

Dealing with Econometrics

Dealing with Econometrics:

*Real World Cases with
Cross-Sectional Data*

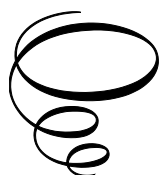
By

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 $wages = f(education, experience, skills)$.

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?

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

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

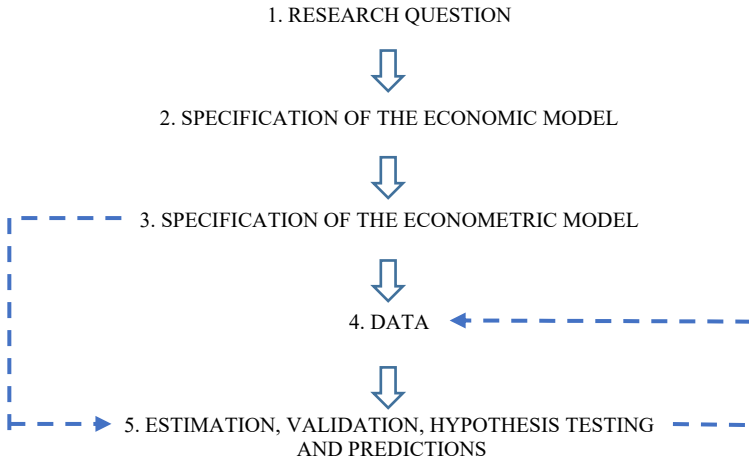
The data used in econometrics are not experimental data. They are collected by passively observing the real world.

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

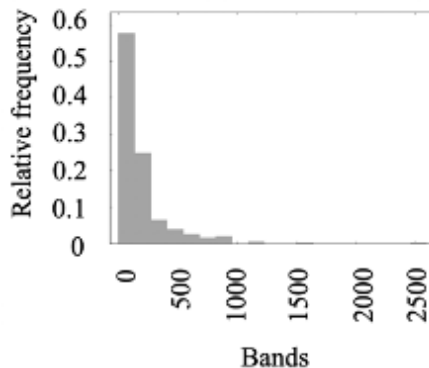
The examples and exercises in this book will be based on the analysis of cross-sectional data sets of samples of information on individuals, countries, firms and other entities collected at a given point in time. One of the basic assumptions is that the data have been collected by extracting a random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

$$Me(x_i) = x_{(n+1)/2} \quad \text{if } n \text{ is odd,}$$

$$Me(x_i) = (x_{(n/2)} + x_{(n/2)+1})/2 \quad \text{if } n \text{ is even.}$$

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.

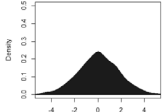
However, the range and the standard deviation depend on the units of measurement of the variable being analysed, making it difficult to compare the representativeness of measures of position in two data sets expressed in different units. As a solution, one could calculate the coefficient of variation (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:

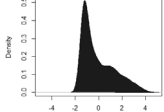
$$CV_x = \frac{s_x}{|\bar{x}|} \text{ if } \bar{x} \neq 0$$

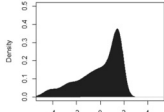
Third, measures of shape describe the distribution of the data set in terms of skewness and kurtosis.

- 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

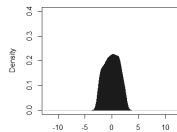
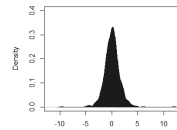
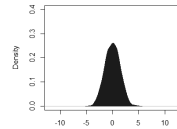


Table 1 presents a summary of the univariate descriptive statistics for the apartments' price and size. As can be seen, the sampled apartments in Valencia have an average selling price of 195,642 euros, with a standard deviation of 234,993 euros. In relative terms, this standard deviation is 120% of 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.

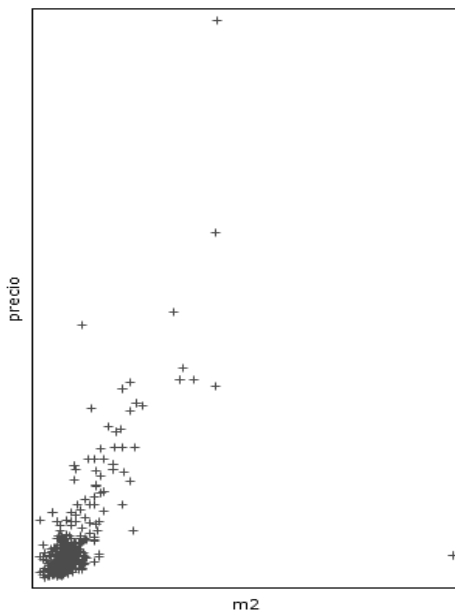
1.3.2. Multivariate descriptive analysis

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.

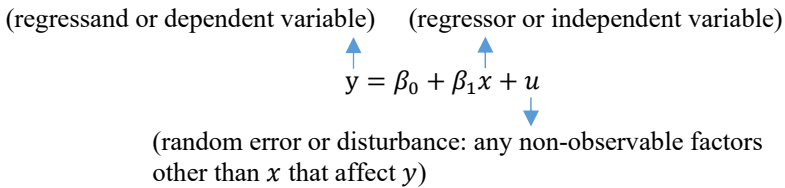
Table 3. Matrix of correlations

<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**:



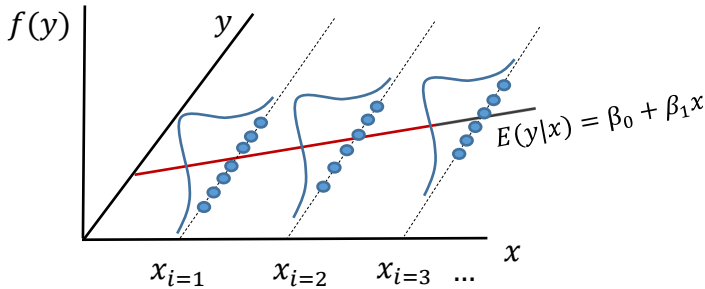
On the one hand, the constant parameter β_0 indicates the value taken by y when $x = 0$. On the other, the slope parameter β_1 provides information about how much y changes given one unit change in x , when other factors that may influence y are relegated to the disturbance term.⁵ For this latter to 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: $E(y|x) = \beta_0 + \beta_1 x$.

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: $y_i = \beta_0 + \beta_1 x_i + u_i$
3. We estimate the **sample regression function (SRF)** for the model: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
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

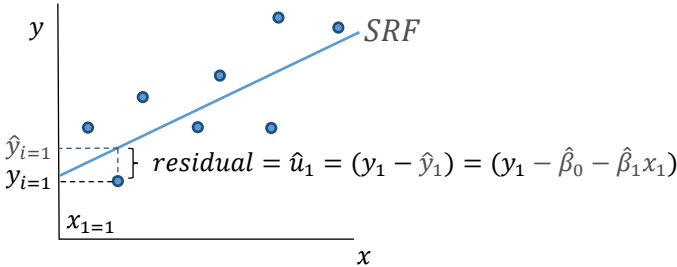
- s_{xy} is the sample covariance between x and y, defined as $s_{xy} = \widehat{Cov}(x_i, y_i) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$

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

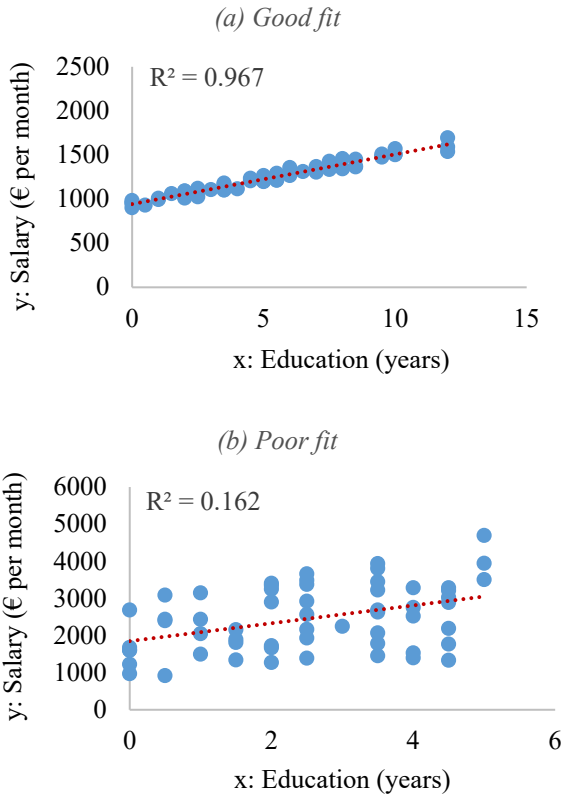
ESS is the explained sum of squares, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$,

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