BASIC COURSE OF CLASSICAL ECONOMETRICS

RESEARCH GROUP:
ECACEN QUIRON

School of Administrative, Accounting, Economics, and Business Sciences

BASIC COURSE OF CLASSICAL ECONOMETRICS

Cristian Orlando Avila Quiñones - UNAD
Nilton Marques de Oliveira - UFT

Research Group: ECACEN QUIRON COL0103217

Basic Course of Classical Econometrics

Authors:
Cristian Orlando Avila Quiñones - UNAD
Nilton Marques de Oliveira - UFT

Research Group: ECACEN QUIRON COL0103217

330.015 195
A958

Basic Course of Classical Econometrics / Avila Quiñones, Cristian Orlando, Nilton Marques de Oliveira. [1.a. ed.]. Bogotá: Sello Editorial UNAD/2024. (School of Administrative, Accounting, Economic, and Business Sciences. - ECACEN)

e-ISBN: 978-6287786-21-9

1. Econometrics 2. Regression Models 3. Matrix Algebra 4. Linear Regression I. Avila Quiñones, CristianOrlando II. Marqués de Oliveira, Nilton III.

E-ISBN: 978-628-7786-21-9
School of Administrative, Accounting, Economics, and Business Sciences

Layout:
Nancy Barreto B.
Hipertexto - Netizen

UNAD University Press
Universidad Nacional Abierta y a Distancia -UNAD
Calle 14 sur No. 14-23
Bogotá D.C.
December 2024

How to cite this book: Avila Quiñones, C. y Marques De Oliveira, N. (2024). *Basic course of Classic Econometrics*. UNAD, University Press. DOI PENDIENTE.

CONTENTS

CHAPTER 1
INTRODUCTION TO ECONOMETRICS

CHAPTER 2
SAMPLE CALCULATION

CHAPTER 3
MINIMUM REQUIREMENTS FOR MATRIX ALGEBRA

CHAPTER 4
ESTIMATION OF ECONOMIC MODELS

CHAPTER 5
TIME SERIES

CHAPTER 6
CASE STUDY–CORRUPTION RISK IN COLOMBIA

PRESENTATION

This work is intended, first, to serve as a guide for any student entering the broad field of econometric methods. Second, to arouse the interest of students in taking advantage of their knowledge of mathematics and statistics to deepen it and apply it to economic science. For this purpose, a basic but rigorous introduction is presented to quantitative methods, matrix algebra, sample statistics, and linear regression as foundations in the analysis of cross-sectional and time series data, delving into the problems of multicollinearity, heteroscedasticity, and autocorrelation in simple regression models that ran in the ordinary least squares (OLS). Likewise, it is intended to provide professors with a material that facilitates classroom teaching and is strictly brief on the minimum requirements to adequately read the econometrics of Gujarati and Novales.

ACKNOWLEDGMENTS

First of all, thanks to the Lord and my family for being the main support for any project. To my beloved and beautiful wife, Prof. Dr. Lina María Grajales, thank you for your time and dedication in helping me achieve this conquest. Thank you for coming into my life. Thank you for our daughter and greatest pride Valentina. Secondly, to my professors, Dr. WILLIAM CARDENAS MAHECHA, Dr. LEONARDO DUARTE, and Dr. ADOLFO JUNCA RODRIGUEZ, who contributed with their postgraduate classes at the National University of Colombia and their instruction in the realization of this work. Thanks to professors SEGUNDO ABRAHAN SANABRIA, WALDECY RODRIGUES, and MARTHA MISAS ARANGO, professors at the Pedagogical and Technological University of Colombia, the Federal University of Tocantins – Brazil, and Universidad de La Sabana – Colombia respectively, for believing in this work.

PREFACE

Questions that, for centuries, has occupied the minds of philosophers, thinkers, and resear chers from Socrates and Aristotle to Foucault in modern times is How do we learn? and how should we teach to make learning happen? The mystery surrounding the acquisition of knowledge and the best pedagogical practice to maximize effective learning has resulted in many theories well-known today by experts in Educational Psychology. As a society, we must ask ourselves: What kind of people do we want to educate through higher education and specifically through the teaching of econometrics?

Econometrics, in a broad sense, is the association of Economic Theory with Mathematical Economics and Statistics. Thus, it can be defined as the statistical methodology used in economic problems posed through a mathematical formulation with stochastic components. The question that arises is how to teach Econometrics so that learning occurs? At least one of the possible answers is simple: through the construction of a bidirectional bridge between theory and practice. In general, the objective of an econometrics professor focuses on students learning theory, its importance, and how it should be applied in a practical context.

This text introduces the basic concepts of econometrics and time series in a simple way, so students understand the step by step through the development of examples and, in this way, a bridge can be built between theory and practice.

The set of concepts, principles, and rules that make up econometric theory requires that what is taught in the classroom be clearly formulated, developed with rigor, and explained through

a mathematical language that involves the use of differential calculus, matrix algebra, and linear algebra, among others; and a statistical language focused on probability theory and statistical inference. This text allows students to understand and reinforce what they have seen in the classroom and allows those interested to get close to the basic principles of econometrics, a discipline that is becoming increasingly relevant and useful in various work contexts.

Martha Misas Arango
Universidad de La Sabana, Chía, Colombia
Highly Prestigious Teacher
International School of Economics and Administrative Sciences
Department of Economics
Puente del Comun University Campus
Ad Portas Building, 2nd Floor.
Chia, Cundinamarca – Colombia

CHAPTER 1

INTRODUCTION TO ECONOMETRICS

1.1 What is Econometrics?

Econometrics could be defined literally as *economic measurement*. However, this definition is quite limited, since econometrics has a broad scope; therefore, it is defined as a separate science because it is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics.

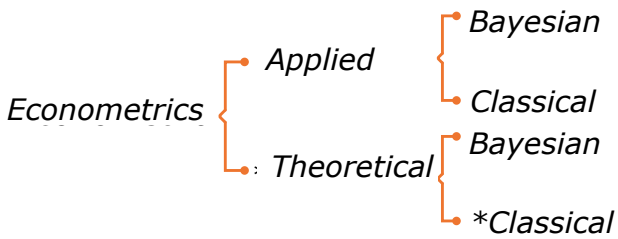
Economic theory formulates hypotheses or qualitative assertions, but these postulations have no significant numerical support. Mathematical economics expresses economic theory in equations but without empirically verifying the theory. Economic statistics is primarily concerned with collecting, processing, and presenting statistical data (graphs, tables, etc.). It captures information but does not apply it to validate the hypotheses or theories of economic science. Finally, mathematical statistics

provides several of the tools used by econometricians, but special methods are needed because the unique nature of most of these economic figures are not generated from a result of a controlled experiment.

Every one of them requires econometrics as a separate discipline of study, where the objective of an econometric exercise is to test a theory, confront an economic hypothesis, or predict or forecast.

1.2 Types of econometrics

Econometrics can be divided into two main groups: applied econometrics and theoretical econometrics. The approach in each group can be classical or Bayesian.



Applied econometrics is used in the special field of economics, including the function of production, investment, consumption, etc. While the development of applicable and appropriate methods to measure economic relationships that are already specified according to the econometric model is theoretical econometrics the latter is what we work on in this book since the widely used method is Ordinary Least Squares (OLS). The approach is classic, which predominates in empirical research.

1.3 Methodology of Classical Theory

1. Choosing the field of application.
2. Specification of an initial model taking into account theory and knowledge of the subject.
3. Data search and editing.

4. Application of a computer program to estimate the model parameters.

5. Hypothesis tests to verify the fit of the model and Interpretation of results.

6. Model refinement process.

To illustrate these steps, for example, we can look for a way to explain the behavior of consumption, understanding consumption as personal spending during a year. This consumption could be food consumption, clothing consumption, or any other consumption that interests us. Therefore, we base on Gujarati's[1] econometrics, which considers the well-known Keynesian theory of Consumption, which is widely used for its ease of understanding in any introductory econometrics course.

1.3.1 Choosing the field of application

John Maynard Keynes postulated the Marginal Propensity to Consume (MPC), which is greater than zero but less than one ($0 < MPC < 1$). In psychological terms, it is expected that a man or woman will increase his or her level of consumption when he or she has experienced a growth in income, but this increase in consumption is not in the same amount as the increase in income.

1.3.2 Specification of an initial model considering theory and knowledge of the subject

Keynes does not specify the precise functional relationship between consumption and income, for this, a mathematical economist can state the following:

$$Y = \beta_1 + \beta_2 X \quad 0 < \beta_2 < 1$$

1 Despite the great universe of texts on econometrics, the author considers Damodar Gujarati's Econometrics from McGraw Hill, 3rd edition, to be the best for the purpose of this text.

Where Y is consumption expenditure. It is defined as the dependent variable or:

- Endogenous variable
- Variable explained
- Response variable
- Returning
- Predicted variable

X is income. It is defined as the explanatory variable or:

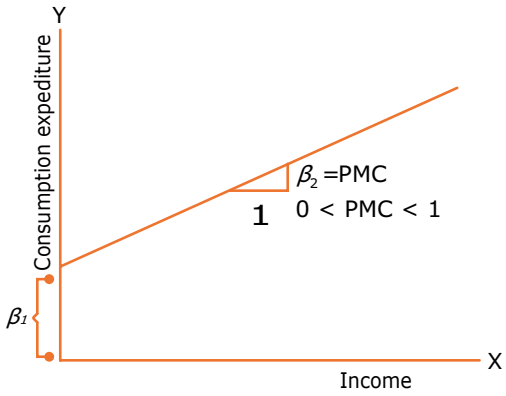
- Exogenous variable.
- Independent variable.
- Control variable.
- Regressor.
- Predictor variable.

(Being a purely personal matter, the terminology used in this text to define Y and X are dependent variable and explanatory variable, respectively).

and where β_1 *y* β_2 are defined as the model parameters, where β_1 is the coefficient of the intercept and β_2 is the coefficient of the slope that measures the MPC (see image 1).

Likewise, if the data starts at the origin, it does not have the intercept coefficient and the model must not carry β_1

Image 1



The equation:

$$Y = \beta_1 + \beta_2 X \qquad \textbf{(1)}$$

states that consumption is linearly related to income (this equation is defined by economy as consumption function). It is also emphasized that a model is simply a set of equations, and since this model has only one equation, it is called a single-equation model; if it had more equations, it would be a multi-equation model.

1.3.3 Data search and editing

The consumption function (1) assumes a deterministic or exact relationship between consumption and income, which makes it limited (generally the relationships between economic variables are inexact)[2]. For this purpose, a random variable (stochastic[3]) must be included in the deterministic consumption function:

2 If we apply a census or a survey to a group of the population, with a sample of 1000 Colombian heads of household and we were to graph these data, it is to be expected that the data would be dispersed to the straight line of image 2, given that the consumption–income relationship proposed is not exact, there are variables that interfere in its behavior (as in the case of the members in charge of the household, age, culture, etc.).

3 Random or stochastic variables are variables that have probability distributions, as their Greek root indicates stokhos (center of the target). For example: The outcome of throwing a tejo [metal disk] from court to court is a stochastic process, in other words, it is a process that allows for errors. The random variable can be either discrete or continuous, e.g. If we throw two dice, which are numbered 1–6, we define X as the sum of the numerical values that the dice acquire, taking X the value of some of the following results: 2,3,4,5,6,7,8,9,10,11, or 12, would be discrete. If it has the possibility of taking any value within an interval of values, it is continuous.

$$Y = \beta_1 + \beta_2 X + u \quad \textbf{(2)}$$

Where *u* is the error term or stochastic disturbance term, which represents all those factors affecting *Y* (consumption expenditure) that do not explain *X* (income), i.e. u represents those factors that are not explicitly considered in the model.

The sources of the error term are:

- Randomness in human responses
- Effect of many omitted variables.
- Measurement errors of *Y* variable
- Unavailability of information.
- Incorrect functional form

Equation (1) $Y = \beta_1 + \beta_2 X$ is an economic model and equation (2) $Y = \beta_1 + \beta_2 X + u$ is an econometric model, technically an example of a linear regression model (see Table 1, Difference between an Economic and an Econometric Model).

Table 1. Difference between an Economic and Econometric Model

Economic Model	Econometric model
Is intending to be general. E.g.of an economic model *Qt = F(L,K)*. It proposes exact relationships. e.g.It = *I* non-labor + Wages	It requires a precise statistical specification of its component varibales and a defined functional form. μ

Likewise, as the text is mainly related to single-equation and linear models, it is relevant to specify the presence of linearity in the parameters and variables.

- Linearity in the variables;
- Linearity in the parameters.

1.3.4 Application of a computer program to estimate the model parameters

In order to estimate β_1 *and* β_2 (i.e., the numerical values) in the consumption function $Y = \beta_1 + \beta_2 X$, it is necessary to have the respective data. To this end, the aggregate personal consumption expenditure of the Colombian economy (as *Y* variable) and GDP (as *X* variable) are taken, always in real terms; thus, data are based on 2010 (they are measured in constant prices). Likewise, it is crucial to determine for our exercise that the data are of the time series type and the variables are quantitative (that is why both variables are expressed in billions of 2010 pesos, see table 1)[4].

1.3.4.1 Types of Information

$$Y = \beta_1 + \beta_2 X + u$$

- Cross-sectional: If information is collected on one or more variables at the same moment in time, the information is defined as cross-sectional, e.g., a cross-sectional sample. The 1993 or 2005 population censuses in Colombia, or a survey.

- Time series[5]: is a set of observations on the values that a variable takes through time. Likewise, this information must be collected in defined time intervals, for example Annual (GDP), semi-annual, quarterly (inflation), monthly (unemployment rates), weekly, or daily. Recalling our initial example, it is a time series and the information is quantitative.

- Panel. Combination of cross-sectional and time series data.

4 These data are graphed in Appendix 1, Figure 13.

5 In Chapter 5, an introduction to time series econometrics is made, which bases much of its empirical work on the assumption that the series are stationary. Basically, a time series is stationary if the value of its mean and variance do not change over time. For example, if we have the data of the unemployment variable for the time periods (1975-1980) and (1985-2000), in which we calculate individually mean, variance and covariance to the 15 observations of the two periods and they are equal, the unemployment time series is stationary. Therefore, whenever working with time series, their stationarity must be checked.

1.3.4.2 Types of Variables

- Quantitative: e.g. income, prices, money supply, etc.
- Qualitative: dichotomous or categorical variables. For example: Male or female, migrant or immigrant, employed or unemployed, married or single, professional or non-professional.
- Proxi.

Table 1. Information on Y (personal consumption expenditure) and X (GDP) from 2005 to 2018 in billions of 2010 pesos.

Año	X	Y
2005	182.228	151.476
2006	205.836	161.161
2007	231.215	172.738
2008	257.229	179.775
2009	271.379	181.446
2010	293.773	190.805
2011	334.297	203.377
2012	360.131	214.144
2013	385.951	222.684
2014	412.602	232.983
2015	435.202	240.188
2016	467.160	243.992
2017	497.669	249.031
2018	529.191	257.779

Source: Own calculations based on DANE 2019 and World Bank–Indicators 2019.

After having the time series data from Table 1, expressed on an annual basis, we proceed to estimate the parameters of our consumption function. To obtain these estimated parameters, the statistical technique known as regression analysis is used, which is the primary tool for the estimation process.

1.3.5 Regression Analysis

Regression analysis deals with the description and evaluation of the relationships between a given variable (called dependent

or explained or endogenous) and one or more additional variables (called independent, explanatory, or exogenous).

It is reaffirmed that, to perform this analysis, a functional relationship between the variables must be proposed. The linear relationship is the functional form most used in practice due to its analytical simplicity. If there is only one independent variable, it reduces to a straight line:

$$\hat{Y} = \beta_1 + \beta_2 X \qquad \textbf{(3)}^6$$

Recalling, the parameter β_1, known as the "sorted at the origin," tells us how much *Y* is when *X* = 0. The parameter β_2, known as the "slope," tells us how much *Y* increases for each one-unit increase in *X.* Our problem consists of obtaining estimates of these coefficients from a sample of observations on the *Y* and *X* variables. In the regression analysis, these estimates are obtained by means of the least squares method.

When applied by the OLS method, we obtain:

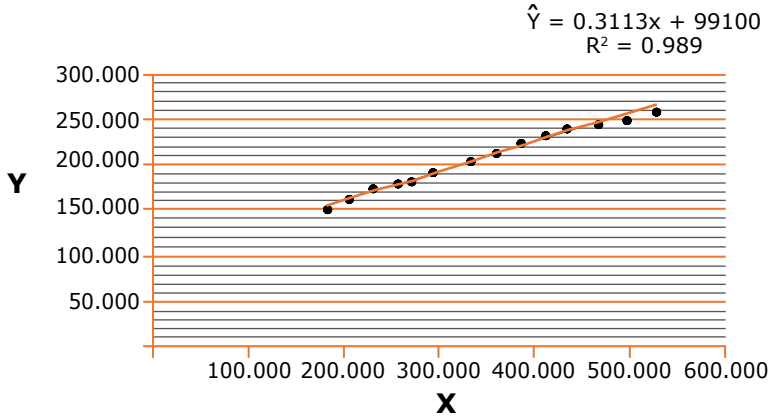
$\hat{Y}$=99099,6+0,3113X Estimate result[7]

However, we must be clear that our first example developed is just a **simple regression** (two variables). To visualize the degree of relationship that exists between the *X* versus *Y* variables, as a basic step in the analysis, a scatter diagram (which can be done in Excel) should be made. This diagram is a representation in a Cartesian coordinate system of the observed numerical data. In the resulting diagram, the X-axis measures GDP (2005-2018), and the Y-axis measures aggregate personal consumption expenditure. Each point in the diagram shows the data pair (GDP as a measure of aggregate income and aggregate personal consumption expenditure) that corresponds to a given year. As expected, there is a positive relationship between these variables: a higher amount of aggregate income corresponds to a higher level of aggregate personal consumption expenditure.

6 The circumflex symbol (^) over Y indicates that Y is an estimate value.

7 Although the reader must not be concerned about how these estimated values were found and their proper interpretation, it is emphasized that the ordinary least squares estimation method is formally addressed in Chapter 4. Likewise, Chapter 3 conducts a broad induction to matrix algebra, required for the maximum understanding of the OLS method. However, for those curious readers, this regression is developed with an econometric package known (gretl) because it is easy to acquire (free on the Web). It can be found in Appendix 1.

Image 2. Simple Regression



Therefore, for the 2005–2018 period, the slope (MPC) is 0.31, i.e., in this sample period, an increase of one Colombian peso in real income leads on average to an increase of 31 cents of a peso in real consumption expenditure[8].

Finally, an important question that arises in regression analysis is the following: What percentage of the total variation in *Y* is due to the *X* variation? In other words, what is the proportion of the total variation in *Y* that can be "explained" by the variation in *X*? The coefficient of determination is the statistic measuring this proportion or percentage, which is calculated by means of an econometric package (see Appendix 1) or in the case of our simple regression; Excel in Windows Vista (see Appendix 2).

The R^2= 0.98 means that the variation in GDP explains 98% of the variation in aggregate personal consumption expenditure. However, up to this point we have only considered the example with simple regression, i.e., proposing a single relationship of explanation of consumption expenditure by income, but as we have stated, this relationship is not entirely correct because there are other variables involved in it. Therefore, it is necessary more than one X_1, when working with $X_1 + X_2 + X_3 .... X_n$ (more than one independent variable). A multiple regression case is a rather complicated and laborious calculation as it requires the use of specialized computer programs.

8 Recalling that the relationship between consumption and income is inexact, it is always said that an increase or decrease is on average, as can be deduced from Appendix 1, Figure 2, which shows the regression line obtained from

1.3.5.1 Role of Econometric Models

From the *n* data of the chosen sample period (t = 1, 2, 3,....n) the knowledge of the coefficients β_1, β_2, β_3 allows us to perform an:

- Structural analysis,
- Prediction $Y_{t+1} = B_1 + B_2X_{t+1} + B_3Z_{t+1}$
- Policy evaluation or simulation of effects

1.3.5.2 Multiple Regression Model–Specification

$$Y_i = F(X_{i1}, X_{i2,}....X_{ik}) + U_i$$
$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + ...\beta_k X_{ik} + U$$

- i = 1, 2, ...n

Matrix expressión

- $Y = X\beta + U$ **(4)**

Modelos no lilenales

$$Y = \beta_1 + \beta_2 e^{\beta_3 X}$$
$$Y = \beta_1 X_2^{\beta_2} X_3^{\beta_3}$$

1.3.6 General Linear Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_K X_K + \varepsilon$$

"It applies generally to inherently linear equations."

For example: $Y = \lambda_1 X_2^{\lambda_2} X_3^{\lambda_3} e^*$

becomes: $\log Y = \alpha_1 + \alpha_2 \log X_2 + \alpha_3 \log X_3 + \varepsilon$

Where: $\alpha_1 = \log \lambda_1 \quad \alpha_2 = \lambda_2 \quad \alpha_3 = \lambda_3 \quad \varepsilon = \log e^*$

"All logarithmic, semilogarithmic and reciprocal models are inherently linear." Alpha Chiang, 2007.

A special case of an inherently linear model is the interaction model:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 (X_2 X_3) + \varepsilon$$

In this case the effect of X_2 on Y is given by:

$$\beta_2 + \beta_4 X_3$$

The effect of variable X_2 on Y depends on the level of variable X_3.

1.3.7 Functional Forms of Regression Models

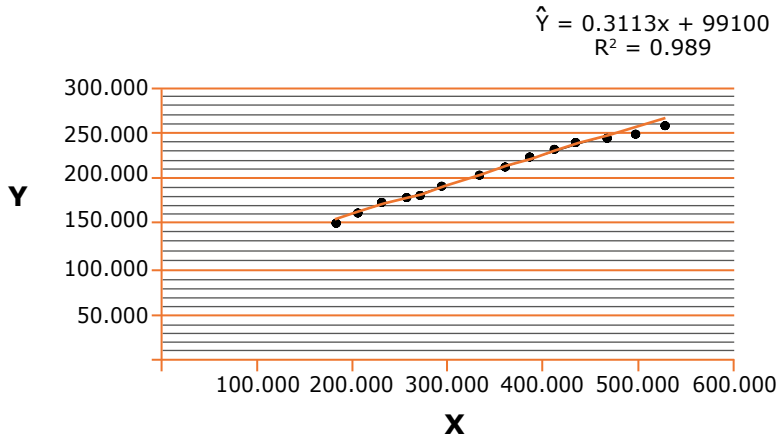
If we recall the data from our example in Table 1: Information on Y (personal consumption expenditure). Y x (GDP), 2005-2018 in billions of 2010 pesos.

We start with 14 data of the endogenous and exogenous variable (2005–2018), we seek to find the regression that gives us the best relationship in the R². which is ***Log-Log***:

1.3.7.1 The Linear–Linear Model

The data without any alteration are only graphed and the regression option is requested and the R² is presented.

Recall Image 2. Simple Regression



X	Y
182.228	151.476
205.836	161.161
231.215	172.738
257.229	179.775
271.379	181.466
293.773	190.805
334.297	203.377
360.131	214.144
385.951	222.684
412.602	232.983
435.202	240.188
467.160	243.992
497.669	249.031
529.191	257.779

1.3.7.2 *The Log-Linear Model*

Logarithms were applied to the endogenous variable:

Image 3. Simple Log-Lin Regression

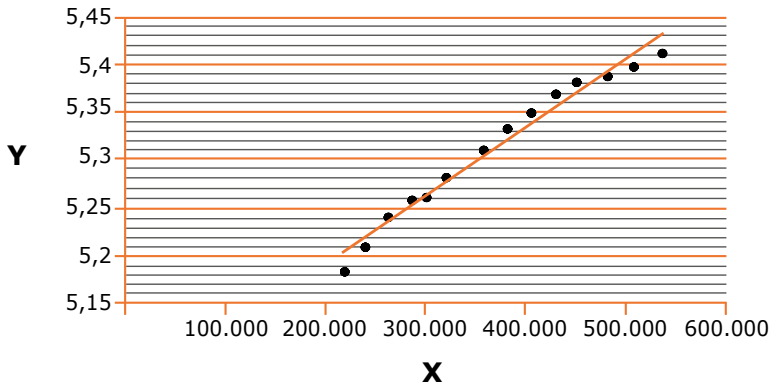
Y = 7E-07 x + 2.0799

Esto sería una expresión en la que:

7E-07 significa 7 × 10^{-7}, que es igual a 0.0000007

Entonces, Y = 0.0000007x + 2.0799

R^2 = 0.9752

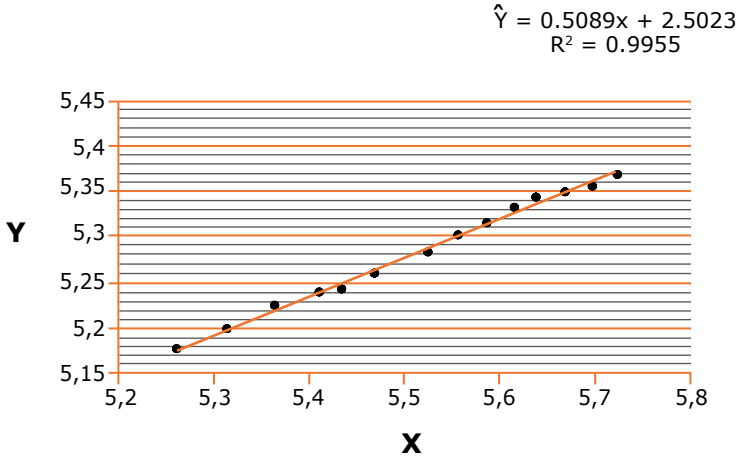


X	Y
182.228	5,18034521
205.836	5,20725892
231.215	5,23738699
257.229	5,25472856
271.379	5,25879628
293.773	5,28058919
334.297	5.30830274
360.131	5.33070677
385.951	5,34768994
412.602	5,36732454
435.202	5,38055131
467.160	5,38737503
497.669	5,39625293
529.191	5,41124719

1.3.7.3 The Log–Log Model

Logarithms were applied to the endogenous and exogenous variables:

Image 4. Simple Log-Lin Regression

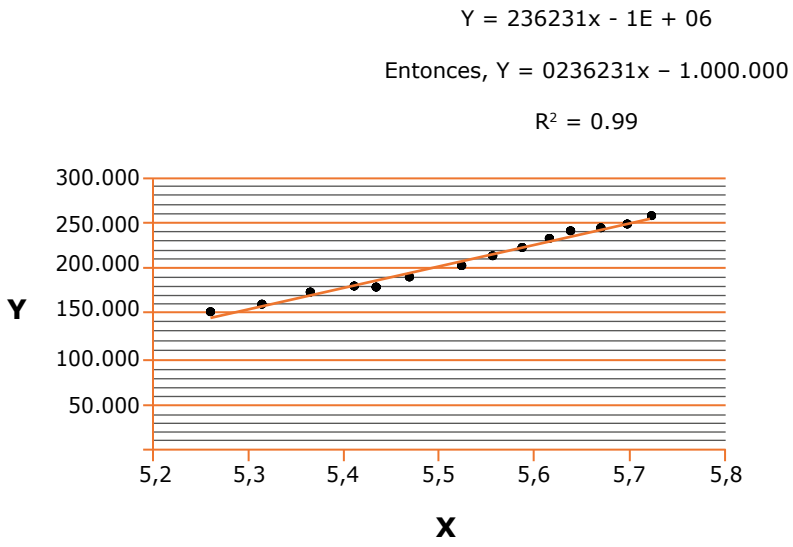


X	Y
5,26061538	5,18034521
5,31352085	5,20725892
5,3640169	5,23738699
5,41031934	5,25472856
5,43357654	5,25879628
5,46801152	5,28058919
5,52413279	5,33070677
5,58653269	5,34768994
5.61553092	5,36732454
5,63869128	5,38055131
5,66946575	5,38737503
5,696941	5,39625293
5,72361209	5,41124719

1.3.7.4 The lin–log model

Logarithm was applied only to the exogenous variable, for this, the endogenous variable returned to the initial one:

Image 5. Simple Lin-Log Regression



X	Y
5,26061538	151.476
5,31352085	161.161
5,3640169	172.738
5,41031934	179.775
5,43357654	181.466
5,46801152	190.805
5,52413279	203.377
5,55564996	214.144
5.58653269	222.684
5,61553092	232.983
5,63869128	240.188
5,66946575	243.992
5,696941	249.031
5,72361209	257.779

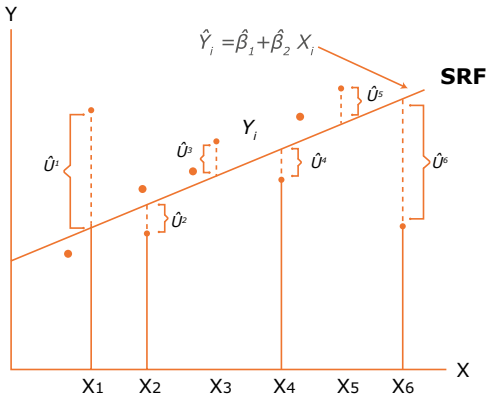
1.3.8 Multiple Regression Model

The aim of regression models is to estimate the population regression function (PRF) based on the sample regression function (SRF) as accurately as possible.

The population regression function (PRF) can be written as $Y_i = \beta_1 + \beta_2 X_i + \mu$ **(5)**

The sample regression function (SRF), stochastic form $\hat{\ }\ \hat{\ }\ \hat{\ }$ **(6)**

Image 6. The SRF



Note that the aim is to estimate the information of each of X_i that do are not explain by Y, i.e., to estimate μ of each X_i

1.3.9 Model Assumptions

1) Y=Xβ+U Linearity of the regression model
2) E(Ui|Xi) = 0
3) $\left.\begin{array}{l} VAR(U_i \mid X) = \sigma_\mu^2 \quad for\ i = 1.....n \\ COV(U_i U_j) = 0 \quad for\ i \neq j \end{array}\right\} \quad \mathrm{E}(UU' \mid X) = \sigma_\mu^2 \mathrm{I}_n$
4) Minimum sample
5) X fixed nxk with rank k
6) $\mu \to N$
7) Cov (UiXj)=0
8) No multicollinearity.

CHAPTER 2

SAMPLE CALCULATION[9]

Calculating the sample size is one of the aspects to be specified in the preliminary stages of the commercial research. It determines the degree of credibility that we will give to the results obtained.

A widely used formula that provides guidance on sample size calculation for aggregate data is as follows:

$$n = \frac{k^2 * p * q * N}{(e^2 * (N-1) + k^2 * p * q} \qquad \textbf{(2.1)}[10]$$

***N*:** is the size of the population or universe (total number of possible respondents).

***k*:** is a constant that depends on the confidence level we assign. The confidence level indicates the probability that the results of our research are true: 95.5% confidence is the same as saying that we can be wrong with a probability of 4.5%.

9 To conduct the basic sample calculation, there are several free simulators available on the web, see the link: https://www.feedbacknetworks.com/cas/experiencia/sol-preguntar-calcular.html.

10 Basic equation of the sample calculation. https://www.feedbacknetworks.com/cas/experiencia/sol-preguntar-calcular.html

2.1 The Most Used K Values and their Confidence Levels are:

This method is very attractive because it uses the Internet and provides convenience for both the interviewer and the respondent.

Trust level	1,15	1,28	1,44	1,65	1,96	2	2,58
Nivel de confianza	75%	80%	85%	90%	95%	95,5%	99%

***e*:** is the desired sampling error. The sampling error is the difference that may exist between the result we obtain by asking a sample of the population and the one we would obtain if we asked the whole population. Examples:

- **Example N° 1:** If the results of an electoral survey indicate that a party obtains 50% of the votes and the estimated error is 4%, the actual percentage of votes is estimated to be in the interval 46-54% (50% +/- 4%).
- **Example N° 2:** If the survey results state that 100 people would buy a product and there is a sampling error of 10%, there will be between 90 and 110 people who will buy it.
- **Example N° 3:** If we conduct an employee dissatisfaction survey with a sampling error of 5% and 30% of the respondents are dissatisfied, it means that between 35% and 25% (30% +/- 5%) of the company's total employees will be dissatisfied.

***p*:** is the rate of individuals in the population who possess the characteristic under study. This data is generally unknown. It is usually assumed that p=q=0.5 is the safest option.

***q*:** is the rate of individuals who do not possess that characteristic, i.e., it is 1-p.

***n*:** is the sample size (number of surveys that will be conducted).

Several examples:

- Example 1: In order to test the percentage of people in Colombia who watch a certain television program. If the

population of the country is 48 million people, and estimating that 30% of the population watches it (p = 0.3 and q = 0.7), with a confidence of 95% that determines that k= 1.96 and assuming a sampling error of 5% (e = 0.05). A sample of 323 people would be needed...

- Example 2: In order to conduct a customer satisfaction survey for a Nissan March vehicle of which 20,000 units (N) have been sold, with a 90% confidence that k = 1.65, a sampling error of 5% (e = 0.05) and considering that 60% will be satisfied (p = 0.6 and q = 0.4), a sample of 258 customers would be needed.

Now, in the case of stratified sampling, it must be ensured that we choose a sufficient number of elements from each group. This type of sampling does not take the population as a whole but instead splits it into several groups with different characteristics (e.g., age 20-35, 35-50, 50-65 and over 65).

In any case, practical criteria based on experience or simple logic are usually used to calculate the sample size. Some of the most used methods are as follows:

1. The available budget for the research.
2. Experience in similar studies.
3. The representativeness of each group considered: choosing from each group an adequate number of respondents so that the results are indicative of the group's opinion.

CHAPTER 3

MINIMUM REQUIREMENTS FOR MATRIX ALGEBRA

- Concept, notation, elements, and order of a matrix
- Types of matrices
- Operations between matrices.

3.1 Concept

Ordered set of numeric or alphanumeric elements arranged in rows and columns. Rows are usually named with the letter m and columns with the letter n. Rows increase from top to bottom and columns from left to right. A matrix is named with the capital letter of the Spanish alphabet, and round brackets ( ) or square brackets [ ] are used to place the elements.

Example: A = (a_{ij}), where *a* = indicates the place occupied by the element in the matrix where *i* = is the row and *j* = is the column.

columns *n*

$$A = (a_{ij}) = \begin{Bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \cdot & \dots & \vdots \\ \mathbf{a_{m1}} & \mathbf{a_{m2}} & \mathbf{a_{m3}} & & \mathbf{a_{mn}} \end{Bmatrix} \; m \textbf{ rows}$$

$$\mathbf{M_{xn}}$$

You always name the row first and then the column.

3.1.1 The order of the matrix indicates the order of the row and columns that make up the matrix:

Example: $A = \begin{vmatrix} 23 & 9 & 7 & 94 \\ 45 & 21 & 34 & 2 \\ 6 & 65 & 23 & 10 \end{vmatrix}$

The order is = 3x4

That is, a matrix of 3 rows by 4 columns.

3.2 Types of Matrices

Depending on the rows and columns, matrices can be:

3.2.1 Rectangular Array

Those where the number of rows is different from the number of columns.

$A = (a_{ij})_{mxn}$ $\quad A = \begin{vmatrix} 2 & 6 \\ -3 & 8 \\ 9 & 1 \end{vmatrix}$ 3x2 $\quad B = \begin{vmatrix} 14 & 3 & 0 \\ 5 & -6 & 4 \end{vmatrix}$ 2x3

As particular cases we would have:

3.2.1.1 Row Vector:

$[a_{11} \quad a_{12} \quad a_{13}]_{1xn} \quad a = [\,2 \quad 5 \quad 3]_{1x3}$

3.2.1.2 Column Vector:

$$A = \begin{vmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a \end{vmatrix} \quad B = \begin{vmatrix} 36 \\ -7 \\ 12 \\ 5 \end{vmatrix}$$

mx1 4x1

3.2.1.3 Unit Vector:

One element is the unit, and the rest are zero.

$$A = \begin{vmatrix} 0 \\ 0 \\ 0 \\ 1 \end{vmatrix} \quad B = [\,1 \quad 0 \quad 0\,]_{1 \times 3}$$

4x1

3.2.1.4 Sum Vector:

Row or column vector whose elements are the unit.

$$A = [\,1 \quad 1 \quad 1\,]_{1 \times 3} \qquad B = \begin{vmatrix} 1 \\ 1 \\ 1 \end{vmatrix}$$

3x1

3.2.1.5 Null Vector:

The vectors are zero.

$$A = \begin{vmatrix} 0 \\ 0 \\ 0 \end{vmatrix} \qquad B = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad 1 \times 3$$

3x1

3.2.2 Square Matrices:

In general, this is the name given to the number of rows and columns that are equal.

$A = (a_{ij})_{nxn}$

13	24	3
12	32	1
4	5	8

3x3

Types of matrices:

3.2.2.1 Upper Triangular:

Matrices whose elements above the main diagonal are zeros.

$$A = \begin{vmatrix} \mathbf{a_{11}} & a_{12} & a_{13} & a_{14} \\ a_{21} & \mathbf{a_{22}} & a_{23} & a_{24} \\ a_{31} & a_{32} & \mathbf{a_{33}} & a_{34} \\ a_{41} & a_{42} & a_{43} & \mathbf{a_{44}} \end{vmatrix} \qquad B = \begin{vmatrix} \mathbf{2} & 0 & 0 & 0 \\ 6 & \mathbf{1} & 0 & 0 \\ 13 & -5 & \mathbf{7} & 0 \\ 9 & 8 & 23 & \mathbf{8} \end{vmatrix}$$

4x4

4x4

3.2.2.2 Lower Triangular:

Matrices whose elements below the main diagonal are zeros.

$$A = \begin{vmatrix} \mathbf{9} & 74 & 0 & 16 \\ 0 & \mathbf{3} & -2 & 8 \\ 0 & 0 & \mathbf{7} & 4 \\ 0 & 0 & 0 & \mathbf{-5} \end{vmatrix}_{4x4}$$

3.2.2.3 Symmetric Matrix:

It is a square matrix in which the elements below and above the main diagonal are equal.

$$A = \begin{vmatrix} 7 & -1 & 4 & 0 \\ -1 & 8 & 5 & -3 \\ 4 & 5 & 2 & 6 \\ 0 & -3 & 6 & 9 \end{vmatrix}_{4x4}$$

Attention! $A_{12}=a_{21}$, $a_{23} = a_{32}$. . .

3.2.2.4 Antisymmetric Matrix:

The elements either above or below the main matrix are reciprocal elements, that is:

$$A = \begin{vmatrix} 0 & -1 & 4 & -10 \\ 1 & 0 & -5 & 3 \\ -4 & 5 & 0 & -6 \\ 10 & -3 & 6 & 0 \end{vmatrix}_{4x4}$$

The important thing is that if on one side it is +, on the other side it must be –.

3.2.2.5 Diagonal Matrix:

Elements above or below the main diagonal are zeros.

$$A = \begin{vmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{vmatrix}_{3x3}$$

3.2.2.6 Scalar Matrix:

It is a diagonal matrix whose elements are all equal on the main diagonal.

$$A = \begin{vmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix}_{4x4}$$

3.2.2.7 Identity Matrix:

A scalar matrix whose component elements equal to 1

$$A = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}_{3x3}$$

3.3 Operations between Matrices

3.3.1 Transposition:

It is an operation carried out on the elements of a matrix. It basically consists of exchanging the rows by the columns; the result of a matrix with the exchanged order, that is:

$$A = (a_{ij})\ mxn \rightarrow A = (a_{ij})\ nxm$$

$$A = \begin{vmatrix} 2 & 3 & 4 \\ 1 & 5 & 8 \\ -2 & 4 & 6 \\ 0 & 1 & 7 \end{vmatrix}_{4x3} \qquad A' = \begin{vmatrix} 2 & 1 & -2 & 0 \\ 3 & 5 & 4 & 1 \\ 4 & 8 & 6 & 7 \end{vmatrix}_{3x4}$$

3.3.2 Sum:

A minimum of two matrices are required. It occurs between matrices of the same order, and the operation is performed term by term.

$$A = \begin{vmatrix} 2 & 5 & 7 \\ 4 & -1 & 9 \\ 6 & 8 & 5 \end{vmatrix}_{3x3} \quad B = \begin{vmatrix} 9 & 8 & 13 \\ 1 & 7 & 26 \\ -6 & 4 & 3 \end{vmatrix}_{3x3} \quad A + B = \begin{vmatrix} 11 & 13 & 20 \\ 5 & 6 & 35 \\ 0 & 12 & 8 \end{vmatrix}_{3x3}$$

Subtraction would be A + (-B) or B + (-A).

3.4 Product between Matrices:

- Scaling of a matrix
- Product between matrices
- Properties

3.4.1 Scaling of a Matrix is any real number

Matrix A = (a_{ij}), $\alpha = 5$, α x $A = B_{mxn}$

$$A = \begin{vmatrix} 1 & 0 & 5 \\ 4 & 2 & 1 \end{vmatrix}_{Mxn} \qquad \alpha \text{ x } A = 5 \text{ x } \begin{vmatrix} 1 & 0 & 5 \\ 4 & 2 & 1 \end{vmatrix}_{2x3} = B = \begin{vmatrix} 5 & 0 & 25 \\ 20 & 10 & 5 \end{vmatrix}_{2x3}$$

3.4.2 Products between Matrices

a) Product of a row vector by a column vector

b) Product of a column vector by a row vector

c) Row vector multiplied by matrix

d) Matrix multiplied by column vector

e) Matrix multiplied by matrix

In order to obtain the product between matrices, the two matrices must be "conformable" (the number of columns in the 1st matrix must be equal to the rows in the 2nd matrix). If this condition is met, the operation is performed by multiplying each of the elements of the rows in the matrix, which premultiplies by each of the elements of the column of the matrix that post multiplies.

"If the number of columns of the matrix I premultiply is equal to the number of rows of the matrix I post multiply."

Given a matrix A = (a_{ij}) and a matrix B = (b_{ij})

$$A_{mxn} x B_{nxp} = C_{mxp}$$

3.4.2.1 Product of a Row Vector Multiplied by a Column Vector

$$A = \begin{vmatrix} 1 & 4 & 3 & 7 \end{vmatrix}_{1x4} \qquad B = \begin{vmatrix} 0 \\ 5 \\ 2 \\ -1 \end{vmatrix}_{4x1} = 19$$

$$Cn = ((1x0) + (4x5) + (3x2) + (7x-1)) = 19$$

The product of a row vector and a column vector is a scalar.

3.4.2.2 Product of a Column Vector Multiplied by a Row Vector

$$A = \begin{vmatrix} 0 \\ 5 \\ 2 \\ -1 \end{vmatrix}_{4x1} \qquad B = \begin{vmatrix} 1 & 4 & 3 & 7 \end{vmatrix}_{1x4} = \begin{vmatrix} 0 & 0 & 0 & 0 \\ 5 & 20 & 15 & 35 \\ 2 & 8 & 6 & 14 \\ -1 & -4 & -3 & -7 \end{vmatrix}_{4x4}$$

$$A_{4x1} \times B_{1x4} = C_{4x4}$$

The product of a column vector multiplied by a row vector is a matrix. Multiply element by element, row by column

3.4.2.3 Row Vector Multiplied by Matrix

$$A = \begin{vmatrix} 1 & 4 & 3 & 7 \end{vmatrix} \quad B = \begin{vmatrix} -2 & 0 & 5 \\ 1 & 3 & 6 \\ 2 & 4 & 7 \\ 5 & -3 & 1 \end{vmatrix}$$

1x4 4x3

$$C = \begin{vmatrix} C_{11} & C_{12} & C_{13} \end{vmatrix}$$ 1x3

$$C = \begin{vmatrix} 43 & 3 & 57 \end{vmatrix}$$ 1x3

$$A \times B = \begin{matrix} ((1x-2) + (4x1) + (3x2) + (7x5)) = 43 \\ ((1x0) + (4x3) + (3x4) + (7x-3)) = 3 \\ ((1x5) + (4x6) + (3x7) + (7x1)) = 57 \end{matrix}$$

The result of multiplying a row vector by a matrix is a row vector.

3.4.2.4 Matrix Multiplied by Column Vector

$$A = \begin{vmatrix} 10 & 9 & 7 \\ 4 & 8 & 2 \\ 1 & 0 & -4 \end{vmatrix} \quad x\ B = \begin{vmatrix} 1 \\ 2 \\ 3 \end{vmatrix} = \begin{vmatrix} C_{11} \\ C_{12} \\ C_{13} \end{vmatrix} = \begin{vmatrix} ((10x1)+(9x2)+(7X3) \\ ((4x1)+(8x2)+(2x3)) \\ ((1x1)+(0x2)+(-4x3)) \end{vmatrix} = \begin{vmatrix} 49 \\ 26 \\ 11 \end{vmatrix}$$

3x3 3x1 3x1 3x1

The result of multiplying a matrix by a column vector is a column vector.

3.4.2.5 Matrix multiplied by Matrix

$$A = \begin{vmatrix} 1 & 0 & 5 \\ 2 & 4 & 3 \end{vmatrix} \quad A\ x\ B = \begin{vmatrix} 1 & 0 & 5 \\ 2 & 4 & 3 \end{vmatrix} \quad x \quad \begin{vmatrix} 2 & 4 & 2 \\ 1 & 3 & 5 \\ 0 & 6 & 2 \end{vmatrix} = \begin{vmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \end{vmatrix}$$

2x3 2x3 3x3 2x3

$$B = \begin{vmatrix} 2 & 4 & 2 \\ 1 & 3 & 5 \\ 0 & 6 & 2 \end{vmatrix}$$

3x3

$$A_{2x3} x\ B_{3x3} = C_{2x3} \quad C = \begin{vmatrix} 2 & 34 & 12 \\ 8 & 38 & 30 \end{vmatrix}$$ 2X3

$C_{11} = ((1x2) + (0x1) + (5x0)) = 2$
$C_{12} = ((1x4) + (0x3) + (5x6)) = 34$
$C_{13} = ((1x2) + (0x5) + (5x2)) = 12$
$C_{21} = ((2x2) + (4x1) + (3x0)) = 8$
$C_{22} = ((2x4) + (4x3) + (3x6)) = 38$
$C_{23} = ((2x2) + (4x5) + (3x2)) = 30$

3.4.3 Properties

A x 0 = 0 Matrix 0
A x I = A Identity Matrix. I
A x B ≠ B x A

Some exceptions where commutativity property occurs:

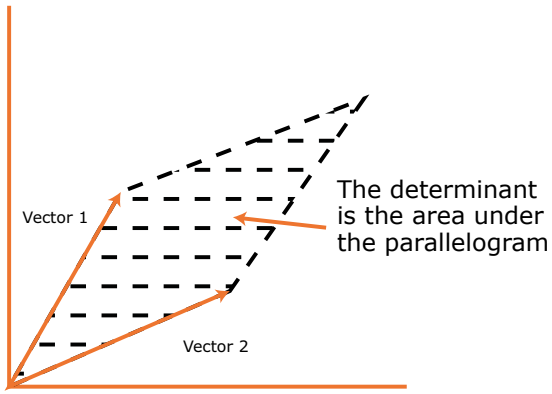
- A^{-1} The reverse
- $A \times I = I \times A$
- $A \times A^{-1} = A^{-1} \times A = I$
- $A \times A' = A' \times A$
- $(A \times B)' = B'A'$
- $A \times (B \times C) = (A \times B) \times C$

3.5 Determinant of a Matrix

The determinant is a numerical value obtained over the elements of a matrix, with notation: $A(a_{ij})n \; x \; m$, the determinant is |A|.

It is only defined for square matrices, the calculation will allow us to a) determine whether or not this matrix has an inverse, or all square matrices have an inverse, singularity, or non-singularity, b) establish the rank of a matrix, and c) determine whether or not a system of simultaneous linear equations (SELS in Spanish) has a solution.

Image 3.1. The determinant ...



For a matrix of order 2x2 A(a_{ij}) is calculated:

$$A = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}_{2x2} \qquad |A| = ((a_{11} \times a_{22}) - (a_{21} \times a_{12}))$$

$$A = \begin{vmatrix} 7 & 1 \\ 3 & 2 \end{vmatrix}_{2x2} \qquad |A| = ((7X2) - (3X1) = 11$$

$$|A| = 11$$

If the matrix is of higher order (3x3), then it is developed by the Sarrus method or by the Laplace method (or cofactor expansion).

3.5.1 Sarrus method (diagonalization method)

Consists of repeating the first two columns of the matrix and drawing diagonals from left to right and right to left. The determinant shall be equal to the sum of the product of six terms; the terms that go from left to right are added and those that go from right to left are subtracted.

$$A=\begin{vmatrix} 1 & 5 & 4 \\ 2 & 3 & 0 \\ -1 & 6 & 2 \end{vmatrix}_{3x3} \qquad \begin{vmatrix} 1 & 5 & 4 & 1 & 5 \\ 2 & 3 & 0 & 2 & 3 \\ -1 & 6 & 2 & -1 & 6 \end{vmatrix}$$

$$|A| =((1x3x2)+(5x0x-1)+(4x2x6)-(-1x3x4)-(6x0x1)-(2x2x5))$$

$$|A| = 6 + 0 + 48 + 12 - 0-20$$

$$|A| = 46 \neq 0$$

3.5.2 Laplace Method or Cofactor Expansion

Laplace states that a row must be fixed and successively block one by one the columns of the matrix until the matrices of order 2x2 are obatined. To these matrices, the determinant is calculated as stated above, and it is multiplied by its numerical position and sign (only of the fixed row), *i.e.*, if we have the following matrix:

$$\begin{vmatrix} 13 & 24 & 3 \\ 12 & 32 & 1 \\ 4 & 5 & 8 \end{vmatrix}_{3x3} \quad (-) \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}_{(-)} \quad (-) \rightarrow \begin{vmatrix} 13 & -24 & 3 \\ -12 & 32 & -1 \\ 4 & -5 & 8 \end{vmatrix}_{3x3}$$

If the sum of the subscripts is an even number, the cofactor takes the sign of the smaller number; however, if the sum of the subscripts is an odd number, a change of sign with respect to the sign of the lesser number occurs.

That is, in the first row and column (a11), the numeric position is + 13 (the sign of the smaller number is positive) and the sum of the subscripts is a1+1 = 2 (even), therefore, its sign remains the same.

In the first row and second column (a12), the numeric position is + 24 (the sign of the smaller number is positive) and the sum of the subscripts is a1+2 = 3 (odd), therefore, its sign must change to -24.

In the first row and third column (a13), the numeric position is + 3 (the sign of the smaller number is positive) and the sum of the subscripts is a1+3 = 4 (even), therefore, their sign remains the same.

WITH THE 1ST ROW FIXED

Blocking the 1st Column

13	(-) 24	3
12	32	1
4	5	8

a_{11} 3x3

13 x((32x8) -(5x1))

13 x 251

= 3263

WITH THE 1ST ROW FIXED

Blocking the 2nd Column

13	(-) 24	3
12	32	1
4	5	8

a_{22} 3x3

-24x((12x8) -(4x1))

-24 x 92

= -2208

WITH THE 1ST ROW FIXED

Blocking the 3rd Column

13	(-) 24	**3**
12	32	1
4	5	8

a_{33} 3x3

3x((12x5) -(4x32))

3 x -68

= -204

|A| = 3263 + (-2208) + (-204)

|851|

If row 1 is fixed and column 1 is blocked, the determinant would be 251, which is multiplied by its numerical position and sign: 13 plus the determinant when blocking column 2 by its numerical position (24) and sign (-), plus the determinant when blocking column 3 by its numerical position and sign (3), finally it is obtained that the determinant of the matrix of order 3x3 is 851.

Similarly, Laplace states that the determinant can be found by fixing any of the rows of the matrix, and the result should be the same **|**A**|** = 851.

Fixing the second row, we would obtain:

WITH THE 2nd ROW FIXED

Blocking the 1st Column

13	24	3
(-) 12	32	(-) 1
4	5	8

3x3

-12x((24x8) -(5x3))

= - 2124

WITH THE 2nd ROW FIXED

Blocking the 2nd Column

13	24	3
(-) 12	32	(-) 1
4	5	8

3x3

32x((13x8) -(4x3))

= 2944

WITH THE 2nd ROW FIXED

Blocking the 3rd Column

13	24	3
(-) 12	32	(1)
4	5	8

3x3

-1x((13x5) -(4x24))

= 31

|A|= -2124 + 2944 +31

|851|

Fixing the third row, we would obtain:

WITH THE 3rd ROW FIXED

Blocking the 1st Column

13	24	3
12	32	1
4	(-) 5	8

3x3

4x((24x1) -(32x3))

=–288

WITH THE 3rd ROW FIXED

Blocking the 2nd Column

13	24	3
12	32	1
4	(-) 5	8

3x3

-5x((13x1) -(12x3))

= 115

WITH THE 3rd ROW FIXED

Blocking the 3rd Column

13	24	3
12	32	1
4	(-) 5	8

3x3

8x((13x32)-(12x24))

= 1024

$|A|$ = -288 + 115 + 1024

|851|

3.5.3 Properties of Determinants

- If two rows or two columns are identical the determinant is zero; similarly, if a row or column is a multiple of another, its determinant is zero. Likewise, if the elements of a row or column are zero, its determinant is zero.
- If a row or column is multiplied by a scalar, the determinant will be *k* times the value of the scalar.
- If all the rows and columns of a matrix (transpose) are exchanged, its determinant does not change. $|A| = |A'|$.
- If two rows and two columns of a matrix are exchanged, the sign of the determinant changes, but not its numerical value.
- If a multiple of another row or column is added to or subtracted from a row and colum n, the value of its determinant is not altered.

3.6 Inverse Matrix

The inverse is an operation conducted or set on a square matrix and it replaces the division operation into scalars.

The necessary condition is that the matrix be $A = (a_{ij})n_x n$.

The sufficient condition is that the matrix is non-singular $|A| \neq 0$

There are two methods to calculate the inverse (Gauss-Jordan method and the adjugate matrix). The latter will be addressed due to its simplicity and familiarity with the requirements.

3.6.1 Adjugate Matrix

It is based on the Laplace expansion or cofactor expansion, discussed in section 3.5.2.

$$A^{-1} = \frac{1}{/A/} adjoint.$$

note that, if the determinant is zero, there is no inverse.

For this purpose, we continue with the previous example where the determinant is |851| ≠ 0; therefore, it is a non-singular matrix and has an inverse:

13	24	3
12	32	1
4	5	8

3x3

THE 1ST FIXED ROW

Blocking the 1st Column

13	24	3
12	32	1
4	5	8

3x3

32	1
5	8

2x2

THE 1ST FIXED ROW

Blocking the 2nd Column

13	24	3
12	32	1
4	5	8

3x3

12	1
4	8

2x2

THE 1ST FIXED ROW

Blocking the 3rd Column

13	24	3
12	32	1
4	5	8

3x3

12	32
4	5

2x2

These three 2x2 matrices are part of the first row of the cofactor matrix.

THE 2nd FIXED ROW

Blocking the 1st Column

13	24	3
12	32	1
4	5	8

3x3

THE 2nd FIXED ROW

Blocking the 2nd Column

13	24	3
12	32	1
4	5	8

3x3

THE 2nd FIXED ROW

Blocking the 3rd Column

13	24	3
12	32	1
4	5	8

3x3

24	3
5	8

2x2

13	3
4	8

2x2

13	24
4	5

2x2

These three 2x2 matrices are part of the second row of the cofactor matrix.

THE 3rd FIXED ROW

Blocking the 1st Column

13	24	3
12	32	1
4	5	8

3x3

THE 3rd FIXED ROW

Blocking the 2nd Column

13	24	3
12	32	1
4	**5**	8

3x3

THE 3rd FIXED ROW

Blocking the 3rd Column

13	24	3
12	32	1
4	5	**8**

3x3

24	3
32	1

2x2

13	3
12	1

2x2

13	24
12	32

2x2

Finally, these three 2x2 matrices are part of the third and last row of the cofactor matrix.

The cofactor matrix is:

$$\begin{vmatrix} 32 & 1 \\ 5 & 8 \end{vmatrix} \quad \begin{vmatrix} 12 & 1 \\ 4 & 8 \end{vmatrix} \quad \begin{vmatrix} 12 & 32 \\ 4 & 5 \end{vmatrix}$$

$$\begin{vmatrix} 24 & 3 \\ 5 & 8 \end{vmatrix} \quad \begin{vmatrix} 13 & 3 \\ 4 & 8 \end{vmatrix} \quad \begin{vmatrix} 13 & 24 \\ 4 & 5 \end{vmatrix}$$

$$\begin{vmatrix} 24 & 3 \\ 32 & 1 \end{vmatrix} \quad \begin{vmatrix} 13 & 3 \\ 12 & 1 \end{vmatrix} \quad \begin{vmatrix} 13 & 24 \\ 12 & 32 \end{vmatrix}$$

(-)

$$(-)\begin{bmatrix} 251 & 92 & -68 \\ 177 & 92 & -31 \\ -72 & -23 & 128 \end{bmatrix}(-)$$

(-)

3x3

This is the **cofactor matrix**

$$\begin{bmatrix} 251 & -92 & -68 \\ -177 & 92 & 31 \\ -72 & 23 & 128 \end{bmatrix}$$

3x3

The transpose to the cofactor matrix is performed and the adjoint is obtained.

$$C=\begin{bmatrix} 251 & -92 & -68 \\ -177 & 92 & 31 \\ -72 & 23 & 128 \end{bmatrix} \longrightarrow$$

3x3

Adjoint

$$\begin{bmatrix} 251 & -177 & -72 \\ -92 & 92 & 23 \\ -68 & 31 & 128 \end{bmatrix}$$

3x3

Recalling equation $A^{-1} = \frac{1}{/A/}\ adjoint$ **(3.1)**

We have:

$$A^{-1} = \frac{1}{/851/}x \begin{vmatrix} 251 & -177 & -72 \\ -92 & 92 & 23 \\ -68 & 31 & 128 \end{vmatrix} = \begin{vmatrix} 0{,}294947 & -0{,}20799 & -0{,}08461 \\ -0{,}10811 & 0{,}108108 & 0{,}027027 \\ -0{,}07991 & 0{,}036428 & 0{,}150411 \end{vmatrix}$$

A^{-1}

3.6.2 Properties of the Inverse

- The inverse of a square matrix, if it exists, is unique.

$A = (a_{ij})nxn$

- The result of a matrix postmultiplied or premultiplied by its inverse is the identity matrix.

$A\,x\,A^{-1} = A^{-1}x\,A = I$

- The inverse of an inverse matrix is equal to the original matrix

$(A^{-1})^{-1} = A$

- The inverse of a transpose matrix is equal to the transpose of the inverse matrix.

$(A')^{-1} = (A^{-1})$

- The inverse of the product of two matrices is equal to the product of two inverses.

$(A\,x\,B)^{-1} = B^{-1}x\,A^{-1}$

3.6.3 Economic Applications of the Inverse

The economic applications of the inverse matrix include a) the input-output matrix, b) systems of simultaneous linear equations, optimization of differential functions, and c) estimation of economic models (the reason for this chapter of matrix algebra and the basis to develop the following chapter).

3.6.3.1 The Input-Output Matrix

Analysis was first proposed by Leontief and applied to the U.S. economy. It is a model that sought to estimate the future production of different industries given a change in the final demand for the products they produced.

It established the following assumptions:

- Each industry or sector produces a single product, and none of them produces the same product as the other.
- It states that, in each industry and sector, the total value of production is equal to the total value of inputs used.
- Given that technological changes occur in the medium and long term, estimates made on the input-output matrix are only valid in the short term.

It also indicates that the columns of the matrix represent the proportion of input used by each industry, and the rows represent the production of each industry.

a_{11} the amount of input used by industry 1 but consumed by industry 1.

a_{21} the amount of input used by industry 1 but produced by industry 2.

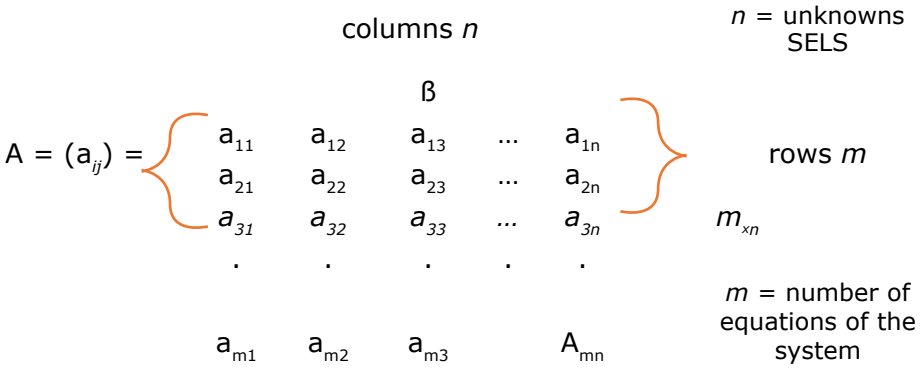
a_{24} the amount of input used by industry 4 but produced by industry 2.

3.6.3.2 Systems of Simultaneous Linear Equations (SELS)

$$
\begin{array}{ccccccc}
a_{11}X_1 & a_{12}X_2 & a_{13}X_3 & \dots & a_1X_n & = & b_1 \\
a_{21}X_2 & a_{22}X_2 & a_{23}X_3 & \dots & a_2X_n & = & b_2 \\
a_{31}X_3 & a_{32}X_2 & a_{33}X_3 & \dots & a_3X_n & = & b_3 \\
. & . & . & . & . & . & \\
a_mX_1 & a_mX_2 & a_mX_3 & & a_{mn}X_n & = & B_m \\
\downarrow & & \downarrow & & \downarrow & &
\end{array}
$$

Coefficients — unknown SELS — independent terms

In order to find the value of the unknown variables, this SELS must take the whole system of linear equations:



Conditions for determining whether SELS has a solution:

• The SELS is consistent if the rank of a coefficient matrix is equal to the rank of an expanded matrix, where the expanded matrix is the same matrix with the vector in independent terms.

$$r(A)^{-1} = r(Ax)$$

If the rank of the coefficient matrix is unequal to the rank of the expanded matrix, then it has no solution.

• If the system is consistent, it has a solution. It can be a unique solution or a multiple solution.

The unique solution is when the rank of *A* is equal to the rank of *B* and equals *n*, in which *n* is the number of unknowns in the system.

A SELS is complete when the number of equations is equal to the number of unknowns.

The multiple solution occurs when the rank of the coefficient matrix is equal to the rank of the expanded matrix, but the value of that rank is smaller than n or the number of unknowns.

3.6.4 Solution Methods

There are four solution methods; all based on Gaussian, Jordan, Inverse, and Cramer, where:

Gauss Jordan	They allow for the solution of Equations *nxm* of any order.
Inverse Cramer	They are only suitable for squared systems

Inverse and Cramer are only for nxn square matrices.

If we have the following system of equations:

3.6.4.1 The solution with Cramer's rule is:

$$\begin{aligned} 7x_1 - x_2 - x_3 &= 0 \\ 10x_1 - 2x_2 + x_3 &= 8 \\ 6x_1 + 3x_2 - 2x_3 &= 7 \end{aligned}$$

$$\underset{}{A} \times \underset{}{X} = Y$$

$$A = \underset{3x3}{\begin{vmatrix} 7 & -1 & -1 \\ 10 & -2 & 1 \\ 6 & 3 & -2 \end{vmatrix}} \times \underset{3x1}{\begin{vmatrix} X_1 \\ X_2 \\ X_3 \end{vmatrix}} = \underset{3x1}{\begin{vmatrix} 0 \\ 8 \\ 7 \end{vmatrix}}$$

Column Y is replaced by each of the columns of the original matrix, generating three new matrices (A_1, A_2, and A_3). The determinant of each matrix ($|A_1|$, $|A_2|$, and $|A_3|$) is calculated and it is divided by the determinant of the original matrix $|A|$; thus, the unknowns x_1, x_2, and x_3 are obtained.

$$\underset{}{A_1} \quad \underset{}{A_2} \quad \underset{}{A_3}$$

$$\underset{3x3}{\begin{vmatrix} \mathbf{0} & -1 & -1 \\ \mathbf{8} & -2 & 1 \\ \mathbf{7} & 3 & -2 \end{vmatrix}} \quad \underset{3x3}{\begin{vmatrix} 7 & \mathbf{0} & -1 \\ 10 & \mathbf{8} & 1 \\ 6 & \mathbf{7} & -2 \end{vmatrix}} \quad \underset{3x3}{\begin{vmatrix} 7 & -1 & \mathbf{0} \\ 10 & -2 & \mathbf{8} \\ 6 & 3 & \mathbf{7} \end{vmatrix}}$$

$$\tilde{x}_1 = \frac{|A_1|}{|A|}$$

$$\tilde{x}_2 = \frac{|A_2|}{|A|}$$

$$\tilde{x}_3 = \frac{|A_3|}{|A|}$$

We find the determinant of the original matrix **|A|** =

WITH THE 1ST ROW FIXED

Blocking the 1st Column

7	(-) -1	-1
10	-2	1
6	3	-2

a_{11} 3x3

7x((-2x-2) -(3x1))

7 x 1

= 7

WITH THE 1ST ROW FIXED

Blocking the 2nd Column

7	(-) -1	-1
10	-2	1
6	3	-2

a_{22} 3x3

(-) -1x((10x-2) -(6x1))

1 x -26

=−26

WITH THE 1ST ROW FIXED

Blocking the 3rd Column

7	(-) -1	-1
10	-2	1
6	3	-2

a_{33} 3x3

-1x((10x3) -(6x-2))

-1 x 42

=−42

|A| = 7 + (-26) + (-42)

|-61|

The determinant of the new matrices ($|A_1|$,$|A_2|$, and $|A_3|$) is found=$|A_1|$=

WITH THE FIXED 1ST ROW

Blocking the 1st Column

0	(-) -1	-1
8	-2	1
7	3	-2

a_{11} 3x3

0x((-2x-2) -(3x1))

0 x 1

= 0

WITH THE FIXED 1ST ROW

Blocking the 2nd Column

0	(-) -1	-1
8	-2	1
7	3	-2

a_{22} 3x3

(-) -1x((8x-2) -(7x1))

1 x -23

=−23

WITH THE FIXED 1ST ROW

Blocking the 3rd Column

0	(-) -1	-1
8	-2	1
7	3	-2

a_{33} 3x3

-1x((8x3) -(7x-2))

-1 x 42

=−38

$|A_1|$ = 0 + (-23) + (-38)

|-61|

$|A_2|=$

WITH TH FIXED Pª ROW

Blocking the 1st Column

7	(-) 0	-1
10	8	1
6	7	-2

a_{11} 3x3

7x((8x-2) -(7x1))
7 x -23
= -161

WITH THE FIXED Pª ROW

Blocking the 2nd Column

7	(-) 0	-1
10	8	1
6	7	-2

a_{22} 3x3

(-) 0x((10x-2) -(6x1))
0 x -26
= 0

WITH THE FIXED Pª ROW

Blocking the 3rd Column

7	(-) 0	-1
10	8	1
6	7	-2

a_{33} 3x3

-1x((10x7) -(6x-8))
-1 x 22
=–22

[A2] = -161 + (0) + (-22)
|-183|

$|A_3|=$

WITH THE FIXED Pª ROW

Blocking the 1st Column

7	(-) -1	0
10	-2	8
6	3	7

$a1_1$ 3x3

7x((-2x7)-(3x8))
7 x -38
= -266

WITH THE FIXED Pª ROW

Blocking the 2nd Column

7	(-) -1	0
10	-2	8
6	3	7

a_{22} 3x3

(-) -1x((10x7)-(6x8))
1 x 22
= 22

WITH THE FIXED Pª ROW

Blocking the 3rd Column

7	(-) -1	0
10	-2	8
6	3	7

a_{33} 3x3

0x((10x3) -(6x-2))
0 x 42
= 0

$[A_3]$ = -266 + (22) + (0)
|-244|

We replace the determinants found and the value of our unknowns $x_1=1$, $x_2=3$, and $x_3=4$ are found.

$$\tilde{x}_1 = \frac{|A_1|}{|A|} = \frac{-61}{-61} = 1$$

$$\tilde{x}_2 = \frac{|A_2|}{|A|} = \frac{-183}{-61} = 3$$

$$\tilde{x}_3 = \frac{|A_3|}{|A|} = \frac{-244}{-61} = 4$$

3.6.4.2 The solution with the inverse is:

$$\begin{aligned} 7x_1 - x_2 - x_3 &= 0 \\ 10x_1 - 2x_2 + x_3 &= 8 \\ 6x_1 + 3x_2 - 2x_3 &= 7 \end{aligned}$$

$$A = \begin{vmatrix} 7 & -1 & -1 \\ 10 & -2 & 1 \\ 6 & 3 & -2 \end{vmatrix} \times \begin{vmatrix} X_1 \\ X_2 \\ X_3 \end{vmatrix} = \begin{vmatrix} 0 \\ 8 \\ 7 \end{vmatrix}$$

A (3x3) × X (3x1) = Y (3x1)

As in the previous example, we remember that the determinant is |-61| ≠ 0, therefore, it is a non-singular matrix and has an inverse:

We continue with the cofactor matrix

THE FIXED Pª ROW

Blocking the 1st Column

7	-1	-1
10	-2	1
6	3	-2

3x3

-2	1
3	-2

2x2

THE Pª FIXED ROW

Blocking the 2nd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

10	1
6	-2

2x2

THE Pª FIXED ROW

Blocking the 3rd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

10	-2
6	3

2x2

These three 2x2 matrices are part of the cofactor matrix's first row.

THE 2nd FIXED ROW
Blocking the 1st Column

7	-1	-1
10	-2	1
6	3	-2

3x3

-1	-1
3	-2

2x2

THE 2nd FIXED ROW
Blocking the 2nd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

7	-1
6	-2

2x2

THE 2nd FIXED ROW
Blocking the 3rd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

7	-1
6	3

2x2

These three 2x2 matrices are part of the cofactor matrix's second row.

THE FIXED 3rd ROW
Blocking the 1st Column

7	-1	-1
10	-2	1
6	3	-2

3x3

-1	-1
-2	1

2x2

THE FIXED 3rd ROW
Blocking the 2nd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

7	-1
10	1

2x2

THE FIXED 3rd ROW
Blocking the 3rd Column

7	-1	-1
10	-2	1
6	3	-2

3x3

7	-1
10	-2

2x2

Finally, these three 2x2 matrices are part of the cofactor matrix's third and last row.

The cofactor matrix is:

$$\begin{vmatrix} -2 & 1 \\ 3 & -2 \end{vmatrix} \quad \begin{vmatrix} 10 & 1 \\ 6 & -2 \end{vmatrix} \quad \begin{vmatrix} 10 & -2 \\ 6 & 3 \end{vmatrix}$$

$$\begin{vmatrix} -1 & -1 \\ 3 & -2 \end{vmatrix} \quad \begin{vmatrix} 7 & -1 \\ 6 & -2 \end{vmatrix} \quad \begin{vmatrix} 7 & -1 \\ 6 & 3 \end{vmatrix}$$

(-)

1	-26	42
5	-8	27
-3	17	-4

(-) 3x3

This is the **cofactor Matrix**

1	26	42
-5	-8	-27
-3	-17	-4

3x3

The transpose of the cofactor matrix and the adjoint is obtained.

$$\begin{vmatrix} -1 & -1 \\ -2 & 1 \end{vmatrix} \quad \begin{vmatrix} 7 & -1 \\ 10 & 1 \end{vmatrix} \quad \begin{vmatrix} 7 & -1 \\ 10 & -2 \end{vmatrix}$$

C=

1	26	42
-5	-8	-27
-3	-17	-4

3x3 →

Adjoint

1	-5	-3
26	-8	-17
42	-27	-4

3x3

Recall the **(3.1)** equation:

$$A^{-1} = \frac{1}{/A/} \, adjoint$$

We have:

$$A^{-1} = \underset{-0,01639344}{\frac{1}{|-61|}} x \begin{vmatrix} 1 & -5 & -3 \\ 26 & -8 & -17 \\ 42 & -27 & -4 \end{vmatrix} = \overset{A^{-1}}{\begin{vmatrix} -0,01639 & 0,081967 & 0,04918 \\ -0,42623 & 0,131148 & 0,278689 \\ -0,68852 & 0,442623 & 0,065574 \end{vmatrix}}$$

$$(A)^{-1} x(Y) = x_n \quad \overset{A^{-1}}{\begin{vmatrix} -0,01639 & 0,081967 & 0,04918 \\ -0,42623 & 0,131148 & 0,278689 \\ -0,68852 & 0,442623 & 0,065574 \end{vmatrix}} \underset{3x3}{x} \overset{Y}{\begin{vmatrix} 0 \\ 8 \\ 7 \end{vmatrix}} = \underset{1x3}{\begin{vmatrix} x_1 \\ x_2 \\ x_3 \end{vmatrix}} = \overset{x}{\underset{1x3}{\begin{vmatrix} 1 \\ 3 \\ 4 \end{vmatrix}}}$$

Final Exercises, Chap. 3

1. Define if the matrix is singular.

a.

$$\begin{vmatrix} 2 & 32 & 4 \\ 4 & 32 & 8 \\ 6 & 5 & 12 \end{vmatrix}$$

b.

$$\begin{vmatrix} 13 & 2 & -6 \\ -3 & 5 & 8 \\ 23 & 10 & 4 \end{vmatrix}$$

c.

$$\begin{vmatrix} 23 & 10 & 6 \\ 3 & 32 & 8 \\ 46 & 20 & 12 \end{vmatrix}$$

2. Find the determinant of the following matrices:

a.

$$\begin{vmatrix} 1 & 2 & -3 \\ -2 & 1 & 9 \\ 3 & -9 & 1 \end{vmatrix}$$

b.

$$\begin{vmatrix} 87 & 37 & 648 \\ 92 & 43 & 732 \\ 11 & -55 & 0 \end{vmatrix}$$

c.

$$\begin{vmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{vmatrix}$$

d.

$$\begin{vmatrix} -23 & -15 & -10 \\ -4 & -72 & -31 \\ -7 & -1 & -2 \end{vmatrix}$$

3. Graph the cofactor matrix and the respective solution.

a.

$$\begin{vmatrix} 3 & -15 & 11 \\ 4 & -36 & 9 \\ 8 & -1 & -2 \end{vmatrix}$$

b.

$$\begin{vmatrix} 9 & -6 & -1 \\ 2 & 5 & 4 \\ 3 & -15 & 7 \end{vmatrix}$$

4. Find the inverse matrix.

a.

$$\begin{vmatrix} 23 & -2 & 8 \\ 235 & 47 & 98 \\ 16 & 24 & -30 \end{vmatrix}$$

b.

$$\begin{vmatrix} 7 & 34 & 56 \\ 64 & 12 & 9 \\ 82 & -4 & 24 \end{vmatrix}$$

c.

$$\begin{vmatrix} 62 & 34 & -72 \\ 4 & -3 & -5 \\ -17 & 9 & 63 \end{vmatrix}$$

d.

$$\begin{vmatrix} -96 & -267 & -16 \\ -107 & 2 & -12 \\ -32 & 1 & 0 \end{vmatrix}$$

e.

$$\begin{vmatrix} 29 & 0 & 1 \\ 0 & 8 & 9 \\ 14 & 3 & 0 \end{vmatrix}$$

5. **Solve the following SELS:**

a. $3x_1 + 5x_2 = 6$
$12x_1 - 2x_2 + 8x_3 = 2$
$21x_1 + 3x_2 - 7x_3 = 21$

b. $2x_1 + 14x_2 + 8x_3 = 3$
$-6x_1 - 2x_2 = 54$
$37x_1 + 24x_2 + 17x_3 = 9$

c. $21x_2 + 93x_3 = 1$
$x_1 + 33x_2 + 67x_3 = 10$
$78x_1 + 45x_2 + 84x_3 = 2$

CHAPTER 4

ESTIMATION OF ECONOMIC MODELS

Our first task is to estimate the population regression function (PRF) based on the sample regression function (SRF) as accurately as possible. To this end, there are several ways of calculating the SRF, but the most widely used is the ordinary least squares method (as regards for regression analysis). Moreover, it should be remembered that the OLS method will allow us to analyze only single-equation linear models and is accurate for the purpose of this text.

4.1 Parameter Estimation by OLS

The OLS method is attributed to Carl Friedrich Gauss, a German mathematician. It has very attractive statistical properties for certain assumptions to be studied:

Remember the PRF

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \qquad \textbf{(5)}$$

we cannot observe it directly, this function must be estimated from the SRF.

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \quad \textbf{(6)}$$

$$Y_i = \hat{Y}_i + \hat{u}_i \quad \textbf{(4.1)}$$

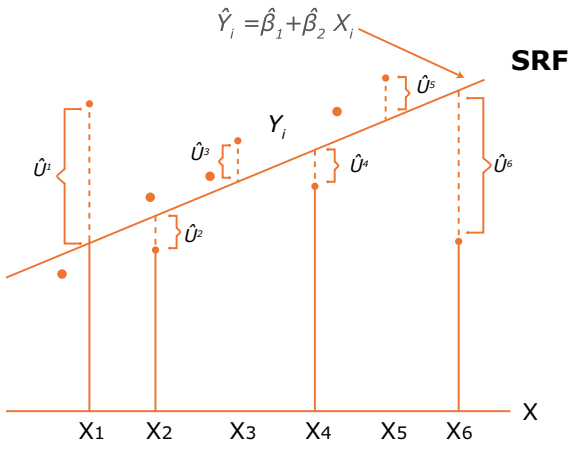
where Y_i is the estimated value of (conditional mean) Y_i. However, how is SRF itself determined? Therefore, (4.1) is expressed as:

$$\hat{u}_i = Y_i - \hat{Y}_i \quad \text{remember} \quad \textbf{(3)}$$

we replace = $Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$ **(4.2)**

which reflects that the residuals ($\hat{u}_i$) are simply the differences between the observed and estimated values of Yi. In this way, the criterion is set to select the SFR in such a way that the sum of the residuals is the smallest possible $\Sigma \hat{u}_i$(this criterion is not very efficient, although it is attractive, see Image 4.1).

Image 4.1. The SRF



By adopting the criterion of minimizing $\Sigma \hat{u}_i$, Image 4.1 shows that the residuals $\hat{u}_2$, $\hat{u}_3$, $\hat{u}_4$, $\hat{u}_5$ as well as the residuals $\hat{u}_1$ and $\hat{u}_6$ receive the same weight in the sum ($\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4 + \hat{u}_5 + \hat{u}_6$), although the first four are much closer to the SRF than the last two.

That is, all residues are given the same weight regardless of how far or close the observations are to the SRF. It is also possible that the total sum of residues is zero[11] even though the $\hat{u}_i$ are scattered around the SRF.

Therefore, we can avoid this problem if we use the OLS method, which states that the SFR can be determined in such a way that it is as small as possible, where $\hat{u}_i^2$ are the squared residuals. Hence, the more distant residuals are given more weight. The situation that arises in the previous example with u_1 and u_6, in which in the minimization of the $\Sigma \hat{u}_i$, the u_i regardless of their degree of dispersion, the sum is small and it could not arise in the OLS process. Moreover, the greater the u_i, the greater the $\Sigma \hat{u}_i$.

It is based on a model:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + \hat{u}_i$$

And in matrix form:

$$Y = X\hat{\beta} + \hat{u}$$

That is to say:

$$\begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \\ : \\ Y_n \end{vmatrix} = \begin{vmatrix} 1 & X_{12} & X_{13} & \dots & X_{k1} \\ 1 & X_{22} & X_{23} & \dots & X_{k2} \\ 1 & X_{32} & X_{33} & \dots & X_{k3} \\ : & : & : & : & : \\ 1 & X_{2n} & X_{3n} & \dots & X_{kn} \end{vmatrix} \times \begin{vmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ : \\ \hat{\beta}_k \end{vmatrix} + \begin{vmatrix} \hat{u}_1 \\ \hat{u}_1 \\ \hat{u}_1 \\ : \\ \hat{u}_1 \end{vmatrix}$$

$Y_{n \times 1}$ $\quad$ $n \times k$ $\quad$ $K \times n$ $\quad$ $n \times 1$

11 If we were to assume that have the values of 8,-3, +2, -2, +3, and -8, respectively, the sum of these residuals is zero, even though two residuals are found with a higher degree of dispersion around the SFR.

The goal is to reduce errors

$$Minimizing \sum \hat{u}^2 = Minimizing(\hat{u}'\hat{u})$$

$$\hat{u}'\hat{u} = \begin{bmatrix}\hat{u}_1 & \hat{u}_2 & \hat{u}_3 & \dots & \hat{u}_n\end{bmatrix} * \begin{bmatrix}\hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \vdots \\ u_n\end{bmatrix} = \hat{u}_1^2 + \hat{u}_2^2 + \hat{u}_3^2 + \dots + \hat{u}_n^2 = \sum \hat{u}^{2`}$$

And we assume that $Y = X\hat{\beta} + \hat{u}$

if we solve $\hat{u} \rightarrow \hat{u} = Y - X\hat{\beta}$

Minimizing $\sum \hat{u}^2 = \min(\hat{u}'\hat{u}) = \min\left[Y - X\hat{\beta}\right]'\left[Y - X\hat{\beta}\right]$

recalling the property $(A+B)' = A' + B'$

we have,

$$\min\left[Y - X\hat{\beta}\right]'\left[Y - X\hat{\beta}\right] = \min\left[Y' - \hat{\beta}'X'\right]\left[Y - X\hat{\beta}\right]$$

we develop the product by multiplying term by term

$$\min\left[Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}\right]$$

Since $Y'X\hat{\beta} = [\,Scale\,] = \hat{\beta}'X'Y$

$$\Rightarrow \min\left[Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}\right]$$

To minimize, we derive with respect to betas

$$\frac{\partial(Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta})}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta}$$

where $\frac{\partial a'\hat{\beta}}{\partial\hat{\beta}} = a$ *and a is* $X'Y$ $\quad \frac{\partial\hat{\beta}'a\hat{\beta}}{\partial\hat{\beta}} = 2a\hat{\beta}$

$$\Rightarrow X'Y = X'X\hat{\beta}'$$

We solve $\hat{\beta}$

To do this, we remember that there is no matrix division, so we multiply by the inverse on both sides to reduce terms,

$$(X'X)^{-1}X'Y = (X'X)^{-1}X'X\hat{\beta} \qquad \textit{where, } (X'X)\text{-}1\ X'X = 1$$
$$(X'X)^{-1}X'Y = I\hat{\beta} \qquad \textit{where, } I\beta\hat{} = \beta\hat{}$$
$$\rightarrow \hat{\beta} = (X'X)^{-1}X'Y \qquad \textbf{(4.3)}$$

4.2 Properties of the Parameter's Estimators

- They are linear $\Rightarrow \hat{\beta} = (X'X)^{-1}X'Y$
  $\hat{\beta}_{kx1} = A_{kxn}Y_{nx1}$
- They are unbiased $E(\hat{\beta}) = \beta$
- They have minimum variance $Cov(\hat{\beta}) = \sigma_{\mu}^{2}(X'X)^{-1}$ **(4.4)**

Recall equation **(4.3)**

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

If we substitute *Y* **(4)**

$$Y = X\beta + u$$

we have

$$\hat{\beta} = (X'X)^{-1}X'(X\beta + u)$$
$$\hat{\beta} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u$$
$$\textit{recalling} \quad (X'X)^{-1}X'X = I$$
$$\textit{where} \quad I\beta = \beta$$
$$\hat{\beta} = I\beta + (X'X)^{-1}X'u$$
$$\hat{\beta} - \beta = (X'X)^{-1}X'u$$

For the property of unbiasedness:

$$= E\left[\left(\hat{\beta}-\beta\right)\left(\hat{\beta}-\beta\right)'\right]$$
$$= E\left[\left((X'X)^{-1}X'u\right)\left((X'X)^{-1}X'u\right)'\right]$$
$$= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \quad \textit{under the assumption } E(uu') = \sigma_u^2 I$$

$$= (X'X)^{-1}X'E(uu')X(X'X)^{-1}$$
$$= (X'X)^{-1}X'(\sigma_u^2 I)X(X'X)^{-1}$$

$$= (X'X)^{-1}(X'X)\sigma_u^2(X'X)^{-1}$$
$$= I\sigma_u^2(X'X)^{-1}$$

$$Cov(\hat{\beta}) = \sigma_\mu^2(X'X)^{-1} \qquad \textbf{(4.4)}$$

where σ_μ^2 *is determined:*

$$\sigma_u^{\,2} = \frac{\hat{u}'\hat{u}}{n-k} = \frac{\Sigma\hat{u}_i^{\,2}}{n-k} = \frac{y'y-\hat{\beta}'x'y}{n-k} \qquad \textbf{(4.5)}$$

4.3 Coefficient of Determination

- It measures the percentage by which the exogenous variables explain the variation of the endogenous variable.
- Variance decomposition

$$\sum_1^n y_i^2 = \sum_1^n \hat{y}_i^2 + \sum_1^n \hat{\mu}_i^2 \qquad \textbf{(4.6)}$$

- SST=ESS+RSS

$$\sum_1^n y_i^2 = \sum (Y_i - \bar{Y})^2 = \quad \textit{TOTAL SUM OF SQUARES} \qquad \textbf{(4.7)}$$

$$\sum_1^n \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \quad \textit{EXPLAINED SUM OF SQUARES} \qquad \textbf{(4.8)}$$

$$\sum_1^n \hat{\mu}_i^2 = \sum (Y_i - \hat{Y}_i)^2 = \textit{RESIDUAL SUM OF SQUARES (RSS)} \qquad \textbf{(4.9)}$$

4.3.1 If the model has an independent term, the R^2 is calculated:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \qquad 0 < R^2 < 1 \tag{4.10}$$

in any case

$$1 = R^2 + \frac{RSS}{TSS} \Rightarrow R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{\mu}_i^2}{\sum(Y_i - \bar{Y})^2} \tag{4.11}$$

In its matrix expression:

$$STC = \sum_1^n y_i^2 = \sum(Y_i - \bar{Y})^2 = \sum Y_i^2 - 2\bar{Y}\sum Y_i + n\bar{Y}^2 = Y'Y - n\bar{Y}^2 \tag{4.12}$$

$$SEC = \hat{\beta}'X'Y - n\bar{Y}^2 \tag{4.13}$$

$$R^2 = \frac{\hat{\beta}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2} \tag{4.14}$$

4.3.2 Adjusted Coefficient of Determination

$$\tilde{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-k}(1 - R^2) \tag{4.15}$$

4.4 Simple and Partial Correlation Coefficient

4.4.1 Simple Correlation Coefficient: measures the degree of linear association between *X* and *Y*

$$r_{xy} = \sqrt{R^2} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} \qquad -1 < r_{xy} < 1 \tag{4.16}$$

4.4.2 Partial Correlation Coefficient

$$r_{yx_2,x_3,x_4,\ldots x_k} = \frac{\sum u_1 u_2}{\sqrt{\sum u_1^2 \sum u_2^2}} \qquad \textbf{(4.17)}$$

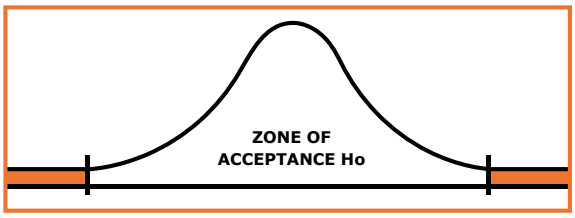
4.5 Interval Estimation

$$P\left[\hat{\beta}_i - t_{\frac{\alpha}{2}} ee(\hat{\beta}_i) \le \beta_i \le \hat{\beta}_i + t_{\frac{\alpha}{2}} ee(\hat{\beta}_i)\right] = 1-\alpha \qquad \textbf{(4.18)}$$

4.6 Statistical Significance Tests for Parameters

Type of Hypothesis	H_0	H_I	Decision	Graphs
Two-tailed	$\beta_i = \beta_{i^0}$	$\beta_i \neq \beta_{i^0}$	$/t/ > t_{\alpha/2}$	1
Left-tailed	$\beta_i \ge \beta_{i^0}$	$\beta_i < \beta_{i^0}$	$t < -t_\alpha$	2
Right tailed	$\beta_i \le \beta_{i^0}$	$\beta_i > \beta_{i^0}$	$t > t_\alpha$	3

Image 4.2



4.7 Tests of Significance

- The model is available:

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + u_i$$

4.7.1 Individual Significance Test

$$H_0 = \beta_i = 0 \qquad H_i = \beta_i \neq 0 \qquad t_{n-k} = \frac{\hat{\beta}_i - \beta_i}{ee(\hat{\beta}_i)} \tag{4.19}$$

4.7.2 Overall Significance Test

$$H_0 : \beta_2 = \beta_3 = \ldots \beta_k = 0 \qquad F = \frac{ESS/(k-1)}{RSS/n-k} \rightarrow F(k-1, n-k) \tag{4.20}$$

4.7.3 Significance Test for a Subset of Parameters

$$H_o : \beta_2 = \beta_3 = 0$$

$$H_i : al\ menos\ un\ \beta_i \neq 0 \quad i = 2,3 \qquad Y_i = \beta_1 + \beta_4 x_4 + \beta_5 x_5 \rightarrow SCR^*$$

4.7.4 Restricted Model

$$Reject\ Ho\ if\ \frac{(RSS^* - RSS)/(k_1)}{RSS(n-k)} > f(k_1, n-k, \alpha) \tag{4.21}$$

Where k_1 is the number of parameters of H_o

4.8 Hypothesis Testing for a Set of Linear Restrictions

The matrices R of size j x k are defined, where *j* is the number of linear constraints and a *r* vector of size (jx1). R contains the

coefficients of each of the linear constraints and *r* contains the values to which these constraints are equal.

$$H_{0i}: R\beta = r \qquad F = \frac{\left(R\hat{\beta} - r\right)'\left[R(X'X)^{-1}R'\right]^{-1}\left(R\hat{\beta} - r\right)}{J\sigma_U^2} \rightarrow F(J, n-k)$$
$$H_i: R\beta \neq r$$

4.9 Prediction

After parameter estimation and structural analysis, the most common use of regression is prediction.

$$\hat{Y}_{t+1} = \hat{\beta}_1 + \hat{\beta}_2 X_{2\,t+1} + \hat{\beta}_3 X_{3_{t+1}} + ... + \hat{\beta}_k X_{k\,t+1}$$

$$\hat{Y}_{t+1} = X_{t+1}\hat{\beta} \quad \textit{point prediction}$$

$$\textit{where } X_{t+1} \equiv \left[1 \quad X_{2\,t+1} \quad X_{3t+1}.....X_{k\,t+1}\right]$$

interval prediction

$$\text{var}(Y_{t+1} - \hat{Y}_{t+1}) = Var\left(e_{t+1}\right) = \hat{\sigma}^2\left(1 + X_{t+1}\left(X'X\right)^{-1}X'_{t+1}\right)$$

$$\hat{Y}_{t+1} \pm t_{\alpha/2+} DS\left(e_{t+1}\right)$$

4.10 Testing Structural Hypotheses of the Model

- Small samples
- Structural change
- Misspecification
- Multicollinearity

4.10.1 Small Samples

The number of data (observations) must be greater than the number of model parameters (n>k) so the model can have a solution.

For operational purposes, a minimum of about 15 data are needed to have some guarantee in the estimation of three or four parameters.

4.10.2 Structural Change

One of the model's structural assumptions is the constancy of the regression model parameters throughout the observation period and that it is maintained for the prediction horizon. When this condition is not met, the model is said to suffer from structural change.

4.10.2.1 How to Identify Structural Change

The chow test (steps) is applied to identify it.

1. Estimate $Y = X\hat{\beta} + \hat{\mu}$ *and obtain SSR*

2. Estimate $Y^* = X^*\hat{\beta}^* + \hat{\mu}^*$ and obtain SSR*

3. Estimate $Y^{**} = X^{**}\hat{\beta}^{**} + \hat{\mu}^{**}$ and obtain SSR**

4. Obtain F calculated $F = \dfrac{\left(RSS - \left(RSS^* + RSS^{**}\right)\right)/k}{\left(RSS^* + RSS^{**}\right)/n-2k}$ **(4.21)**

5. Contrast the calculated F vs. Tabulated F with K and n-2k gl, under the hypothesis

Ho: B=B*=B**(there is no structural change)

4.10.2.2 To Solve Structural Change by:

- Running switching regressions to each subsample
- Readapting the simple regression model
- Including a dummy variable in the regression

$Y_t = \mu + \alpha D_t + X_t^*\hat{\beta} + U_t$ *affects the intercept*

$Y_t = \mu + X_t^*\hat{\beta} + \alpha D_t X_t^* + U$ *affects the slopes*

4.10.3 Misspecification

When a model is proposed, a correct specification is assumed; however this is difficult to fulfill well because:

- The functional form is not correct
- Relevant variables are omitted
- Non-relevant variables are included.

4.10.4 Multicollinearity (MC)

4.10.4.1 Perfect MC

is when there is an exact relationship between exogenous variables. In this case the determinant of X'X is equal to zero, the inverse $(X'X)^{-1}$ cannot be found and therefore, the parameters (β) cannot be estimated

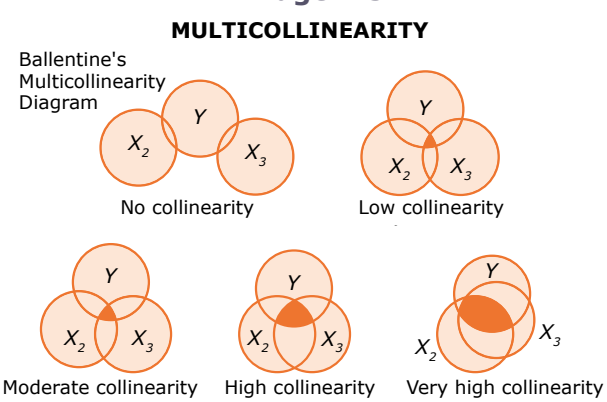
$$Remembering\ \frac{1}{|A|}\ x^{o}\ adjoint\ (X'X)^{-1},\ but\ if\ |A| = 0.$$

$$\frac{1}{|0|}\ x^{o}\ adjoint = (0) \rightarrow\ no\ inverse$$

4.10.4.2 Approximate MC

occurs when there is an approximate relationship between exogenous variables. In this case, the determinant of X'X is close to zero, causing the estimators to be distorted.

Image 4.3



4.10.4.3 How to Identify Multicollinearity

- From the consequences.
- Calculate the simple correlation coefficients–r_{xy} between exogenous variables and highlight those cases where $r_{xy} > 0.9$.
- Run auxiliary regressions between the exogenous variables of the model and highlight those cases in which the R^2 of the auxiliary regression is greater than the R^2 of the original model.

4.10.4.4 Treatment of Dummy Variables

Dummy variables receive the values 1 and 0. They are used when there are variables in the models that are not directly quantifiable but essential in explaining the endogenous variable.

Ex. 1. Socioeconomic stratus
2. The time of year.

Ex. $Wage = \beta_1 + \beta_2 Exp + \beta_3 D_1 + \beta_4 D_2 + \beta_5 D_3 + \beta_6 D_4 + u$

Care must be taken when including dummy variables, because one can fall into the dummy variable trap: m, m-1 multicollinearity,

1	7	1	0	0	0
1	10	0	0	1	0
1	25	0	1	0	0
1	20	0	0	0	1
1	18	1	0	0	0
1	16	0	0	1	0
1	10	0	1	0	0

If you have m-categories, the sum of the dummies is equal to the intercept column. Therefore, one category must be omitted.

For example in the case of educacion, omit the dummy for primary school.

By eliminating the D_1 primary dummy, everything is referenced to primary

$$Wage = 500 + 52Exp + 12D_2 + 20D_3 + 50D_4 + u$$

D2→The high school individual earns $12,000 more than the elementary school individual.

4.11 Testing Hypotheses on Random Perturbation

Heteroscedasticity and autocorrelation.

4.11.1 Heterocedasticity:

The variance of errors is not constant throughout the sample.

$$VAR(U_i \mid X) = \sigma^2_{\mu_i} \quad para \quad i = 1.....n \Big\} \quad \mathrm{E}(UU' \mid X) = \sigma^2_{\mu}\Omega_n$$

4.11.1.1 How to identify Heteroscedasticity

- Graph
- The Park test
- Glejser test
- Goldfeld-Quandt test
- White Test
- Spearman's rank correlation test

4.11.1.2 Possible Solutions of the Heteroscedasticity

- Application of generalized least squares (GLS)

$$\hat{\beta}_{MCG} = (X\Omega^{-1}X)^{-1}X\Omega^{-1}Y$$

- Transformation of the model by the P matrix and apply OLS.

Y*=PY

X*=PX the new model will be Y* = X*B+U*.

In this, the P matrix has the variable's elements on the diagonal, making the heteroscedasticity proportional. The rest are zero.

4.11.2 Autocorrelation:

There is a relationship between the errors of one period and those of another.

$$COV(U_t U_{t-i}) \neq 0 \qquad for\ i \neq 0\} \quad \mathrm{E}(UU' \mid X) = \sigma_\mu^2 \Omega_n$$

4.11.2.1 How to Identify Autocorrelation

- Graph
- Durbin Watson test
- Durbin's h test
- Box Pierce
- Ljung Box
- Kruskal-Wallis test
- Run test
- Chi-square test of independence

4.11.2.2 Durbin Watson Test

It is used to detect AR(1)

$$d = \frac{\sum_{t=2}^{n} \left(\hat{U}_t - \hat{U}_{t-1}\right)^2}{\sum_{t=1}^{n} \hat{U}_t^2} \qquad \textbf{(4.22)}$$

It is defined as the ratio of the sum of the squared differences of successive residuals over the SSR.

Assumptions on which it is based:

- The regression model includes intercept

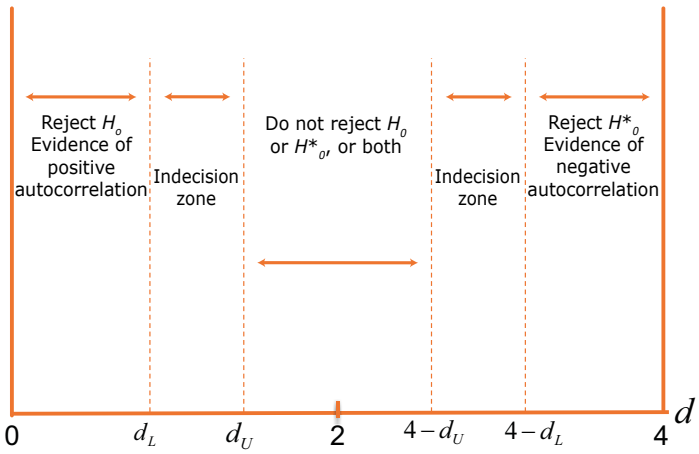
- Explanatory variables X are fixed in repeated samples.
- Errors are generated by an AR(1) process, i.e. $u_t=\rho u_{t-1} +\varepsilon_t$
- The model does not include lagged values of the dependent variable.
- Given the definition, it is possible to prove that, $d \approx 2(1-\hat{\rho})$ but as:

$-1 < \rho < 1$, then $0 < d < 4$.

Decision rules:

- If 0<d<dl, AR(1) positive
- If du<d<4-du, no AR(1)
- If 4-dl<d<4, AR(1) negative
- If dl<d<du or 4-du<d<4-dl, it cannot be concluded.

Image 4.4. Durbin Watson



However, the Durbin-Watson (d) test has the drawback that when the statistic falls within the inconclusive range (or 'region of ignorance'), one cannot definitively conclude the presence of autocorrelation.

4.11.2.3 Possible Solutions to the Autocorrelation

- Test new explanatory variables or reconsider the model's functional form.
- Apply GLS or transform the model using the P matrix Both Ω and P are matrices that depend on ρ.

4.11.3 Non-normality

- The assumption is that residues are normally distributed.
- Failure to meet this condition causes statistical inference tests to lose relia bility.
- The main tests to detect it are the Jarque-Bera and Shapiro-Wilk tests.
- Possible solutions: To apply logarithm to all variables.

EXAMPLES

WORKSHOP[12]

1) The following information is a sample of seven soft drink industries, where K is the number of machines used in the process, L is the number of workers, and Q is the number of units produced (all expressed in logarithms).

	K	L	Q
	2,08	3,14	4,66
	2,2	2,64	4,4
	1,39	3,64	4,29
	0,69	4,57	4,05
	1,79	2,4	4,2
	1,79	3,76	4,59
	1,1	4,53	4,41
Sums	11,04	24,68	30,6
averages	1,57714286	3,52571429	4,37142857

$(X'X)^{-1}$		
27,1065029	-6,9220181	-4,55131445
-6,9220181	1,98670246	1,07459204
-4,55131445	1,07459204	0,81019874

12 To see the solution of this exercise in Excel, see Appendix 4.

a) Interpreter B_2 and B_3

b) Calculate R^2 and adjoint R^2 and interpret them

c) Establish confidence intervals and significance tests for B_2 and B_3 and interpret them

d) To perform the overall significance test and interpret it

e) Test the hypothesis that these industries have, a cobb douglas production function with constant returns to scale.

Solution

$Y_i = \beta_1 X_2^{\beta_2} X_3^{\beta_3}$ $\ln Y_i = \beta_0 + \beta_2 \ln X_2 + \beta_3 \ln X_3$ *where* $\beta_0 = \ln \beta_1$

Recalling the (4.3) equation to estimate the betas and $\hat{\beta} = (X'X)^{-1} X'Y$ that we already have the inverse matrix $(X'X)^{-1}$, we would have:

$$\boldsymbol{X} =$$

X matrix is composed of the intercept; the column of some and by: k and l. The number of machines used in the process multiplied by the number of workers required respectively.

1	2,08	3,14
1	2,2	2,64
1	1,39	3,64
1	0,69	4,57
1	1,79	2,4
1	1,79	3,76
1	1,1	4,53

Therefore, the X' = Due to the properties of the transpose (where the rows are transformed into columns)

$$\begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2,08 & 2,2 & 1,39 & 0,69 & 1,79 & 1,79 & 1,1 \\ 3,14 & 2,6 & 3,64 & 4,57 & 2,4 & 3,76 & 4,53 \end{vmatrix}_{3x7}$$

X'Y = X' multiplied by Y (the number of units produced)
X' Y X'Y

$$\underset{X'}{\begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2,08 & 2,2 & 1,39 & 0,69 & 1,79 & 1,79 & 1,1 \\ 3,14 & 2,6 & 3,64 & 4,57 & 2,4 & 3,76 & 4,53 \end{vmatrix}_{3x7}} \times \underset{Y}{\begin{vmatrix} 4,66 \\ 4,4 \\ 4,29 \\ 4,05 \\ 4,2 \\ 4,59 \\ 4,41 \end{vmatrix}_{7x1}} = \underset{X'Y}{\begin{vmatrix} 30,6 \\ 48,7155 \\ 107,6882 \end{vmatrix}_{3x1}}$$

Remember, you multiply term by term and row by column.

(1x4.66)+(1x4.4)+(1x4.29)+(1x4.05)+(1x4.2)+(1x4.59)+(1x4.41) = 30.6 and so on.

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

$$\underset{(X'X)^{-1}}{\begin{vmatrix} 27,1065029 & -6,9220181 & -4,55131445 \\ -6,9220181 & 1,98670246 & 1,07459204 \\ -4,55131445 & 1,07459204 & 0,81019874 \end{vmatrix}_{3x3}} \times \underset{X'Y}{\begin{vmatrix} 30,6 \\ 48,7155 \\ 107,6882 \end{vmatrix}_{3x1}} = \underset{\hat{\beta}}{\begin{vmatrix} 2,1266 \\ 0,6903 \\ 0,3279 \end{vmatrix}_{3x1}}$$

a) Interpretation of the regression:

$$\hat{Y} = 2,1266 + 0,6903X_1 + 0,3279X_2$$

B_2: Taking into account the result of the capital coefficient of elasticity of 0.6903. It implies that for a 1% increase in machinery used in the process, the product (measured in units of soft drinks produced) increases on average by about 0.69% while keeping the number of workers constant.

B_3: This indicates that with a 1% increase in the number of workers (taking into account the results of the work elasticity coefficient of 0.3279) the number of units produced increases on average by about 0.33% while keeping constant the number of machines used in the process.

b) Recall equation (4.10) and equation (4.15),

$$R^2 = \frac{ESS}{TSS} \qquad \tilde{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} = 1 - \frac{n-1}{n-k}(1-R^2)$$

To find R^2 and adjusted R^2, the TSS, ESS, and RSS must be found. Equations 4.12, 4.13, and 4.23 respectively:

$$ESS = \hat{\beta}'X'Y - n\bar{Y}^2 \qquad \textbf{(4.13)}$$

$\hat{\beta}$				$X'Y$		$n\,(-)\,\bar{Y}^2$
2,1266	0,6903	0,3279	x	30,6	=	134,01323 - (7 x 19,1094) =
		1x3		48,7155		(134,0132 - 133,7658) 0,247514
				107,6882		ESS
				3x1		

We transpose the *Betas*.

$$TSS = Y'Y - n\bar{Y}^2 \qquad \textbf{(4.12)}$$

$Y'Y$								Y		
4,66	4,4	4,29	4,05	4,2	4,59	4,41	X	4,66	=	134,0384 - (7 x 19,1094) =
						1x7		4,4		(134,0384 - 133,7658) =
								4,29		0,272684
								4,05		TSS
								4,2		
								4,59		
								4,41		
								7x1		

We transpose Y.

TSS = ESS + RSS by clearing RSS we have:

TSS–ESS = RSS

$$Y'Y - n\overline{Y}^2 - \hat{\beta}'X'Y - n\overline{Y}^2$$
$$= Y'Y - \hat{\beta}'X'Y \qquad \textbf{(4.23)}$$

0.272684 – 0.247514 = 0.02377565 *RSS*

Now:

$$R^2 = \frac{ESS}{TSS} \qquad \textbf{(4.10)} \qquad R^2 = \frac{0{,}247514}{0{,}272684} = 0{,}907695$$

As the model has an independent term its:

$$\widetilde{R}^2 = 1 - \frac{\frac{RSS}{n-k}}{\frac{TSS}{n-1}} \qquad \widetilde{R}^2 = 1 - \frac{\frac{0{,}02377565}{7-3}}{\frac{0{,}272684}{7-1}} = 0{,}869213 \qquad \textbf{(4.15)}$$

The coefficient of determination R^2 shows how well the regression line fits the data, $R^2 = 0.9076$. It indicates us that approximately 91% of the variation of the units produced in the seven soft drink industries is explained by the following variables: number of machines and number of workers. Considering that the R^2 is between 0 and 1 ($0 < R^2 < 1$) this variation is quite acceptable.

c) Significance Test and Confidence Intervals for B_2 and B_3

In order to perform the hypothesis tests, the standard errors of betas are required. Therefore, we must find the Var-Cov matrix of the betas (equation **4.4***), which, as seen in the introductory chapter of matrix algebra, has in its diagonal the variances of the betas and the square root of each one of them. Recall the equation:*

$$\sqrt{\text{var}}\hat{\beta}_i = ee(\hat{\beta}_i)$$

They are the standard errors. $Var - Cov(\hat{\beta}) = \sigma_\mu^2 (X'X)^{-1}$ **(4.4)**

The inverse matrix is obtained from the matrix $Var - Cov(\hat{\beta}) = \sigma_{\mu}^2 (X'X)^{-1}$, but the variance of the residuals must be found:

$$\sigma_u^{\ 2} = \frac{\hat{u}'\hat{u}}{n-k} = \frac{\Sigma \hat{u}_i^{\ 2}}{n-k} = \frac{y'y - \hat{\beta}'x'y}{n-k} = \frac{RSS}{n-k} \qquad \sigma_u^{\ 2} = \frac{0{,}02377565}{7-3} = 0{,}005944$$

$(X'X)^{-1}$ $\qquad$ $Var - Cov(\hat{\beta}) = \sigma_{\mu}^2 (X'X)^{-1}$

$$0{,}005944 \times \begin{vmatrix} 27{,}1065029 & -6{,}9220181 & -4{,}55131445 \\ -6{,}9220181 & 1{,}98670246 & 1{,}07459204 \\ -4{,}55131445 & 1{,}07459204 & 0{,}81019874 \end{vmatrix}_{3x3} = \begin{vmatrix} \mathbf{0{,}16112105} & -0{,}04114448 & -0{,}02705301 \\ -0{,}04114448 & \mathbf{0{,}01180896} & 0{,}00638738 \\ -0{,}02705301 & 0{,}00638738 & \mathbf{0{,}00481582} \end{vmatrix}_{3x3}$$

$$\sqrt{\text{var}}\,\hat{\beta}_i = ee(\hat{\beta}_i) \qquad \sqrt{0{,}16112105} = ee(\hat{\beta}_1) = 0{,}401399$$

$$\sqrt{0{,}01180896} = ee(\hat{\beta}_2) = 0{,}1086682$$

$$\sqrt{0{,}00481582} = ee(\hat{\beta}_3) = 0{,}069396$$

d) Significance Test of the Coefficients of Regression. A significance test is a procedure by which sample results are used to verify the truth or falsehood of a null hypothesis (H_o). **(4.19)**

Under the normality assumption we have:

$$t = \frac{\hat{\beta}_2 - \beta_2}{e\,(\hat{\beta}_2)} \qquad t = \frac{\alpha}{2} = 2.776$$

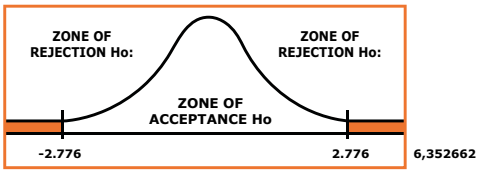
$\hat{\beta}_2 =$ 0.69033233 α = 5%.

$ee(\hat{\beta}_2) =$ 0.1086682 gl = 4
(look in the t-table in the appendix for an example)

Let us pose the following:

$$H_0 = \beta_2 = \beta_2{}^* = 0$$
$$H_i = \beta_2 \neq 0$$

$$t = \frac{0{,}69033233}{0{,}1086682} = 6{,}352662$$



H_o = is rejected. B_2 is statistically significant, since the test's statistic value fell in the rejection zone.

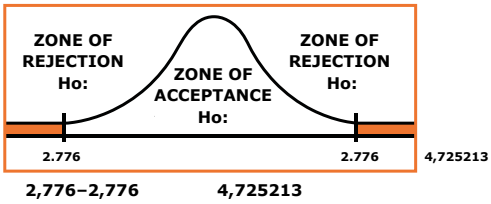
For B_3:

$$t = \frac{\hat{\beta}_3 - \beta_3}{ee(\hat{\beta}_3)} \qquad t = \frac{\alpha}{2} = 2.776 \qquad \hat{\beta}_3 = 0.32791085 \quad \alpha = 5\%.$$

$$ee(\hat{\beta}_3) = 0.069396 \text{ gl} = 4$$

Let us pose the following:

$$H_0 = \beta_3 = \beta_3 * = 0$$
$$H_i = \beta_3 \neq 0$$



$$t = \frac{0{,}32791085}{0{,}069396} = 4{,}725213$$

H_o = it is rejected. B_3 is statistically significant, since the test statistic value fell in the rejection zone.

Intervals of significance:

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2} ee(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} ee(\hat{\beta}_2)\right] = 1 - \alpha \qquad \textbf{(4.18)}$$

$$t = \frac{\alpha}{2} = 2{,}776 \qquad \hat{\beta}_2 = 0.69033233 \quad \alpha = 5\%.$$

$$ee(\hat{\beta}_2) = 0.1086682 \text{ gl} = 4$$

We propose:

$$H_0 = \beta_2 = \hat{\beta}_2$$
$$H_i = \beta_2 \neq \hat{\beta}_2$$

$$\Pr[(0{,}69033233) - (2.776) * (0{,}1086682) \leq \beta_2 \leq (0{,}69033233) +$$

$$(2.776) * (0{,}1086682)] = 95\%$$

$$0{,}38869 \leq \beta_2 \leq 0{,}991995$$

0.69033233

We accept the null hypothesis.

Given the 95% confidence coefficient in the long term, for 95 out of 100 cases, the intervals such as 0.38869 and 0.991995 will contain the true value of B2.

B_3

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2} ee(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} ee(\hat{\beta}_2)\right] = 1 - \alpha$$

$$t = \frac{\alpha}{2} = 2{,}776 \qquad \hat{\beta}_2 = 0.32791085 \ \alpha = 5\%.$$

$$ee(\hat{\beta}_2) = 0.069396 \ gl = 4$$

We propose:

$$H_0 = \beta_2 = \hat{\beta}_2$$
$$H_i = \beta_2 \neq \hat{\beta}_2$$

$$\Pr\left[(0{,}69033233) - (2.776) * (0{,}1086682) \le \beta_2 \le (0{,}69033233) + (2.776) * (0{,}1086682)\right] = 95\%$$

$$0{,}32791085$$

We accept the null hypothesis.

Given the 95% confidence coefficient in the long run, for 95 out of 100 cases, intervals such as 0.135268 and 0.520554 will contain the true value of B_2.

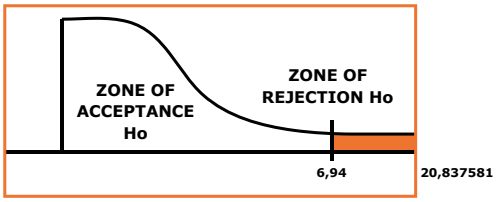
Significance test of the parameters as a whole:

$$F = \frac{ESS/(k-1)}{RSS/n-k} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad \textbf{(4.20)}$$

we replaced:

$$F = \frac{0{,}247514/(3-1)}{0{,}02377565/7-3} = \frac{0{,}9076/(3-1)}{(1-0{,}9076)/(7-3)} = 20{,}837581$$

Under the assumption that $U_i \to N(0, \sigma_u^2)$ we propose the null hypothesis:



6.94...20.837581

$$H_o : \beta_2 = \beta_3 = 0$$
$$H_i : \beta_2, \beta_3;\ at\ least\ one\ \ \beta_i \neq 0$$

If Fc > Ft, the null hypothesis must be rejected (see table f – appendix, example (6.94)).

In our case the $F_{calculated}$ >$F_{in\ tables}$; where F(k-1, n-k) α = 5% 1–α = 95%.

20,837581 > 6,94

Contrary to the *p-value* of the F obtained, it allows us to reject the null hypothesis because it is sufficiently low.

$$H_o : \beta_2 = \beta_3 = 0$$

2) We have the results of the following regression:

LINREG Y
CONSTANT X1 X2

Dependent Variable Y–Estimation by Least Squares			
Annual Data From 1978:01 To		1992:01	
Usable Observations	15	Degrees of Freedom	12
R**2 0.851461 Adjusted		R**2 0.826705	
Sum of Squared Residuals		411886.37547	
F.			
Significance Level of		F 0.00001074	
Durbin-Watson Statistic		1.517636	
Q(3-0) 5.679080			
Significance Level of Q		0.12831149	

	Variable	Coeff	Std Error	T-Stat	Signif
1.	Constant	523.7710	211.3178466	2.47859	0.02903281
2.	X_1	17.02586	2.1929287	7.76399	0.00000510
3.	X_2	-8.11345	2.1364034	-3.79772	0.00254035

Y = coffee exports (thousands of dollars)

X1 = external price of coffee (US cents per pound)

X_2 = price of other soft commodities (cents per pound)

a) Interpret B_2 and B_3 and the R^2.

b) Perform the significance test and the confidence interval for B_2 and B_3, interpret t (alpha/2) = 2.179

c) Test the significance of the parameters as a whole and interpret f (2.12 alpha=5%) = 3.89

d) What is the q statistic used for and what can be concluded for the model proposed?

e) Describe two problems or restrictions of the Durbin-Watson statistic (d)

f) Briefly describe a solution to the problems of Durbin-Watson statistic

g) What can be concluded about AR (1) for the proposed model?

SOLUTION

$$\hat{Y} = 523{,}7710 + 17{,}02586X_1 - 8{,}11345X_2$$

Mathematically B_2 is the slope of the line.

Economically B_2 is the average growth of coffee exports during the 01-1978 to 01-1992 period.

a) Interpretation of the regression process:

If during the sample period X_1, X_2 had been 0, the average coffee exports would be US USD524,000.

The partial regression coefficient 17.02586 means that by keeping X_2 (the price of other soft products) constant, the

growth observed in coffee exports during the period 01-1978 to 01-1992 on average was 0.17%. Likewise, by keeping the external price of coffee constant, the value of -8.11345 implies that, during the same period of time, total coffee exports fell by approximately 0.08%.

The coefficient of determination R^2 shows how well the regression line fits the data. $R^2 = 0.85$ indicates that approximately 85% of the variation of total coffee exports during the period 01-1978 to 01-1992 is explained by the variables: external price of coffee and price of other soft products. Considering that the R^2 is between 0 and 1 ($0 < R^2 < 1$), this variation is quite acceptable.

b) Significance Test and ConfidenceIntervals for B_2 and B_3

Significance test of coefficients regression. It is a procedure by which sample results are used to verify the truth or falsehood of a null hypothesis (H_o).

Under the normality assumption we have:

$$t = \frac{\hat{\beta}_2 - \beta_2}{ee(\hat{\beta}_2)} \qquad t = \frac{\alpha}{2} = 2.179$$

$$\hat{\beta}_2 = \quad 17.02586 \qquad \alpha = 5\%.$$

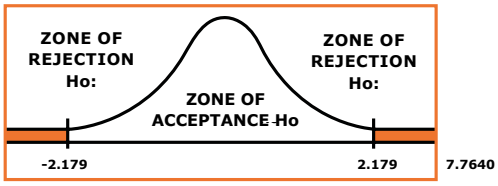
$$ee(\hat{\beta}_2) = \quad 2.1929287 \qquad gl = 12$$

Let us pose the following:

$$H_0 = \beta_2 = \beta_2{}^* = 0$$

$$H_i = \beta_2 \neq 0$$

$$t = \frac{17.02586}{2.1929287} = 7.7640$$



H_o = is rejected. B_2 is statistically significant, since the test statistic value fell in the rejection zone.

For B_3:

$$t = \frac{\hat{\beta}_3 - \beta_3}{ee(\hat{\beta}_3)} \qquad t = \frac{\alpha}{2} = 2.179$$

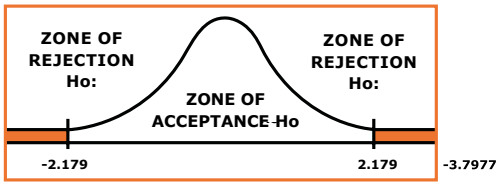
$$\hat{\beta}_3 = -8.11345 \ \alpha = 5\%.$$

$$ee(\hat{\beta}_3) = 2.1364034 \ gl = 12$$

Let us pose the following:

$$H_0 = \beta_3 = \beta_3^* = 0$$
$$H_i = \beta_3 \neq 0$$

$$t = \frac{-8.11345}{2.1364034} = -3.7977$$



H_o = It is rejected. B_3 is statistically significant, since the test statistic value fell in the rejection zone.

Significance intervals:

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2} ee(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} ee(\hat{\beta}_2)\right] = 1 - \alpha$$

$$t = \frac{\alpha}{2} = 2.179 \qquad \hat{\beta}_2 = 17.02586 \qquad \alpha = 5\%$$

$$ee(\hat{\beta}_2) = 2.1929287 \qquad gl = 12$$

We propose:

$$H_0 = \beta_2 = \hat{\beta}_2$$
$$H_i = \beta_2 \neq \hat{\beta}_2$$

$$\Pr[(17.2586) - (2.179) * (2.1929287) \leq \beta_2 \leq (17.2586) + (2.179) * (2.1929287)] = 95\%$$

17.2586

We accept the null hypothesis.

Given the 95% confidence coefficient in the long run, for 95 out of 100 cases, intervals such as 12.2475 and 21.8043 will contain the true value of B_2.

B_3

$$\Pr\left[\hat{\beta}_3 - t_{\alpha/2} ee(\hat{\beta}_3) \leq \beta_3 \leq \hat{\beta}_3 + t_{\alpha/2} ee(\hat{\beta}_3)\right] = 1 - \alpha$$

$$t = \frac{\alpha}{2} = 2.179 \qquad \hat{\beta}_3 = -8.11345 \qquad \alpha = 5\%$$

$$ee(\hat{\beta}_3) = 2.1364034 \qquad gl = 12$$

We propose:

$H_0 = \beta_3 = \hat{\beta}_3$

$H_i = \beta_3 \neq \hat{\beta}_3$

$$\Pr[(-8.11345) - (2.179) * (2.1364034) \leq \beta_3 \leq (-8.11345) + (2.179) * 2.1364034] = 95\%$$

-8.11345

We accept the null hypothesis.

Given the 95% confidence coefficient in the long run, for 95 out of 100 cases, intervals such as -12.7687 and 1.8314 will contain the true value of B_2.

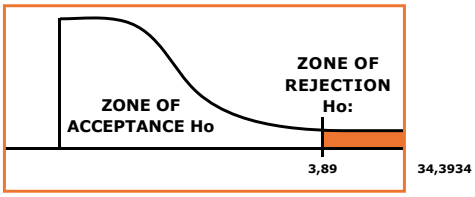
Significance test of the parameters as a whole:

$$F = \frac{ESS/(k-1)}{RSS/n-k} \quad \frac{R^2/(k-1)}{(1-\bar{R}^2)/(n-k)} \qquad \textbf{(4.20)}$$

Under the assumption that $U_i \rightarrow N(0, \sigma_{u_i}^2)$ the null hypothesis is set:

$H_o : \beta_2 = \beta_3 = 0$

$H_i : \beta_2, \beta_3;\ at\ least\ a \quad \beta_i \neq 0$



If Fc > Ft the null hypothesis must be rejected.

In our case the $F_{calculated} > F_{in\ tables}$; where F(k-1, n-k)

α = 5% 1– α = 95%.

34,3934 > 3,89

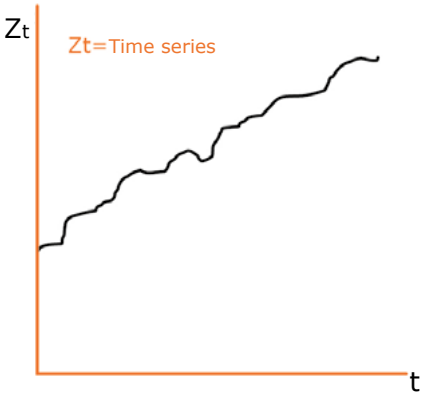
Now in contrast to the *p-value* of the F obtained, it allows us to reject the null hypothesis because it is sufficiently low.

$H_o : \beta_2 = \beta_3 = 0$

CHAPTER 5

TIME SERIES

Image 5.1



• A time series is a set of ordered measurements over time of a variable of interest.

• With time series, the historical behavior of a variable is analyzed through a mathematical function.

• They are used for forecasting. not structural analysis.

• It is assumed that data are available at regular time intervals (hours, days, months, quarters, years...) and it is desired to use the possible "inertia" in the behavior of the series to forecast its future evolution.

5.1 Ways to Analyze a Series

- By decomposition of the series or deterministic.
- A second approach is stochastic.

5.2 Components of a Series

5.2.1 Trend Component: increase or decrease behavior over a period of time.

Image 5.2a

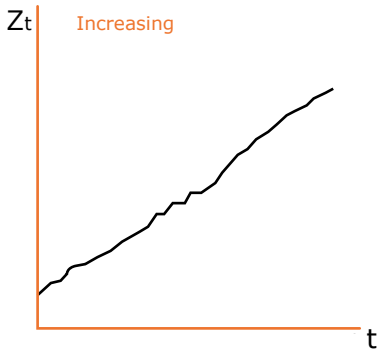


Image 5.2b

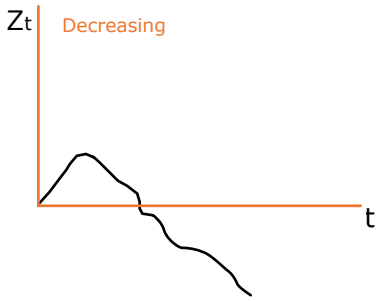
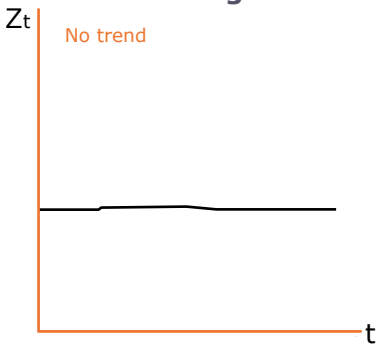


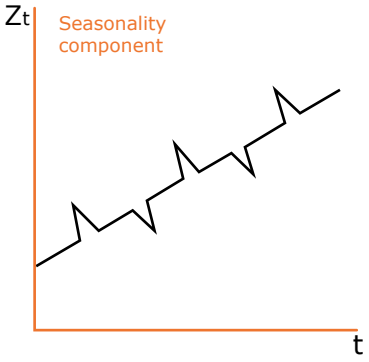
Image 5.2c



5.2.2 Components of seasonality:

Behavior repeated at regular time intervals and is represented in data less than annual, such as: output, inflation, money supply, etc.

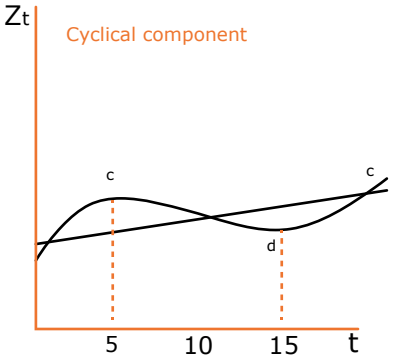
Image 5.3



5.2.3 Cyclic Component:

It is repeated over long periods of time, growth or depreciation behavior of economies, fluctuations of the series with respect to its trend.

Image 5.4

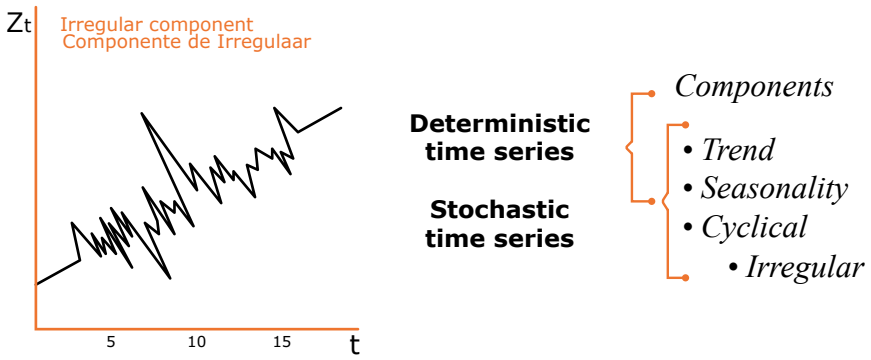


5.2.4 Irregular Component:

It occurs due to the multitude of factors that affect the series and is difficult to represent through a mathematical function. This

component results into stochastic analysis, since the irregularity manages to hide the other components.

Image 5.5

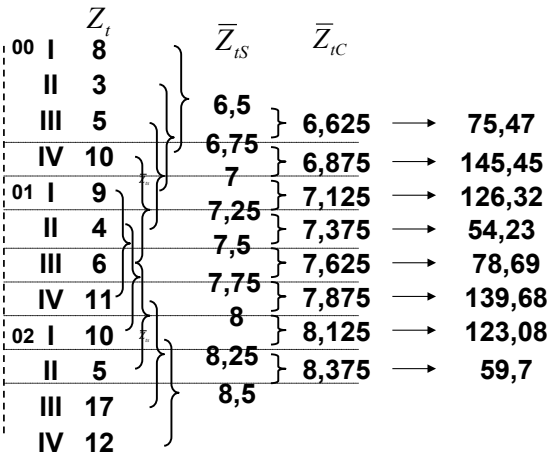


5.3 Moving Averages for Smoothing of Series

- The irregular component in some time series can be so large that it hides any underlying regularity, making any interpretation of the time chart difficult. To avoid this, the series is smoothed by using moving averages.
- In the case of quarterly data, the 4-point moving average is used and in the case of monthly data, 12-point moving averages are used.

5.3.1 Steps to calculate seasonal indices, deseasonalization of series, and forecasting

- Calculate the moving average for the time series ($\bar{Z}_{tS}$). Given that data are quarterly in our example, we average every four data and lose four observations (two at the beginning and two at the end). Therefore, to perform monthly averages, the average is every 12 data and we also lose 12 observations (six at the beginning and six at the end).
- Center the moving average, if necessary. $\bar{Z}_{tC}$



• Calculate the ratio $\frac{Real\ value}{Moving\ average} * 100 = \frac{X_t}{X_t^*} * 100$

Example for the 3er quarter of 00:

$$\frac{5}{6{,}625} * 100 = 75{,}47$$

For the 3rd quarter of 01:

$$\frac{6}{7{,}625} * 100 = 78{,}689$$

• Organize the ratios $\frac{X_t}{X_t^*} * 100$ according to periodicity (quarters, months), and calculate the average of each quarter or month (Indexo).

	I	II	III	IV	
00	-	-	75,470	145,45	
01	126,32	54,24	78,69	139,68	
02	123,08	59,7	-	-	
	124,7	56,97	77,08	142,565	**401,315**

Seasonal indices →

← *The index is adjusted because it must be 400*

How much the variable behavior changes with respect to its mean value.

To eliminate outliers without affecting the seasonal indices, the maximum and minimum data (before taking the average) are eliminated.

- Adjust the $_{\text{adjusted}}\text{I}_{\text{in}}\text{dex}_{\text{o}} = I_o \frac{\sum I_r}{\sum I_o}$

$$I_A = I_o \frac{\sum I_r}{\sum I_o} \qquad I_A = I_I + I_I + I_{III} + I_V = 400$$

$$I_I = \frac{400\ x\ 124{,}7}{401{,}31} = 124{,}29 \qquad I_{III} = \frac{400\ x\ 77{,}080}{401{,}31} = 76{,}83$$

	I_I	I_{II}	I_{III}	I_{IV}	
Seasonal indices $\Rightarrow$	124,7	56,97	77,080	142,565	**401,315**
Adjusted indices $\Rightarrow$	124,29	56,78	76,83	142,10	**400**

- Finally, the original series is adjusted as:

$$\textit{Adjusted value} = \textit{original value} * \left(\frac{100}{\textit{Adjusted index}}\right)$$

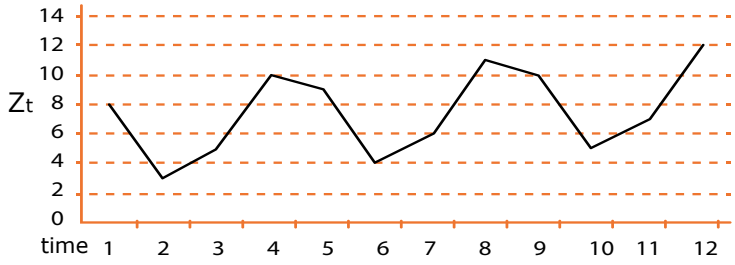
Once the seasonal variation is removed, the trend line can be calculated without seasonal variation, which can be projected into the future. Finally, the seasonal component is included by multiplying the projection by the corresponding component of the index.

$$I_I = 8*\left(\frac{100}{124{,}29}\right) = 6{,}436 \qquad I_{II} = 3*\left(\frac{100}{56{,}78}\right) = 5{,}28$$

$$I_{III} = 5*\left(\frac{100}{76{,}83}\right) = 6{,}81 \qquad I_{IV} = 10*\left(\frac{100}{142{,}095}\right) = 7{,}04$$

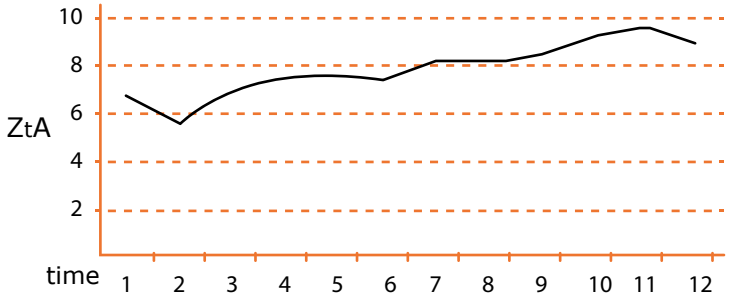
8	6,436
3	5,283
5	6,508
10	7,037
9	7,241
4	7,044
6	7,810
11	7,741
10	8,046
5	8,805
7	9,111
12	8,445

Image 5.6. Original Series



With the series adjusted, we extract the trend.

Figure 5.7. Adjusted series



5.4 Time Series viewed as Stochastic Processes

• We define a time series as a variable $Z_i = Z(t_i)$ where $i = t_i = 1, 2,...n$ indicates the different moments in time for a series of length n and where all intervals between observations are equal (i.e. all are referred to days, months, or years).

• Stochastic process is a succession of random variables $\{Z_t\}$ $t = -\infty.....+\infty$ each data in the series is considered as a random variable.

5.4.1 Stochastic Processes

• In the context of stochastic models, any time series is considered to be generated by a stochastic process.

• The values of the time series $\xi_1, \xi_2, \ldots \xi_n$ can thus be considered as sample realizations of the theoretical variables $Z_1, Z_2, \ldots Z_n$, with a probability of occurrence deduced from a_n assumed joint distribution function $p(\xi_1, \xi_2, \ldots \xi_n)$.

Theoretical variable	*Probability of occurrence*
Z_1	ξ_1
Z_2	ξ_2
Z_3	ξ_3
Z_4	ξ_4
⋮	⋮
Z_n	ξ_n

5.4.2 Simplifying Assumptions

No distinction will be made between a random variable Z_t and its observed value ξ_τ which will be also denoted by Z_t.

• The process is considered to be exactly stationary, i.e.:

$$F(Z_{t1,}\, Z_{t2,\ldots}\, Z_{tn}) = F(Z_{t1}+k_{,}\, Z_{t2}+k, \ldots Z_{tn}+k)$$

Stationary → the function does not change over time.

• If it is admitted that probability distributions are normal, it would be sufficient to know means and variances-covariances for their characterization.

5.4.3 Use of Lag Operators

• Lag operator is denoted by the letter B.

• This is defined by the relation $BZ_t = Z_{t-1}$ for all t ($\forall_t$). By successive application of the operator B we obtain:

Z_t	Z_{t-1}	Z_{t-2}
8		
3	8	
9	3	8
7	9	3
3	7	9
5	3	7
6	5	3

Note that for each lag, the loses an observation

$$B^2 Z_t = B(B Z_t) = B(Z_{t-1}) = Z_{t-2}$$

$$B^3 Z_t = B(B^2 Z_t) = B(Z_{t-2}) = Z_{t-3}$$

$$B^k Z_t = B(B^{k-1} Z_t) = B(Z_{t-(k-1)}) = Z_{t-k}$$

Generalizing:

$$B^k Z_t = Z_{t-k} \qquad for \quad K = 0,1,2\ldots y \; \forall_t$$

5.4.4 Delay Polynomials

The use of delay polynomials is of particular importance because they allow to express in a concise and simple way some of the models that have proven to be most useful in practice to represent real phenomena.

5.5 Most Used Time Series Models

5.5.1 Autoregressive:

When the Z_t series is a function of itself lagged 1, 2, 3 ... *p* periods, it is defined as autoregressive.

$$Z_t = \phi Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t \qquad \rightarrow AR(p)$$

We assume that the mean has been subtracted from all variables (μ).

$\phi \rightarrow$ is the coefficient to estimate.

$$Z_t - \phi Z_{t-1} - \phi Z_{t-2} - \ldots - \phi Z_{t-p} = a_t$$

Including the lag operator B, we have:

$$Z_t - \phi B Z_t - \phi_2 B^2 Z_t - \ldots - \phi B^p Z = a_t$$

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) Z_t = a_t \qquad \rightarrow (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)(Z_t - \mu) = a_t$$

Lag polynomials.

$$\phi_p(B)(Z_t - \mu) = a_t$$

$$\phi_p(B) Z_t = a_t$$

5.5.2 Moving Averages:

This occurss when the series is a function of the residuals, its behavior in previous periods

$$Z_t = a_t - \theta a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad \rightarrow MA(q)$$

$\theta \rightarrow$ is the coefficient to estimate. We introduce the lag operator B.

$$Z_t = a_t - \theta B a_{t-1} - \theta_2 B^2 a_{t-2} - \ldots - \theta_q B^q a_{t-q}$$

$$Z_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q) a_t$$

$$Z_t = \theta_q(B) a_t \qquad Z_t - \mu = \theta_q(B) a_t$$

5.5.3 Autoregressive of Moving Average:

This occurs when the series will be lagged 1, 2, 3 ... *p* periods, and the residuals will be lagged 1, 2, 3 ... *q* periods.

$$Z_t = \phi Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t - \theta a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad \rightarrow ARMA(p,q)$$

$$\phi_p(B) Z_t = \theta_q(B) a_t$$

Subtracting the mean of all variables:

$$\phi_p(B)(Z_t - \mu) = \theta_q(B) a_t$$

5.5.4 Difference Operator

Image 5.8a

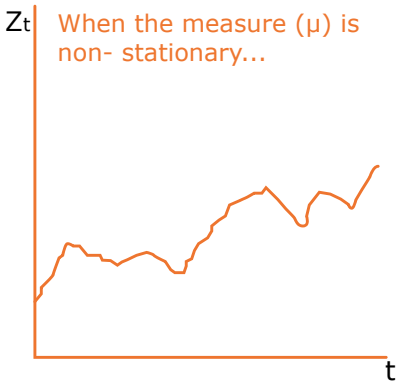
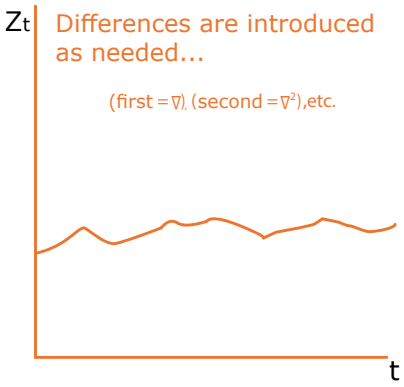


Image 5.8b



This operator is used to express relationships of the type $Y_t = Z_t - Z_{t-1}$

We define ∇ by $\nabla \ Z_t = Z_t - Z_{t-1}$ for all $_t$

The variable Y_t can be written as $Y_t = \nabla \ Z_t$
The relationship linking ∇ with B is $\nabla Zt = Zt - B\ Zt$

$$\nabla Z_t = (1-B)\ Z_t \qquad \nabla = (1-B)$$

• When the series is stationary at the first difference (∇), the series is defined to be of order *1*.

$$\nabla^2 = (1-B)^2 = 1 - 2B + B^2$$

$$\nabla 2\ Z_t = (1 - 2B + B^2)\ Z_t = Z_t - 2\ Z_{t-1} + Z_{t-2}$$

$$\nabla(\nabla Z_t) = \nabla(\ Z_t - Z_{t-1}) = Z_t - Z_{t-1} - Z_{t-1} + Z_{t-2}\ \nabla$$

• When the series is stationary at the second difference (∇), the series is defined to be of order 2.

Note that differentiating the series increases its variance.

5.6 Equations in Stochastic Differences

The term difference *Eq* is used to note the discrete equivalent of differential equations, involving variables over time.

To write the discrete equivalent of:

$$\frac{\partial Z_t}{\partial t} \Leftrightarrow \frac{\nabla Z}{\nabla t}$$

5.6.1 First Order Difference Equation

$$Z_t = a_o + a_1 Z_{t-1}$$
$$Z_t - a_1 Z_{t-1} = a_o$$
$$(1 - a_1 B) Z_t = a_o$$

by multiplying both sides by $(1 - a_1 B)^{-1}$

$$Z_t = \frac{a_o}{1 - a_1} + s a_1^t \qquad Z_o = \frac{a_o}{1 - a_1} + s$$

5.6.1.1 First Order Difference Equation (singular solution)

$$Z_t = \underbrace{\frac{a_o}{1 - a_1}}_{\substack{\text{complementary solution} \\ \text{deviation from the} \\ \text{intertemporal equilibrium}}} + \underbrace{(Z_o - \frac{a_o}{1 - a_1}) a_1^t}_{\substack{\text{particular solution} \\ \text{indicates equilibrium} \\ \text{trajectory}}}$$

In general terms, we have:

$$Z_t = \underbrace{A}_{\substack{\text{particular} \\ \text{solution}}} + \underbrace{C a_1^t}_{\text{constant}}$$

where the temporal trajectory of a_1^t *will be:*

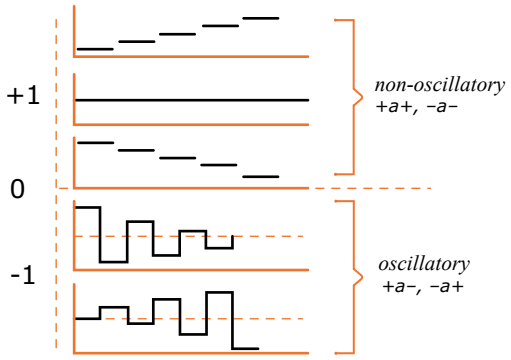
non-oscillatory $\rightarrow \; a_1 > 0$

oscillatory $\rightarrow \; a_1 < 0$

divergent $\rightarrow \; |a_1| > 0$

convergent $\rightarrow \; |a_1| < 0$

Image 5.9

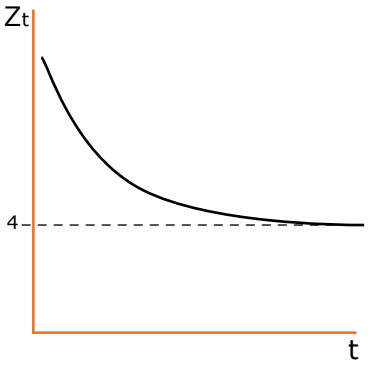


$$non\text{-}oscillatory \rightarrow a_1 > 0$$
$$oscillatory \rightarrow a_1 < 0$$
$$divergent \rightarrow |a_1| > 0$$
$$convergent \rightarrow |a_1| < 0$$

In this, the *A* point is where the trajectory converges– equilibrium level.

Example: Consider difference *Eq* (1- 0.5B) Zt = 2
Together with the initial condition Z0 = 50

Image 5.10



$$Z_t = \frac{a_o}{1-a_1} + sa_1^t \rightarrow Z_t = \underbrace{\frac{2}{1-0{,}5}}_{4} + s(0.5)^t$$

$$Z_0 = 50 + s \rightarrow Z_0 = 4 + s = 50 \rightarrow s = 46$$

The solution to this *Eq* will be:

Since |a1|<1 then the process tends to stabilize at point 4 when *t* is large.

t	1	2	3	4	5	6	7	8	9	10	11
Zt	27	15.5	9.75	6.87	5.44	4.72	4.36	4.18	4.09	4.04	4.02

5.6.2 Second Order Difference Equation

$$Z_t = a_o + a_1 Z_{t-1} + a_2 Z_{t-2}$$

$$Z_t - a_1 Z_{t-1} - a_2 Z_{t-2} = a_o$$

$$(1 - a_1 B - a_2 B^2) Z_t = a_o$$

whose general solution is:

Inverse square root 1 *Inverse square root 2*

$$Z_t = \frac{a_0}{1 - a_1 - a_2} + s_1 g_1^t + s_2 g_2^t$$

$$Z_t = \frac{a_0}{(1 - g_1)(1 - g_2)} + s_1 g_1^t + s_2 g_2^t$$

Yes, $g_1 y\, g_2$ *are real and different*

$$Z_t = \frac{a_0}{(1 - g)^2} + s_1 g^t + s_2 t g^t$$

Yes, $g_1 y\, g_2$ *are real and equal*

$$Z_t = \frac{a_0}{(1 - g_1)(1 - g_2)} + r^t[(s1 + s2)\cos(\theta t) + i(s2 - s1) sen(\theta t)]$$

Yes, $g_1 y\, g_2$ *are complex*

Example. Consider difference *Eq* $(1 - 0.9B + 0.2B^2)\, Z_{t=3}$ Together with the initial conditions $Z_0 = 0$ and $Z_1 = 50$

$$Z_t = \underbrace{0.9}_{a_1} Z_{t-1} - \underbrace{0.2}_{a_2} Z_{t-2} + \underbrace{3}_{a_0}$$

$$\frac{a_0}{1 - a_1 - a_2} = \frac{3}{1 - 0.9 + 0.2} = 10$$

$$s_1 = -450 \qquad s_2 = 440$$

$$Z_0 = 10 + s_1 (0{,}4)^0 + s_2 (0{,}5)^0 = 0$$

$$Z_1 = 10 + s_1 (0{,}4)^0 + s_2 (0{,}5)^2 = 50$$

$$x = \frac{b \pm \sqrt{b^2 - 4(a * b)}}{2(a * b)}$$

$$x = \frac{0{,}9 \pm \sqrt{0{,}9^2 - 4(0{,}2 * 0{,}9)}}{2(0{,}2 * 0{,}9)}$$

The solution to this *Eq* will be:

$$Z_1 = 10–450(0,4)'+440(0,5)'$$

Since $|g_1|<1$ and $|g_2|<1$ then the process tends to stabilize at point 10 when *t* is large.

5.6.3 Difference Equation in General Case

$$(1 - a_1B - a_2B^2 -.....- a_pB^p)Z_t = a_o$$

Whose general solution is:

$$Z_t = \frac{a_0}{(1-g_1)(1-g_2).....(1-g_p)} + s_1g_1^t + s_2g2^t + ...s_pg_p^t$$

5.7 Stationary Processes

A stochastic process is said to be stationary if its mean and variance are constant over time and if the value of the covariance between two periods depends only on the distance or lag between them and not on the time at which it was calculated.

5.7.1 Stationarity Conditions

- Stationary mean. The mean must be equal to the expected value.

$\mu zt = E(Z_t) = E(Z_{t+m}) = \mu$ for all t and m

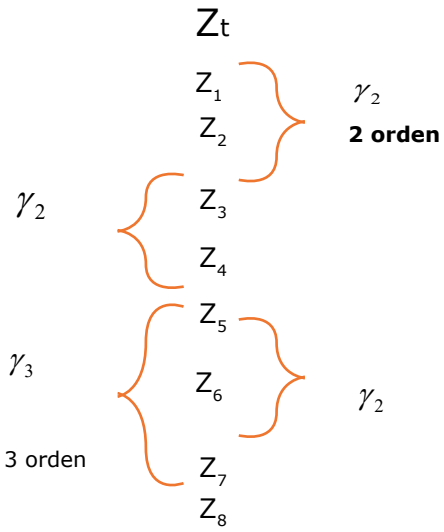
- Stationary Variance

$Var\ (Z_t) = E[(Z_{t-\mu})^2] = [(Z_{t+m} - \mu)^2] = \gamma o$

- Covar depends on the delay k.

$Cov(Z_t, Z_{t+k}) = E[(Z_t-\mu)\ (Z_{t+k} -\mu)] = E[(Z_{t+m}-\mu)\ (Z_{t+k+m} -\mu)] = cov(Z_{t+m}, Z_{t+k+m}) = \gamma k$

Example



$Cov\,(Z_1, Z_3) = Cov\,(Z_3, Z_5) = Cov\,(Z_5, Z_8)$

If the series is non-stationary, it is very difficult to represent it with a mathematical model.

5.7.2 Simple Correlation Coefficient ρ

• They are used to avoid the influence of units of measurement are denoted as ρ0, ρ1, ρ2....

$$\rho k = \frac{E[(Z_t - \mu)(Z_{t+k} - \mu)]}{\sqrt{E[(Z_t - \mu)^2]E[(Z_{t+k} - \mu)^2]}} = \frac{\mathrm{cov}(Z_t Z_{t+k})}{\sigma_{z_t}\sigma_{z_{t+k}}}$$

• As in the stationary process σz=σz+k

$$\text{Thus } \rho k = \frac{COV(Z_t, Z_{t+k})}{VAR(Z_t)} = \frac{\gamma_K}{\gamma_O}$$

therefore ρ0 = 1 k = 0, ±1, ±2 ...

Image 5.11a

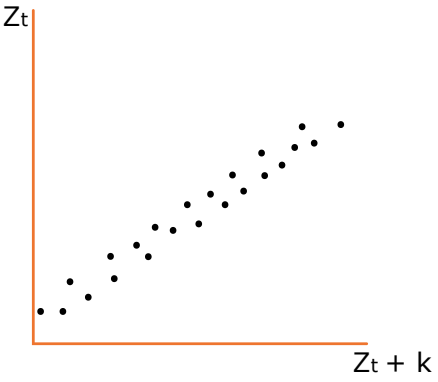
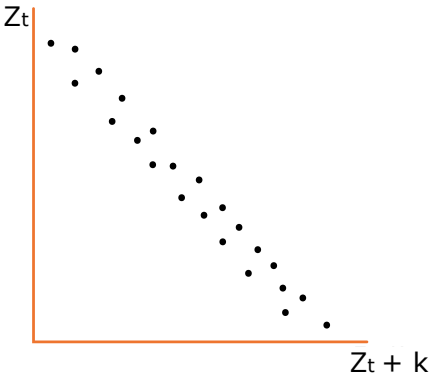


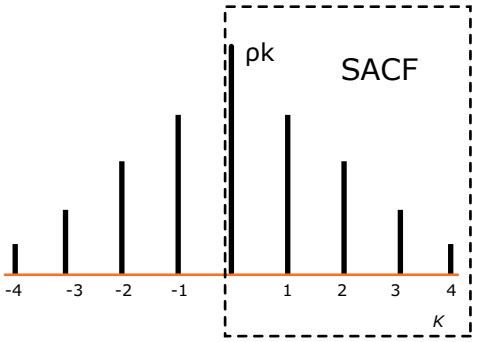
Image 5.11b



5.8 Simple Autocorrelation Function–SACF

SACF or correlogram is the graphic of the autocorrelation coefficients as a delay function.

Image 5.12



Considering ρk = ρ-k, when graphing ρk for different values of k, it is sufficient to consider only positive values of k.

5.8.1 White Noise Process

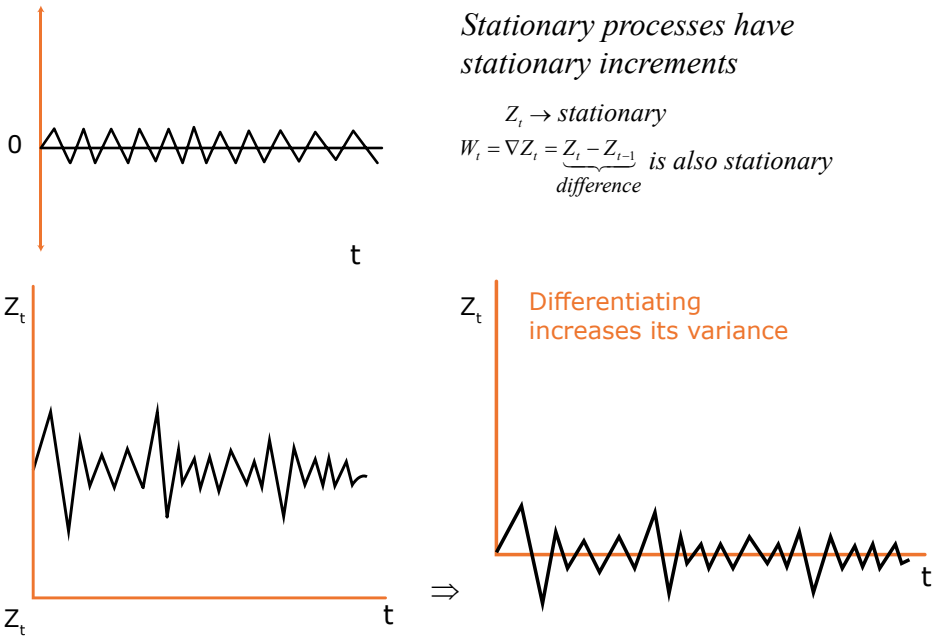
A very important stationary process is defined by:

- $E(Z_t) = 0$
- $Var(Z_t) = \sigma 2$ Homoscedasticity.
- $Cov(Z_t\ Z_{t-k}) = 0\ \forall\ k \neq 0$ No autocorrelation

The following are required for the model resiudes

If the residuals do not meet these conditions, the process is called white noise process. The model is missing something.

Image 5.13a, b, and c



5.8.2 Homogeneous Process–Integrated of 1st orde

Most of the series are not stationary, they increase their variance, but they can be re-differentiated.

When a series is not stationary and with the first difference it becomes stationary, the process is said to be homogeneous of 1st order or I (1).

$\{ Z_t \}$, t= 1,….$n \rightarrow W_t$

$W_t = Z_t - Z_{t-1} \rightarrow W_t = I\ (1)$

If two differences are necessary to make the series stationary, the process is homogeneous of 2nd order or I (2).

Image 5.14a

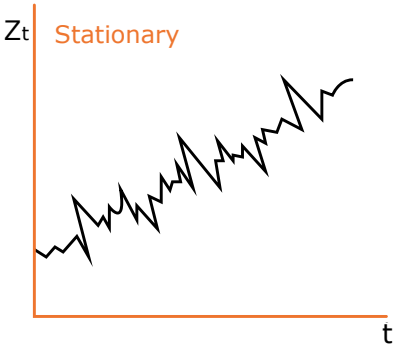
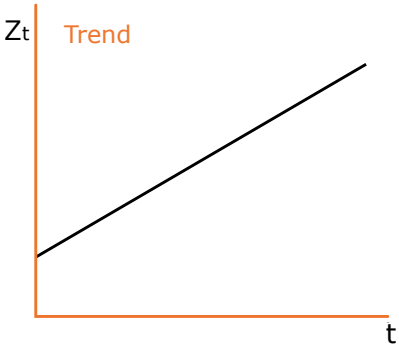


Image 5.14b



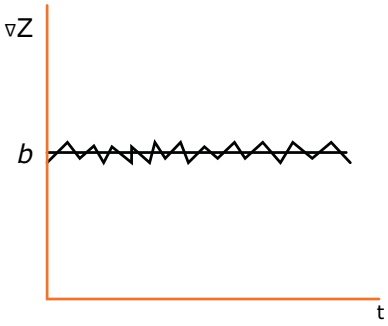
5.8.3 Integrated of 1st Order

Image 5.15

$$\nabla Z_t = a + b_t + \varepsilon_t - (a + b\,(t-1) + \varepsilon_{t-1})$$

$$= \cancel{a} + \cancel{b_t} + \varepsilon_t - \cancel{a} + \cancel{b_t} + b - \varepsilon_{t-1}$$

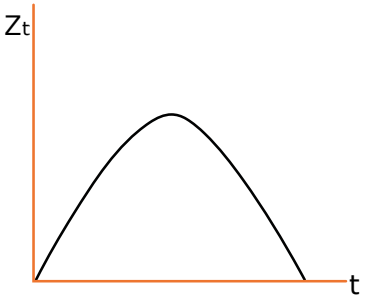
$$\nabla Z_t = b_t + \varepsilon_t - \varepsilon_{t-1} \;=\; b + \nabla\varepsilon_t$$

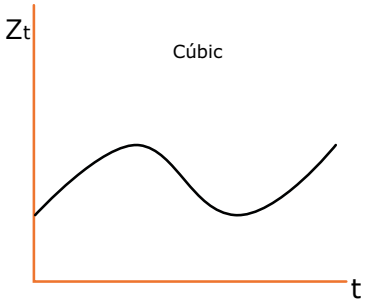


Where *b* is the constant and $\nabla\varepsilon_t$ is stationary.

If the series is the result of a quadratic function + ε double differentiation is required.

Image 5.16a Image 5.16b





If a series is the result of a polynomial of order *h* + a stationary process (ε), that series requires *h* differentiations to make it stationary.

5.8.4 Autoregressive Processes

The series is represented in itself lagged 1, 2, 3,..., p periods, i.e., it is basically a linear regression equation, with the special feature that the dependent variable Z in period t depends on its own values observed in periods prior to t, and weighted according to the autoregressive coefficients $\phi 1$ $\phi 2$..... ϕp.

$$Z_t = \phi Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t \qquad \rightarrow AR(p)$$

Where $\phi \rightarrow$ are the importance weights of each of the lags.

5.8.4.1 First-order AR processes [AR (1)]

$$Z_t = \phi Z_{t-1} + a_t$$

$$1 - \phi Z_{t-1} = a_t$$

$$Z_t - \phi B Z_t = a_t$$

$$(1 - \varphi B) Z = a_1$$

$$Z_t = (1 - \phi B)^{-1} a_t$$

$$a_t \begin{cases} E(a_t) = 0 & \forall_t \\ Var(a_t) = \sigma_a^2 & \forall_t \\ Cov(a_t\ a_{t-K}) = 0 & \forall_t \\ Cov(a_t\ a_{t-K}) = \begin{cases} \sigma_a^2 & k = 0 \\ 0 & k > 0 \end{cases} \end{cases}$$

$$Z_t = Z_t - u$$

where a_t is a white noise

$$\frac{\sigma_z^2}{(1-\varphi^2)} = \sigma_z^2 = \gamma_o \quad -1 < \phi < 1$$

$$\gamma_k = \varphi_k \gamma_o$$

$$\rho_k = \frac{\gamma_k}{\gamma_o} = \frac{\phi^k \gamma_o}{\gamma_o} = \phi^k$$

$$\frac{1}{1-\varphi} = (1 + \varphi B + \varphi^2 B^2 + \varphi^3 B^3 + \ldots) a_t$$

$$Z_t = a_t + \varphi\, a_{t-1} + \varphi^2 a_{t-2} + \varphi^3 a_{t-3} + \ldots \quad MA(\infty)$$

$$AR(1) \;=\; MA(\infty)$$

Equivalent

$$E(Z_t) = 0 \rightarrow \forall_t\ E(a_t) = 0$$

$$Var(Z_t) = (\sigma_a^2 + \varphi^2 \sigma_a^2 + \varphi^4 \sigma_a^2\ \varphi^6 \sigma_a^2 \ldots)$$

$$Var(Z_t) = \frac{\overbrace{1 + \phi^2 + \phi^4 + \phi^6}}{(1 - \phi^2)} = \frac{\sigma_a^2}{(1 - \phi^2)}$$

$$Cov(Z_t\ Z_{t-k})$$

$E(\varphi Z_{t-k})$

$E(Z_t Z_{t-k}) =$

5.8.4.2 Second-order AR processes [AR (2)]

$Z_t = \varphi 1 Z_{t-1} + \varphi 2 Z_{t-2} + a_t$ $\qquad Z_t = Z_{t-\mu}$

$(1 - \varphi 1 B - \varphi 2 B2) Zt = a_t$ $\qquad a_t$ is White Noise.

- $k>0\ \gamma_k = \varphi_1 \gamma_{k-1} + \varphi_2 \gamma_{k-2}$
- $K=0\ \gamma_0 = \varphi_1 \gamma_1 + \varphi_2 \gamma_2 + \sigma_a^2$

$\rho_k = \varphi_1 \rho_{k-1} + \varphi_2 \rho_{k-2}$ $\qquad k > 0$

$\rho_1 = \varphi_1 + \varphi_2 \rho_1$ for k=1

$\rho_2 = \varphi_1 \rho_1 + \varphi_2$ for k=2

$$\sigma_z^2 = \frac{(1-\phi_2)\sigma_a^2}{(1+\phi_2)(1-\phi_1-\phi_2)(1+\phi_1-\phi_2)}$$

Conditions for an AR process (2) to be stationary:

$-1 < \varphi 2 < 1$ $\qquad \varphi 1 + \varphi 2 < 1$

$\varphi 2 - \varphi 1 < 1$

5.9 Partial Autocorrelation Function (PACF)

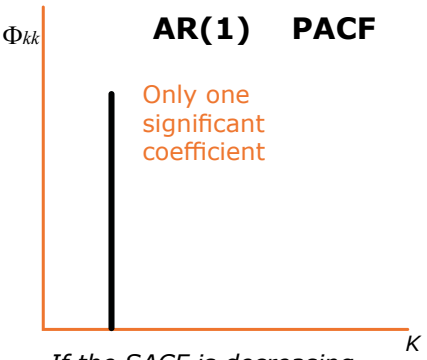
$$Z_t = \phi Z_{t-1} + \phi_2 Z_{t-2} + \ldots + \phi_p Z_{t-p} + a_t$$

The partial autocorrelation between Z_{t-2} and Z_t eliminates the effects of Z_{t-1}, thus the partial autocorrelation between Z_t and Z_{t-5} eliminates the effects of Z_{t-1}, Z_{t-2}, Z_{t-3}, and Z_{t-4}.

Hence, in an AR(1) process the partial autocorrelation between Z_t and Z_{t-2} is equal to zero. In an AR(2) process, the partial autocorrelation between Z_t and Z_{t-3} is equal to zero, etc.

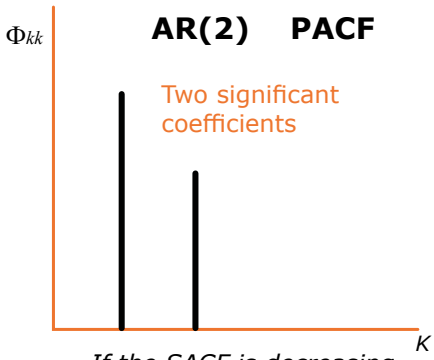
From the above definition, it is inferred that an AR(ρ) process will have the first ρ non-zero partial autocorrelation coefficients and, therefore, in the PACF the number of non-zero coefficients indicate the order of the process.

Image 5.17a



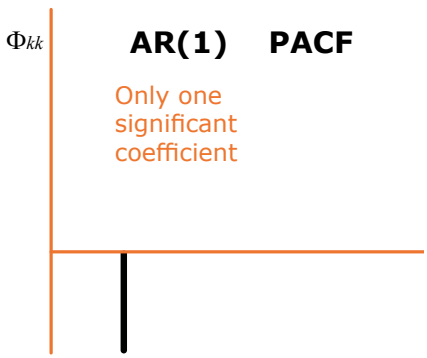
If the SACF is decreasing, it is an AR (1).

Image 5.17b



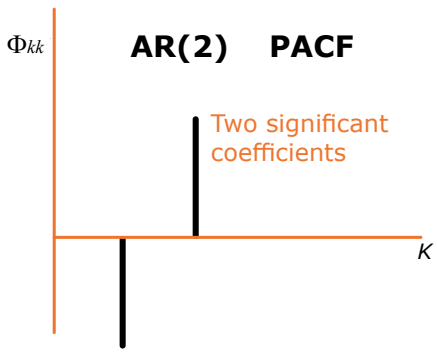
If the SACF is decreasing, it is an AR (2).

Image 5.18a



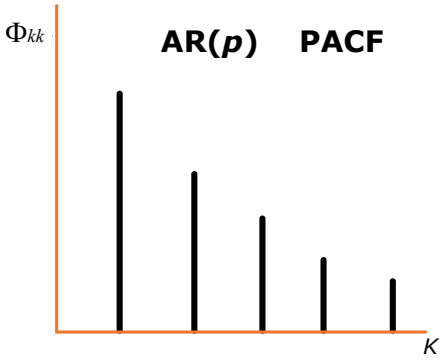
If the SACF is decreasing, it is an AR (1).

Image 5.18b



If the SACF is decreasing, it is an AR (2).

Finally: Image 5.19



The most direct way to find PACFs is through the regression equation

- $Z_t = \varphi_{11} Z_{t-1} + e_t$ — 1st Order
- $Z_t = \varphi_{21} Z_{t-1} + \varphi_{22} Z_{t-2} + e_t$ — 2nd Order
- $Z_t = \varphi_{k1} Z_{t-1} + \varphi_{k2} Z_{t-2} + \ldots\ldots\ldots + \varphi_{kk} Z_{t-k} + e_t$ — K Order

Coefficients φ_{11}, φ_{22}, ... φ_{kk} are the partial correlation coefficients

Image 5.20a

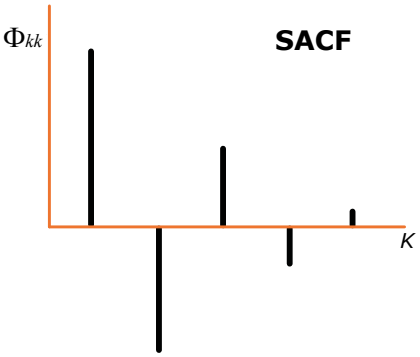


Image 5.20b

Image 5.21a

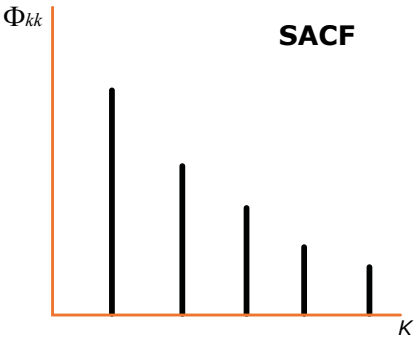
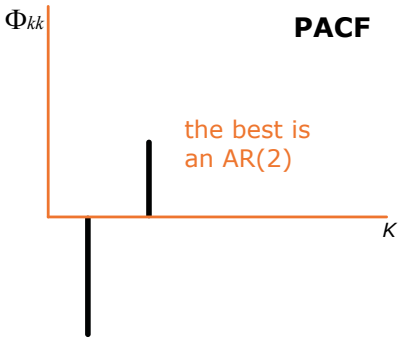


Image 5.21b



5.9.1 Significance of φ kk

- t is used to test the significance of φ kk.

$$t = \frac{\hat{\phi}_{kk}}{\frac{1}{\sqrt{N}}} = \sqrt{N}\hat{\phi}_{kk} \qquad \hat{\phi}_{kk} \to N(0, 1/N)$$

Standard deviation

In econometric packages, what is within the bands is not significant, and what falls outside is significant.

The calculated value is contrasted with a critical value tc = 2 for α = 0.05.

5.9.2 Moving Average Processes (MA model)

MA models represent a stochastic process $\{Z_t\}$ whose values can be dependent on each other as a weighted finite sum of independent random shocks $\{a_t\}$ i.e.

$$Z_t = a_t - \theta a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q} \qquad \to MA(q)$$

Where $\theta_1, \theta_2, \ldots \theta q$ are weightings (moving average parameters) associated with the random shocks in periods t_{-1}, t_{-2},.....t_{-q} respectively.

5.9.2.1 MA (1) Process

$Z_t = a_t - \theta\, a_{t-1}$

$Z_t = (1 - \theta B)_{at}$

If $|\theta| < 1$ there exists inverse operator $(1 - \theta B)^{-1}$

$\gamma_k = \theta_{\sigma a}{}^2$ if k=1 $\qquad$ $\rho_k = \frac{\gamma_k}{\gamma_o} = \frac{-\vartheta}{1+\vartheta^2}$ si $k = 1$

$\gamma_k = 0$ if $k > 1$ $\qquad$ $\rho_k = 0$ si $k > 1$

Based on this, it is found that SACF of an MA(1) process is equal to zero for k > 1

5.9.2.2 MA (2) Process

$Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$

Y is stationary for all values of θ_1 and θ_2. It is invertible only if the characteristic roots of the equation $(1-\theta_1 B-\theta_2 B_2) = 0$ fall outside the unit circle, that is if it satisfies that:

$$\theta_2 + \theta_1 < 1$$
$$\theta_2 - \theta_1 < 1$$
$$-1 < \theta_2 < 1 \qquad (1 - \theta_1 B - \theta_2 B^2) = 0$$

5.9.2.3 ARMA Processes

The autoregressive moving-average (ARMA) process (p,q) is represented by

$$(1-\phi_1 B-\phi_2 B^2 -\ldots-\phi_p B^p) Z_t = (1-\theta_1 B-\theta_2 B^2 -\ldots-\theta_q B^q)\, a_t$$

or in compact set notation

$$\phi_p (B) Z_t = \theta_q (B)\, a_t$$

This generalization arises from the fact that the time series observed in practice often have characteristics of both AR and MA processes.

5.9.2.3.1 ARMA Process (1, 1)

The simplest process is the ARMA (1,1) which is written as

$Z_t - \varphi_1 Z_{t-1} = a_t - \theta_1 a_{t-1}$

$(1- \varphi_1 B) Z_t = (1-\theta_1 B) a_t$

where φ_1 different from θ_1 and $|\varphi| < 1$ for the process to be stationary and $|\theta| < 1$ for it to be invertible.

Image 5.22a

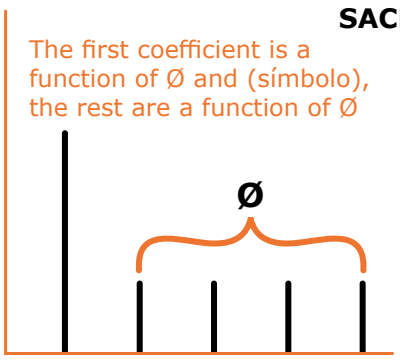
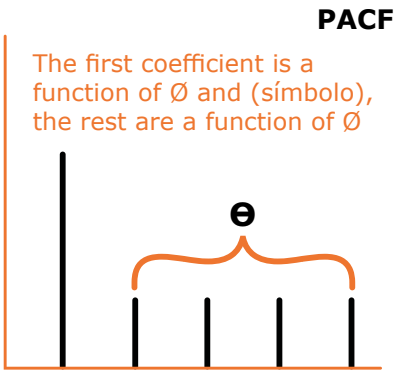


Image 5.22b



	SACF	PACF
AR(ρ)	Many non-zero coefficients that decrease with the delay as a mixture of exponentials and sinusoidal	First non-zero ρ coefficients and the remaining zero coefficients
MA(q)	First non-zero q coefficients an d the remaining zero coefficients	Many non-zero coefficients that decrease with the lag as a mixture of exponentials and sinusoidal
ARMA(ρ,q)	Initial q coefficients that depend on the order by the MA part. Then, there is a decrease dictated by the AR part	Initial ρ values that depend on the order of the AR(ρ) part and followed by decreases that depend on the MA part.

5.10 Processes for Non-stationary Series

- In the discussion of the AR, MA, and ARMA processes we relied on the assumptions that the time series were stationary. However, many of the observable series in economics are non-stationary.
- A special case of non-stationary processes is the random walk.

Image 5.23a

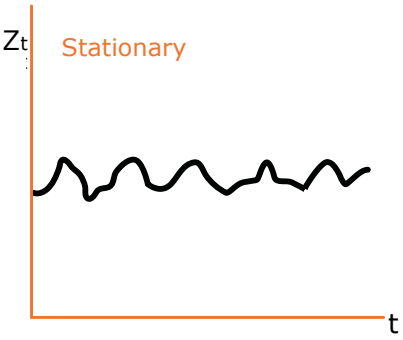


Image 5.23b

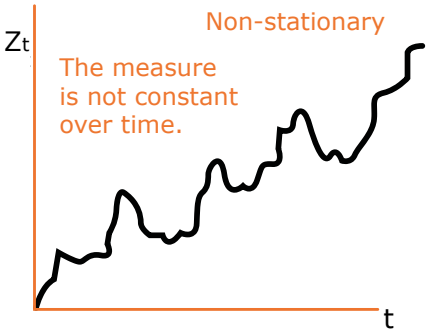
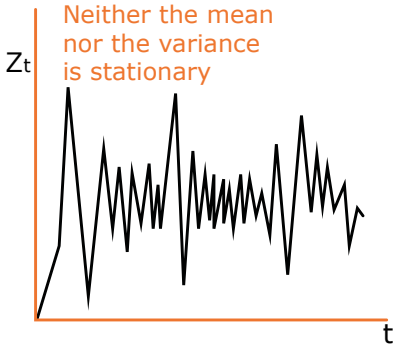


Figure 5.23c



5.10.1 Random Walk

In the $Z_t=\varphi Z_{t-1}+a_t$ equation if $|\varphi|=1$, the process is not stationary, but neither is it explosive. It becomes a homogeneous process of first order (since its first difference is $Z_t-Z_{t-1}=a_t$ if it is a stationary process) which is called random walk.

5.10.2 Autoregressive Integrated Moving Average (ARIMA) Process

It is a process of type:

$$(1-\phi B-\phi_2 B^2 \ldots-\phi_p B^p)(1-B)^d Z_t=(1-\theta B-\theta_2 B^2-\ldots..-\theta_q B^q)at$$

which we will call the ARIMA (p,d,q) process.

In this notation, *p* is the order of the stationary in the AR part, *d* is the number of unit roots (order of homogeneity of the process) or number of required differentiations, and *q* is the order of the moving average part.

We use the $\nabla = 1 - B$ (difference operator) when a series is non-stationary, we conduct a differential process (∇^d) and make it stationary.

$$W_t = (1 - B)Z_t$$

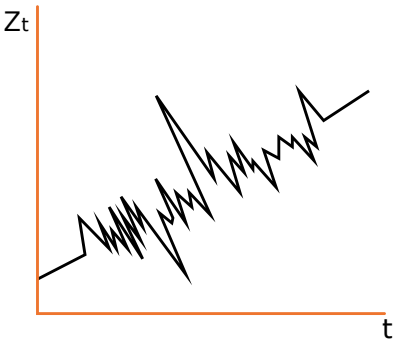
$$\underbrace{\frac{1}{(1\text{-}B)}}_{(1+B+B\text{^}2+B\text{^}3\ldots)} W_t = Z_t$$

$$W_t + W_{t-1} + W_{t-2} + W_{t-3} \ldots = Z_t$$

5.10.3 Construction of Time Series Models by the Box-Jenkins method (1970)

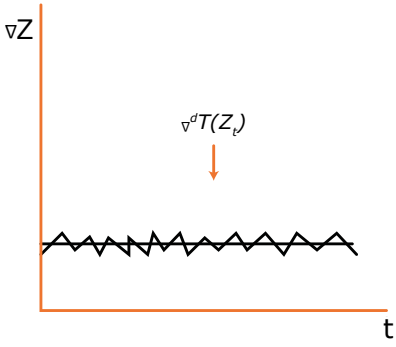
a) Identification of a possible model within the ARIMA (p,d,q) models; that is, determination of the *p*, *d*, and *q* values that specify the appropriate ARIMA model for the series under study.

Image 5.24a



⇒

Image 5.24b



b) Estimation of the parameters involved in the model, through non-linear estimation techniques.

c) Diagnosis or check of the model, in order to verify whether the basic assumptions made regarding the residues are true.

5.10.3.1 Identification

To conduct the identification is necessary to:

- Decide which transformation must be applied to convert the underlying process into a stationary process. Determine the transformation to make the T variance stationary and/or the number of differentiations to station the *d* mean.
- Determine a model for the stationary process, i.e., the p and q orders of its ARMA (p,q) representation.

5.10.3.2 Estimation

For an ARMA (p, q) model, which is the most general form, the parameter estimation will be discussed.

$\Phi = (\varphi 1, \varphi 2, \varphi 3, ... \varphi p)$
$\Theta = (\theta 1, \theta 2, \theta 3, ...\theta q)$

For pure RAs, OLS can be used.

Non-linear least squares (NLS) estimation techniques are used for ARIMA.

Box-Jenkins suggests a nonlinear estimation method (MCNL in Spanish) based on Marquardt algorithm (1963).

(RATS→ estimates by the Gauss-Newton method).

5.10.3.3 Diagnostic or Check-up

Verify whether the model residues meet the white noise conditions.

$E(a_t) = 0$
$Var(a_t) = \sigma^2$
$Cov(a_t\ a_{t-k}) = 0$ for all *k* different from zero

The most important assumption for measuring the validity of an ARIMA model relates with the assumption that the random errors are independent, i.e. they are not autocorrelated.

5.10.4 Ways to Check-up the Model

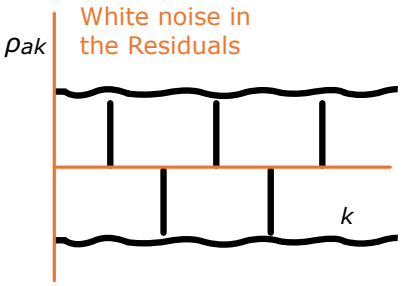
1. The SACF of residuals, the SACF of residual for an ARIMA model will ideally have autocorrelation coefficients that may be statistically equal to zero. The Ljung-Box and Box-Pierce tests are used to verify this condition.
2. Figure of the residudes.
3. Overestimation technique.

5.10.4.1 SACF on Waste

$SACF \rightarrow Z_t$

$SACF \rightarrow \hat{a}_t$

Image 5.25



$\rho_{ak} \quad H_O = \rho_{a1} = \rho_{a2} = \rho_{a3} = \ldots = \rho_{an} = 0$

$H_i = at\ least\ one\ \rho_{a1}\ is \neq 0$

5.10.4.1.1 LJUNG-BOX

The null hypothesis

$H_o : \rho_1(a) = \rho_2(a) = \rho_3(a) = \ldots\ldots\ldots \rho_K(a) = 0 \quad Q^* = n(n+2)\sum_{k=1}^{K} \frac{\hat{\rho}_k^2(a)}{n-k}$

It is checked with the test

where *n* is the number of observations used to estimate the model *n* = (*N-d-p*)

The Q* statistic approximately follows a distribution χ^2_{k-m} where *m* is the number of estimated parameters in the ARIMA model (*p+q*).

5.10.4.1.2 Using Residuals to Modify the Model

Suppose an AR(1) model is identified and estimated.

$$(1-\varphi B)Z_t = b_t \qquad \textbf{(1)}$$

where bt is not RB, and suppose that the SACF of the bt residuals have a significant coefficient followed by the rest that are approximately equal to zero. This suggests an MA(1) for b_t.

$$b_t = (1-\theta B)a_t \qquad \textbf{(2)}$$

Where at is not autocorrelated; using 2 to substitute b_t in 1. The result is an ARMA (1,1) for Z_t

$$(1-\varphi B)Z_t = (1-\theta B)a_t$$

5.10.4.2 Graph of Residuals Versus Time

Image 5.26

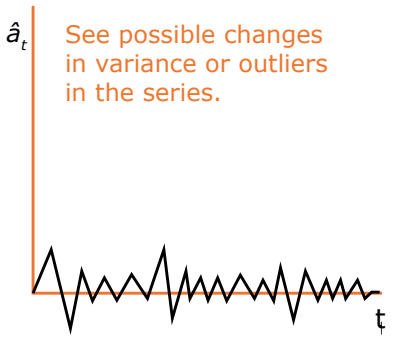


Image 5.27a

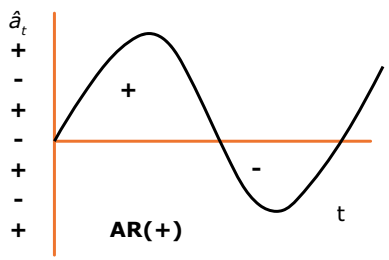
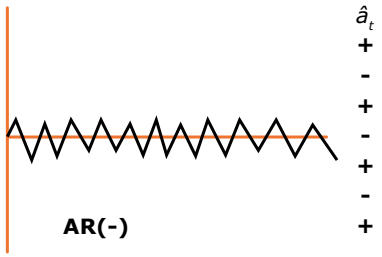


Image 5.27b



5.10.4.3 Overestimation Technique

$$ARMA(p,q) \rightarrow ARMA(p+1,q)$$
$$\rightarrow ARMA(p,q+1) \quad \textit{Overestimate on moving average part}$$

However, to overestimate them together *(p,q)* is not advised.

5.10.5 Model Selection Criteria

- AKAIKE Information Criterium (AIC)
- SCHWARTZ Information Criterium (SBC) or Bayesian Information Criterion (BIC)

Which one has the best fit? The selection criterion is the smallest value because it represents the best fit (it is parsimonious).

Parsimony → To explain much with few exogenous variables.

In time series, the more variables there are, the ↑R². Therefore, is not parsimonious.

- They are calculated as:
- AIC = Nln(SRC)+2K
- SBC = Nln(SRC)+Kln(N)
- K = number of estimated parameters *p* + *q* + *cte*
- N = number of usable observations

In order to have a selection criterion, it is necessary that the number of observations to be compared be the same. AR (1) → 99, AR (2) → 98.

$$Boxjenk\ (AR=1)\ Z_t\ \ it\ is\ estimated\ starting\ from\ the\ second\ observation.$$

$$Boxjenk\ (AR=2)\ Z_t\ /\ residuals$$

Graph 1

zt

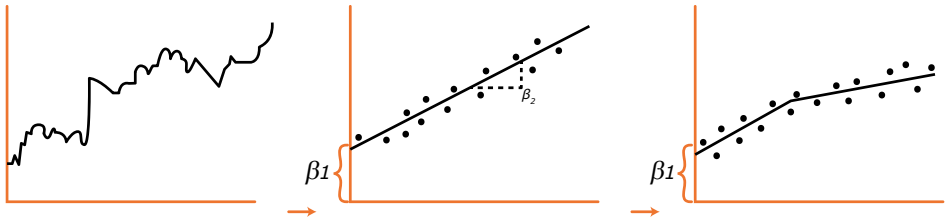
Correlate (Partial = name of the partial correlations) Zt/ Name of the simple correlations.

Boxjenk (AR = 1, de t = 1, MA = 1)Zt / residuals.

Correlate (θstat) residues.

The assumption of the regression model is structural stability, but an exogenous change modifies the parameters.

Image 5.28a **Image 5.28b** **Image 5.28c**



Structural change also occurs in time series, for which the Chow test is applied.

$$Boxjenk\ (AR=2)\ Z_t\ /\ residuals$$

5.11 Seasonality in Time Series

They are series that, apart from a long-term trend and/or cycle, show fluctuations that are annually repeated.

Seasonality makes the series non-stationary because it changes to mean. If the seasonality is approximate / constant it can be eliminated by taking seasonal differences.

Seasonal difference operator ∇_s^D *Number of seasonal differences. Periodicity of the series in which it is repeated.*

$$\nabla_s^D = (1-B^s)^D$$
$$\nabla_{12}^1 Z_t = Z_t - Z_{t-12}$$
$$(1-B^{12})\,Z_t$$

D x S remarks are lost

5.11.1 General form of a Seasonal ARIMA

5.11.1.1 SARIMA (p, d, q) (P, D, Q)S

$$\Phi_P(B^s)\phi_p(B)\nabla^d\nabla_s^D Z_t = \Theta_Q(B^s)\theta_q(B)a_t \rightarrow \text{SARIMA (p, d, q) (P, D, Q)s}$$

$$\phi_p(B) = 1-\phi_1 B-\phi_2 B^2-\ldots-\phi_p B^p \qquad \theta_q(B) = 1-\theta_1 B-\theta_2 B^2-\ldots..-\theta_q B^q$$
$$\Phi_P(B^s) = 1-\Phi_s B^s-\Phi_{2s} B^{2s}-\ldots..-\Phi_{Ps} B^{Ps} \qquad \Theta_Q(B^s) = 1-\Theta_s B^s-\Theta_{2s} B^{2s}-\ldots..-\Theta_{Qs} B^{Qs}$$

$\nabla^d = (1-B)^d \rightarrow$ *Regular difference operator*

$\nabla_S^D = (1-B^S)^D \rightarrow$ *Seasonal difference operator*

Example. *(p,d,q) (P,D,Q)s*

- SARIMA (1,1,2)(2,1,1,1)4 *pPdDD = 1Q*

$$\overbrace{(1-\phi B)}\underbrace{(1-\Phi_4 B^4-\Phi_8 B^8)}\overbrace{(1-B)}\underbrace{(1-B^4)}Z_t = (1-\overbrace{\theta_1 B-\theta_2 B^2}^{})\underbrace{(1-\Theta_4 B^4)}a_t$$

- SARIMA (0,2,1)(2,1,2)12

$$(1-\Phi_{12}B^{12}-\Phi_{24}B^{24})(1-B)^2(1-B^{12})Z_t = (1-\theta B)(1-\Theta_{12}B^{12}-\Theta_{24}B^{24})a_t$$

- SARIMA (2,1,1)(1,0,2)S

$$(1-\phi_1 B\ -\phi_2 B^2)(1-\Phi_s B^s)\underbrace{(1-B)(1-B^s)}Z_t = (1-\theta_1 B)(1-\Theta_s B^s-\Theta_{2s}B^{2s})a_t$$

Seasonal lag operator

5.11.2 Identification of Seasonal Processes

The identification process is similar to that for non-seasonal processes.

To select the values of D that make the series stationary, it is recommended to graph the sample SACF of:

$$\{T(Z_t)\}, \{\nabla T(Z_t)\}, \{\nabla^2 T(Z_t)\}, \{\nabla_s T(Z_t)\}, \{\nabla\nabla_s T(Z_t)\}, \{\nabla^2\nabla_s T(Z_t)\}$$

5.12 Forecasting with Time Series Models

It aims to predict *h* periods from an estimated ARMA (p, q) model that has information available up to *T* period for the variable Z_t.

The objective is to obtain a prediction that is as close as possible to the true future value.

One way to achieve this objective is to minimize the mean squared error (MSE) between the true value and the predicted value, based on the information available up to *T* period.

Specifically, it seeks to minimize the expected value

$$E_T\left[(Z_{T+h} - \hat{Z}_{T+h})^2\right]$$

$$\hat{Z}_{T+h}$$

It is shown that the predictor that minimizes the mean squared error is the expectation of the conditional distribution

$$\hat{Z}_{T+h} = E_T\left[Z_{T+h}\right]$$

5.12.1 Prediction with an AR(1) model

$$\hat{Z}_{T+1} = \phi_0 + \phi_1 Z_T$$

$$\hat{Z}_{T+2} = \phi_0 + \phi_1 \hat{Z}_{T+1}$$

:

$$\hat{Z}_{T+h} = \phi_0 + \phi_1 \hat{Z}_{T+h-1}$$

5.12.2 Prediction with an MA(1) model

$$\hat{Z}_{T+1} = \mu + \theta_1 a_T$$
$$\hat{Z}_{T+2} = \mu$$
$$:$$
$$\hat{Z}_{T+h} = \mu$$

5.12.3 Prediction with an ARMA(1,1) model

$$\hat{Z}_{T+1} = \phi_0 + \phi_1 Z_T + \theta_1 a_T$$
$$\hat{Z}_{T+2} = \phi_0 + \phi_1 \hat{Z}_{T+1}$$
$$\hat{Z}_{T+3} = \phi_0 + \phi_1 \hat{Z}_{T+2}$$

5.12.4 Variance of prediction error

5.12.4.1 AR (1) Model

$$\text{var}(\hat{a}_{t+1}) = \sigma_a^2$$
$$\text{var}(\hat{a}_{t+2}) = \sigma_a^2(1+\phi^2)$$
$$.$$
$$\text{var}(\hat{a}_{t+h}) = \sigma_a^2(1+\phi^2+\phi^4+.....+\phi^{2(h-1)})$$
$$.$$

5.12.4.2 MA (1) Model

$$\text{var}(\hat{a}_{t+1}) = \sigma_a^2$$
$$\text{var}(\hat{a}_{t+2}) = \sigma_a^2(1+\theta^2)$$
$$.$$
$$\text{var}(\hat{a}_{t+h}) = \sigma_a^2(1+\theta^2)$$

5.12.4.3 ARMA (1,1) Model

$$\text{var}(\hat{a}_{t+1}) = \sigma_a^2$$

$$\text{var}(\hat{a}_{t+2}) = \sigma_a^2(\phi+\theta)^2+1)$$
$$.\text{var}(\hat{a}_{t+3}) = \sigma_a^2(1+(\phi+\theta)^2+(\phi^2+\theta)^2)$$

Interval prediction:

Z_{t+h}+/-1.96 (SD(error of prediction))

- Model selection from prediction
- Out of sample forecasting

5.13 Prediction in Non-Stationary Series

$$W_t = \Delta Z_t = Z_t - Z_{t-1}$$
$$Z_t = W_t + Z_{t-1}$$
$$\hat{Z}_{t+1} = \hat{W}_{t+1} + Z_t$$

$$W_t = \Delta^2 Z_t = Z_t - 2Z_{t-1} + Z_{t-2}$$
$$Z_t = W_t + Z_{t-1}$$
$$\hat{Z}_{t+1} = \hat{W}_{t+1} + 2Z_t - Z_{t-1}$$

$$W_t = \ln(Z_t)$$
$$\hat{Z}_{t+1} = Exp(\hat{W}_{t+1} + \frac{1}{2}\text{var}(a_{t+1}))$$

5.13.1 Unit Root

If |φ| = 1, there is a unit root, and the process is non-stationary. $Z_t = \phi Z_{t-1} + a_t \quad (1)$

One of the tests to detect unit root is the Dickey-Fuller (DF) test, which proposes the following models:

Where it is contrasted
The Ho:δ = 0 (unit root).
where δ = φ-1

$$\nabla Z_t = \delta Z_{t-1} + a_t \quad (2)$$
$$\nabla Z_t = \beta_1 + \delta Z_{t-1} + a_t \quad (3)$$
$$\nabla Z_t = \beta_1 + \beta_2 t + \delta Z_{t-1} + a_t \quad (4)$$

These *Eqs.* are contrasted with tabulated values τ, τ_μ y τ_t respectively.

To protect the model from the AR problem, the ΔZ_{t-i} term is added to the right side of *Eqs.* This test is called the augmented Dickey-Fuller test (ADF).

$$\nabla Z_t = \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \nabla Z_{t-i} + a_t \quad (5)$$

$$\nabla Z_t = \beta_1 + \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \nabla Z_{t-i} + a_t \ (6)$$

$$\nabla Z_t = \beta_1 + \beta_2 t + \delta Z_{t-1} + \sum_{i=1}^{m} \alpha_i \nabla Z_{t-i} + a_t \ (7)$$

These *Eqs.* are contrasted with the tabulated values τ, τ_μ y τ_t respectively.

5.13.2 Phillips-Perron (PP)

established a test for unit root with serial correlation of the errors which considers the same Dickey-Fuller *Eqs.* (2), (3), and (4). However, the PP test calculates some Zρ and Z_t values which contrast with the values tabulated by DF τ, τ_μ y τ_t respectively.

5.13.3 Seasonal Unit Roots

∇ $\sum$ ∇

DF tu

5.13.4 Phillips-Perron Test in The Presence Of Structural Changes

$$H_1 : Z_t = \mu + Z_{t-1} + \gamma D_P + a_t$$
$$A_1 : Z_t = \mu + \beta t + \gamma D_S + a$$

$$H_2 : Z_t = \mu + Z_{t-1} + \gamma D_S + a_t$$
$$A_2 : Z_t = \mu + \beta t + \gamma D_T + a_t$$

$$H_3 : Z_t = \mu + Z_{t-1} + \gamma_1 D_P + \gamma_2 D_S + a_t$$
$$A_1 : Z_t = \mu + \beta t + \gamma_2 D_S + \gamma_3 D_T + a_t$$
$$D_S \begin{cases} 1 & IF \quad t \geq \tau + 1 \\ 0 & IF \quad t < \tau + 1 \end{cases}$$
$$D_P = \nabla D_S = \begin{cases} 1 & IF \quad t = \tau + 1 \\ 0 & OTHERWISE \end{cases}$$
$$D_T = \begin{cases} t - \tau & IF \quad t > \tau \\ 0 & OTHERWISE \end{cases}$$

5.13.4.1 Phillips-Perron Test Step by Step

- Eliminate the data's trend according to Ha. Example under specification 1.

$$Z_t = \mu + \beta t + \gamma D_S + \hat{Z}_t$$

- Estimate the following equation:

$$\hat{Z}_t = \phi \hat{Z}_{t-1} + \sum_{i=1}^{m} \alpha_i \nabla \hat{Z}_{t-i} + a_t \rightarrow to\ test\ RB$$

- Calculate the statistic *t* for the Ho:Φ=1

$$t = \frac{\phi - 1}{DS(\phi)}$$

- Compare the calculated *t* with the values tabulated by Phillips-Perron test.

5.14 Intervention Analysis

- Event exogenous to the behavior of the variable under study.
- Generic expression of an intervention model:

$$T(Z_t) = \varepsilon_{I,t} + N_t$$
$$\phi(B)\nabla^d N_t = \theta_0 + \theta(B) a_t$$
$$\delta_r(B)\nabla^b \varepsilon_{I,t} = \omega_s(B) P_{i,t}$$
$$T(Z_t) = \frac{\omega_s(B)}{\delta_r(B)\nabla^b} P_{i,t} + \frac{\theta_0 + \theta(B) a_t}{\phi(B)\nabla^d}$$

5.14.1 Intervention and transfer function analysis

$$Z_t = a_0 + A(B)Z_{t-1} + C_0 X_{t-b} + D(B)a_t \rightarrow \textit{intervention model}$$
$$Z_t = a_0 + A(B)Z_{t-1} + C(B)X_t + D(B)a_t \rightarrow \textit{transfer function model}$$

Xt is the intervening variable or dummy in the first equation and it would be an exogenous variable in the transfer function.

CHAPTER 6

CASE STUDY–CORRUPTION RISK IN COLOMBIA

Corruption is one of the problems that most limit a country's development, widens productive inequality between its regions, and, thus, stagnates its economic growth. In recent years, corruption measurements in Colombia have shown figures above the international average. Therefore, the main objective of this study was to analyze the efficiency of public spending at the regional level in Colombia to determine the risk of corruption in the State[13].

Thus, the study analyzed the corruption risk of Colombia's 32 departments. Moreover, an alternative measure, the Golden & Picci Corruption Risk Index (GRI&P), developed in this chapter, was applied.

Golden & Picci (2005) proposed an alternative measure to quantify corruption, IG&P, based on observed data, not on opinions. Their proposal is based on the relationship between the goods and/or services provided the State provides and the cumulative payment made for them. In other words, the index is constructed as the ratio between the provision of goods and services and their accumulated investment.

13 Basing the exercise on Avila and Oliveira (2023).

In addition to this endogenous variable, the transparency index of public entities (ITEP in Spanish) and the open government index (IGA in Spanish) were used to obtain a more robust model. Throughout the chapter, we present the endogenous variables and our proposal of exogenous variables to determine the degree of explanation of the variables that we consider to be determinant in the risk of corruption, which are supported by Avila et al (2022).

Finally, statistically, the results showed that, according to the IG&P results, there is a direct relationship between inefficiency and ineffectiveness in fulfilling the government's duties, socioeconomic development, and revenues from exploiting of natural mineral resources.

6.1 Methodology

6.1.1 Construction of the Golden and Picci (GI&P) Index for Colombia

It is important to note that the GI&P is constructed as the ratio between the provision of goods and services and their accumulated investment. Initially, the numerator of the GI&P is calculated, corresponding to the provision of goods and services, and three groups of variables were used as a reference: education, health, and basic sanitation.

In terms of service provision[14], for the education group, the variables are taken into account: 1) coverage in primary and secondary education in official schools, with information from the Ministry of National Education, and 2) classroom-space public schools, with data available in the comprehensive performance index of the National Planning Department (DNP). For the second health group, the variables are considered: 1) number of hospital beds in public hospitals, as reported by the Ministry of Health's hospital information system; 2) infant mortality, estimated by National Administrative Department of Statistics (DANE in Spanish); and 3) total coverage of the subsidized regime, according to the

14 The selection of the variables included in the construction of the index was made considering their relationship with investment expenses and the availability of information.

Ministry of Health. For the third group, sanitation, the variables are considered: 1) water supply coverage and 2) sewerage coverage. All data are for the year 2018.

Once these data are available, they are normalized by dividing the variables by the total population. In this case, each variable was divided by the estimated population of the municipality receiving the service in 2018, as calculated by DANE. That is, each variable is expressed in its 1,101 Colombian municipalities.

Once the variables of the three groups are normalized, they are standardized/typed in an index from 0 to 100, where 100 corresponds to the best data[15], and, therefore, all the data are organized in a relatively. This step is necessary to have the same unit comparable from one group to another since each can be in a different unit of measurement. Once each variable has been standardized, the arithmetic average of the variables within each group is extracted. However, for the proportional weight of each sub-index to be correct (Education-*I Educ*, Health-*I Sal,* and Basic Sanitation-*I SB*), its percentage share within the goods and services provision index *(IP Bs and Sv*) must first be determined. Therefore, it was essential to have the accumulated investment *(Inv Ac*) during the fifteen years for each of these items by the municipality and thus determine their percentage of participation to calculate the *IP Bs* and *Sv*.

$$IP\ Bs\ y\ Sv = \left[\left(IEduc * \left(\frac{\%\ Inv\ E}{100}\right)\right) + \left(I\ Sal * \left(\frac{\%\ Inv\ AcS}{100}\right)\right) + \left(ISB * \left(\frac{\%\ Inv\ AcSB}{100}\right)\right)\right]/3 \quad \textbf{(6.1)}$$

Finally, the standardization of each group is expressed for each municipality as a proportion of the national average. Thus, a value greater than 1 implies that the provision of that good or service per person is greater than the national average, and a value less than 1 implies that the provision is less than the national average. It is important to note that this measure is taken for each of the defined variable group.

In addition, to construct the denominator of the IG&P, a cumulative investment indicator must be created. For this purpose, we used data on national transfers to municipalities for education, health, and basic sanitation, which are generally the primary investment objectives in all Colombian municipalities. For all 1,101

15 In the case of the health group, with the "infant mortality" variable, the scale is inverse; the best data clearly corresponds to the one with the lowest value.

Colombian municipalities, transfer data were taken from the National Planning Department (DNP in Spanish) from 2004 to 2018 and expressed in constant 2018 values. Once this conversion is made, they are added together, which gives a cumulative investment value per municipality. With this data, we proceeded to normalize with the total population, and, as for the index of provision of goods and services, we expressed the data as a proportion of the national average. Thus, a value greater than one implies per capita investment in a given group of goods or services is higher than the national average that for a given municipality.

With the provision and expenditure indicators for each group of variables at the municipal level (1,101 municipalities), the IG&P is calculated from the ratio of the two results: each of the indices is arithmetically averaged to obtain a single municipal indicator and a subsequent departmental indicator (32). The results show that higher values reflect a better execution of the earmarked investment. Thus, an indicator of 0.9 means that, with the same money per person, that department only achieved 90% of the expenditure execution it should have achieved given its investment, expressed as a proportion of the average. The fact that a department has an indicator lower than unity would imply the existence of greater spaces for inefficiency, ineffectiveness, and corruption risk factors. Similarly, departments with higher values of the indicator, e.g., 1.2, mean that the department achieved an additional 20% of service provision results relative to its investment relative to the national average. Likewise, it would face the lowest corruption risks compared to the rest.

After each of the IG&P ratings are obtained, they are classified into different levels of corruption risk:

Table 6.1. IG&P Scale-Corruption risk Classification:

Levels		Range
Very high		0.36-0.41
High		0.42-0.67
Medium		0.68 -0.86
Moderate		0.87-1.08
Low		1.09-1.31
Very low		1.32-2.04

Source: Authors' own elaboration (2020), based on Golden & Picci (2005).

However, emphasis is placed on the existence of two corruption risk indexes used. Consequently, this study uses them alternatively as endogenous variables to contribute to more robust results.

6.2.2 Proposed Endogenous Variables

6.2.2.1 Transparency Index of Public Entities (ITEP)

The first indicator to be used is the Transparency Index of Public Entities (ITEP), a civil society initiative to contribute to the prevention of corruption in the administrative management of the State and measured by the Organización Transparencia por Colombia (Transparency for Colombia Organization in English), TI's national chapter. The Municipal Transparency Index, ITM 2015-April 2016, evaluated 28 municipal capitals, except for Bogota, Cali, and Medellín[16].

The ITEP evaluates three vital characteristics in public administration to control corruption risks:

- *Visibility*: The ability of an entity to make public its policies, procedures and decisions in a sufficient, timely, clear, and appropriate manner.
- *Institutionality*: The ability of an entity to ensure that public servants and the administration as a whole comply with norms and standards established for management processes.

Control and sanction: The capacity to generate control and sanction actions through internal processes, through the action of control bodies and spaces for citizen participation.

These measurement factors group sixteen indicators, which in turn are composed of sub-indicators and variables focused on key processes for institutional management. The factors of visibility, control, and sanction each have a weight of 30% on the final score of the index, while the institutional factor weighs 40%. The indicators for each factor are also weighted differently.

16 These capitals have a special methodology adapted to their institutional conditions because they are big cities.

For the calculation of the index, each measurement unit has a specific score ranging from zero (0) to one hundred (100). One hundred (100) is the highest possible score.

$$\mathbf{ITEP} = \left(Vis * \left(\frac{30}{100}\right)\right) + \left(Inst * \left(\frac{40}{100}\right)\right) + \left(Con\ y\ Sanc * \left(\frac{30}{100}\right)\right) \textbf{(6.2)}$$

After obtaining each of the ratings, they are classified into different levels of corruption risk:

Table 6.2. ITEP-Classification of corruption risks:

Levels	Range	
Low	89.5	100
Moderate	74.5	89.4
Medium	60	74.4
High	44.5	59.9
Very high	0	44.4

Source: Transparency Index of Public Entities 2015-2016. Corporación Transparencia por Colombia, 2017

6.1.2.2 Open Government Index (IGA in Spanish)

The second indicator corresponds to the open government index[17] (IGA in Spanish), calculated by the Attorney General's Office (PGN in Spanish). It is a composite indicator that determines the level of information reporting and the state of progress in implementing some standards that seek to promote the strengthening of territorial public management. In other words, it measures the level of compliance with reports and some standards considered strategic to prevent corruption and/or inefficiencies in public management by grouping 24 indicators into eight categories and three dimensions, which makes it possible to obtain simplified information on some of the activities carried out by the entities concerning their management and results (see Table 6.3).

17 According to the Organization for Economic Co-operation and Development (OECD), an open government is one that has four main characteristics: transparency and accessibility, participation, accountability and open public data.

Table 6.3. OGI dimensions

1. INFORMATION ORGANIZATION - OI	1.1. Internal Control - IC	1. Standard Model of Internal Control – (MECI in Spanish)
		2. Internal Accounting Control
	1.2. Document management - DM	3. Archives Law
2. INFORMATION EXPOSURE EI	2.1. Visibility of Contracting VC	4. Contract Publication
		5. Annual Acquisition Plan
	2.2. Territorial Core Competencies - (CBT in Spanish)	6. Single Information System - (SUI in Spanish)
		7. Beneficiary System Social Programs - (SISBEN in Spanish)
		8. Integrated Enrollment System
		9. Hospital Information System – (SIHO in Spanish)
	2.3. Administrative and Financial Management Systems - (SGAF in Spanish)	10. Single Territorial Form - (FUT in Spanish)
		11. Royalties
		12. Budget Execution System- (SICEP in Spanish)
		13. Public Employment Information and Management System - (SIGEP in Spanish)
		14. Asset Management and Information System - (SIGA in Spanish)
3. INFORMATION DIALOGUE DI	3.1. E-government E-GOV	15. GEL Open Government
		16. Gel Services
		17. SICEP Open Data
		18. SICEP Publicity
		19. Single System of Information on Procedures - (SUIT in Spanish)
	3.2. Transparency and Accountability (TyRC in Spanish)	20. SICEP Anticorruption
		21. SICEP Risk Map
		22. SICEP Control and Monitoring
		23. SICEP Accountability
	3.3. Citizen Service - (AC in Spanish)	24. SICEP Citizen Service

Source: Authors' own elaboration (2020), based on PGN 2017.

The three dimensions give rise to a methodology that should be observed in any area of public administration and responds to a tool for articulating and implementing the practices and techniques of good governance:

• *Information Organization (OI)*: is composed of the indicators of the internal control category (CI) and the indicator of the document management category (GD).

$$OI = \left(CI * \left(\frac{12}{20}\right)\right) + \left(GD * \left(\frac{8}{20}\right)\right) \quad \textbf{(6.3)}$$

• *Information Exposure (EI)*: is composed of indicators from the visibility of contracting(VC), the territorial core competencies (CBT), and the administrative and financial management systems (SGAF) categories.

$$EI = \left(VC * \left(\frac{24}{50}\right)\right) + \left(CBT * \left(\frac{6}{50}\right)\right) + \left(SGAF * \left(\frac{20}{50}\right)\right) \quad \textbf{(6.4)}$$

• *Information Dialogue (DI)*: is composed of indicators from the e-government (GE), the transparency and accountability (TyRC), and the citizen service (AC) categories.

$$DI = \left(GE * \left(\frac{16}{30}\right)\right) + \left(TyRC * \left(\frac{8}{30}\right)\right) + \left(AC * \left(\frac{6}{30}\right)\right) \quad \textbf{(6.5)}$$

Finally, the open government index is calculated as follows:

$$\mathbf{IGA} = \left(OI * \left(\frac{20}{100}\right)\right) + \left(EI * \left(\frac{50}{100}\right)\right) + \left(DI * \left(\frac{30}{100}\right)\right) \quad \textbf{(6.6)}$$

Likewise, the correct interpretation of this index is as follows: when the IGA reaches higher values, there is greater compliance with anti-corruption regulations, which implies lower risks of corrupt practices in local administrations.

Table 6.4. IGA-Corruption Risk Classification:

Levels	Range	
Low	89,5	100
Moderate	74,5	89,4
Medium	60	74,4
High	44,5	59,9
Very high	0	44,4

Source: Open Government Index 2016-2017. PGN Republic of Colombia.

Like the IG&P, the IGA and ITEP are only proxy variables for the possible corruption risks faced by a department, but they undoubtedly also provide precious information in the fight against this scourge.

6.2.3 Proposed Exogenous Variables

The variables previously considered by the literature as potential determinants of corruption and the determining factors in the decision to engage in a corrupt act already described in Chapter 2: Theoretical Framework are considered in the empirical exercise in Chapter 5, together with others proposed by this thesis (in blue and highlighted). Table 6.5 specifies the analyzed endogenous variables -IG&P, ITEP, and IGA- and exogenous variables -determinants-. For the exogenous variables, based on Castañeda (2016), a classification proposal is presented according to the state of the art for the expected signs of the respective statistical relationships and the respective sources of the official data, together with an additional brief theoretical justification for each variable as a determining factor in the behavior of individuals in the face of corruption scenarios.

6.2.3.1 Socioeconomic Variables

6.2.3.1.1 Education

According to Brunetti and Weder (2003) and Van Rijckeghem and Weder (1997), there is a negative relationship between education and corruption[18]. One might conjecture that a society with higher average levels of schooling also exhibits fewer opportunities for corruption to flourish. Therefore, education -as the primary sphere of social interaction- favors the future political participation of individuals, either as politicians or overseers (Eicher, et al., 2009; Glaeser and Saks, 2006) and generates capacities to anticipate the implications of corruption (Galston, 2001; Delli Carpini and Keeter, 1996).

The reason for the success of the corrupt act lies in not being discovered or punished. This is based on the anti-corruption

18 Although few and dissimilar, other results coincide in determining a direct relationship between the education and corruption variables (Frechette, 2001).

measures effectively applied by the State and on social control. The measures used are the average years of schooling received over the course of life by people over 24 years of age and access to learning and knowledge.

6.2.3.1.2 GDP per capita

According to Besley and Persson (2009), adequate material conditions theoretically encourage society to report acts of corruption. In addition, GDP per capita partially captures the degree of institutional development and, therefore, the State's capacities to limit the incidence of acts of corruption[19]. In summary, there is a negative association between GDP per capita and corruption (Kunicova and Rose-Ackerman, 2005; Persson et al., 2003; and Brunetti and Weder, 2003). However, Braun and Di Tella (2004) and Frechette (2001) find a positive association.

One of the most common ways used in the literature to measure a society's economic development level is through the gross domestic product (GDP) per capita. Although, there are more sophisticated options, such as the Human Development Index (HDI), formulated by the United Nations Development Program (UNDP), which considers three dimensions: health, education, and standard of living. It closely correlates with GDP per capita and the education variable already explained; nevertheless, it gathers more information on socioeconomic development than the set of previously specified variables; thus justifying its use for this research.

6.2.3.1.3 HDI

According to the UNDP, HDI is a synthetic measure used to assess long-term progress in three basic dimensions of human development: a long and healthy life, access to knowledge and a decent standard of living. Life expectancy is taken as the indicator to measure how long and healthy life is. The level of knowledge is measured through the average years of schooling among the adult population, i.e., the average years of schooling received over a lifetime by people aged 25 and older. Moreover, access to

19 According to Castañeda (2016), as in most cases, the order of causality is subject to debate. It can be argued that a lower level of corruption encourages investment and makes available a higher amount of resources to finance public policies that support economic growth and a better income distribution.

learning and knowledge through the expected years of schooling of children of school starting age, i.e., the total number of years of schooling a child of that age can expect to receive if current patterns of enrollment rates by age are maintained throughout the child's life. The standard of living is measured through gross national income (GNI) per capita, expressed in 2017 international dollars, revalued at purchasing power parity (PPP) conversion rates.

6.2.3.1.4 Unemployment

According to Rehman and Naveed (2007), there is a positive relationship between the level of unemployment and corruption. Passive agents who are responsible for corruption through their decisions also participate in illicit acts. The passive agents (officials and clients), those who observe the illegal actions without being the beneficiaries, are faced with two alternatives: to be indifferent and ask for some reward in exchange for their silence, if possible, or to report them. As previously explained in the theoretical framework, these individuals' decisions are influenced by their financial situation. For example, in an economic context of high unemployment, it is likely that the witness of a corrupt act will prefer to remain silent because the expected cost of reporting it would be high, given the risk of having to leave his or her position due to the pressure that could be exerted by those harmed by the report, especially if they are senior officials or managers.

6.2.3.1.5 Natural Returns and Mining GDP

According to Castañeda (2012, 2016), the abundance of natural resources creates opportunities for rent-seekers – at least in contexts of weak institutions – and favors corruption due to less political control by citizens, consistent with the findings of Leite and Weidmann (1999). According to Persson (2008), a country's mineral wealth is associated with an environment in which a corrupt person is less likely to be discovered and effectively punished, as evidenced by Gamarra (2006) for the Colombian case. Following Avila and Oliveira (2023) in Chapter 1, which is dedicated to the analysis of the nature of corruption, this environment is a contextual condition that influences the decision to engage in a corrupt act. To measure this condition, the mining GDP is proposed, which groups to a large extent the economic impact of additional resources from extraction, commercialization, and compensation royalties in the different territorial entities.

6.2.3.1.6 Unsatisfied Basic Needs (UBN)

According to DANE, the unsatisfied basic needs index (UBN) is a measure of poverty, given that, with the help of some simple indicators in its measurement -inadequate housing, critically overcrowded housing, housing with inadequate services, housing with high economic dependency and housing with school-age children not attending school-, it determines whether the basic needs of the population are covered. Those groups not reaching a pre-set minimum threshold are classified as poor. Therefore, given that the poorest departments are the most dependent on national government resources, the risk of corruption will be positively associated with greater unsatisfied basic needs in the population.

6.2.3.2 Political and Institutional Variables

6.2.3.2.1 Opposition

Regarding the functioning of the political system, in his analysis of partisan competition, Castañeda (2016) includes, in a novel way, the existence of opposition to the government. Opposition to the government in power can serve as an additional control factor, limiting the possible scenarios for corruption and to a greater extent, when there are few political parties. However, the greater variety of political groups will not be equivalent to a greater counterweight to the power of the current government because evidence suggests that even political parties in opposition, numerous or not, do not pursue the welfare of society but rather private profit (Stigler, 1971). Therefore, a negative relationship is expected between the ratio of votes obtained by the opposition in local government elections[20] and the degree of corruption risk (IG&P).

6.2.3.2.2 Electoral Districts

According to Persson et al. (2003) and Castañeda (2016), the size of electoral districts and the method of converting votes into public office positions affect the degree of accountability[21] -representativeness, commitment, or responsibility- between

20 Consequently, this factor is captured by taking the proportion of votes obtained by the opposition in the 2015 gubernatorial election.

21 The optimal way to work in an organization.

elected and electors, which helps explain differences in the incidence of corruption between countries at similar levels of development. That is, the larger the size of the electoral district, the lower the risk of corruption, so a small electoral district will tend to have a higher probability of corruption, similar to what occurs with closed-list electoral systems. However, Alfano et al. (2012) suggest taking this conclusion with reservation.

Even though the literature studied provides a solid basis to select the indicators for the variables described and implicit in the agent-principal-client model, it is not possible to establish with certainty that their coefficients will be statistically significant since the possible corrupt individuals are not only officials elected by popular vote. To better characterize corruption in the political context, the indices of political stability and absence of violence (psav), rule of law (rl), and voice and accountability (v&a), which could be taken from World Bank (WB) government statistics, should be incorporated into the national analysis. However, it should be remembered that the research objective is administrative corruption, rather than political corruption.

6.2.3.2 Demographic Variables

6.2.3.2.1 Population Density

According to Alt and Lassen (2003) and Knack and Azfar (2003), a higher population density generates opportunities for association among the corrupt for the complicit diversion of public resources, which would suggest a positive and statistically significant relationship between the two variables[22].

6.2.3.2.2 Rural Population

Finally, the ratio of rural people is expected to be positively related to corruption. On the contrary, urban population, generally more schooled and better informed through the media and because of their greater access to social networks, is expected to be more critical about the actions of public officials (Elbahnasawy and Revier, 2012).

22 However, other research emphasizes the lower control and management costs that economies of scale would entail in the public sector, which, in the end, would imply a greater probability that acts of corruption would be uncovered and that the statistical relationship between corruption and population density would be negative.

Table 6.5. Description of Potential Endogenous and Exogenous Variables

Category	Variable		Description	Fountain
Endogenous	IG&P		Índice Golden E. Picci (0-100)	Cálculos propios con base en DNP, MEN, MinSalud, DANE, Gobernaciones y alcaldias municipales.
	ITEP		Índice de Transparencia de las Entidades Públicas — (0-100)	Organización Transparencia por Colombia que es el capítulo local de Transparency International
	IGA		Índice de Gobierno Abierto (0-100)	Procuraduría General de la Nación (PGN)
Exogenous	Socio-económicas	IDH (-)	Vida larga y saludable (esperanza de vida) Acceso al conocimiento (Educación - promedio de años de escolarización y acceso al aprendizaje) Vida digna (Ingreso Nacional Bruto -INB per cápita).	PNUD
		Desempleo (+)	Desempleo (% fuerza laboral)	DANE
		PIB minero (+)	PIB por la explotación de recursos naturales	DANE
		NBI (+)	Pobreza	DANE
	Politicas e Institucionales	Oposición (-)	Participación en las votaciones legislativas de todos los partidos de oposición.	Cálculos propios con base en RGN, Gobernaciones y alcaldias municipales.
		Distrito_ elec (-)	Tamaño medio de un distrito electoral.	Cálculos propios con base en RGN, Gobernaciones y alcaldias municipales.
	Demo-gráficas	Dens_ pob(?)	Densidad poblacional en miles.	DANE
		Pob_rural (+)	Población rural como proporción de la población total.	DANE

Source: Adapted from Castañeda (2016), based on the literature reviewed and cited.
Note: When it is necessary to report the range in which a variable takes values, a parenthesis is inserted on the right side of the name with the respective limits.

It is preferred to perform some regressions that include all the variables for which observatio ns are available and then gradually exclude those that are not statistically significant, at least at 90% confidence (cleaned models*), and reduce the probability

of eventual multicollinearity problems, defining a parsimonious set of determinants.

$$Corrupción_{it} = \beta_0 + \beta_1 IDH\ \beta_2 Desempleo + \beta_3 PIB\ Min + \beta_4 NBI_{it} + \beta 5\ Opos\ gob_{it} + \beta_6 Distribución\ elec_{it} + \beta 7\ Dens\ pob_{it} + \beta_{10} Pob\ rural_{it} + \mu_{it}$$

In the equation, the subscripts *(it*) refer to department *(i)* in year *(t*), and is the respective error term.

6.3 Results and Discussion

6.3.1 Departmental IG&P Results

Before presenting the results obtained for the 32 departments and the Capital District from measuring the GI&P, which, as explained in the methodology, is the ratio between the provision of goods and services (education, health, and basic sanitation) and their accumulated investment and that are also the pillars to generate socioeconomic development in any territorial entity. It is worth noting that the procedure initially required 58,353 basic data, summarized in Table 6.6. Likewise, to deflate, normalize, and standardize them -as also foreseen by the methodology- the final data used were significantly multiplied to calculate the departmental IG&P in 2018.

Table 6.6. Amount of Data Initially used to Calculate IG&P, 2018

Net average coverage	1.101	
Number of public establishments	1.101	
Health coverage	1.101	
Number of beds	1.101	
Infant mortality	1.101	
Water supply coverage	1.101	
Sewerage coverage	1.101	
	7.707	
		15 years

Education	1.101	16.515
Health	1.101	16.515
Basic sanitation	1.101	16.515
		49.545
Population	1.101	
	1.101	
TOTAL of initial basic data:		58.353

Source: Authors' own elaboration.

Therefore, it was considered pertinent to present the intermediate results of each numerator and denominator required for the final calculation of the Golden & Picci corruption risk index (see Table 6.7). An exception must also be made: when reviewing the international literature on the subject and contrasting it with the original proposal by Miriam Golden and Lucio Picci (2005), a severe error in its calculation is noticed, given that, for the calculation of the index of provision of goods and services, several measurements and scientific publications do not take into account the percentage share of investment in each sub-index. In short, they replicated the exercise wrongly.

In turn, 33 territorial entities examined, seven are at very low risk of corruption, eleven are at low risk of corruption, nine are at moderate risk of corruption, three are at medium risk of corruption, two are at high risk of corruption, and one is at very high risk of corruption. Moreover, a low corruption risk index for the seven territorial entities in 2018 means that Bogota D. C. and the departments of Cundinamarca, Santander, Valle del Cauca, Atlántico, Antioquia, and Meta achieved more than 30% additional results in the provision of goods and services, in relation to their investment relative to the national average. While a high and very high corruption risk index for the three departments of Guaviare, Guainía, and Vichada means that, with the same money per person, each department achieved only 67%, 50%, and 39% of the expenditure execution they should have achieved given its investment, expressed as a ratio of the mean. Therefore, by having an indicator lower than unity, these three departments face greater corruption risks, given that the results imply the existence of greater spaces for inefficiency, ineffectiveness, and increasing corruption risk factors. See Table 6.7.

Table 6.7. Golden & Picci Index for Colombia 2018

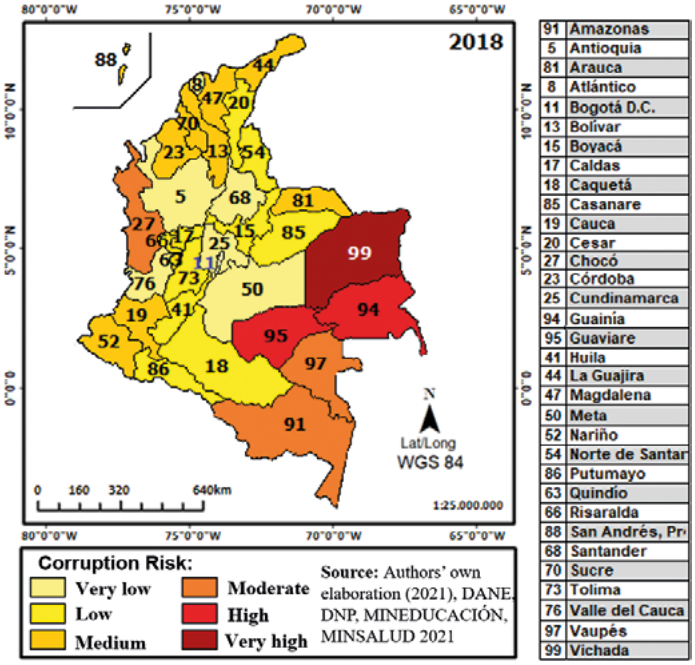
COLOMBIAN DEPARTMENTS		PROVISION OF GOODS AND SERVICES							CUMULATIVE INVESTMENT Constant Prices 2018		G&P Index
		Ed	%	H	%	BS	%	I 1	2004-2018	I 2	
1	Bogotá D.C.	0,69	71,67	1,14	23,69	1,89	4,64	0,86	$ 24.328.359.680.623	0,42	2,04
2	Cundinamarca	0,97	69,65	0,77	22,76	0,87	7,59	0,92	$ 13.780.118.749.793	0,60	1,52
3	Santander	1,12	70,01	0,96	23,70	0,79	6,30	1,06	$ 12.587.807.620.550	0,74	1,44
4	Valle del Cauca	0,78	66,99	0,91	28,16	1,27	4,85	0,84	$ 20.810.871.025.509	0,59	1,42
5	Atlántico	0,74	63,77	1,07	30,16	1,16	6,07	0,86	$ 12.317.819.986.698	0,62	1,39
6	Antioquia	0,79	67,15	0,98	26,88	0,91	5,97	0,85	$ 31.595.292.579.302	0,63	1,35
7	Meta	0,95	64,95	0,90	28,18	0,93	6,87	0,94	$ 5.726.558.950.237	0,70	1,33
8	Caldas	0,97	69,12	1,00	25,71	1,03	5,17	0,98	$ 5.868.381.284.554	0,75	1,31
9	Cesar	0,93	64,06	1,25	30,48	1,26	5,46	1,05	$ 7.893.069.009.970	0,84	1,25
10	Risaralda	0,84	70,51	0,89	24,84	1,00	4,65	0,86	$ 5.134.285.695.281	0,69	1,24
11	Quindío	0,90	69,01	0,88	26,40	1,51	4,59	0,93	$ 3.273.492.654.204	0,77	1,20
12	Norte de Santander	1,02	65,59	0,92	28,66	0,61	5,75	0,97	$ 9.629.156.455.058	0,82	1,17
13	Huila	1,09	64,65	1,05	29,32	1,17	6,03	1,08	$ 7.978.971.335.384	0,93	1,17
14	Tolima	1,12	68,26	0,86	25,90	0,97	5,84	1,04	$ 9.471.855.346.207	0,91	1,15
15	Boyacá	1,33	68,71	0,78	23,00	0,71	8,29	1,15	$ 9.824.076.010.189	1,03	1,12
16	Caquetá	1,52	61,39	1,13	32,08	1,01	6,53	1,36	$ 3.880.151.162.355	1,23	1,10
17	Casanare	1,06	65,85	0,86	27,67	0,91	6,48	0,99	$ 2.987.021.314.356	0,91	1,09
18	Putumayo	1,42	66,28	1,02	27,44	0,91	6,28	1,28	$ 3.205.524.743.765	1,18	1,09
19	La Guajira	0,96	64,18	1,05	29,48	1,46	6,34	1,02	$ 6.760.126.627.342	0,98	1,04
20	Cauca	1,04	67,01	0,89	26,84	0,55	6,14	0,97	$ 10.853.572.092.329	0,95	1,03
21	Arauca	1,05	64,69	1,20	29,31	1,14	5,99	1,10	$ 2.211.713.347.966	1,08	1,02
22	Sucre	0,92	64,89	1,40	29,51	1,04	5,60	1,07	$ 7.474.450.103.725	1,05	1,01
23	Córdoba	0,91	65,23	1,04	28,60	0,82	6,17	0,94	$ 13.430.808.880.792	0,96	0,98
24	San Andrés, Providencia y Santa Catalina (Archipiélago)	0,94	65,43	0,82	25,47	0,83	9,10	0,90	$ 444.280.356.260	0,93	0,97
25	Nariño	1,02	64,54	0,83	28,57	0,83	6,89	0,95	$ 12.651.696.552.600	0,99	0,96
26	Magdalena	0,81	66,87	1,11	27,52	0,77	5,61	0,89	$ 9.921.954.063.492	0,94	0,94
27	Bolívar	0,72	64,04	0,95	29,38	0,57	6,57	0,78	$ 14.328.553.316.019	0,88	0,88
28	Chocó	1,08	67,47	0,83	24,65	1,10	7,87	1,02	$ 5.079.674.576.148	1,21	0.84
29	Vaupés	1,17	60,98	1,46	29,00	1,74	10,02	1,31	$ 534.775.907.357	1,67	0,78
30	Amazonas	1,08	58,41	1,47	34,63	1,32	6,96	1,23	$ 1.002.687.695.525	1,67	0,74
31	Guaviare	1,27	58,15	0,94	35,42	1,14	6,43	1,15	$ 1.104.378.256.645	1,70	0,67
32	Guainía	0,86	56,60	1,05	34,93	0,27	8,46	0,88	$ 665.453.962.479	1,77	0,50
33	Vichada	0,93	36,80	0,60	59,73	0,52	3,47	0,72	$ 1.557.345.568.279	1,84	0,39
National average											1,095

Source: Authors' own elaboration (2021), based on data from DANE (2021), National Planning Department (DNP in Spanish) (2021), the Ministry of Education (Mineducación in Spanish) (2021), and the Ministry of Health (Minsalud in Spanish) (2021)

The territorial entity with the lowest risk of corruption in Colombia is its capital, Bogota. This could be attributed, to a large extent, to the variety of agencies and control entities that exercise their functions in the Capital District, to the antiquity of its institutions and even to its history and tradition as a city; additionally, to the fact that it houses 80% of the higher education institutions and that, due to its importance as the seat of central power, it is under continuous public scrutiny, more than any other territorial entity.

On the other hand, it is evident that the five territorial entities among those with the highest risk of corruption (medium, high, and very high) are part of the group of new departments created in the 1991 Political Constitution of Colombia (see Figure 6.1) and all belong to the Amazon region, characterized by the lowest levels of development, given that these departments have an HDI below the national average, including the lowest in the country, particularly for the departments of Guainía and Vaupés (see Table 6.8). Finally, the result obtained of a low corruption risk index for Colombia in 2018 should be highlighted, reflecting the relatively satisfactory scope of some of the modifications and correctives put in place in the fight for public efficiency and transparency.

Figure 6.1. Golden & Picci 2018 Corruption Risk Index



Source: Authors' own elaboration (2021), DANE, DNP, MINEDUCACIÓN, MINSALUD 2021

Table 6.8. 2018 HDI in the Colombian Departments

	2018	
Department	IDH	Population
Bogotá	0,806	8 181 047
Valle del Cauca	0,785	4 755 760
San Andrés	0,781	78 413
Atlántico	0,780	2 546 138
Quindío	0,778	574 960
Caldas	0,772	993 870
Santander	0,772	2 090 854
Meta	0,770	1 016 672
Cundinamarca	0,767	3 250 238
Antioquia	0,766	6 407 977
Risaralda	0,755	967 780
Boyacá	0,754	1 281 979
Bolívar	0,749	2 171 558
Guaviare	0,749	115 829
Vichada	0,748	77 276
Casanare	0,743	375 258
Norte de Santander	0,744	1 391 366
Tolima	0,741	1 419 957
Sucre	0,737	877 024
Arauca	0,735	270 708
Cesar	0,724	1 065 637
Magdalena	0,721	1 298 562
Huila	0,720	1 197 049
Nariño	0,716	1 809 301
Cauca	0,714	1 416 145
Putumayo	0,713	358 896
Amazonas	0,712	78 830
Caquetá	0,712	496 262
Córdoba	0,711	1 788 648
La Guajira	0,693	1 040 193
Chocó	0,691	515 166
Guainía	0,664	43 446
Vaupés	0,635	44 928
COLOMBIA	0,761	49834727

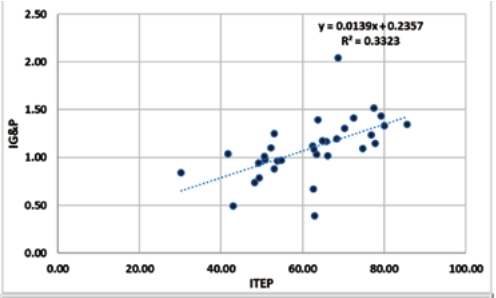
Source: UNDP 2019, IDH 2018.

6.2.2 Transparency Index of Public Entities (ITEP)

The first indicator to be used as an alternative endogenous variable of the IG&P is the transparency index of public entities (ITEP). As explained in the methodology, it is a civil society initiative seeking to contribute to the prevention of acts of corruption in the administrative management of the State. It is measured by Organización Transparencia por Colombia, TI's national chapter. The ITEP evaluates three vital characteristics in public administration to control corruption risks: visibility, institutionality, and control and sanction.

Constructed based on data from other sources, it is noteworthy that the IG&P measurements are closely related to the figures reported by the ITEP. As shown in Figure 6.2, there is a positive relationship between the two estimates, which supports the idea that the efficiency, effectiveness, integrity, and corruption risk of departmental administrations in Colombia belong together.

Figure 6.2. Relationship between IG&P and ITEP




```
Modelo 1: estimaciones MCO
utilizando las 33 observaciones 1-33
Variable dependiente: I_G_P

 VARIABLE      COEFICIENTE       DESV.TÍP.         ESTAD T   VALOR P
  const            0.235733        0.223193        1.056   0.29905
  ITEP             0.0139168       0.00354293      3.928   0.00045 ***

  Media de la var. dependiente = 1.09458
  Desviación típica de la var. dependiente. = 0.310197
  Suma de cuadrados de los residuos = 2.05586
  Desviación típica de los residuos = 0.257523
  R-cuadrado = 0.332321
  R-cuadrado corregido = 0.310783
  Grados de libertad = 31
  Log-verosimilitud = -1.02407
  Criterio de información de Akaike (AIC) = 6.04815
  Criterio de información Bayesiano de Schwarz (BIC) = 9.04116
  Criterio de Hannan-Quinn (HQC) = 7.05521
```

Source: Authors' own elaboration. ITEP Data 2017

As with the IG&P, it was considered relevant to present the intermediate results of each numerator and denominator required for the final ITEP calculation (see Table 6.9).

According to the ITEP 2017 report, the corruption risks haunting in public management in Colombia are not few. While important steps are taken regarding measures and actions for management visibility and ensuring access to information, the decisions and actions related to hiring, public employment, and the fight against corruption still do not have the expected results. 61.5/100 is the average rating for the governorates. That is, a mean ITEP for Colombia.

No departmental administration is at low corruption risk: seven are at moderate corruption risk, twelve are at medium corruption risk, ten are at high corruption risk, and three are at very high corruption risk (see Table 3.6).

Table 6.9. Transparency Index of Public Entities, 2017.

RANKING	GOBERNACIÓN	CATEGORÍA	REGIÓN	VISIBILIDAD	INSTITUCIONALIDAD	CONTROL Y SANCION	ITD	NIVEL DE RIESGO
	Ninguna	-	-	-	-	-	-	BAJO
1	Gobernación de Antioquia	Especial	Occidente	91.3	85.9	79.5	85.6	MODERADO
2	Gobernación de Meta	Segunda	Orinoquía	80.1	77.1	84.0	80.1	MODERADO
3	Gobernación de Santander	Segunda	Centro Oriente	86.3	72.9	80.6	79.2	MODERADO
4	Gobernación de Tolima	Tercera	Centro Oriente	84.8	70.1	81.2	77.8	MODERADO
5	Gobernación de Cundinamarca	Especial	Centro Oriente	85.5	77	70.1	77.5	MODERADO
6	Gobernación de Risaralda	Segunda	Occidente	81.4	72.4	78	76.8	MODERADO
7	Gobernación de Casanare	Tercera	Orinoquía	84.9	74.5	64.9	74.7	MODERADO
8	Gobernación del Valle del Cauca	Primera	Pacífica	84.4	67.8	66.8	72.5	MEDIO
9	Gobernación de Caldas	Segunda	Occidente	75.7	62.5	75.3	70.3	MEDIO
10	Gobernación del Quindío	Tercera	Occidente	77.2	64	65.3	68.3	MEDIO
11	Gobernación de Arauca	Cuarta	Orinoquía	82.3	60.7	57	66.1	MEDIO
12	Gobernación del Huila	Segunda	Centro Oriente	80.9	58.9	60.1	65.9	MEDIO
13	Gobernación de Norte de Santander	Segunda	Centro Oriente	75.2	58.6	63	64.9	MEDIO
14	Gobernación del Atlántico	Primera	Caribe	67.9	67.9	53.7	63.7	MEDIO
15	Gobernación del Cauca	Tercera	Pacífica	81.1	44.8	70.9	63.5	MEDIO
16	Gobernación del Vichada	Cuarta	Orinoquía	65.7	65.9	55.9	62.9	MEDIO
17	Gobernación del Putumayo	Cuarta	Amazonía	91.5	57.5	41.3	62.8	MEDIO
18	Gobernación del Guaviare	Cuarta	Orinoquía	79.9	56	54.2	62.6	MEDIO
19	Gobernación de Boyacá	Primera	Centro Oriente	53.4	73.4	56.7	62.4	MEDIO
20	Gobernación San Andrés	Tercera	Caribe	67.1	54.4	43	54.8	ALTO
21	Gobernación de Nariño	Primera	Pacífica	71.9	42	51.9	53.9	ALTO
22	Gobernación del Cesar	Tercera	Caribe	59.5	49.8	51.1	53.1	ALTO
23	Gobernación de Bolívar	Segunda	Caribe	54.3	54.9	49.3	53	ALTO
24	Gobernación de Caquetá	Cuarta	Amazonía	47.2	49.5	61	52.2	ALTO
25	Gobernación de Córdoba	Segunda	Caribe	60.2	45.9	47.8	50.8	ALTO
26	Gobernación de Sucre	Tercera	Caribe	63.4	42.7	48.2	50.6	ALTO
27	Gobernación del Vaupés	Cuarta	Orinoquía	64.1	38.4	49.3	49.4	ALTO
28	Gobernación del Magdalena	Tercera	Caribe	48.4	53.7	43.9	49.2	ALTO
29	Gobernación del Amazonas	Cuarta	Amazonía	64.2	40.2	43.2	48.3	ALTO
30	Gobernación del Guainía	Cuarta	Orinoquía	46.7	40.7	42.4	43	MUY ALTO
31	Gobernación de La Guajira	Cuarta	Caribe	66.4	32.3	29.5	41.7	MUY ALTO
32	Gobernación del Chocó	Cuarta	Pacífica	40.1	18.1	36.4	30.2	MUY ALTO
PROMEDIO NACIONAL GOBERNACIONES				70.72	57.20	57.98	61.49	MEDIO

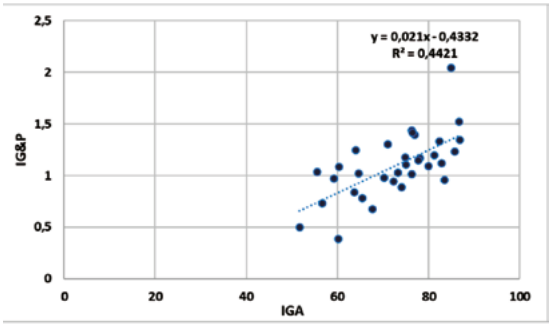
Source: Authors' own elaboration (2021), based on TI data (2017).

6.2.3 Open Government Index (IGA)

The second indicator to be used as an alternative endogenous variable of the IG&P is the IGA, which, to recall the methodological explanation, is a composite indicator that determines the level of information reporting and the state of progress in implementing some regulations aimed at promoting the strengthening of territorial public management. In other words, it measures the level of compliance with reports and some standards considered strategic to prevent corruption and/or inefficiencies in public management, grouping 24 indicators in eight categories and three dimensions, which makes it possible to generate simplified information on some of the activities carried out by the entities in relation to their management and results.

As is also the case with the ITEP, even with information from other sources, the IGA estimates are closely related to those of IG&P (see Figure 6.3). This positive relationship confirms the concomitance between the efficiency, effectiveness, integrity, and corruption risk of departmental administrations in Colombia.

Figure 6.3. Relationship of the IG&P to the IGA




```
Modelo 2: estimaciones MCO
utilizando las 33 observaciones 1-33
Variable dependiente: I_G_P

 VARIABLE     COEFICIENTE       DESV.TÍP.          ESTAD T   VALOR P
  const          -0.433202         0.310928       -1.393    0.17345
  IGA             0.0210381        0.00424424      4.957    0.00002 ***

  Media de la var. dependiente = 1.09458
  Desviación típica de la var. dependiente. = 0.310197
  Suma de cuadrados de los residuos = 1.71769
  Desviación típica de los residuos = 0.235392
  R-cuadrado = 0.44215
  R-cuadrado corregido = 0.424154
  Grados de libertad = 31
  Log-verosimilitud = 1.94126
  Criterio de información de Akaike (AIC) = 0.117488
  Criterio de información Bayesiano de Schwarz (BIC) = 3.1105
  Criterio de Hannan-Quinn (HQC) = 1.12455
```

Source: Authors' own elaboration based on IMF's data.

As was done with the IG&P and ITEP, it was considered relevant to present the intermediate results of each numerator and denominator required for the final calculation of the IGA (see Table 6.10).

No sectional public administration is at low corruption risk, but sixteen are at moderate corruption risk, twelve are at medium corruption risk -similar to the ITEP- and four are at high corruption risk. According to the IGA, there are no public entities located in very high corruption risk (see Table 6.10). The public entity with the lowest departmental corruption risk is Antioquia, a result that coincides with the ITEP's. In other words, the Antioquia department requires greater compliance with anti-corruption regulations, which implies lower risks of corrupt practices in departmental administrations. However, it should be noted that no government entity obtained a low or very high corruption risk.

Table 6.10. Open Government Index (IGA)

Ranking	GOVERNMENT	INTERNAL CONTROL	DOCUMENT MANAGEMENT	INFORMATION ORGANIZATION	VISIBILITY OF PROCUREMENT	BASIC TERRITORIAL COMPETENCIES	ADMINISTRATIVE MANAGEMENT SYSTEMS	INFORMATION EXPOSURE	TRANSPARENCY AND E-GOVERNMENT	TRANSPARENCY AND ACCOUNTABILITY	CITIZEN SERVICE	INFORMATION DIALOGUE	IGA	RISK LEVEL
None		-	-	-	-	-	-	-	-	-	-	-	-	LOW
1	ANTIOQUIA GOVERNMENT	91,40	77,69	85,91	86,23	100	95,06	91,42	91,61	60,60	75,0	80,02	86,9	MODERATE
2	CUNDINAMARCA GOVERNMENT	90,48	79,93	86,26	98,93	96,15	94,94	97,0	79,84	61,44	55,0	69,96	86,7	MODERATE
3	RISARALDA GOVERNMENT	87,38	78,77	83,93	98,94	100	75,48	89,68	89,68	65,57	75,0	80,31	85,7	MODERATE
4	NARIÑO GOVERNMENT	88,77	34,55	67,08	97,51	100	88,95	94,38	80,19	72,75	70,0	76,17	83,5	MODERATE
5	BOYACA GOVERNMENT	70,73	58,53	65,85	96,77	100	95,48	96,64	78,82	72,31	50,0	71,32	82,9	MODERATE
6	META GOVERNMENT	83,34	74,67	79,87	98,00	100	83,36	92,39	69,46	66,92	63,33	67,56	82,4	MODERATE
7	QUINDIO GOVERNMENT	85,00	84,71	84,88	86,92	100	75,55	83,94	85,11	64,20	60,0	74,51	81,3	MODERATE
8	CASANARE GOVERNMENT	82,46	35,95	63,86	98,52	100	86,44	93,87	77,22	69,21	40,0	67,64	80,0	MODERATE
9	HUILA GOVERNMENT	90,83	57,14	77,35	95,83	100	62,13	82,85	68,29	80,26	63,33	70,49	78,0	MODERATE
10	TOLIMA GOVERNMENT	75,97	57,59	68,62	97,85	100	71,88	87,72	61,78	75,20	71,67	67,34	77,8	MODERATE
11	ATLANTICO GOVERNMENT	84,20	50,02	70,53	65,02	100	95,39	81,37	76,62	80,88	56,67	73,76	76,9	MODERATE
12	VALLE DEL CAUCA GOVERNMENT	85,39	46,78	69,95	83,73	100	75,18	82,26	74,66	78,32	53,33	71,37	76,5	MODERATE
13	SUCRE GOVERNMENT	68,77	40,90	57,62	98,69	100	84,52	93,18	72,47	56,51	35,0	60,72	76,3	MODERATE
14	SANTANDER GOVERNMENT	86,87	67,61	79,17	56,65	98,80	81,04	71,47	95,44	66,25	68,33	82,24	76,2	MODERATE
15	CAQUETA GOVERNMENT	70,91	26,09	52,98	89,17	100	94,86	92,74	67,81	61,23	40,0	60,49	75,1	MODERATE
16	N. DE SANTANDER GOVERNMENT	79,80	34,43	61,65	85,98	100	75,96	83,65	75,16	65,83	56,67	68,97	74,8	MODERATE
17	BOLIVAR GOVERNMENT	78,06	58,73	70,33	84,54	96,35	80,72	84,43	59,93	64,96	50,00	59,28	74,1	MEDIUM
18	CAUCA GOVERNMENT	83,36	41,05	66,44	84,37	100	76,13	82,95	67,26	66,04	40,00	61,48	73,2	MEDIUM
19	MAGDALENA GOVERNMENT	78,46	46,38	65,63	99,04	97,14	65,42	85,37	52,52	66,67	46,67	55,12	72,3	MEDIUM
20	CALDAS GOVERNMENT	73,20	42,36	60,87	86,43	100	86,38	88,04	56,05	50,44	31,67	49,68	71,1	MEDIUM
21	CORDOBA GOVERNMENT	64,83	54,15	60,56	73,95	96,77	80,18	79,18	60,57	71,16	51,67	61,61	70,2	MEDIUM
22	GUAVIARE GOVERNMENT	68,88	0,0	41,33	93,37	100	71,13	85,27	63,64	48,57	46,67	56,23	67,8	MEDIUM
23	VAUPES GOVERNMENT	57,98	31,80	47,51	95,76	100	58,48	81,36	57,86	50,47	33,33	50,98	65,5	MEDIUM
24	ARAUCA GOVERNMENT	64,30	64,49	64,38	44,75	100	72,50	62,48	79,37	67,67	40,0	68,38	64,6	MEDIUM
25	CESAR GOVERNMENT	77,25	43,58	63,78	40,46	100	82,70	64,50	73,01	56,07	48,33	63,56	64,1	MEDIUM
26	CHOCO GOVERNMENT	52,34	31,65	44,06	99,54	83,33	60,96	82,16	45,79	60,69	28,33	46,27	63,8	MEDIUM
27	PUTUMAYO GOVERNMENT	68,89	15,60	47,57	99,62	100,00	60,90	84,17	53,52	1,56	0,0	28,96	60,3	MEDIUM
28	VICHADA GOVERNMENT	71,21	24,68	52,60	49,04	100	71,70	64,22	55,15	70,22	51,67	58,47	60,2	MEDIUM
29	SAN ANDRES GOVERNMENT	68,12	63,78	66,39	47,33	N/A	63,87	55,60	65,90	56,05	51,67	60,43	59,2	HIGH
30	AMAZONAS GOVERNMENT	70,80	40,83	58,81	88,18	100	52,80	75,45	44,48	1,56	0,0	24,14	56,7	HIGH
31	LA GUAJIRA GOVERNMENT	58,03	8,26	38,12	38,75	71,84	80,10	59,26	62,40	61,93	55,0	60,79	55,5	HIGH
32	GUAINIA GOVERNMENT	65,28	0,0	39,17	98,99	N/A	51,99	75,49	37,48	1,56	0,00	20,41	51,7	HIGH
NATIONAL AVERAGE OF GOVERNMENTS		75,73	46,02	63,85	83,09	98,01	76,63	82,02	68,10	59,16	47,14	61,52	72,2	MEDIUM

Source: Authors' own elaboration (2021), based on data from PGN 2017.

6.3 Results of the estimates of corruption indexes

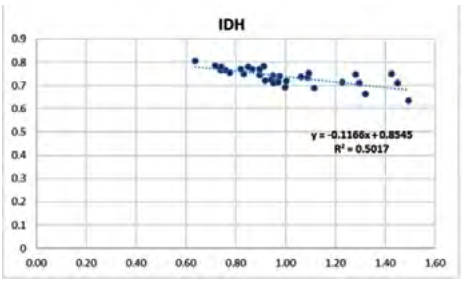
Initially, an exception must be made: the HDI and the UBN index are specialized measures composed of several dimensions and should, therefore, be treated methodologically with greater care. As already mentioned about the HDI, taking measures of education -years of schooling or coverage rates, among others- and/or GDP per capita would generate a high correlation, so initially a brief analysis of these two indexes is presented before including them in the respective models.

HDI

As also discussed earlier in the methodology, the HDI is a synthetic measure used to assess long-term progress in three basic dimensions of human development: a long and healthy life, access to knowledge and a decent standard of living. According to UNDP (2019), for 2018, Colombia is classified within the 54 "high human development countries", among a total of 189, which, on average, have an HDI of 0.750. This position, although it represents a good relative performance, reflects the existence of large gaps and challenges for the country in relation to dimensions such as life expectancy, years of schooling and income level.

Figure 6.4 shows a negative relationship between the IG&P estimates and the HDI, which confirms that the higher the risk of corruption in departmental administrations, the lower the levels of development of society.

Figure 6.4. Relationship between GI&P and HDI




```
Modelo 1: estimaciones MCO
utilizando las 33 observaciones 1-33
Variable dependiente: IDH

 VARIABLE      COEFICIENTE        DESV.TÍP.          ESTAD T    VALOR P
  const            0.854541        0.0213340         40.055   <0.00001 ***
  I_G_P           -0.116552        0.0208625         -5.587   <0.00001 ***

  Media de la var. dependiente = 0.738121
  Desviación típica de la var. dependiente. = 0.0366007
  Suma de cuadrados de los residuos = 0.0213611
  Desviación típica de los residuos = 0.0262501
  R-cuadrado = 0.501695
  R-cuadrado corregido = 0.48562
  Grados de libertad = 31
  Log-verosimilitud = 74.3294
  Criterio de información de Akaike (AIC) = -144.659
  Criterio de información Bayesiano de Schwarz (BIC) = -141.666
  Criterio de Hannan-Quinn (HQC) = -143.652
```

Source: Authors' own elaboration.

UBN

The geographic distribution of the indices reveals a relationship with the areas of greatest poverty in the country. Taking the UBN index as a measure, the poorest departments are those that at the same time face the greatest corruption problems. While Vaupés and Vichada have the highest values of the UBN index -68.94 and 67.76, respectively-, they also have the lowest values of the IG&P (0.78 and 0.39). On the other hand, Bogota has the lowest poverty problems and is the territorial entity with the highest GI&P value (2.04).

As shown in Table 6.11, the results of the estimations for the 33 Colombian territorial entities in the main model (1) have the expected signs and are statistically significant, as found in the economic literature. A major exception is the higher poverty rates (UBN) which, although statistically very significant, did not obtain the indicated sign. As mentioned in the methodology, the poverty variable should have a positive relationship with higher corruption risks, given that an increase in any of the five simple indicators– among which are inadequate housing, critically overcrowded housing, housing with inadequate public services, housing with high economic dependency and housing with school-age children not attending school–generates corruption risk.

However, HDI, GDP of natural mineral resources and population density are the most statistically significant variables with the expected signs required for this study. As already taught, the 2018 IG&P evidenced that there are eleven departments that, with the same money per person, reached less than 100% of the expenditure execution that each territorial entity should have reached with its investment, expressed as a proportion to the average. These eleven departments, by presenting an indicator below unity, face greater risks of corruption, given that the results imply the existence of greater spaces for inefficiency and ineffectiveness in the fulfillment of the State's duties. In the result of the HDI coefficient, model 1 shows that such efficiency losses in the fulfillment of the State's duties will be assumed by citizens with stagnation or regression in their development levels, which means that their lives will probably not be dignified or healthy, due to their low per capita income and restricted access to knowledge. See Table 6.11.

As shown in Avila and Oliveira (2023) through Chapter 4, with the economic base indicators for Colombian departments, and in Chapter 5, with the analysis of national exports and imports, there is a strong dependence of the productive activities of Colombian territorial entities on the mining sector, as well as a large weight of minerals in total exports -58.5%-. It was to be expected that the GDP of the mining sector would be the variable with the highest degree of association in explaining the risk of corruption.

Finally, model 1 coincides with the results of Alt and Lassen (2003) and Knack and Azfar (2003), obtaining the same positive and statistically significant relationship that shows how a higher population density generates opportunities for association among the corrupt to, in conspiracy, divert public resources.

Among the three corruption *proxy* variables, the IG&P regression presents the best fit of the regression line to the data set, with 86% of the corruption risk being explained by the variables in the main model (1). ITEP and IGA had 76 % and 62 %, respectively.

Table 6.11. Estimated Results of the Corruption Risk Models

	Modelo 1 - IG&P		Modelo 2 - ITEP		Modelo 3 - IGA	
VARIABLES	**COEFICIENTE**	**VALOR P**	**COEFICIENTE**	**VALOR P**	**COEFICIENTE**	**VALOR P**
Constante	2.97628	0.00127 ***	-10.6559	0.84845	66.9065	0.18682
IDH 2018	-2.09003	0.10568 ***	84.6921	0.23959	23.0124	0.73949
Desempleo	0.00150310	0.77905	1.04353	0.00043 ***	-0.0449705	0.89155
PIB Sector Minero	1.46893E-05	0.00050 ***	0.000565321	0.08171 *	0.000130941	0.76299 *
NBI	-0.00863219	0.00128 ***	-0.477459	0.00032 ***	-0.324217	0.02234 **
Oposición	-0.262198	0.47096	0.813570	0.94981	-14.1360	0.41693
Tam. Dist electoral	0.0867162	0.00876 ***	3.61135	0.00616 ***	3.39900	0.02181 **
Densidad Poblac.	0.000100567	0.00463 ***	-0.00340965	0.02308 **	-0.00173903	0.25165
Población Rural	-0.00454537	0.02718 **	0.215249	0.06127 *	0.0585707	0.59607
	R^2 = 0.86		R^2 = 0.76		R^2 = 0.61	
	R^2_{ajus} = 0.81		R^2_{ajus} = 0.68		R^2_{ajus} = 0.49	

Note: *significant at 10%, **significant at 5% and *** significant at 1%.
Source: Authors' own elaboration.

In summary, after analyzing the three corruption *proxy* variables based on the IG&P, the ITEP and the IGA, it was found that, for the first indicator, the variables associated with higher corruption risks are low human development indexes, high income caused by the extraction of natural mineral resources and opportunities for association among the corrupt to divert public resources. For the ITEP, unemployment is precisely the variable most associated with corruption risks, together with the rural nature of the population and the mining origin of income. Finally, with the IGA, low opposition to the government in power in each

of the territorial entities is the variable with the highest association to corruption risk, as is the income received in the mining sector.

In short, although each of these three proxy variables contains different variables with a higher degree of association to explain the risk of corruption, the only variable with significant and positively associated results in all regressions is the mining sector's GDP.

6.4 Conclusions and Recommendations

The 2018 IG&P showed that of the 33 territorial entities that make up Colombia, seven are at very low corruption risk, eleven at low corruption risk, nine at moderate corruption risk, three at medium corruption risk, two at high corruption risk, and one at very high corruption risk.

It was also evidenced that the five territorial entities with the highest risk of corruption (medium, high, and very high) belong to the group of new departments created in the Political Constitution of Colombia of 1991, and all are in the Amazon region. They characterized by the lowest levels of development, given that these departments have an HDI below the national average and even the lowest in the country, in the cases of the departments of Guainía and Vaupés.

Although the data for each indicator -IG&P, ITEP, and IGA- come from different sources of information, a positive relationship is corroborated between the consolidated estimates for each one. This confirms that the efficiency, effectiveness, integrity, and corruption risks of departmental administrations in Colombia are interrelated.

The analysis of the three corruption proxy variables, IG&P, ITEP, and IGA, showed that the variable with the highest degree of association in explaining corruption risk in all regressions is the mining GDP.

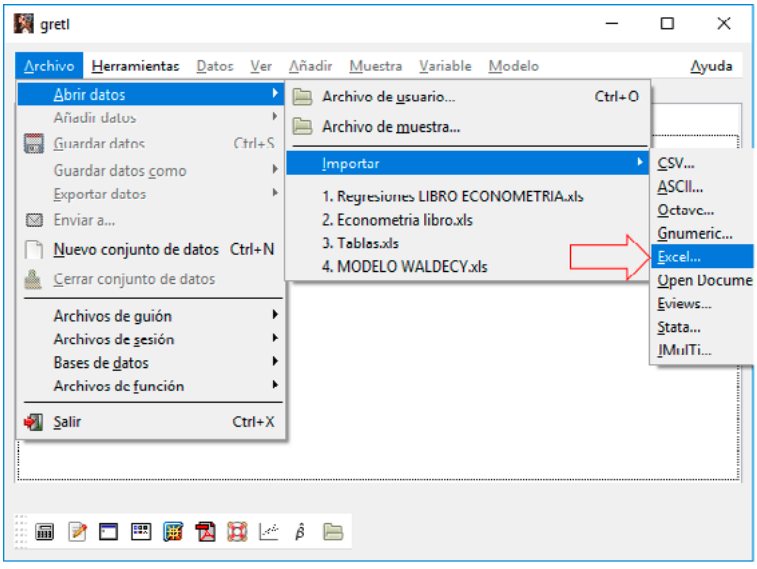
Finally, the regression of model 1 shows that the losses in efficiency in fulfilling the State's duties caused by the risk of corruption will fall on the citizens due to the lower levels of development they will achieve. The chitizens' lives will probably not be dignified or healthy, because of their low per capita income and their restricted access to knowledge.

APPENDIX 1

Gretl is a software package for econometric analysis used in several economics departments of universities worldwide. It is free software (free to distribute and modify under the terms provided at www.gnu.org/copyleft/gpl.html). It can be downloaded from the web at http://gretl.softonic.com/.

As in image 1, the file, which is generally in Excel format, must be imported and is compatible with specialized programs such as Eviews and Stata (econometric programs).

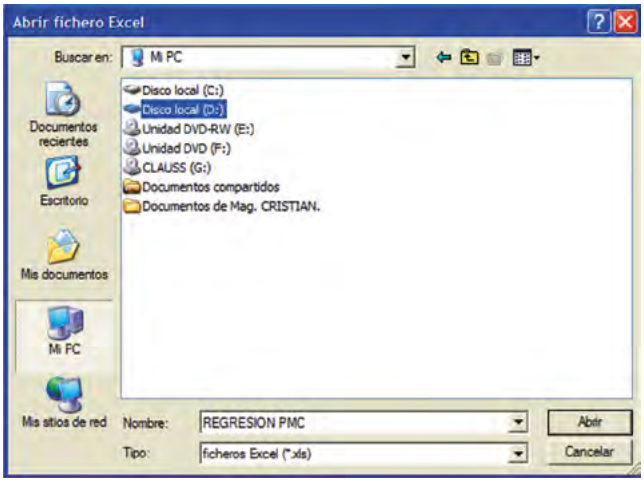
Image 1.



Immediately the window shown in Image 2 opens, where you must specify the location of the document in Excel, it is

recommended to leave the document in my documents or desktop, this way it will be easier to locate it.

Image 2.



Please note that at the moment of importing the database (which in our case is in an Excel file, since it can be in another type of document), the Excel file must be closed.

Note the Excel presentation of the data in Figure 3, which contains Colombian data on aggregate personal consumption expenditure (Y) and Gross Domestic Product (X). These must be in spreadsheet 1 and start in row 1 and column A, without spaces and separated by commas, not periods. This imports the data into the new window shown in image 4.

Image 3.

	A	B	C
1	**Año**	**Y**	**X**
2	**2005**	151.476	182.228
3	**2006**	161.161	205.836
4	**2007**	172.738	231.215
5	**2008**	179.775	257.229
6	**2009**	181.466	271.379
7	**2010**	190.805	293.773
8	**2011**	203.377	334.297
9	**2012**	214.144	360.131
10	**2013**	222.684	385.951
11	**2014**	232.983	412.602
12	**2015**	240.188	435.202
13	**2016**	243.992	467.160
14	**2017**	249.031	497.669
15	**2018**	257.779	529.191
16			

Image 4.

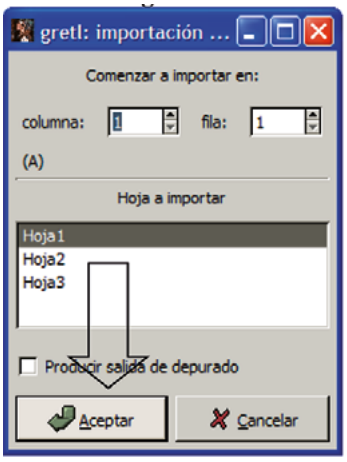
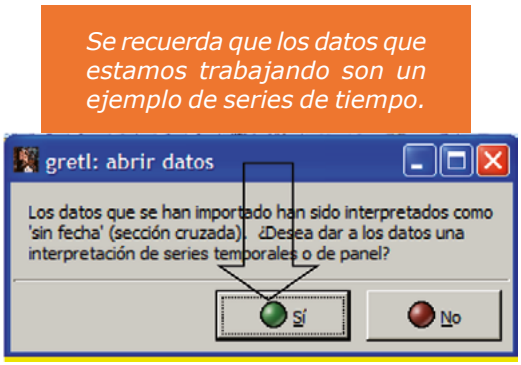


Image 5.



As the structure of the dataset is time series with annual periodicity, we click in the open window ⇒ Forward, as in figure 6 and the time series frequency window opens, we choose–Annual and click ⇒ Forward, as in figure 7.

Image 6.

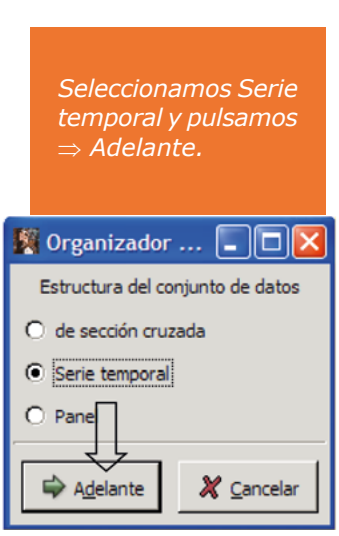
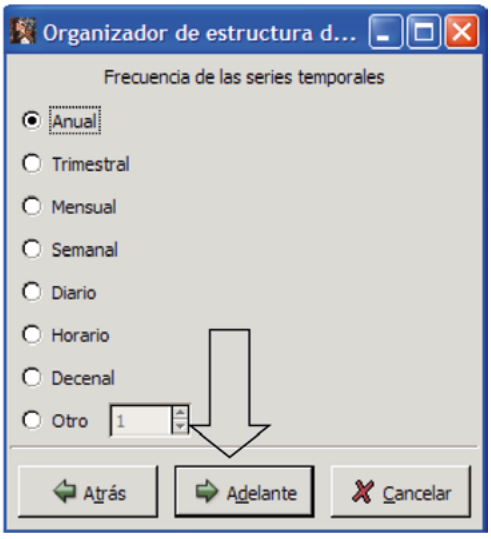


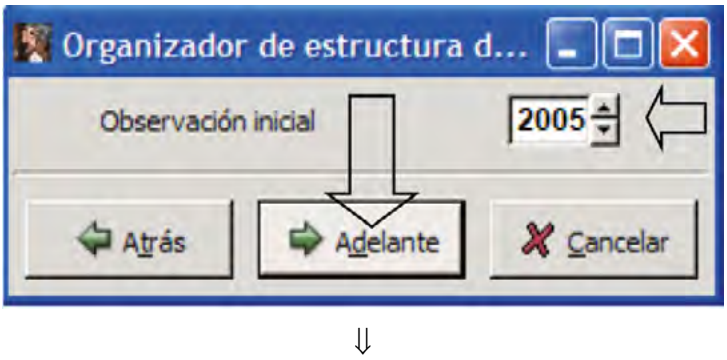
Image 7.



⇓

We define the start of the time series from 2005 to 2018.

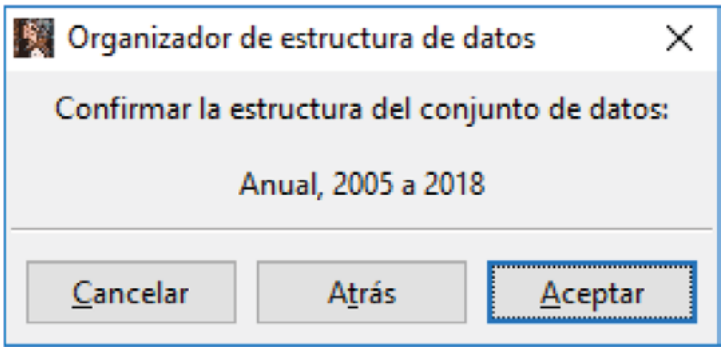
Image 8.



⇓

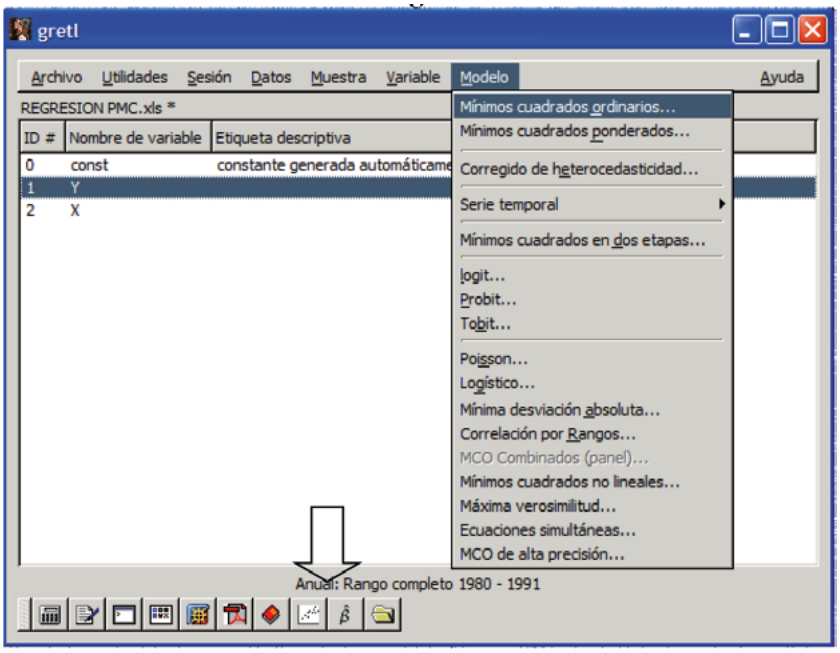
The structure of the data set is confirmed in the window of image 9.

Image 9.



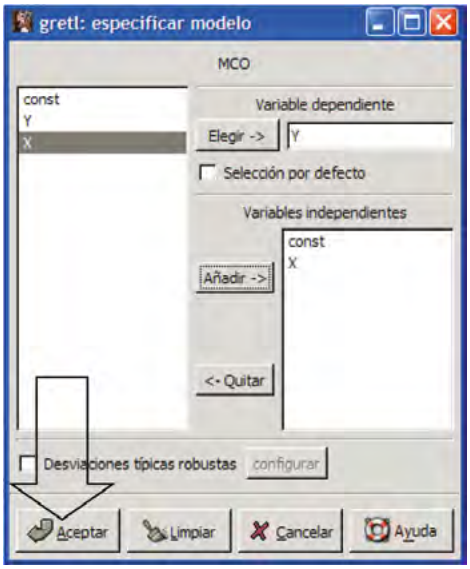
As it is a linear and non-equational model, the linear regression is performed by the ordinary least squares OLS method, which can be seen in image 10, this action can be performed through the "Modelo" (model in English) option, or go through the option $\hat{\beta}$.

Image 10.



Immediately the window of Image 11 opens, where we specify that the dependent variable is Y and the independent variables, X, (aggregate personal consumption expenditure and GDP, respectively, for this exercise). When loading the database, all the variables are displayed on the left side of the window, where we select the variable to *Elegir ->,* as dependent variable, procedure to follow with each of the variables to *Añadir ->,* as independent variables and because the model automatically generates the coefficient of the intercept or constant, given the possible case to estimate a model that does not have this coefficient, we underline*(const*) and select the option *Quitar ->,* finally click on *Aceptar*.

Image 11.



The window shown in Image 12 is the estimation of Model 1, where we obtain the estimated constant = $\hat{\beta}_1$ = -231.795, the slope = $\hat{\beta}_2$= PMC= 0.72. Note the R^2 = 0.99, excellent. However, these interpretations are discussed in Chapter 4.

Likewise, to graph the regression line obtained from the estimation (Figure 13), we choose the option "Gráficos", "Gráfico de variable estimada y observada," against GDP, in the window of Figure 12. (Graph made in Excel in Appendix 2).

Image 12.

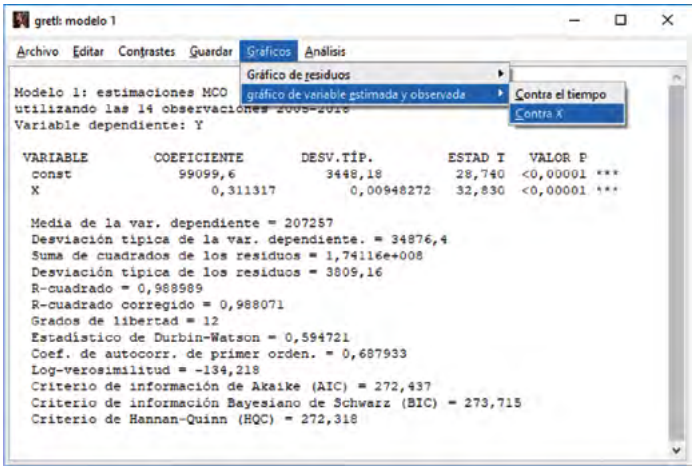
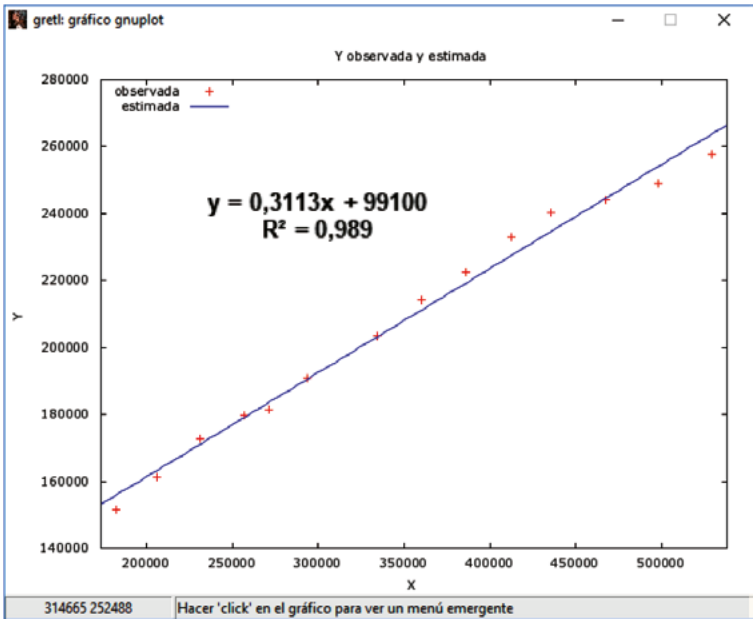


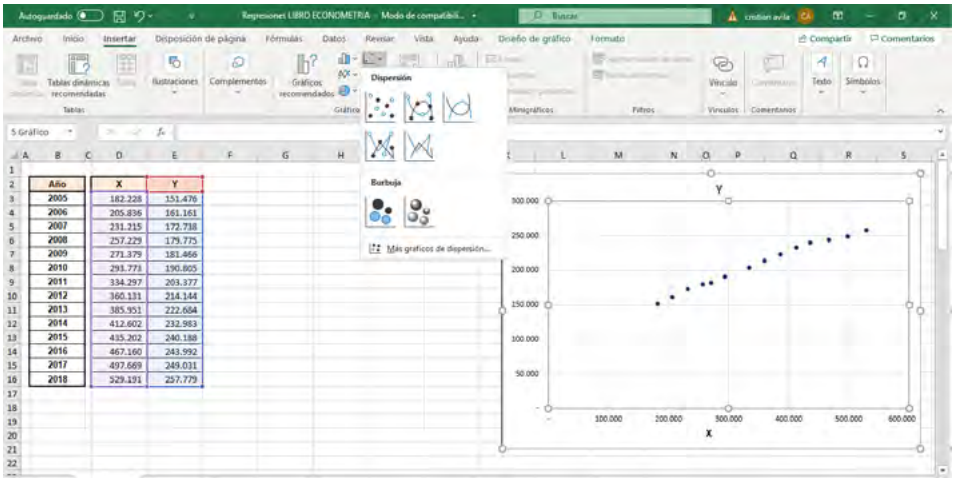
Image 13.


gretl: gráfico gnuplot
Y observada y estimada
observada
estimada
y = 0,3113x + 99100
R² = 0,989
280000
260000
240000
220000
200000
180000
160000
140000
Y
200000
250000
300000
350000
400000
450000
500000
X
314665 252488
Hacer 'click' en el gráfico para ver un menú emergente

APPENDIX 2

To graph a scatter plot, the data are transported to Excel as in Image 1. To this end, the columns (Y and X) are interchanged to; first is the explanatory variable (X) and then the endogenous variable (Y), after that, they are selected with the cursor, then click on *Insert*, and click on *Scatter*.

Image 1.



With the cursor on the graph, a new toolbar appears, click on design, and choose fx

Image 2.

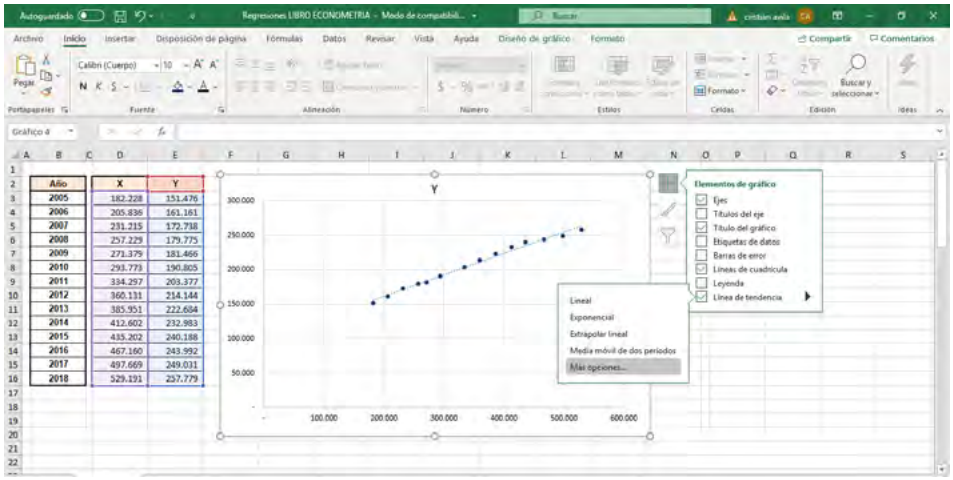
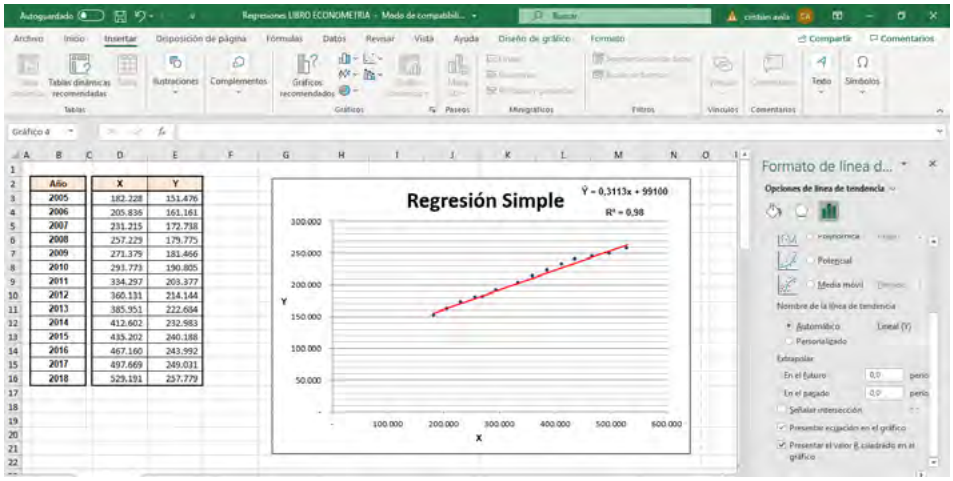


Image 3.

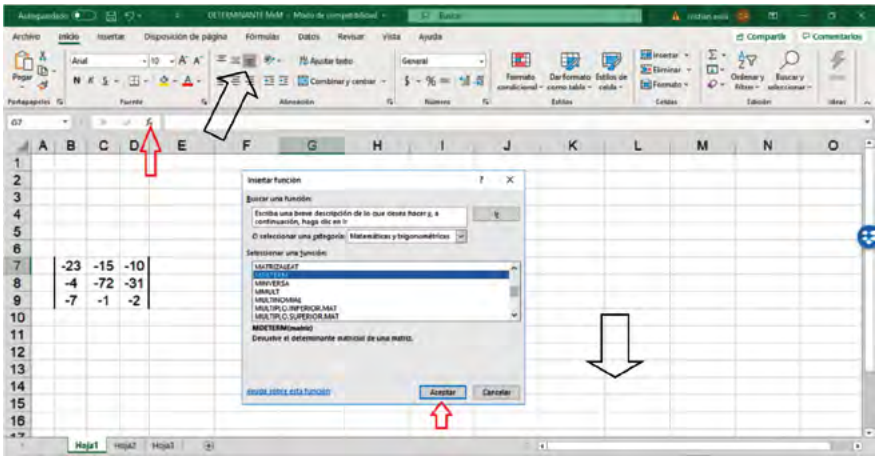


APPENDIX 3

DETERMINANT

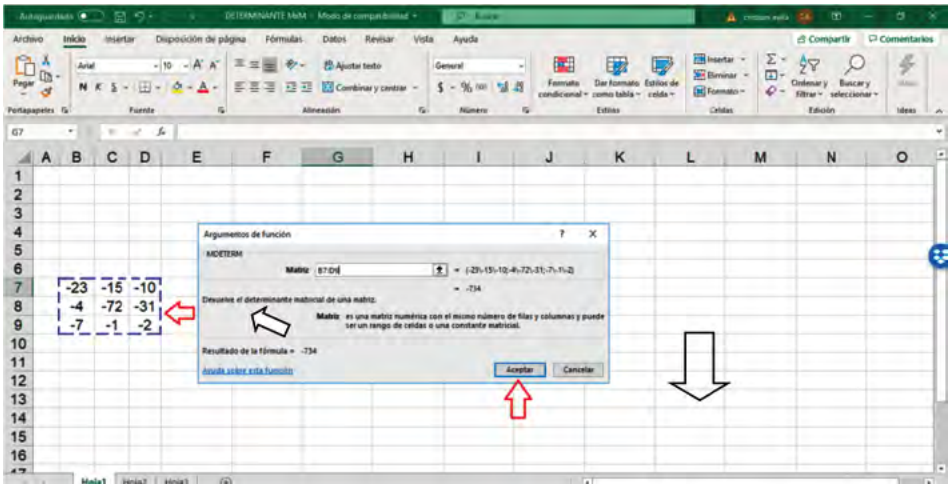
After typing the respective matrix in Excel, the formula format is opened, either by ***fx,*** as the arrow in Image 1 shows, or using the toolbar **insert** = **function**... = by selecting **MDETERM + Accept.**

Image 1.



Immediately a window opens and using the cursor, we must select the **matrix** to be determined. In this example B7:D9 + **Accept** (see Image 2).

Image 2.



The Determinant–743 is obtained.

INVERSA

After verifying that our matrix has -743 as determinant, i.e. that it has an inverse, in **fx function**, we select the category **Mathematics and trigonometry** and choose the option **MINVERSA** (see Image 3).

Image 3.

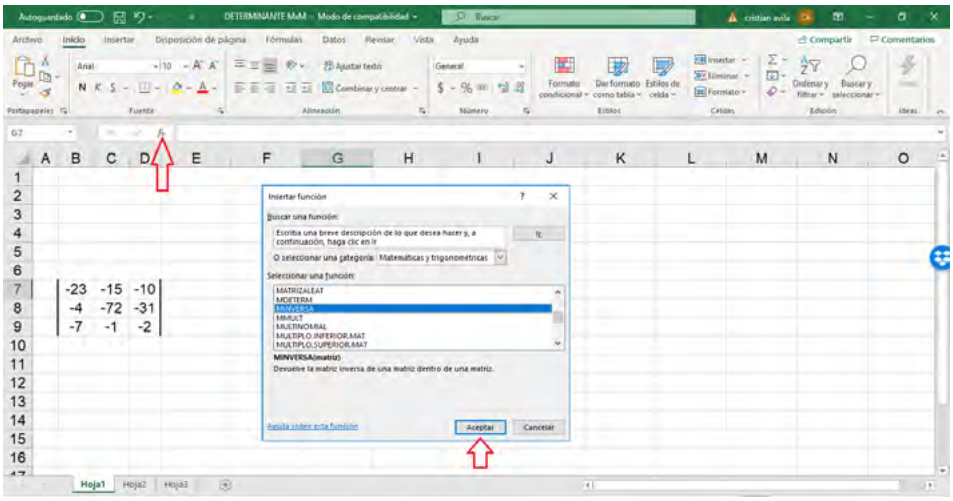
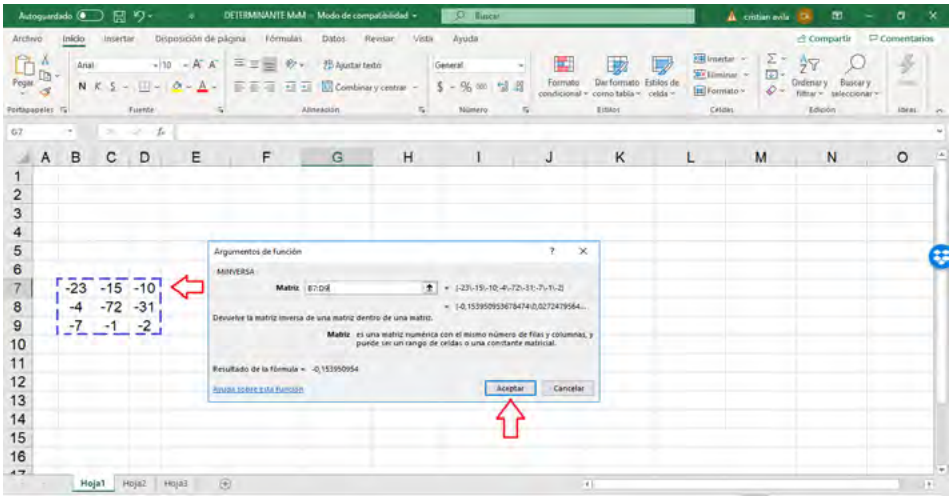


Image 4.


-23 -15 -10
-4 -72 -31
-7 -1 -2
Argumentos de función
MINVERSA
Matriz
Devuelve la matriz inversa de una matriz dentro de una matriz.
Matriz es una matriz numérica con el mismo número de filas y columnas, y puede ser un rango de celdas o una constante matricial.
Resultado de la fórmula = -0,153950954
Aceptar
Cancelar
Hoja1
Hoja2
Hoja3

APPENDIX 4

The following operations belong to the first example in Chapter 4.

TRANSPOSE

Image 1.

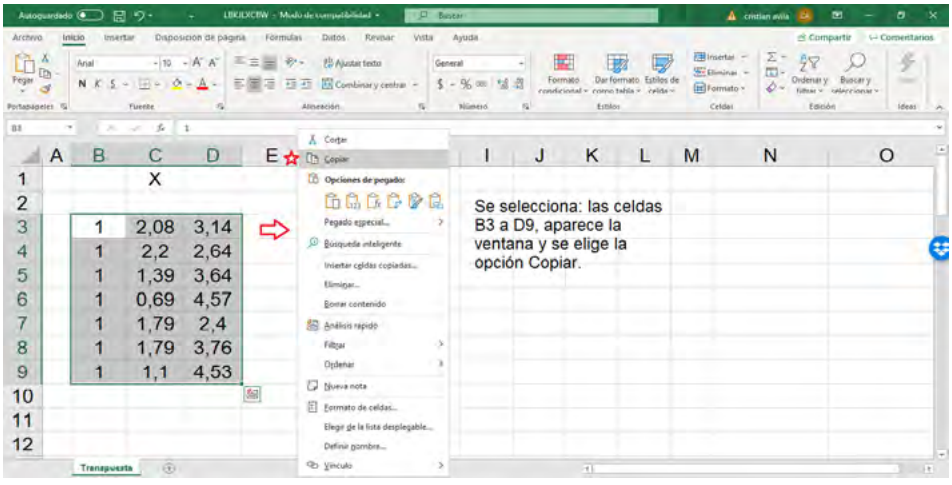


Image 2.

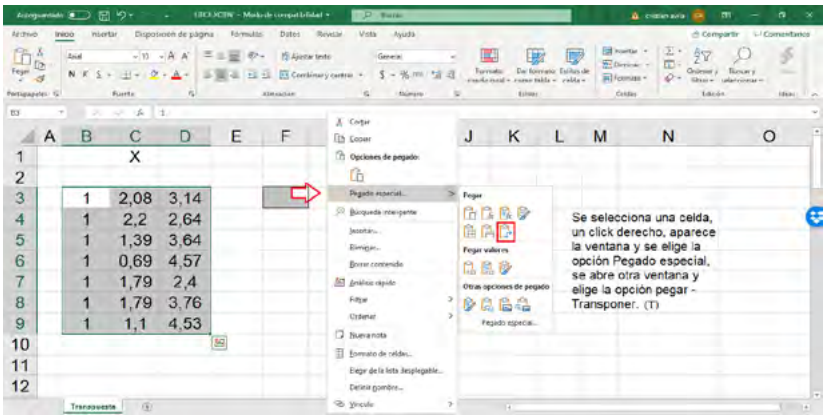


Image 3.

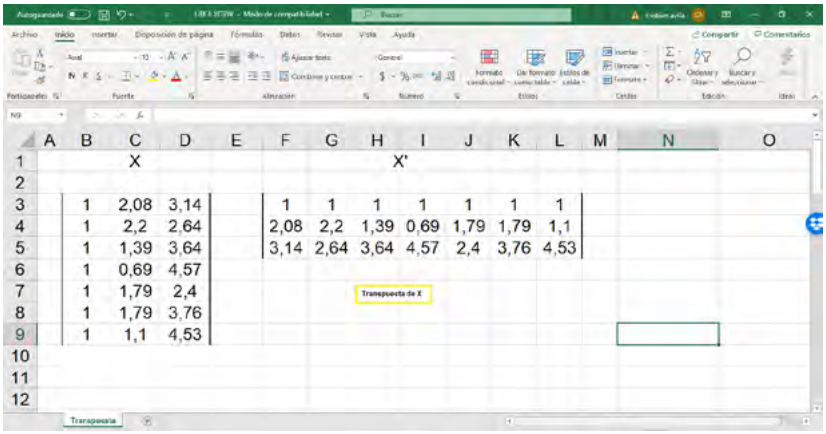


Image 4.

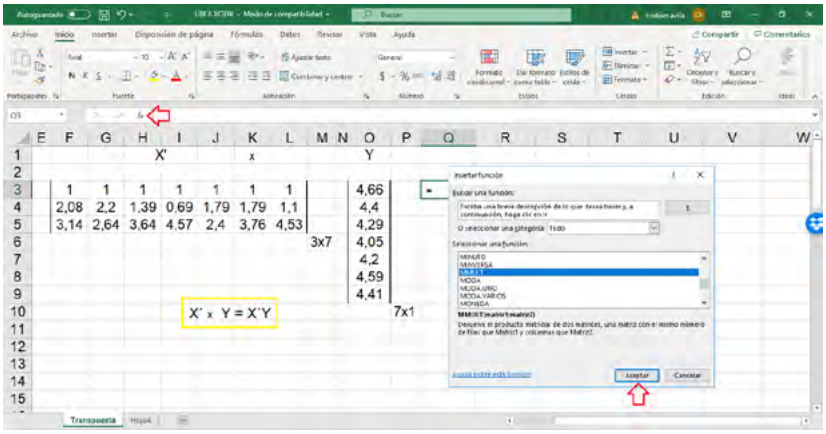


Image 5.

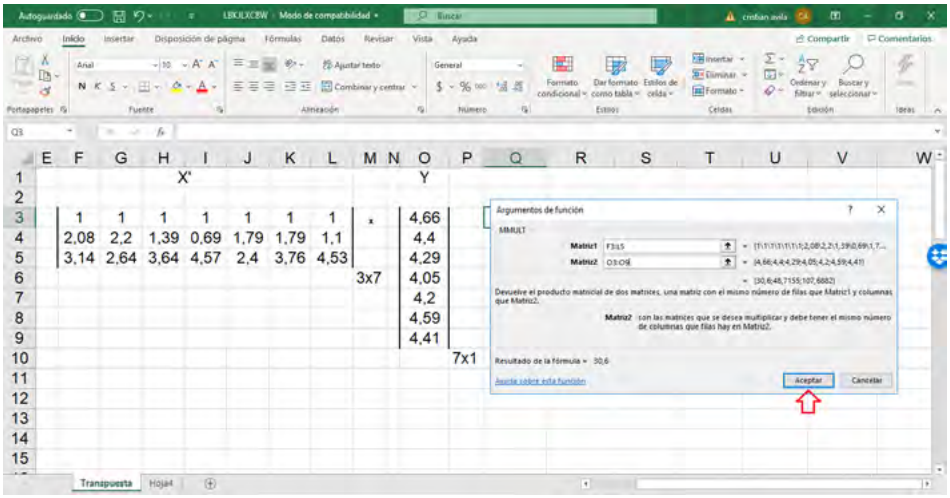
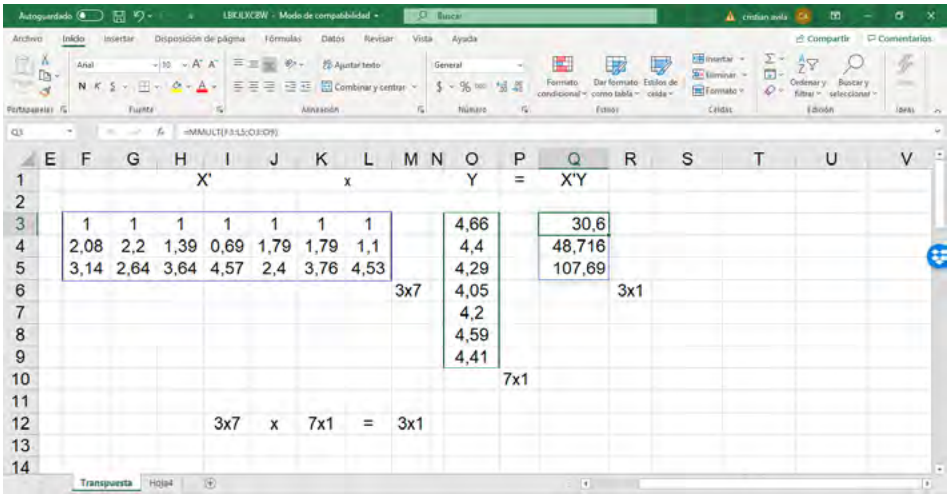


Image 6.



Initially we underline the area exactly proportional to the result of the inverse function, that is, if the initial matrix is of order 2x2, the inverse matrix is 2x2. Thus, the area to underline is 2 lines by 2 columns. For this exercise it is a matrix of order 3x3, therefore, we underline, using the cursor, the same space where we want the result H12:J14 and we choose in ***fx function*** the option **MINVERSE**, then in the window that opens, we select B12:D14 + **Accept**. Finally, we press **f2** and simultaneously **Ctrl.** + **Caps Lock + Enter.**

Image 7.

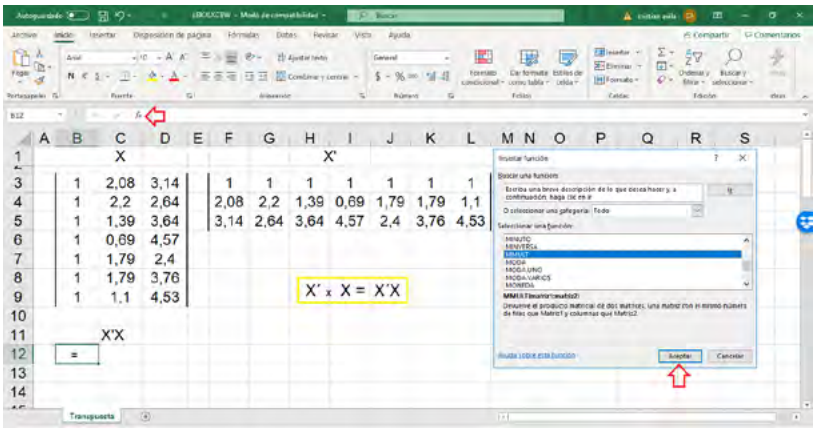


Image 8.

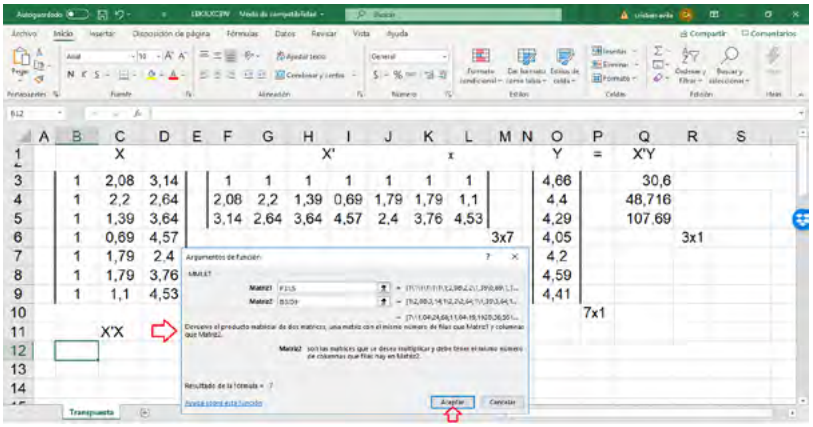


Image 9.

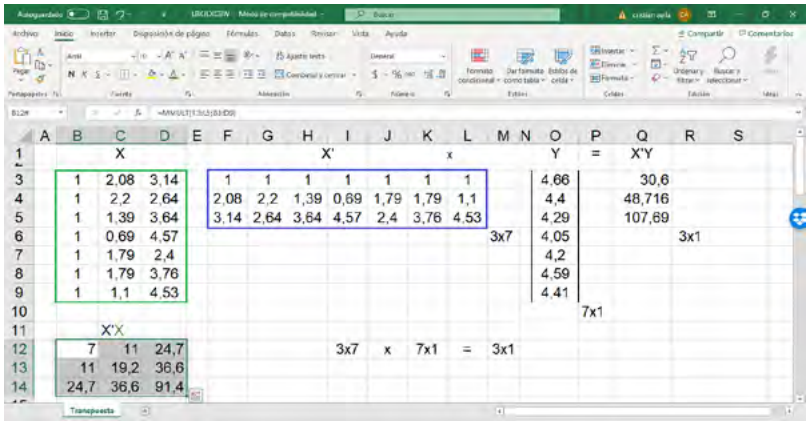


Image 10.

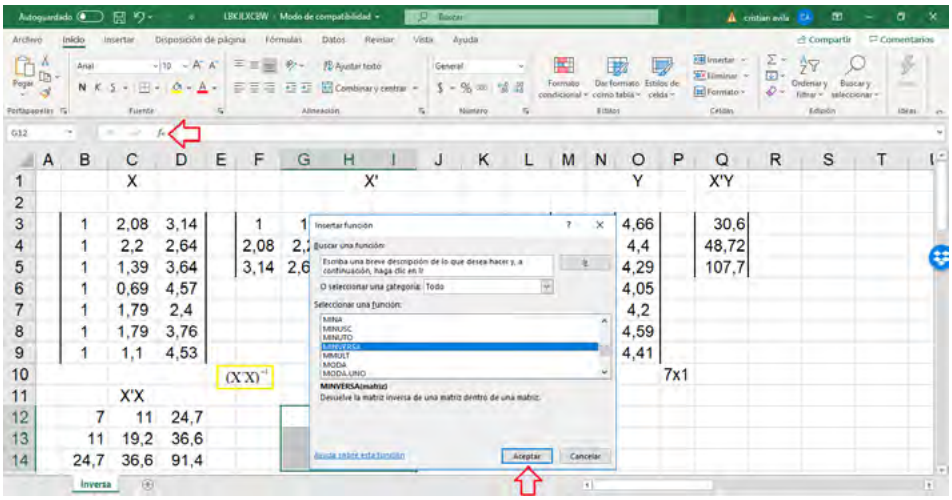
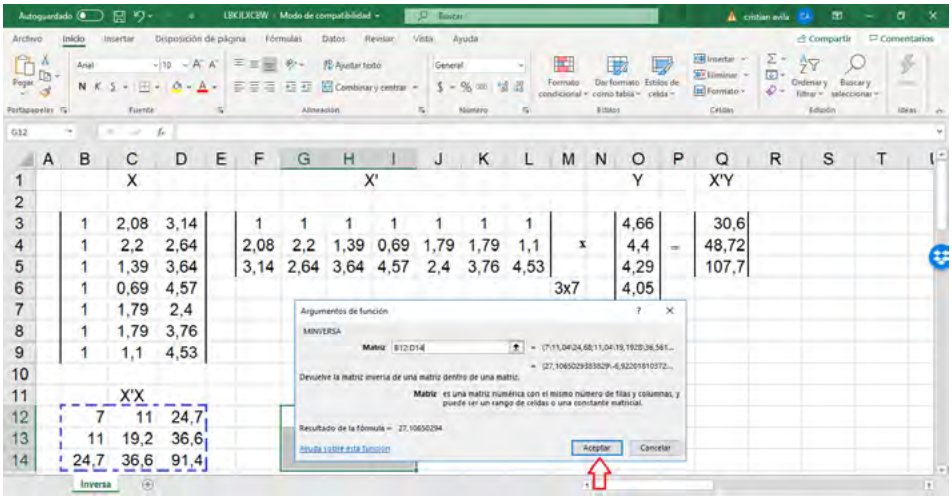
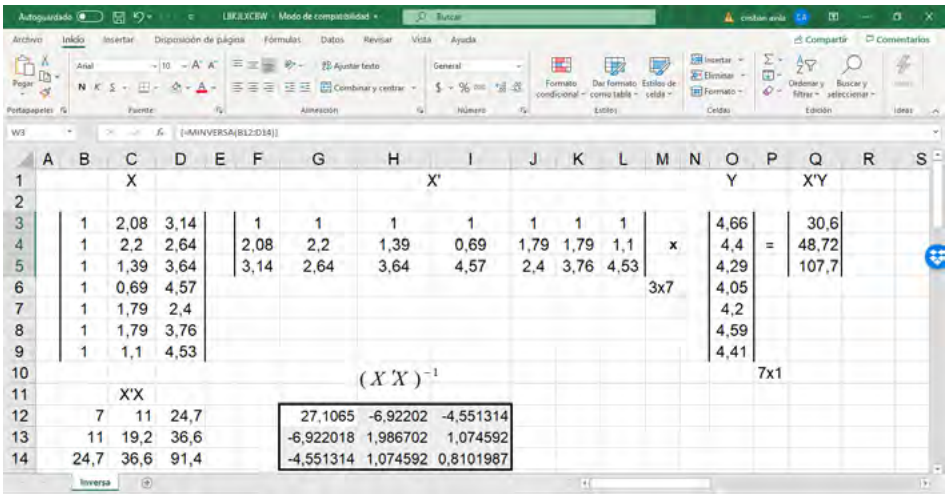


Image 11.



And we obtain the inverse:

Image 12.



$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_t + \hat{\beta}_3 Z_t$$

Now, having obtained the inverse matrix, we estimate the betas:

Image 13.

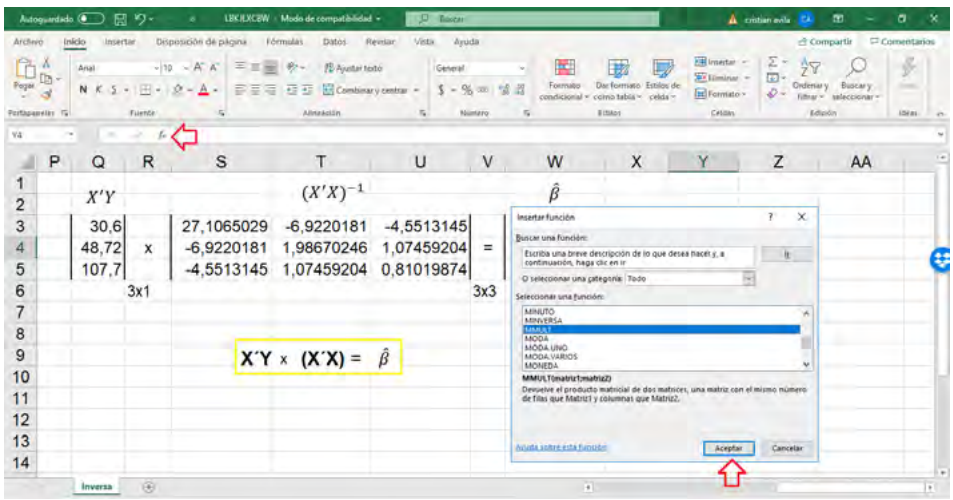


Image 14.

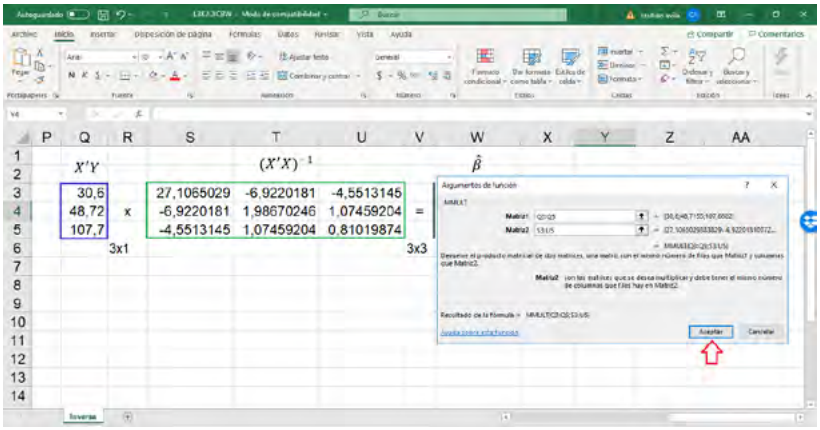


Image 15.

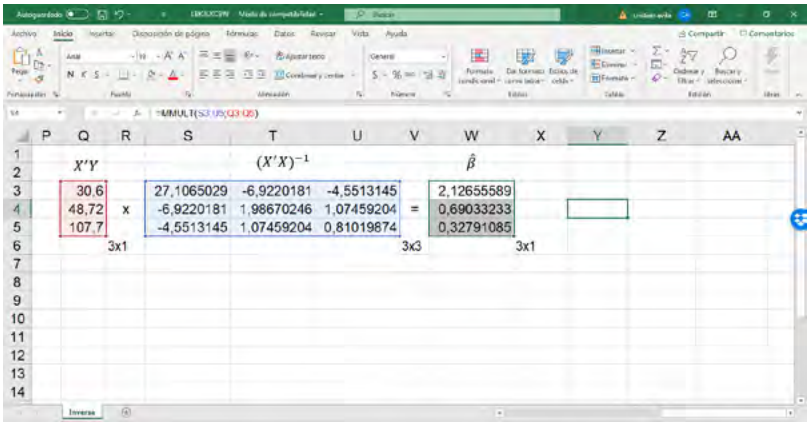
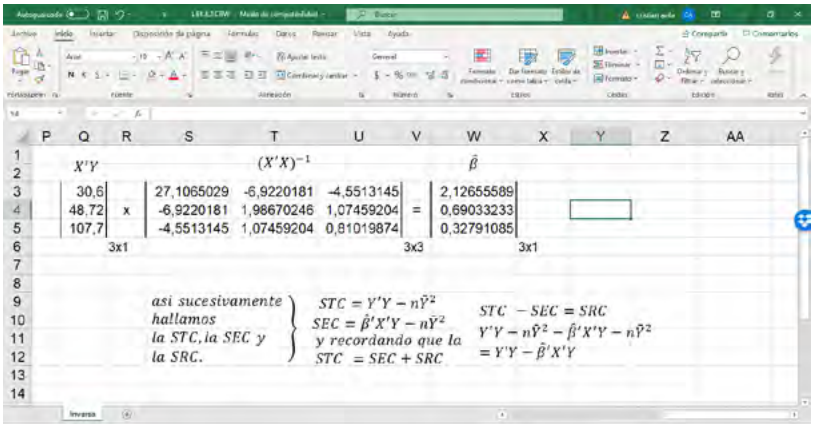


Image 16.



APPENDIX 5

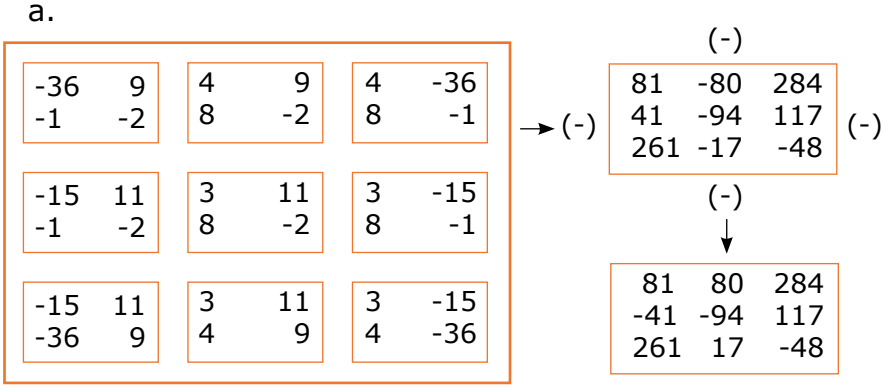
ANSWER SHEET
Exercises of Chapter 3

1. Answer. Exercises: a. = 0, b. = 482 and c = 0.

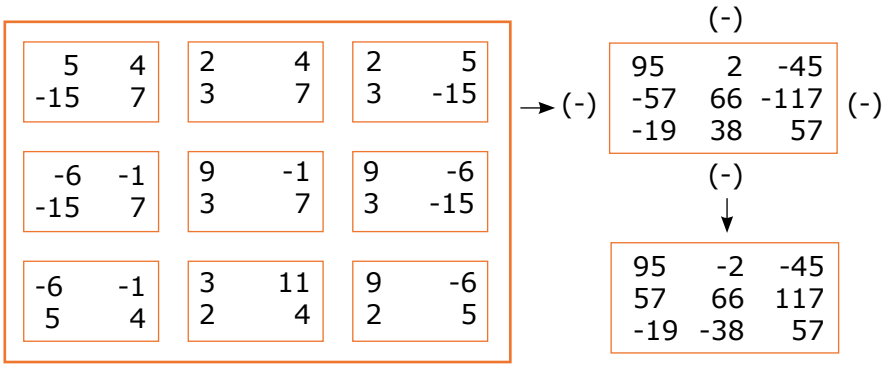
2. Answer. Exercises: a = 95, b = 215.160, c =–4 and d = -734

3. Answer. Exercises: a. and b.

a.



b.



4. Rta. Exercises: a, b, c, d and e.

a.

0.0582	- 0.0020	0.0088
- 0.1333	0.0127	0.0058
- 0.0756	0.0090	- 0.0240

b.

- 0.0034	0.0110	0.0039
0.0085	0.0469	- 0.0373
0.0131	- 0.0299	0.0222

c.

0.0106	0.2063	0.0285
0.0123	- 0.1983	-0.0016
0.0011	0.0840	0.0238

d.

- 0.0001	0.0002	- 0.0314
- 0.0037	0.0050	- 0.0054
0.0004	- 0.0839	0.2793

e.

$$\begin{vmatrix} 0.0302 & -0.0034 & 0.0089 \\ -0.1408 & 0.0156 & 0.2916 \\ 0.1251 & 0.0972 & -0.2592 \end{vmatrix}$$

5. Answer. Exercises: a, b and c.

a. X1= 0 .6911, X2 = 0.7854, X3 = -0.5902

b. X1 =–3.0525, X2 =–17.8425, X3 = 32.3625

c. X1 =–0.1623, X2= 0.5283, X3 =–0.1085

ANNEX 1

DISTRIBUTION t-TABLE

Puntos porcentuales superiores de la distribución *F*

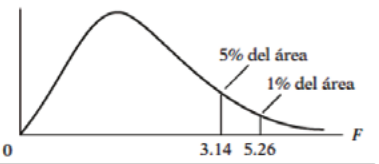
Ejemplo

$\Pr(F > 1.59) = 0.25$

$\Pr(F > 2.42) = 0.10$ para gl $N_1 = 10$

$\Pr(F > 3.14) = 0.05$ y $N_2 = 9$

$\Pr(F > 5.26) = 0.01$



gl para el denominador N_2	Pr	gl para el numerador N_1 1	2	3	4	5	6	7	8	9	10	11	12
	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
1	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90

Taken from: Gujarati (2010), for explanatory purposes only.

ANNEX 2

DISTRIBUTION t-TABLE

Puntos porcentuales superiores de la distribución *F*

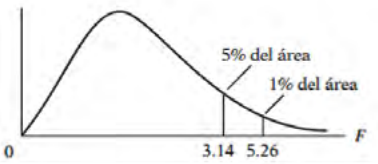
Ejemplo

$Pr(F > 1.59) = 0.25$

$Pr(F > 2.42) = 0.10$ para gl $N_1 = 10$

$Pr(F > 3.14) = 0.05$ y $N_2 = 9$

$Pr(F > 5.26) = 0.01$



gl para el denominador N_2	Pr	gl para el numerador N_1 1	2	3	4	5	6	7	8	9	10	11	12
	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
1	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
2	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
3	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
4	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
5	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
6	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90

Taken from: Gujarati (2010), for explanatory purposes only.

BIBLIOGRAPHY

Acemoglu, D., Johnson, S., & Robinson, J. (2004). Institutions as the fundamental cause of long-run growth. Working Paper Series n.o 10481. Cambridge, MA, National Bureau of Economic Research (NBER). https://doi.org/10.3386/w10481

Acemoglu, D., & Verdier, T. (1998). Property rights, Corruption and the Allocation of Talent: A General Equilibrium Approach. The Economic Journal, 108(450), 1381- 1403. https://doi.org/10.1111/1468-0297.00347

Ades, A., & Di Tella, R. (1996). The Causes and Consequences of Corruption: A Review of Recent Empirical Contributions. IDS Bulletin 27(2): 6-11. https://doi.org/10.1111/j.1759-5436.1996.mp27002002.x

Ades, A., & Di Tella, R. (1999) Competition and corruption. The American Economic Review, 89(4), p 982-994, 1999. https://doi.org/10.1257/aer.89.4.982

Agatiello, O. (2010). Corruption not an end. Management Decision, 48(10), 1456-1468. http://dx.doi.org/10.1108/00251741011090270

Agerberg, M. (2019). The Curse of Knowledge? Education, Corruption, and Politics. Political Behavior, 41(2), 369-399. https://doi.org/10.1007/s11109-018-9455-7

Aidt, T. S. (2009). Corruption, institutions and economic development. Oxford Review of Economic Policy, 25, 271-291. https://doi.org/10.1093/oxrep/grp012

Aidt, T. S. (2003). Economic analysis of corruption: A Survey. The Economic Journal, 113(491), F632-F652. http://dx.doi.org/10.1046/j.0013-0133.2003.00171.x

Alfano, R., Baraldi, L., & Cantabene, C. (2012). Political Competition, Electoral System and Corruption: The Italian Case. MPRA Paper. University Library of Munich. https://mpra.ub.uni-muenchen.de/id/eprint/41480

Almeida dos Santos, R., De Hoyos Guevara, A. J., & Sanchez Amorim, M. C. (2013). Corrupcao nas organizacoes privadas: analise da percepcao moral segundo genero, idade e grau de instrucao. Rausp: Revista de Administracao da Universidade de Sao Paulo, 48(1), 53-66. https://doi.org/10.5700/rausp1073

Alt, J., & Lassen, D. D. (2003). The Political Economy of Corruption in American States. Journal of Theoretical Politics, 15(3), 341-365. https://doi.org/10.1177/0951692803015003006

Andersen, T. B. (2009). E-government as an anti-corruption strategy. Information Economics and Policy, 21(3), 201-210. https://doi.org/10.1016/j.infoecopol.2008.11.003

Anechiarico, F., & Jacobs, J. (1996). The pursuit of absolute integrity. How corruption control makes government ineffective. The University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/P/bo3633806.html

Avila, C. y Oliveira, N. CORRUPCIÓN, un análisis a escala regional en Colombia. Págs. 315, 2023 UNAD, Sello Editorial UNAD. https://doi.org/10.22490/9789586519410

Avila, C., y Oliveira, N. Desarrollo y crecimiento económico, Casanare. Lecciones aprendidas. Universidade Federal do Tocantins (uft), Universidad Nacional Abierta y a Distancia (UNAD) y Cámara de Comercio del Casanare (CCC). Editorial Jotamar, 2018. https://www.academia.edu/43272301/DESARROLLO_Y_CRECIMIENTO_ECON%C3%93MICO_CASANARE_Lecciones_aprendidas

Avila, C., Moreno, C., Barrera, S., Rojas, N. y Oliveira, N. Globalización, localización, competitividad y especialización

productiva, un análisis empírico para Colombia. Págs. 214, UNAD. Sello Editorial UNAD. Colombia, 2022. https://doi.org/10.22490/9789586518369

Avila, C. y Oliveira, N. Curso básico de econometría clásica. Págs. 180, Sello Editorial UNAD. Colombia, 2019. https://doi.org/10.22490/9789586517171

Avila, C., Sanabria, S., & Oliveira, N. Indicadores de subdesarrollo regional: una aplicación para Colombia/ Regional underdevelopment indicators: an application for Colombia. Informe GEPEC, [S. l.], v. 26, n. 1, p. 106–126, 2022. https://doi.org/10.48075/igepec.v26i1.28152

Avila, C., Sanabria, S., y Oliveira, N. Localización y especialización productiva: la región de la Amazonia colombiana. Revista Ra'ega. O Espaço Geográfico em Análise, 52, 60-83, 2021. http://dx.doi.org/10.5380/raega.v52i0.76860

Besley, T., & Persson, T. (2009). The Origins of State Capacity: Property Rights, Taxation, and Politics. American Economic Review 99(4), 1218-1244. https://doi.org/10.1257/aer.99.4.1218

Brunetti, A., & Weder, B. (2003). A Free Press is Bad News for Corruption. Journal of Public Economics 87(7-8), 1801-1824. https://doi.org/10.1016/S0047-2727(01)00186-4

Castañeda, Victor. (2012) "Una revisión de los determinantes de la estructura y el recaudo tributario: el caso latinoamericano tras la crisis de la deuda externa" en Cuadernos de Economía. Vol. 31, núm. 58, pp. 77-112. Bogotá, Universidad Nacional de Colombia.

Castañeda, V. (2016). Una investigación sobre la corrupción pública y sus determinantes. Revista Mexicana de Ciencias Políticas y Sociales, 61(227), 103-136. https://doi.org/10.1016/S0185-1918(16)30023-X

Cepeda, F (1997). La corrupción en Colombia. Bogotá D,C: Fedesarrollo – TM Editores.

Delli Carpini, M., & Keeter, S. (1996). What Americans Know about Politics and why it Matters. New Haven: Yale University Press.

Elbahnasawy, N., & Revier, C. (2012). The Determinants of Corruption: Cross-Country-Panel-Data Analysis. The Developing Economies, 50(4), 311-333. https://doi.org/10.1111/j.1746-1049.2012.00177.x

Frechette, Guillaume, (2001) A Panel Data Analysis of the Time-Varying Determinants of Corruption (Working paper). París, Cirano.

Gamarra Vergara, J. R. (2006). Pobreza, corrupción y participación política: una revisión para el caso colombiano. Documentos de Trabajo sobre Economía Regional n.o 70. Cartagena: Banco de la República.

Galston, W. A. (2001). Political Knowledge, Political Engagement, and Civic Education. Annual Review of Political Science, 4, 217-234. https://doi.org/10.1146/annurev.polisci.4.1.217

Glaeser, E., & Saks, R. (2006). Corruption in America. Journal of Public Economics, 90(6-7), 1053-1072. https://doi.org/10.1016/j.jpubeco.2005.08.007

Golden, M., & Picci, L. (2005). Proposal for a new measure of corruption, illustrated with Italian data. Economics & Politics, 17(1), 37-75. https://doi.org/10.1111/j.1468-0343.2005.00146.x

Leite, C., & Weidmann, J. (1999). Does Mother Nature corrupt? Natural resources, corruption and economic growth. Working Paper n.o 85/1999. International Monetary Fund (IMF). https://www.imf.org/external/pubs/ft/wp/1999/wp9985.pdf

Martínez, M., Avila, C., Sanabria, S., y Oliveira, N.Localización y especialización productiva: el caso de las trece ciudades principales en Colombia. En Revista Brasileira de Gestão e Desenvolvimento Regional. G&DR. V. 15, N. 6, Edição Especial, P. 216-230, nov/2019. Taubaté, SP, Brasil. https://doi.org/10.54399/rbgdr.v15i6.5196

Persson, T., Tabellini, G., & Trebbi, F. (2003). Electoral Rules and Corruption. Journal of the European Economic Association, 1(4), p. 958-989. https://www.jstor.org/stable/40005174

Persson, Anna, (2008) "he Institutional Sources of Statehood: Assimilation, Multiculturalism and Taxation in Sub-Saharan Africa". Tesis de doctorado. Gothenburgo, University of Gothenburg, Department of Political Science.

Piketty, T. (2019). Capital e ideología. Editorial Planeta, S. A.

Piketty, T. (2015). La economía de las desigualdades. Editorial Siglo XXI.

Piketty, T. (2013). El capital en el siglo XXI. Fondo de Cultura Económica (FCE).

Rose-Ackerman, S. (1978). Corruption: A Study in Political Economy. New York: Academic Press.

Rose-Ackerman, S. (1999). Corruption and Government: Causes, Consequences, and Reform. New York: Cambridge University Press. https://doi.org/10.1017/CBO9781139175098

Rose-Ackerman, S. (2002). "Grand" corruption and the ethics of global business. Journal of Banking.

Stigler, G. (1971). The theory of Economic Regulation. The Bell Journal of Economics and Management Science, 2(1), 3-21. https://doi.org/10.2307/3003160

Van Rijckeghem, C., & Weder, B. (1997). Corruption and the Rate of Temptation: Do Low Wages in the Civil Service Cause Corruption? Working Paper n.o 73-1997. Washington: International Monetary Fund. https://www.imf.org/external/pubs/ft/wp/wp9773.pdf

Recomended Bibliography

Avila, C. y Oliveira, N. Curso básico de econometría clásica. Págs. 180, Sello Editorial UNAD. Colombia, 2019. https://doi.org/10.22490/9789586517171

Cramer, J. S., Empirical Econometrics, North-Holland, Amsterdam, 1969.

Chiang, Alpha C., Métodos fundamentales de economía matemática, 4ta. Ed. McGraw-Hill, Nueva York, 2006.

Gujarati, Damodar N., Essentials of Econometrics, 3a. ed., McGraw-Hill, Nueva York, 2006.

Gujarati, Damodar N & Porter, Dawn., *Econometria*, 5a. ed., McGraw-Hill, Mexico, 2010.

Pindyck, Robert. S. y Rubinfeld, Daniel L., Econometría Modelos y Pronósticos, 4a. ed., McGraw-Hill, 2001.

Wei, William. "Time Series Analysis" Univariate and Multivariate Methods USA. 1990.

Wooldridge, Jeffrey M., Introductory Econometrics, 3a. ed., South-Western College Publishing, 2000.

Sello Editorial
Universidad Nacional
Abierta y a Distancia

**UNIVERSIDAD NACIONAL ABIERTA
Y A DISTANCIA (UNAD)**

Sede Nacional José Celestino Mutis
Calle 14 Sur 14-23
PBX: 344 37 00 - 344 41 20
Bogotá, D.C., Colombia

www.unad.edu.co

9 786287 786219