66	2.783	2.385	2.169	2.032	1.937	1.865	1.810	1.765	1.728	1.697
67	2.782	2.384	2.167	2.031	1.935	1.864	1.808	1.764	1.727	1.696
68	2.781	2.382	2.166	2.029	1.934	1.863	1.807	1.762	1.725	1.694
69	2.780	2.381	2.165	2.028	1.933	1.861	1.806	1.761	1.724	1.693

70	2.779	2.380	2.164	2.027	1.931	1.860	1.804	1.760	1.723	1.691
71	2.778	2.379	2.163	2.026	1.930	1.859	1.803	1.758	1.721	1.690
72	2.777	2.378	2.161	2.025	1.929	1.858	1.802	1.757	1.720	1.689
73	2.776	2.377	2.160	2.024	1.928	1.856	1.801	1.756	1.719	1.687
74	2.775	2.376	2.159	2.022	1.927	1.855	1.800	1.755	1.718	1.686
75	2.774	2.375	2.158	2.021	1.926	1.854	1.798	1.754	1.716	1.685
76	2.773	2.374	2.157	2.020	1.925	1.853	1.797	1.752	1.715	1.684
77	2.772	2.373	2.156	2.019	1.924	1.852	1.796	1.751	1.714	1.683
78	2.771	2.372	2.155	2.018	1.923	1.851	1.795	1.750	1.713	1.682
79	2.770	2.371	2.154	2.017	1.922	1.850	1.794	1.749	1.712	1.681
80	2.769	2.370	2.154	2.016	1.921	1.849	1.793	1.748	1.711	1.680
81	2.769	2.369	2.153	2.016	1.920	1.848	1.792	1.747	1.710	1.679
82	2.768	2.368	2.152	2.015	1.919	1.847	1.791	1.746	1.709	1.678
83	2.767	2.368	2.151	2.014	1.918	1.846	1.790	1.745	1.708	1.677
84	2.766	2.367	2.150	2.013	1.917	1.845	1.790	1.744	1.707	1.676
85	2.765	2.366	2.149	2.012	1.916	1.845	1.789	1.744	1.706	1.675
86	2.765	2.365	2.149	2.011	1.915	1.844	1.788	1.743	1.705	1.674
87	2.764	2.365	2.148	2.011	1.915	1.843	1.787	1.742	1.705	1.673
88	2.763	2.364	2.147	2.010	1.914	1.842	1.786	1.741	1.704	1.672
89	2.763	2.363	2.146	2.009	1.913	1.841	1.785	1.740	1.703	1.671
90	2.762	2.363	2.146	2.008	1.912	1.841	1.785	1.739	1.702	1.670
91	2.761	2.362	2.145	2.008	1.912	1.840	1.784	1.739	1.701	1.670
92	2.761	2.361	2.144	2.007	1.911	1.839	1.783	1.738	1.701	1.669
93	2.760	2.361	2.144	2.006	1.910	1.838	1.782	1.737	1.700	1.668

94	2.760	2.360	2.143	2.006	1.910	1.838	1.782	1.736	1.699	1.667
95	2.759	2.359	2.142	2.005	1.909	1.837	1.781	1.736	1.698	1.667
96	2.759	2.359	2.142	2.004	1.908	1.836	1.780	1.735	1.698	1.666
97	2.758	2.358	2.141	2.004	1.908	1.836	1.780	1.734	1.697	1.665
98	2.757	2.358	2.141	2.003	1.907	1.835	1.779	1.734	1.696	1.665
99	2.757	2.357	2.140	2.003	1.906	1.835	1.778	1.733	1.696	1.664
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663
101	2.756	2.356	2.139	2.001	1.905	1.833	1.777	1.732	1.694	1.663
102	2.755	2.355	2.138	2.001	1.905	1.833	1.777	1.731	1.694	1.662
103	2.755	2.355	2.138	2.000	1.904	1.832	1.776	1.731	1.693	1.661
104	2.754	2.354	2.137	2.000	1.903	1.832	1.775	1.730	1.692	1.661
105	2.754	2.354	2.137	1.999	1.903	1.831	1.775	1.729	1.692	1.660
106	2.753	2.353	2.136	1.999	1.902	1.830	1.774	1.729	1.691	1.660
107	2.753	2.353	2.136	1.998	1.902	1.830	1.774	1.728	1.691	1.659
108	2.753	2.352	2.135	1.998	1.901	1.829	1.773	1.728	1.690	1.658
109	2.752	2.352	2.135	1.997	1.901	1.829	1.773	1.727	1.690	1.658
110	2.752	2.351	2.134	1.997	1.900	1.828	1.772	1.727	1.689	1.657
111	2.751	2.351	2.134	1.996	1.900	1.828	1.772	1.726	1.689	1.657
112	2.751	2.351	2.133	1.996	1.899	1.827	1.771	1.726	1.688	1.656
113	2.750	2.350	2.133	1.995	1.899	1.827	1.771	1.725	1.688	1.656
114	2.750	2.350	2.132	1.995	1.898	1.826	1.770	1.725	1.687	1.655
115	2.750	2.349	2.132	1.994	1.898	1.826	1.770	1.724	1.687	1.655
116	2.749	2.349	2.132	1.994	1.898	1.826	1.769	1.724	1.686	1.654
117	2.749	2.349	2.131	1.994	1.897	1.825	1.769	1.723	1.686	1.654

118	2.749	2.348	2.131	1.993	1.897	1.825	1.768	1.723	1.685	1.653
119	2.748	2.348	2.130	1.993	1.896	1.824	1.768	1.722	1.685	1.653
120	2.748	2.347	2.130	1.992	1.896	1.824	1.767	1.722	1.684	1.652
∞	2.706	2.303	2.084	1.945	1.847	1.774	1.717	1.670	1.632	1.599

The critical values have been obtained with EXCEL, using the function F.INV.RT(0.10; degrees of freedom in the numerator; degrees of freedom in the denominator).

Table A5. Critical values (percentiles) for the Chi-square distribution

Degrees of freedom:	Level of significance (α)		
	0.1	0.05	0.01
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	42.980
25	34.382	37.652	44.314
26	35.563	38.885	45.642
27	36.741	40.113	46.963
28	37.916	41.337	48.278
29	39.087	42.557	49.588
30	40.256	43.773	50.892
31	41.422	44.985	52.191
32	42.585	46.194	53.486
33	43.745	47.400	54.776
34	44.903	48.602	56.061
35	46.059	49.802	57.342
36	47.212	50.998	58.619

37	48.363	52.192	59.893
38	49.513	53.384	61.162
39	50.660	54.572	62.428
40	51.805	55.758	63.691
41	52.949	56.942	64.950
42	54.090	58.124	66.206
43	55.230	59.304	67.459
44	56.369	60.481	68.710
45	57.505	61.656	69.957
46	58.641	62.830	71.201
47	59.774	64.001	72.443
48	60.907	65.171	73.683
49	62.038	66.339	74.919
50	63.167	67.505	76.154
51	64.295	68.669	77.386
52	65.422	69.832	78.616
53	66.548	70.993	79.843
54	67.673	72.153	81.069
55	68.796	73.311	82.292
56	69.919	74.468	83.513
57	71.040	75.624	84.733
58	72.160	76.778	85.950
59	73.279	77.931	87.166
60	74.397	79.082	88.379
61	75.514	80.232	89.591
62	76.630	81.381	90.802
63	77.745	82.529	92.010
64	78.860	83.675	93.217
65	79.973	84.821	94.422
66	81.085	85.965	95.626
67	82.197	87.108	96.828
68	83.308	88.250	98.028
69	84.418	89.391	99.228
70	85.527	90.531	100.425
71	86.635	91.670	101.621
72	87.743	92.808	102.816
73	88.850	93.945	104.010
74	89.956	95.081	105.202
75	91.061	96.217	106.393
76	92.166	97.351	107.583

77	93.270	98.484	108.771
78	94.374	99.617	109.958
79	95.476	100.749	111.144
80	96.578	101.879	112.329
81	97.680	103.010	113.512
82	98.780	104.139	114.695
83	99.880	105.267	115.876
84	100.980	106.395	117.057
85	102.079	107.522	118.236
86	103.177	108.648	119.414
87	104.275	109.773	120.591
88	105.372	110.898	121.767
89	106.469	112.022	122.942
90	107.565	113.145	124.116
91	108.661	114.268	125.289
92	109.756	115.390	126.462
93	110.850	116.511	127.633
94	111.944	117.632	128.803
95	113.038	118.752	129.973
96	114.131	119.871	131.141
97	115.223	120.990	132.309
98	116.315	122.108	133.476
99	117.407	123.225	134.642
100	118.498	124.342	135.807

The critical values have been obtained with EXCEL, using the function CHISQ.INV.RT(significance level, degrees of freedom).