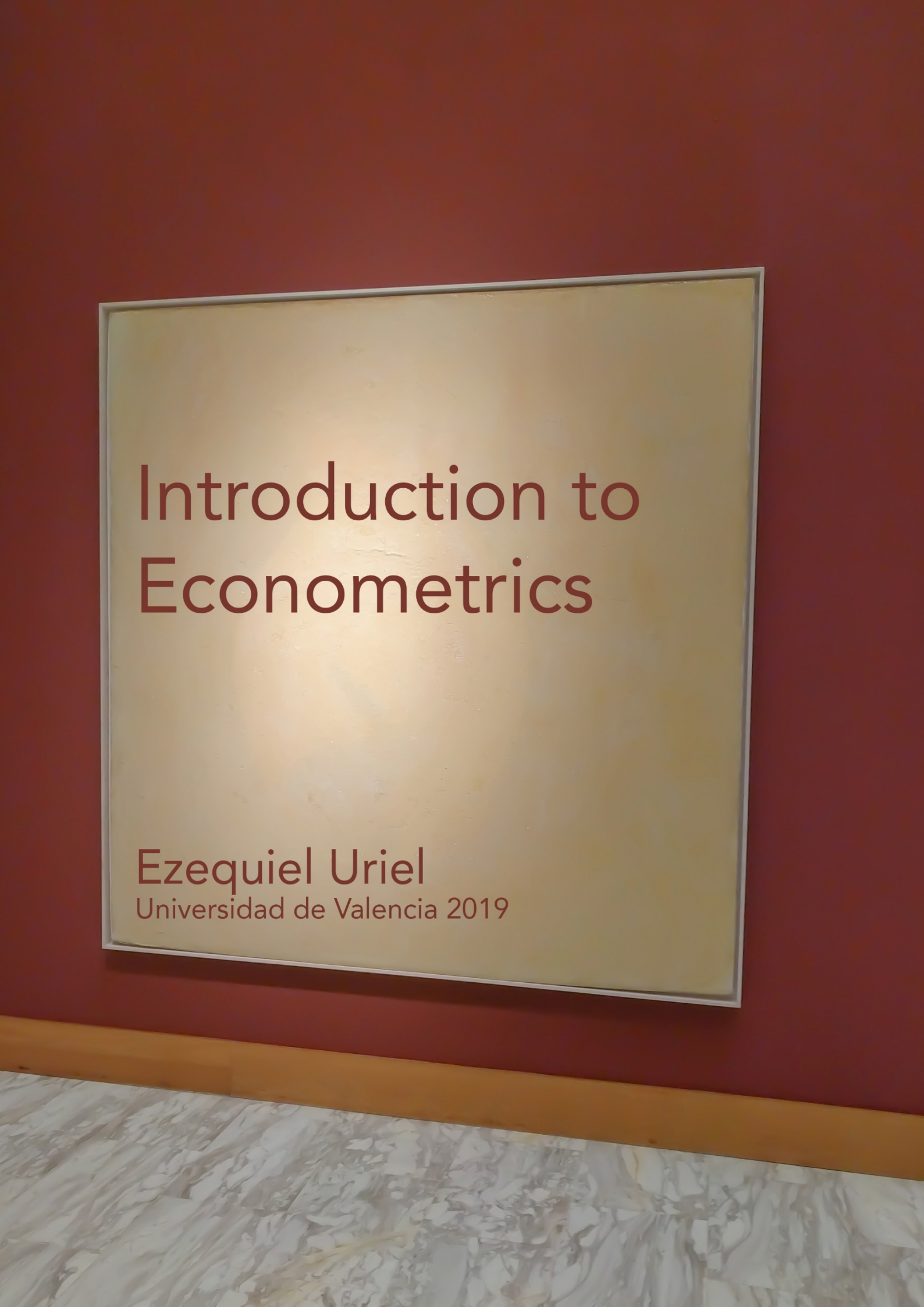

Introduction to
Econometrics
Ezequiel Uriel
Universidad de Valencia 2019

INTRODUCTION TO ECONOMETRICS

Ezequiel Uriel

2019

University of Valencia

I would like to thank the professors Luisa Moltó, Amado Peiró, Paz Rico, Pilar Beneito and Javier Ferri for their suggestions for the errata they have detected in previous versions, and for having provided me with data to formulate exercises. Some students have also collaborated in the detection of errata. In any case, I am solely responsible for the errata that have not been detected.

Summary

1 ECONOMETRICS AND ECONOMIC DATA

1.1 What is econometrics?

First, let us see something about the origin of econometrics as a discipline. The term econometrics is believed to have been crafted by Ragnar Frisch, co-winner of the first Nobel Prize in Economic Sciences in 1969, along with fellow econometrician Jan Tinbergen. Both of them were founders of the Econometric Society in 1933. In section I of the constitution of this society, it is stated that

> "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics. Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems and that are penetrated by constructive and rigorous thinking similar to that which has come to dominate the natural sciences"

In the first issue of *Econometrica* (1933), the Econometric Society journal, Ragnar Frisch gives us an explanation about the meaning of *econometrics*:

> "But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a definitely quantitative character. Nor should econometrics be taken as synonymous with the application of mathematics to economics. Experience has shown that each of these three viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics."

Today, we would also say that econometrics is the combined study of economic models, mathematical statistics, and economic data. Within the field of econometrics, econometric theory can be distinguished from applied econometrics.

Econometric theory concerns the development of tools and methods, and the study of the properties of econometric methods. Econometric theory belongs to the field of statistics.

Applied econometrics is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data. Applied econometrics is mainly used in the field of applied economics.

What are the goals of Econometrics? We are going to examine three:

1) *Knowledge of the real economy*. Econometric methods allow us to estimate economic magnitudes such as the marginal propensity to consume or the elasticity of labor with respect to output. These estimations are located in a determined time and space:

for example, in Spain in the last quarter of the 20th century. In addition to the estimation, in which numerical values are obtained, econometric methods allow us to perform tests of hypothesis; for example, in a production function, is the hypothesis of constant returns to scale admissible?

2) *Economic simulation policy*. Econometrics methods can be used to simulate the effects of alternative policies. For example, with an appropriate econometric model we could see, in quantitative terms, how the different increases in tobacco tax affect the consumption of tobacco.

3) *Prediction or forecasting*. Very often econometric methods are used to predict values of economic variables in the future. By making predictions we try to reduce our uncertainty in the future of the economy. This is not an easy task, since in general the predictions are only satisfactory when there are no drastic changes in the economy. Although it would be useful to be able to predict these drastic changes accurately, both econometric and other alternative methods tend to be imprecise.

1.2 Steps in developing an econometric model

There are three main steps in developing an econometric model: specification, estimation and validation.

While in a first approximation these stages follow a sequential order, in econometric analysis it is generally necessary to go back more than once within this sequence. It is necessary to continuously confront the model with the data and any other information source, in order to obtain an econometric model compatible with the data. The model can be used to analyze reality, offer better predictions or constitute a good basis for making decisions. Now we will describe the steps listed above.

(a) Specification

In this first step, the model or models used must be defined, as well as data to be used in the estimation stage.

In the specification step, we will refer to four elements: the economic model, the econometric model, the statistical assumptions of the model and the data. In this section we will refer to the first three elements; in the following section we will examine different types of data used in econometric analysis.

The first element we need is an economic model. In some cases, a formal economic model is constructed entirely using economic theory. In other cases, economic theory is used less formally in constructing an economic model.

After we have an economic model, we must convert it into an econometric model. We are going to see that with two examples.

EXAMPLE 1.1 Keynesian consumption function

Keynes formulated his well-known consumption function in three propositions:

Proposition 1: Consumption is a function of income, and both variables are measured in real terms. If the variables are measured in real terms, it means that when consumers decide the proportion of income devoted to consumption, they are not affected by monetary illusion.

Analytically, proposition 1 can be expressed in the following way:

$$cons = f(inc) \tag{1-1}$$

Proposition 2: Consumption is an increasing function of income, but an increase in income always causes an increase, to a lesser degree, in consumption.

This proposition implies that marginal propensity to consumption is greater than 0 (it is an increasing function), but it is smaller than 1 (an increase in income always causes an increase, to a lesser degree, in consumption).

Analytically, proposition 2 can be expressed in the following way:

$$0 < \frac{\mathrm{d}\, cons}{\mathrm{d}\, inc} < 1 \tag{1-2}$$

Proposition 3: The proportion of income consumed is smaller when income increases. That is to say, the proportion of the last euro earned devoted to consumption is smaller than the proportion of total income earned devoted to consumption.

Analytically, proposition 3 can be expressed in the following way:

$$\frac{\mathrm{d}con}{\mathrm{d}\, inc} < \frac{cons}{inc} \tag{1-3}$$

In other words, the marginal propensity to consume is smaller than the average propensity to consume.

These three propositions constitute an economic model: the Keynesian consumption function.

To estimate and test this model we must convert it into an econometric model. For this conversion, two requirements must be accomplished.

According to the first requirement, it is necessary to specify the mathematical form of the function. The linear function has been used in this case because, in addition to being simple, it is compatible with the description made by Keynes.

In order to justify the second requirement, it must be taken into account that the model formulated in proposition 1 is deterministic. That is to say, income is the only factor in the determination of consumption. But in real life there are many other factors, other than income, which have an influence on consumption. In an econometric model, all the factors different from the independent variables included are gathered in a variable denominated random disturbance or error (u). The second requirement is the introduction of the term of error in the equation.

In general, all the relevant factors must be introduced explicitly in the econometric model; all the other factors are taken into account in a unique variable: the error or the random disturbance. In the Keynesian consumption function the only relevant factor considered is income.

Taking into account these two requirements, Keynesian consumption function can be expressed in the following way:

$$cons = \beta_1 + \beta_2 inc + u \tag{1-4}$$

This is an econometric model that can be estimated if you have data on consumption and income. Let us see now the other two propositions. In this linear model, the marginal propensity to consumption is the following:

$$\frac{\mathrm{d}\, cons}{\mathrm{d}\, inc} = \beta_2 \tag{1-5}$$

Consequently, proposition 2 in this model is the following:

$$0 < \beta_2 < 1 \tag{1-6}$$

Once the model has been estimated, it is possible to test whether the estimate of β_2 is between 0 and l.

The average propensity to consume in the linear model, considering that the error is equal to 0, is the following:

$$\frac{cons}{inc} = \frac{\beta_1 + \beta_2 inc}{inc} = \frac{\beta_1}{inc} + \beta_2 \tag{1-7}$$

Therefore, proposition 3 implies that

$$\frac{\beta_1}{inc} + \beta_2 > \beta_2 \quad \text{or} \quad \frac{\beta_1}{inc} > 0 \tag{1-8}$$

That is to say,

$$\beta_1 > 0 \tag{1-9}$$

Once the model has been estimated, testing proposition 3 is equivalent to testing whether the intercept is significantly greater than 0.

EXAMPLE 1.2 Wage determination

Economic model:

Formal economic theory - human capital theory- says that education (*educ*), experience (*exper*) and *training* are factors that affect productivity and hence the *wage*. Therefore, an economic model for wage determination could be the following:

$$wage = f(educ, exper, training) \tag{1-10}$$

Incidentally, do you think there is any variable missing in this model?

Econometric model:

The corresponding econometric model, using a mathematical linear form, is the following:

$$wage = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 training + u \tag{1-11}$$

To sum up, to convert an economic model into an econometric model:

a) The form of the function *f*(.) has been specified.

b) A disturbance variable has been included to reflect the effect of other variables affecting wage, but not appearing in the model.

An important element in the specification of the model is the formulation of a set of statistical assumptions, which are used in subsequent steps. These statistical assumptions play a key role in hypothesis testing and, in general, throughout the inference process carried out with the model.

(b) Estimation

In the estimation process we obtain numerical values of the coefficients of an econometric model. To complete this stage, data are required on all observable variables that appear in the specified econometric model, while it is also necessary to select the appropriate estimation method, taking into account the implications of this choice on the statistical properties of estimators of the coefficients. The distinction between estimator and estimate should be made clear. An estimator is the result of applying an estimation method to an econometric specification. On the other hand, an estimate consists of obtaining a numerical value of an estimator for a given sample. For example, applying a very simple estimation method, called *ordinary least squares*, to the specification of the consumption function (1-4) provides expressions which determine the *estimators* $\hat{\beta}_1$ and $\hat{\beta}_2$. Substituting the sample data in these expressions, two numbers are obtained: one for $\hat{\beta}_1$ and one for $\hat{\beta}_2$ which provide *estimates* of the parameters β_1 and β_2.

In general, it is possible to obtain analytical expressions of the estimators, particularly in the case of estimating linear relationships. But in non-linear procedures of estimation it is often difficult to establish their analytical expression.

(c) Validation

The results are assessed in the validation stage, where we assess whether the estimates obtained in the previous stage are acceptable, both theoretically and from the statistical point of view. On the one hand, we analyze, whether estimates of model parameters have the expected signs and magnitudes: that is to say, whether they satisfy the constraints established by economic theory.

From the statistical point of view, on the other hand, statistical tests are performed on the significance of the parameters of the model, using the statistical assumptions made in the specification step. In turn, it is important to test whether the statistical assumptions of the econometric model are fulfilled, although it should be noted that not all assumptions are testable. The violation of any of these assumptions implies, in general, the application of another estimation method that allows us to obtain estimators whose statistical properties are as good as possible.

One way to establish the ability of a model to make predictions is to use the model to forecast outside the sample period, and then to compare the predicted values of the endogenous variable with the values actually observed.

1.3 Economic data

As we have seen, an empirical analysis uses data to test a theory or to estimate a relationship. It is important to stress that in Econometrics we use non-experimental data. Non experimental or observational data are collected by observing the real world in a passive way. In this case, data are not the outcome of controlled experiments.

Experimental data are often collected in laboratory environments in the same way as in natural sciences. Now, we are going to see three types of data which can be used in the estimation of an econometric model: time series, cross sectional data, and panel data.

Time Series

In time series, data are observations on a variable over time. For example: magnitudes from national accounts such as consumption, imports, income, etc. The chronological ordering of observations provides potentially important information. Consequently, ordering matters.

Time series data cannot be assumed to be independent across time. Most economic series are related to their recent histories. Typical examples include macroeconomic aggregates such as prices and interest rates. This type of data is characterized by serial dependence.

Given that most aggregated economic data are only available at a low frequency (annual, quarterly or perhaps monthly), the sample size can be much smaller than in typical cross sectional studies. The exception is financial data where data are available at a high frequency (weekly, daily, hourly, etc.) and so sample sizes can be quite large.

Cross Sectional Data

Cross sectional data sets have one observation per individual and data are referred to a determined point in time. In most studies, the individuals surveyed are individuals (for example, in the Labor Force Survey (EPA) more than 100000 individuals are interviewed every quarter), households (for example, the Household Budget Survey),

firms (for example, industrial firm survey) or other economic agents. Surveys are a typical source for cross-sectional data. In many contemporary econometric cross sectional studies the sample size is quite large.

In cross sectional data, observations must be obtained by random sampling. Thus, cross sectional observations are mutually independent. The ordering of observations in cross sectional data does not matter for econometric analysis. If the data are not obtained with a random sample, we have a sample selection problem.

So far we have referred to micro data type, but there may also be cross sectional data relating to aggregate units such as countries, regions, etc. Of course, data of this type are not obtained by random sampling.

Panel Data

Panel data (or longitudinal data) are time series for each cross sectional member in a data set. The key feature is that the same cross sectional units are followed over a given time period. Panel data combines elements of cross sectional and time series data. These data sets consist of a set of individuals (typically people, households, or corporations) surveyed repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but for a given individual, observations are mutually dependent. Thus, the ordering in the cross section of a panel data set does not matter, but the ordering in the time dimension matters a great deal. If we do not take into account the time in panel data, we say that we are using pooled cross sectional data.

2 THE SIMPLE REGRESSION MODEL: ESTIMATION AND PROPERTIES

2.1 Some definitions in the simple regression model

2.1.1 Population regression model and population regression function

In the simple regression model, the *population regression model* or, simply, the *population model* is the following:

$$y = \beta_1 + \beta_2 x + u \tag{2-1}$$

We shall look at the different elements of the model (2-1) and the terminology used to designate them. We are going to consider that there are three types of variables in the model: *y*, *x* and *u*. In this model there is only one factor *x* to explain *y*. All the other factors that affect *y* are jointly captured by *u*.

We typically refer to *y* as the endogenous (from the Greek: generated inside) variable or dependent variable. Other denominations are also used to designate *y*: left-hand side variable, explained variable, or regressand. In this model all these denominations are equivalent, but in other models, as we will see later on, there can be some differences.

In the simple linear regression of *y* on *x*, we typically refer to *x* as the exogenous (from the Greek: generated outside) variable or independent variable. Other denominations are also used to designate *x*: right-hand side variable, explanatory variable, regressor, covariate, or control variable. All these denominations are equivalent, but in other models, as we will see later, there can be some differences.

The variable *u* represents factors other than *x* that affect *y*. It is denominated error or random disturbance. The disturbance term can also capture measurement error in the dependent variable. The disturbance is an unobservable variable.

The parameters β_1 and β_2 are fixed and unknown.

On the right hand of (2-1) we can distinguish two parts: the systematic component $\beta_1 + \beta_2 x$ and the random disturbance *u*. Calling μ_y to the systematic component, we can write:

$$\mu_y = \beta_1 + \beta_2 x \tag{2-2}$$

This equation is known as the *population regression function* (*PRF*) or *population line*. Therefore, as can be seen in figure 2.1, μ_y is a linear function of *x* with intercept β_1 and slope β_2.

The linearity means that a one-unit increase in *x* changes the *expected value* of *y* - $m_y = E(y)$ - by β_1 units.

Now, let us suppose we have a random sample of size *n* $\{(y_i, x_i)$: $i = 1,\ldots,n\}$ from the studied population. In figure 2.2 the scatter diagram, corresponding to these data, have been displayed.

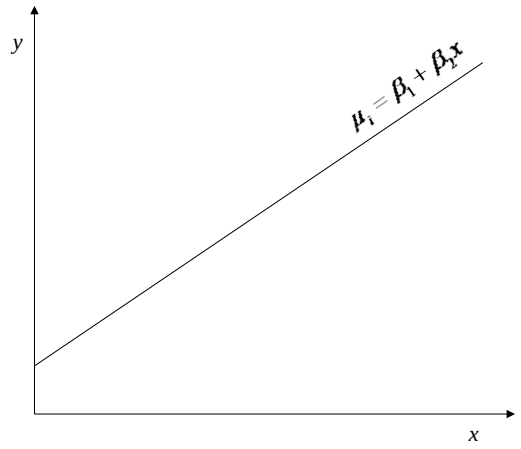


FIGURE 2.1. The population regression function. (PRF)

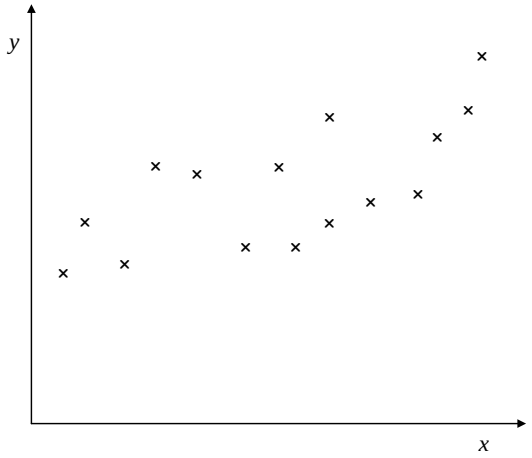


FIGURE 2.2. The scatter diagram.

We can express the population model for each observation of the sample:

$$y_i = \beta_1 + \beta_2 x_i + u_i \quad i = 1, 2, \cdots, n \tag{2-3}$$

In figure 2.3 the population regression function and the scatter diagram are put together, but it is important to keep in mind that although β_1 and β_2 are fixed, they are unknown. According to the model, it is possible to make the following decomposition from a theoretical point of view:

$$y_i = \mu_{yi} + u_i \quad i = 1, 2, \cdots, n \tag{2-4}$$

which is represented in figure 2.3 for the i^{th} observation. However, from an empirical point of view, it is not possible because β_1 and β_2 are unknown parameters and u_i is not observable.

2.1.2 Sample regression function

The basic idea of the regression model is to estimate the population parameters, β_2 and β_1, from a given sample.

The *sample regression function* (*SRF*) is the sample counterpart of the population regression function (*PRF*). Since the *SRF* is obtained for a given sample, a new sample will generate different estimates.

The *SRF*, which is an estimation of the PRF, given by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-5}$$

allows us to calculate the *fitted value* ($\hat{y}_i$) for *y* when $x = x_i$. In the *SRF* $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimators of the parameters β_1 and β_2 . For each x_i we have an observed value (y_i) and a fitted value ($\hat{y}_i$).

The difference between y_i and $\hat{y}_i$ is called the residual $\hat{u}_i$:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \tag{2-6}$$

In other words, the residual $\hat{u}_i$ is the difference between the sample value y_i and the fitted value of $\hat{y}_i$, as can be seen in figure 2.4. In this case, it is possible to calculate the decomposition:

$$y_i = \hat{y}_i + \hat{u}_i$$

for a given sample.

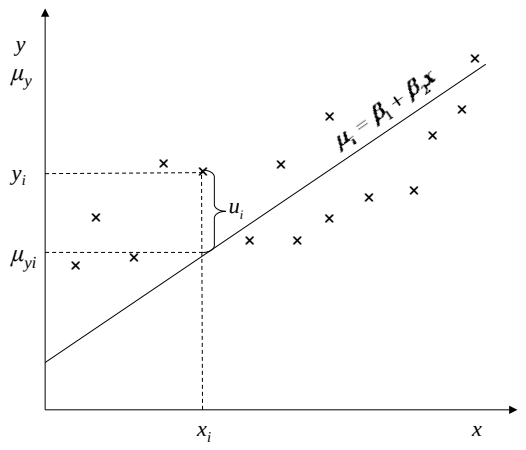


FIGURE 2.3. The population regression function and the scatter diagram.

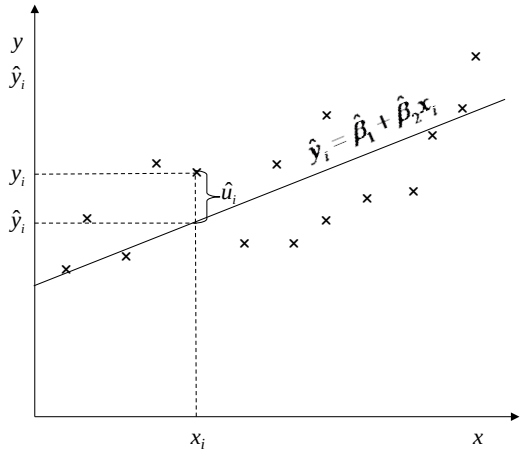


FIGURE 2.4. The sample regression function and the scatter diagram.

To sum up, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{y}_i$ and $\hat{u}_i$ are the sample counterpart of β_1, β_2, μ_{yi} and u_i respectively. It is possible to calculate $\hat{\beta}_1$ and $\hat{\beta}_2$ for a given sample, but the estimates will change for each sample. On the contrary, β_1 and β_2 are fixed, but unknown.

2.2 Obtaining the Ordinary Least Squares (OLS) Estimates

2.2.1 Different criteria of estimation

Before obtaining the least squares estimators, we are going to examine three alternative methods to illustrate the problem in hand. What these three methods have in common is that they try to minimize the residuals as a whole.

Criterion 1

The first criterion takes as estimators those values of $\hat{\beta}_1$ and $\hat{\beta}_2$ that make the sum of all the residuals as near to zero as possible. According to this criterion, the expression to minimize would be the following:

$$\text{Min}\left|\sum_{i=1}^{n}\hat{u}_i\right| \tag{2-7}$$

The main problem of this procedure is that the residuals of different signs can be compensated. Such a situation can be observed graphically in figure 2.5, in which three aligned observations are graphed, (x_1, y_1), (x_2, y_2) and (x_3, y_3). In this case the following happens:

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y_3 - y_1}{x_3 - x_1}$$

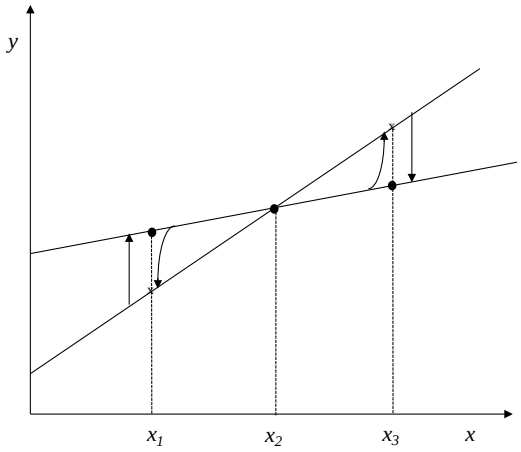


FIGURE 2.5. The problems of criterion 1.

If a straight line is fitted so that it passes through the three points, each one of the residuals will take value zero, and therefore

$$\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$$

This fit could be considered optimal. But it is also possible to obtain $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$, by rotating the straight line - from the point x_2, y_2 - in any direction, as figure 2.5 shows, because $\hat{u}_3 = -\hat{u}_1$. In other words, by rotating this way the result $\left|\sum_{i=1}^{3}\hat{u}_i = 0\right|$ is always obtained. This simple example shows that this criterion is not appropriate for the estimation of the parameters given that, for any set of observations, an infinite number of straight lines exist, satisfying this criterion.

Criterion 2

In order to avoid the compensation of positive residuals with negative ones, the absolute values from the residuals are taken. In this case, the following expression would be minimized:

$$\text{Min} \sum_{i=1}^{n} |\hat{u}_i| \tag{2-8}$$

Unfortunately, although the estimators thus obtained have some interesting properties, their calculation is complicated and requires resolving the problem of linear programming or applying a procedure of iterative calculation.

Criterion 3

A third procedure is to minimize the sum of the square residuals, that is to say,

$$\text{Min } S = \text{Min} \sum_{i=1}^{n} \hat{u}_i^2 \tag{2-9}$$

The estimators obtained are denominated least square estimators (*LS*), and they enjoy certain desirable statistical properties, which will be studied later on. On the other hand, as opposed to the first of the examined criteria, when we square the residuals their compensation is avoided, and the least square estimators are simple to obtain, contrary to the second of the criteria. It is important to indicate that, from the moment we square the residuals, we proportionally penalize the bigger residuals more than the smaller ones (if a residual is double the size of another one, its square will be four times greater). This characterizes the least square estimation with respect to other possible procedures.

2.2.2 Application of least square criterion

Now, we are going to look at the process of obtaining the *LS* estimators. The objective is to minimize the residual sum of the squares (*S*). To do this, we are firstly going to express *S* as a function of the estimators, using (2-6):

Therefore, we must

$$\underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}}\, S = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{t=1}^{T} \hat{u}_t^2 = \underset{\hat{\beta}_1,\hat{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \tag{2-10}$$

To minimize *S*, we differentiate partially with respect to $\hat{\beta}_1$ and $\hat{\beta}_2$:

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i$$

The *LS* estimators are obtained by equaling the previous derivatives to 0:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2-11}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) x_i = 0 \tag{2-12}$$

The equations (2-11) are denominated *normal equations* or *LS first order conditions*.

In operations with summations, the following rules must be taken into account:

$$\sum_{i=1}^{n} a = na$$

$$\sum_{i=1}^{n} ax_i = a\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Operating with the normal equations, we have

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} x_i \tag{2-13}$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2 \tag{2-14}$$

Dividing both sides of (2-13) by *n*, we have

$$\overline{y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{x} \tag{2-15}$$

Therefore

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2 \overline{x} \tag{2-16}$$

Substituting this value of $\hat{\beta}_1$ in the second normal equation (2-14), we have

$$\sum_{i=1}^{n} y_i x_i = (\overline{y} - \hat{\beta}_2 \overline{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i = \overline{y} \sum_{i=1}^{n} x_i - \hat{\beta}_2 \overline{x} \sum_{i=1}^{n} x_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i^2$$

Solving for $\hat{\beta}_2$ we have:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i - \overline{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i} \tag{2-17}$$

Or, as can be seen in appendix 2.1,

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{2-18}$$

Dividing the numerator and denominator of (2-18) by *n*, it can be seen that $\hat{\beta}_2$ is equal to the ratio between the two variables covariance and variance of *x*. Therefore, the sign of $\hat{\beta}_2$ is the same as the sign of the covariance.

Once $\hat{\beta}_2$ is calculated, then we can obtain $\hat{\beta}_1$ by using (2-16).

These are the *LS* estimators. Since other more complicated methods exist, also called least square methods, the method that we have applied is denominated ordinary least square (*OLS*), due to its simplicity.

In the precedent epigraphs $\hat{\beta}_1$ and $\hat{\beta}_2$ have been used to designate generic estimators. From now on, we will only designate *OLS* estimators with this notation.

EXAMPLE 2.1 Estimation of the consumption function

Given the Keynesian consumption function,

$$cons = \beta_1 + \beta_2 inc + u_i$$

we will estimate it using data from six households that appear in table 2.1.

TABLE 2.1. Data and calculations to estimate the consumption function.

Observ.	$cons_i$	inc_i	$cons_i \times inc_i$	inc_i^2	$cons_i - \overline{cons}$	$inc_i - \overline{inc}$	$(cons_i - \overline{cons}) \times (inc_i - \overline{inc})$	$(inc_i - \overline{inc})^2$
1	5	6	30	36	-4	-5	20	25
2	7	9	63	81	-2	-2	4	4
3	8	10	80	100	-1	-1	1	1
4	10	12	120	144	1	1	1	1
5	11	13	143	169	2	2	4	4
6	13	16	208	256	4	5	20	25
Sums	**54**	**66**	**644**	**786**	**0**	**0**	**50**	**60**

Calculating $\overline{cons}$ and $\overline{inc}$, and applying the formula (2-17), or alternatively (2-18), for the data table 2.1, we obtain

$$\overline{cons} = \frac{54}{6} = 9; \quad \overline{inc} = \frac{66}{6} = 11; \quad \text{(2-17): } \hat{\beta}_2 = \frac{644 - 9 \times 66}{786 - 11 \times 66} = 0.8\hat{3}; \quad \text{(2-18): } \hat{\beta}_2 = \frac{50}{60} = 0.8\hat{3}$$

Then by applying (2-16), we obtain $\hat{\beta}_1 = 9 - 0.8\hat{3} \times 11 = -0.1\hat{6}$

2.3 Some characteristics of *OLS* estimators

2.3.1 Algebraic implications of the estimation

The algebraic implications of the estimation are derived exclusively from the application of the *OLS* procedure to the simple linear regression model:

1. The sum of the OLS residuals is equal to 0:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \tag{2-19}$$

From the definition of residual

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \qquad i = 1, 2, \cdots, n \tag{2-20}$$

If we sum up the *n* observations, we get

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0 \tag{2-21}$$

which is precisely the first equation (2-11) of the system of normal equations.

Note that, if (2-19) holds, it implies that

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \tag{2-22}$$

and, dividing (2-19) and (2-22) by *n*, we obtain

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \tag{2-23}$$

2. The OLS line always goes through the mean of the sample ($\bar{x}, \bar{y}$).

Effectively, dividing the equation (2-13) by *n*, we have:

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x} \tag{2-24}$$

3. The sample cross product between each one of the regressors and the OLS residuals is zero.

That is to say,

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \tag{2-25}$$

We can see that (2-25) is equal to the second normal equation,

$$\sum_{i=1}^{n} x_i \hat{u}_i = \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0$$

given in (2-12).

4. The sample cross product between the fitted values ($\hat{y}$) and the OLS residuals is zero.

That is to say,

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{2-26}$$

Proof

Taking into account the algebraic implications 1 -(2-19)- and 3 -(2-25)-, we have

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_i) \hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_i \hat{u}_i = \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 = 0$$

2.3.2 Decomposition of the variance of *y*

By definition

$$y_i = \hat{y}_i + \hat{u}_i \tag{2-27}$$

Subtracting $\bar{y}$ on both sides of the previous expression (remember that $\bar{\hat{y}}$ is equal to $\bar{y}$), we have

$$y_i - \bar{y} = \hat{y}_i - \bar{\hat{y}} + \hat{u}_i$$

Squaring both sides:

$$\left[y_i - \bar{y}\right]^2 = \left[(\hat{y}_i - \bar{\hat{y}}) + \hat{u}_i\right]^2 = (\hat{y}_i - \bar{\hat{y}})^2 + \hat{u}_i^2 + 2\hat{u}_i(\hat{y}_i - \bar{\hat{y}})$$

Summing for all *i*:

$$\sum\left[y_i - \bar{y}\right]^2 = \sum(\hat{y}_i - \bar{\hat{y}})^2 + \sum\hat{u}_i^2 + 2\sum\hat{u}_i(\hat{y}_i - \bar{\hat{y}})$$

Taking into account the algebraic properties 1 and 4, the third term of the right hand side is equal to 0. Analytically,

$$\sum\hat{u}_i(\hat{y}_i - \bar{\hat{y}}) = \sum\hat{u}_i\hat{y}_i - \bar{\hat{y}}\sum\hat{u}_i = 0 \qquad (2\text{-}28)$$

Therefore, we have

$$\sum\left[y_i - \bar{y}\right]^2 = \sum(\hat{y}_i - \bar{\hat{y}})^2 + \sum\hat{u}_i^2 \qquad (2\text{-}29)$$

In words,

Total sum of squares (*TSS*) =

Explained sum of squares (*ESS*)+Residual sum of squares (*RSS*)

It must be stressed that it is necessary to use the relation (2-19) to assure that (2-28) is equal to 0. We must remember that (2-19) is associated to the first normal equation: that is to say, to the equation corresponding to the intercept. If there is no intercept in the fitted model, then in general the decomposition obtained will not be fulfilled (2-29).

This decomposition can be made with variances, by dividing both sides of (2-29) by *n*:

$$\frac{\sum(y_i - \bar{y})^2}{n} = \frac{\sum(\hat{y}_i - \bar{\hat{y}})^2}{n} + \frac{\sum\hat{u}_i^2}{n} \qquad (2\text{-}30)$$

In words,

Total variance=explained variance+ residual variance

2.3.3 Goodness of fit: Coefficient of determination (R^2)

A priori we have obtained the estimators minimizing the sum of square residuals.

Once the estimation has been done, we can see how well our sample regression line fits our data.

The measures that indicate how well the sample regression line fits the data are denominated *goodness of fit* measures. We are going to look at the most well-known measure, which is called *coefficient of determination* or the *R*-square (R^2). This measure is defined in the following way:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{2-31}$$

Therefore, R^2 is the proportion of the total sum of squares (*TSS*) which is explained by the regression (*ESS*): that is to say, which is explained by the model. We can also say that $100\,R^2$ is the percentage of the sample variation in *y* explained by *x*.

Alternatively, taking into account (2-29), we have:

$$\sum (\hat{y}_i - \bar{\hat{y}})^2 = \sum (y_i - \bar{y})^2 - \sum \hat{u}_i^2$$

Substituting in (2-31), we have

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\sum (y_i - \bar{y})^2 - \sum \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum \hat{u}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} \tag{2-32}$$

Therefore, R^2 is equal to 1 minus the proportion of the total sum of squares (*TSS*) that is non-explained by the regression (*RSS*).

According to the definition of R^2, the following must be accomplished

$$0 \le R^2 \le 1$$

Extreme cases:

a) If we have a perfect fit, then $\hat{u}_i = 0 \quad \forall i$. This implies that

$$\hat{y}_i = y_i \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = \sum (y_i - \bar{y})^2 \quad \Rightarrow R^2 = 1$$

b) If $\hat{y}_i = c \quad \forall i$, it implies that

$$\bar{\hat{y}} = c \quad \Rightarrow \hat{y}_i - \bar{\hat{y}} = c - c = 0 \quad \forall i \quad \Rightarrow \sum (\hat{y}_i - \bar{\hat{y}})^2 = 0 \quad \Rightarrow R^2 = 0$$

If R^2 is close to 0, it implies that we have a poor fit. In other words, very little variation in *y* is explained by *x*.

In many cases, a high R^2 is obtained when the model is fitted using time series data, due to the effect of a common trend. On the contrary, when we use cross sectional data a low value is obtained in many cases, but it does not mean that the fitted model is bad.

What is the relationship between the coefficient of determination and the coefficient of correlation studied in descriptive statistics? The coefficient of determination is equal to the squared coefficient of correlation, as can be seen in appendix 2.2:

$$r_{xy}^2 = R^2 \tag{2-33}$$

(This equality is only valid in the simple regression model, but not in multiple regression model).

EXAMPLE 2.2 Fulfilling algebraic implications and calculating R^2 in the consumption function

In column 2 of table 2.2, $\bar{c}ons_i$ is calculated; in columns 3, 4 and 5, you can see the fulfillment of algebraic implications 1, 3 and 4 respectively. The remainder of the columns shows the calculations to obtain

$$TSS = 42 \qquad ESS = 41.67 \qquad RSS = 42 - 41.67 = 0.33 \qquad R^2 = \frac{41.67}{42} = 0.992$$

or, alternatively, $R^2 = 1 - \frac{0.33}{42} = 0.992$

TABLE 2.2. Data and calculations to estimate the consumption function.

Observ.	$\bar{c}ons_i$	$\hat{u}_i$	$\hat{u}_i \times inc_i$	$\bar{c}ons_i \acute{}\ \hat{u}_i$	$cons_i^2$	$(cons_i - \overline{cons})^2$	$\bar{c}ons_i^2$	$(\bar{c}ons_i - \overline{\bar{c}ons})^2$
1	4.83	0.17	1.00	0.81	25	16	23.36	17.36
2	7.33	-0.33	-3.00	-2.44	49	4	53.78	2.78
3	8.17	-0.17	-1.67	-1.36	64	1	66.69	0.69
4	9.83	0.17	2.00	1.64	100	1	96.69	0.69
5	10.67	0.33	4.33	3.56	121	4	113.78	2.78
6	13.17	-0.17	-2.67	-2.19	169	16	173.36	17.36
	54.00	**0.00**	**0.00**	**0.00**	**528**	**42**	**527.67**	**41.67**

2.3.4 Regression through the origin

If we force the regression line to pass through the point (0,0), we are constraining the intercept to be zero, as can be seen in figure 2.6. This is called a regression through the origin.

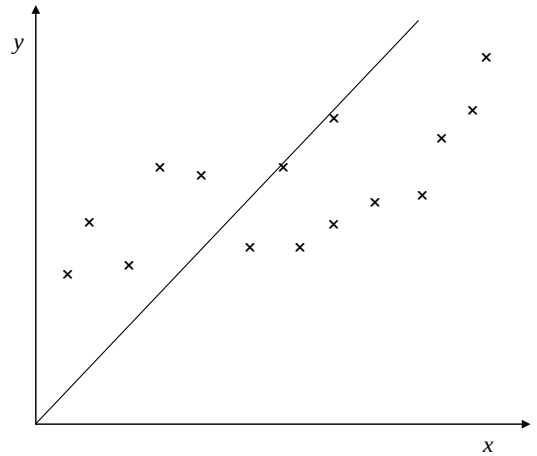


FIGURE 2.6. A regression through the origin.

Now, we are going to estimate a regression line through the origin. The fitted model is the following:

$$\tilde{y}_i = \tilde{\beta}_2 x_i \tag{2-34}$$

Therefore, we must minimize

$$\underset{\tilde{\beta}_2}{\text{Min}}\, S = \underset{\tilde{\beta}_2}{\text{Min}} \sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i)^2 \tag{2-35}$$

To minimize *S*, we differentiate with respect to $\tilde{\beta}_2$ and equal to 0:

$$\frac{dS}{d\tilde{\beta}_2} = -2\sum_{i=1}^{n} (y_i - \tilde{\beta}_2 x_i) x_i = 0 \tag{2-36}$$

Solving for $\tilde{\beta}_2$

$$\tilde{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} \tag{2-37}$$

Another problem with fitting a regression line through the origin is that the following generally happens:

$$\sum (y_i - \bar{y})^2 \neq \sum (\hat{y}_i - \bar{\hat{y}})^2 + \sum \hat{u}_i^2$$

If the decomposition of the variance of *y* in two components (explained and residual) is not possible, then the R^2 is meaningless. This coefficient can take values that are negative or greater than 1 in the model without intercept.

To sum up, an intercept must always be included in the regressions, unless there are strong reasons against it supported by the economic theory.

2.4 Units of measurement and functional form

2.4.1 Units of Measurement

Changing the units of measurement (change of scale) in x

If *x* is multiplied/divided by a constant, $c \neq 0$, then the *OLS* slope is divided/multiplied by the same constant, *c*. Thus

$$\hat{y}_i = \hat{\beta}_1 + \left[\frac{\hat{\beta}_2}{c}\right](x_i \times c) \tag{2-38}$$

EXAMPLE 2.3

Let us suppose the following estimated consumption function, in which both variables are measured in thousands of euros:

$$\hat{cons}_i = 0.2 + 0.85\,' \; inc_i \tag{2-39}$$

If we now express income in euros (multiplication by 1000) and call it *ince*, the fitted model with the new units of measurement of income would be the following:

$$\hat{cons}_i = 0.2 + 0.00085 \times ince_i$$

As can be seen, changing the units of measurement of the explanatory variable does not affect the intercept.

Changing the units of measurement (change of scale) in y

If *y* is multiplied/divided by a constant, $c \neq 0$, then the *OLS* slope and intercept are both multiplied/divided by the same constant, *c*. Thus,

$$(\hat{y}_i \times c) = (\hat{\beta}_1 \times c) + (\hat{\beta}_2 \times c)x_i \tag{2-40}$$

EXAMPLE 2.4

If we express consumption in euros (multiplication by 1000) in model (2-39), and call it *conse*, the fitted model with the new units of measurement of consumption would be the following:

$$\widehat{conse}_i = 200 + 850 \times inc_i$$

Changing the origin

If one adds/subtracts a constant *d* to *x* and/or *y*, then the *OLS* slope is not affected. However, changing the origin of either *x* and/or *y* affects the intercept of the regression.

If one subtracts a constant *d* to *x*, the intercept will change in the following way:

$$\hat{y}_i = (\hat{\beta}_1 + \hat{\beta}_2 \times d) + \hat{\beta}_2(x_i - d) \tag{2-41}$$

If one subtracts a constant *d* to *y*, the intercept will change in the following way:

$$\hat{y}_i - d = (\hat{\beta}_1 - d) + \hat{\beta}_2 x_i \tag{2-42}$$

EXAMPLE 2.5

Let us suppose that the average income is 20 thousand euros. If we define the variable $incd_i = inc_i - \overline{inc}$ and both variables are measured in thousands of euros, the fitted model with this change in the origin will be the following:

$$\widehat{cons}_i = (0.2 + 0.85 \times 20) + 0.85 \times (inc_i - 20) = 17.2 + 0.85 \times incd_i$$

EXAMPLE 2.6

Let us suppose that the average consumption is 15 thousands euros. If we define the variable $consd_i = cons_i - \overline{cons}$ and both variables are measured in euros, the fitted model with this change in the origin will be the following:

$$\widehat{cons}_i - 15 = 0.2 - 15 + 0.85 \times inc_i$$

that is to say,

$$\widehat{consd}_i = -\ 14.8 + 0.85' \ inc_i$$

Note that R^2 is invariant to changes in the units of *x* and/or *y*, and also is invariant to the origin of the variables.

2.4.2 Functional Form

In many cases linear relationships are not adequate for economic applications. However, in the simple regression model we can incorporate nonlinearities (in variables) by appropriately redefining the dependent and independent variables.

Some definitions

Now we are going to look at some definitions of variation measures that will be useful in the interpretation of the coefficients corresponding to different functional forms. Specifically, we will look at the following: proportional change and change in logarithms.

The *proportional change* (or relative variation rate) between x_1 and x_0 is given by:

$$\frac{\Delta x_1}{x_0} = \frac{x_1 - x_0}{x_0} \tag{2-43}$$

Multiplying a proportional change by 100, we obtain a *proportional change in %*. That is to say:

$$100\frac{\Delta x_1}{x_0}\% \tag{2-44}$$

The *change in logarithms* and *change in logarithms in %* between x_1 and x_0 are given by

$$\begin{gathered}\Delta \ln(x) = \ln(x_1) - \ln(x_0) \\ 100\Delta \ln(x)\%\end{gathered} \tag{2-45}$$

The *change in logarithms* is an approximation to the *proportional change*, as can be seen in appendix 2.3. This approximation is good when the proportional change is small, but the differences can be important when the proportional change is big, as can seen in table 2.3.

TABLE 2.3. Examples of proportional change and change in logarithms.

x_1	202	210	220	240	300
x_0	200	200	200	200	200
Proportional change in %	1%	5.0%	10.0%	20.0%	50.0%
Change in logarithms in %	1%	4.9%	9.5%	18.2%	40.5%

Elasticity is the ratio of the relative changes of two variables. If we use proportional changes, the elasticity of the variable *y* with respect to the variable *x* is given by

$$\varepsilon_{y/x} = \frac{\Delta y / y_0}{\Delta x / x_0} \tag{2-46}$$

If we use changes in logarithms and consider infinitesimal changes, then the elasticity of the variable *y* with respect to a variable *x* is given by

$$\varepsilon_{y/x} = \frac{dy / y}{dx / x} = \frac{d \ln(y)}{d \ln(x)} \tag{2-47}$$

In econometric models, elasticity is generally defined by using (2-47).

Alternative functional forms

The *OLS* method can also be applied to models in which the endogenous variable and/or the exogenous variable have been transformed. In the presentation of the model (2-1) we said that the exogenous variable and regressor were equivalent terms. But from now on, a regressor is the specific form in which an exogenous variable appears in the equation. For example, in the model

$$y = \beta_1 + \beta_2 \ln(x) + u$$

the exogenous variable is *x*, but the regressor is ln(*x*).

In the presentation of the model (2-1) we also said that the endogenous variable and the regressand were equivalent. But from now on, the regressand is the specific form in which an endogenous variable appears in the equation. For example, in the model

$$\ln(y) = \beta_1 + \beta_2 x + u$$

the endogenous variable is *y*, but the regressand is ln(*y*).

Both models are linear in the parameters, although they are not linear in the variable *x* (the first one) or in the variable *y* (the second one). In any case, if a model is linear in the parameters, it can be estimated by applying the *OLS* method. On the contrary, if a model is not linear in the parameters, iterative methods must be used in the estimation.

However, there are certain nonlinear models which, by means of suitable transformations, can become linear. These models are denominated linearizables.

Thus, on some occasions potential models are postulated in economic theory, such as the well-known Cobb-Douglas production function. A potential model with a unique explanatory variable is given by

$$y = e^{\beta_1} x^{\beta_2}$$

If we introduce the disturbance term in a multiplicative form, we obtain:

$$y = e^{\beta_1} x^{\beta_2} e^{u} \tag{2-48}$$

Taking natural logarithms on both sides of (2-48), we obtain a linear model in the parameters:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + u \tag{2-49}$$

On the contrary, if we introduce the disturbance term in an additive form, we obtain

$$y = e^{\beta_1} x^{\beta_2} + u$$

In this case, there is no transformation which allows this model to be turned into a linear model. This is a non-linearizable model.

Now we are going to consider some models with alternative functional forms, all of which are linear in the parameters. We will look at the interpretation of the coefficient $\hat{\beta}_2$ in each case.

a) Linear model

The $\hat{\beta}_2$ coefficient measures the effect of the regressor *x* on *y*. Let us look at this in detail. The observation *i* of the sample regression function is given according to (2-5) by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{2-50}$$

Let us consider the observation *h* of the fitted model whereupon the value of the regressor and, consequently, of the regressand has changed with respect to (2-50):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_h \tag{2-51}$$

Subtracting (2-51) from (2-50), we see that *x* has a linear effect on $\hat{y}$:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x \tag{2-52}$$

where $\Delta\hat{y} = \hat{y}_i - \hat{y}_h$ and $\Delta x = x_i - x_h$

Therefore, $\hat{\beta}_2$ is the change in *y* (in the units in which *y* is measured) by a unit change of *x* (in the units in which *x* is measured).

For example, if income increases by 1 unit, consumption will increase by 0.85 units in the fitted function (2-39).

The linearity of this model implies that a one-unit change in *x* always has the same effect on *y*, regardless of the value of *x* considered.

EXAMPLE 2.7 **Quantity sold of coffee as a function of its price. Linear model**

In a marketing experiment[1] the following model has been formulated to explain the quantity sold of coffee per week (*coffqty*) as a function of the price of coffee (*coffpric*).

$$coffqty = \beta_1 + \beta_2 coffpric + u$$

The variable *coffpric* takes the value 1 for the usual price, and also 0.95 and 0.85 in two price actions whose effects are under investigation. This experiment lasted 12 weeks. *coffqty* is expressed in thousands of units and *coffpric* in French francs. Data appear in table 2.4 and in work file *coffee1*.

The fitted model is the following:

$$\widehat{coffqty} = 774.9 - 693.33 coffpric \quad R^2 = 0.95 \quad n = 12$$

TABLE 2.4. Data on quantities and prices of coffee.

week	*coffpric*	*coffqty*
1	1.00	89
2	1.00	86
3	1.00	74
4	1.00	79
5	1.00	68
6	1.00	84
7	0.95	139
8	0.95	122
9	0.95	102
10	0.85	186
11	0.85	179
12	0.85	187

Interpretation of the coefficient $\hat{\beta}_2$: if the price of coffee increases by 1 French franc, the quantity sold of coffee will decrease by 693.33 thousands of units. As the price of coffee is a small

[1]The data of this exercise were obtained from a controlled marketing experiment in stores in Paris on coffee expenditure, as reported in A. C. Bemmaor and D. Mouchoux, "Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment", *Journal of Marketing Research*, 28 (1991), 202–14.

magnitude, the following interpretation is preferable: if the price of coffee increases by 1 cent of a French franc, the quantity sold will decrease by 6.93 thousands of units.

EXAMPLE 2.8 Explaining market capitalization of Spanish banks. Linear model

Using data from Bolsa de Madrid (***Madrid Stock Exchange***) on August 18, 1995 (file *bolmad95*, the first 20 observations), the following model has been estimated to explain the market capitalization of banks and financial institutions:

$$\widehat{marktval} = 29.42 + 1.219 bookval$$

$$R^2=0.836 \qquad n=20$$

where

- ***marktval*** is the capitalization the market value of a company. It is calculated by multiplying the price of the stock by the number of stocks issued.
- *bookval* is the book value or the net worth of the company. The book value is calculated as the difference between a company's assets and its liabilities.
- Data on ***marktval*** and *bookval* are expressed in millions of pesetas.

Interpretation of the coefficient β_2: if the book value of a bank increases by 1 million pesetas, the market capitalization of this bank will increase by 1.219 million of pesetas.

b) Linear-log model

A linear-log model is given by

$$y = \beta_1 + \beta_2 \ln(x) + u \tag{2-53}$$

The corresponding fitted function is the following:

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 \ln(x) \tag{2-54}$$

Taking first order differences in (2-54), and then multiplying and dividing the right hand side by 100, we have

$$\Delta\hat{y} = \frac{\hat{\beta}_2}{100} 100 \times \Delta \ln(x)\%$$

Therefore, if x increases by 1%, then $\hat{y}$ will increase by $(\hat{\beta}_2 / 100)$ units.

c) Log-linear model

A log-linear model is given by

$$\ln(y) = \beta_1 + \beta_2 x + u \tag{2-55}$$

The above model can be obtained by taking natural logs on both sides of the following model:

$$y = \exp(\beta_1 + \beta_2 x + u)$$

. For this reason, the model (2-55) is also called exponential.

The corresponding sample regression function to (2-55) is the following:

$$\widehat{\ln(y)} = \hat{b}_1 + \hat{b}_2 x \tag{2-56}$$

Taking first order differences in (2-56), and then multiplying both sides by 100, we have

$$100 \cdot \Delta \widehat{\ln}(y)\% = 100 \cdot \hat{b}_2 \Delta x$$

Therefore, if x increases by 1 unit, then $\hat{y}$ will increase by $100\,\hat{\beta}_2$ %.

d) Log-log model

The model given in (2-49) is a log-log model or, before the transformation, a potential model (2-48). This model is also called a constant elasticity model.

The corresponding fitted model to (2-49) is the following:

$$\widehat{\ln}(y) = \hat{b}_1 + \hat{b}_2 \ln(x) \qquad (2\text{-}57)$$

Taking first order differences in (2-57), we have

$$\Delta \widehat{\ln}(y) = \hat{b}_2 \Delta \ln(x)$$

Therefore, if x increases by 1%, then $\hat{y}$ will increase by $\hat{\beta}_2$ %. It is important to remark that, in this model, $\hat{\beta}_2$ is the estimated elasticity of y with respect to x, for any value of x and y. Consequently, in this model the elasticity is constant.

In annex 1 in a study case on the Engel curve for demand of dairy, six alternative functional forms are analyzed.

EXAMPLE 2.9 Quantity sold of coffee as a function of its price. Log- log model (Continuation example 2.7)

As an alternative to the linear model the following log-log model has been fitted:

$$\widehat{\ln(coffqty)} = 4.415 - 5.132 \ln(coffpric) \quad R^2 = 0.90 \quad n = 12$$

Interpretation of the coefficient $\hat{\beta}_2$: if the price of coffee increases by 1%, the quantity sold of coffee will decrease by 5.13%. In this case $\hat{\beta}_2$ is the estimated demand/price elasticity.

EXAMPLE 2.10 Explaining market capitalization of Spanish banks. Log-log model (Continuation example 2.8)

Using data from example 2.8, the following log-log model has been estimated:

$$\widehat{\ln(marktval)} = 0.6756 + 0.938 \ln(bookval)$$

$$R^2 = 0.928 \qquad n = 20$$

Interpretation of the coefficient $\hat{\beta}_2$: if the book value of a bank increases by 1%, the market capitalization of this bank will increase by 0.938%. In this case $\hat{\beta}_2$ is the estimated market value/book value elasticity.

In table 2.5 and for the fitted model, the interpretation of $\hat{\beta}_2$ in these four models is shown. If we are considering the population model, the interpretation of β_2 is the same but taking into account that Δu must be equal to 0.

TABLE 2.5. Interpretation of $\hat{\beta}_2$ in different models.

Model	If *x* increases by	then *y* will increase by
linear	1 unit	$\hat{\beta}_2$ units
linear-log	1%	$(\hat{\beta}_2/100)$ units
log-linear	1 unit	$(100\hat{\beta}_2)\%$
log-log	1%	$\hat{\beta}_2\%$

2.5 Assumptions and statistical properties of *OLS*

We are now going to study the statistical properties of *OLS* estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. But first we need to formulate a set of statistical assumptions. Specifically, the set of assumptions that we are going to formulate are called *classical linear model assumptions (CLM).* It is important to note that *CLM assumptions* are simple and that the *OLS* estimators have, under these assumptions, very good properties.

2.5.1 Statistical assumptions of the CLM in simple linear regression

a) Assumption on the functional form

1) The relationship between the regressand, the regressor and the random disturbance is linear in the parameters:

$$y = \beta_1 + \beta_2 x + u \tag{2-58}$$

The regressand and the regressors can be any function of the endogenous variable and the explanatory variables, respectively, provided that among regressors and regressand there is a linear relationship, i.e. the model is linear in the parameters. The additivity of the disturbance guarantees the linear relationship with the rest of the elements.

b) Assumptions on the regressor x

2) The values of x are fixed in repeated sampling:

According to this assumption, each observation of the regressor takes the same value for different samples of the regressand. This is a strong assumption in the case of the social sciences, where in general it is not possible to experiment. Data are obtained by observation, not by experimentation. It is important to remark that the results obtained using this assumption would remain virtually identical if we assume the repressors are stochastic, provided the additional assumption of independence between the regressors and the random disturbance is fulfilled. This alternative assumption can be formulated as:

2) The regressor x is distributed independently of the random disturbance.*

In any case, throughout this chapter and the following ones we will adopt assumption 2.

3) The regressor x does not contain measurement errors

This is an assumption that is not often fulfilled in practice, since the measurement instruments are unreliable in economy. Think, for example, of the multitude of errors that can be made in the collection of information, through surveys on families.

4) The sample variance of x is different from 0 and has a finite limit as n tends to infinity

Therefore, this assumption implies that

$$S_X^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n} \neq 0 \qquad (2\text{-}59)$$

c) Assumptions on the parameters

5) The parameters β_1 and β_2 are fixed

If this assumption is not adopted, the regression model would be very difficult to handle. In any case, it may be acceptable to postulate that the model parameters are stable over time (if it is not a very long period) or space (if it is relatively limited).

d) Assumptions on the random disturbances

6) The disturbances have zero mean,

$$E(u_i) = 0, \quad i = 1, 2, 3, \ldots, n \qquad (2\text{-}60)$$

This is not a restrictive assumption, since we can always use β_1 to normalize $E(u)$ to 0. Let us suppose, for example, that $E(u) = 4$. We could then redefine the model in the following way:

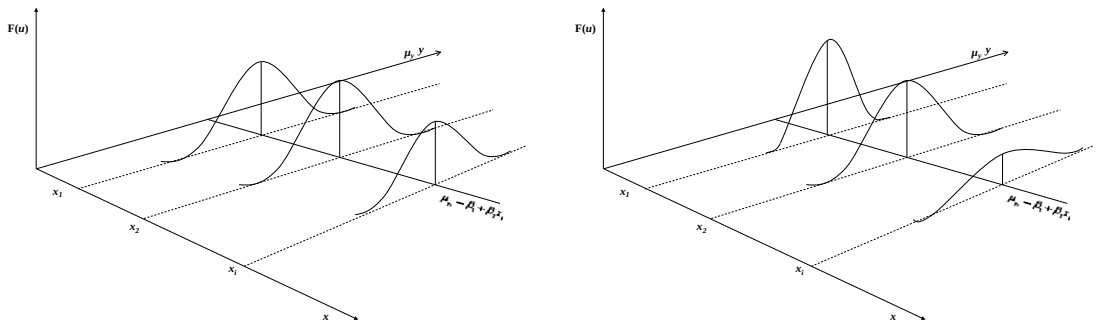
$$y = (\beta_1 + 4) + \beta_2 x + v$$

where $v = u - 4$. Therefore, the expectation of the new disturbance, *v*, is 0 and the expectation of *u* has been absorbed by the intercept.

7) The disturbances have a constant variance

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \qquad (2\text{-}61)$$

This assumption is called the *homoskedasticity* assumption**.** The word comes from the Greek: *homo* (equal) and *skedasticity* (spread). This means that the variation of *y* around the regression line is the same across the *x* values; that is to say, it neither increases or decreases as *x* varies. This can be seen in figure 2.7, part a), where disturbances are homoskedastic.



a) **b)**

FIGURE 2.7. Random disturbances: a) homoskedastic; b) heteroskedastic.

If this assumption is not satisfied, as happens in part b) of figure 2.7, the *OLS* regression coefficients are not efficient. Disturbances in this case are heteroskedastic (*hetero* means different).

8) The disturbances with different subscripts are not correlated with each other (no autocorrelation assumption):

$$E(u_i u_j) = 0 \qquad i \neq j \tag{2-62}$$

That is, the disturbances corresponding to different individuals or different periods of time are not correlated with each other. This assumption of *no autocorrelation* or *no serial correlation*, like the previous one, is testable a posteriori. The transgression occurs quite frequently in models using time series data.

9) The disturbance u is normally distributed

Taking into account assumptions 6, 7 y 8, we have

$$u_i \sim NID(0, \sigma^2) \qquad i = 1, 2, \cdots, n \tag{2-63}$$

where *NID* states for *normally independently distributed*.

The reason for this assumption is that if *u* is normally distributed, so will *y* and the estimated regression coefficients, and this will be useful in performing tests of hypotheses and constructing confidence intervals for β_1 and β_2. The justification for the assumption depends on the Central Limit Theorem. In essence, this theorem states that, if a random variable is the composite result of the effects of an indefinite number of variables, it will have an approximately normal distribution even if its components do not, provided that none of them is dominant.

2.5.2 Desirable properties of the estimators

Before examining the properties of *OLS* estimators under the statistical assumptions of the *CLM*, we pose the following question: what are the desirable properties for an estimator?

Two desirable properties for an estimator are that it is unbiased and its variance is as small as possible. If this occurs, the inference process will be carried out in optimal conditions.

We will illustrate these properties graphically. Consider first the property of unbiasedness. In Figures 2.8 and 2.9 the density functions of two hypothetical estimators obtained by two different methods are shown.

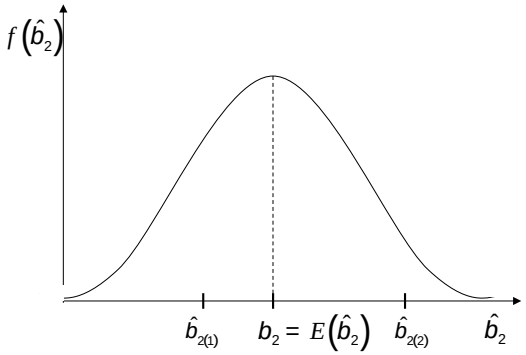


FIGURE 2.8. Unbiased estimator.

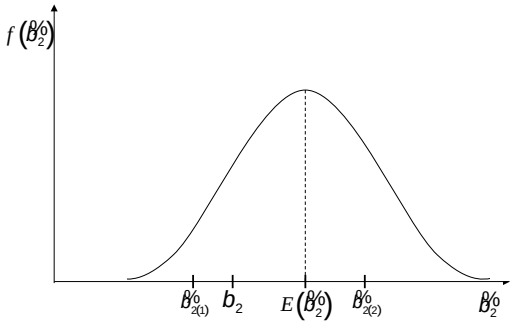


FIGURE 2.9. Biased estimator.

The estimator $\hat{b}_2$ is unbiased, i.e., its expected value is equal to the parameter that is estimated, β_2. The estimator $\hat{b}_2$ is a random variable. In each sample of *y*'s – the *x*'s are fixed in a repeated sample according to assumption 2- $\hat{b}_2$ taking a different value, but *on average* is equal to the parameter β_2, bearing in mind the infinite number of values $\hat{b}_2$ can take. In each sample of *y*'s a specific value of $\hat{b}_2$, that is to say, an *estimation* of $\hat{b}_2$ is obtained. In figure 2.8 two estimations of β_2 ($\hat{b}_{2(1)}$ and $\hat{b}_{2(2)}$) are obtained. The first estimate is relatively close to β_2, while the second one is much farther away. In any case, unbiasedness is a desirable property because it ensures that, on average, the estimator is centered on the parameter value.

The estimator $\tilde{b}_2$ is biased, since its expectation is not equal to β_2. The bias is precisely $E(\tilde{b}_2) - b_2$. In this case two hypothetical estimates, $\tilde{b}_{2(1)}$ and $\tilde{b}_{2(2)}$, are represented in figure 2.9. As can be seen $\tilde{b}_{2(1)}$ is closer to β_2 than the unbiased estimator $\hat{b}_{2(1)}$, but this is a matter of chance. In any case, when it is biased, it is not centered on the parameter value. An unbiased estimator will always be preferable, regardless of what happens in a specific sample, because it has no systematic deviation from the parameter value.

Another desirable property is efficiency. This property refers to the variance of the estimators. In figures 2.10 and 2.11 two hypothetical unbiased estimators, which are also called $\hat{b}_2$ and $\tilde{b}_2$, are represented. The first one has a smaller variance than the second one.

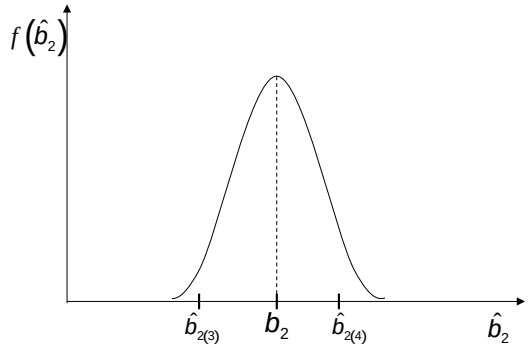


FIGURE 2.10. Estimator with small variance.

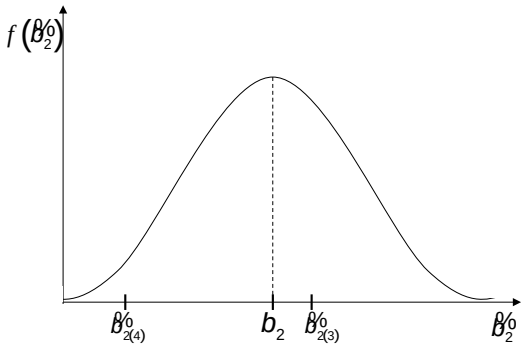


FIGURE 2.11. Estimator with big variance.

In both figures we have represented two estimates: $\hat{b}_{2(3)}$ and $\hat{b}_{2(4)}$ for the estimator with the smallest variance; and $\tilde{b}_{2(3)}$ and $\tilde{b}_{2(4)}$ for the estimator with the greatest variance. To highlight the role played by chance, the estimate that is closer to β_2 is precisely $\tilde{b}_{2(3)}$. In any case, it is preferable that the variance of the estimator is as small as possible. For example, when using the estimator $\hat{b}_2$ it is practically impossible that an estimate is so far from β_2 as it is in the case of $\hat{b}_2$, because the range of $\hat{b}_2$ is much smaller than the range of $\tilde{b}_2$

2.5.3 Statistical properties of *OLS* estimators

Under the above assumptions, the *OLS* estimators possess some ideal properties. Thus, we can say that the *OLS* are the best linear unbiased estimators.

Linearity and unbiasedness of the OLS

The *OLS* estimator $\hat{b}_2$ is unbiased. In appendix 2.4 we prove that $\hat{b}_2$ is an *unbiased* estimator using implicitly assumptions 3, 4 and 5, and explicitly assumptions 1, 2 and 6. In that appendix we can also see that $\hat{b}_2$ is a linear estimator using assumptions 1 and 2.

Similarly, one can show that the *OLS* estimator $\hat{b}_1$ is also unbiased. Remember that unbiasedness is a general property of the estimator, but in a given sample we may be "near" or "far" from the true parameter. In any case, its distribution will be centered at the population parameter.

Variances of the OLS estimators

Now we know that the sampling distribution of our estimator is centered around the true parameter. How spread out is this distribution? The variance (which is a measure of dispersion) of an estimator is an indicator of the accuracy of the estimator.

In order to obtain the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$, assumptions 7 and 8 are needed, in addition to the first six assumptions. These variances are the following:

$$Var(\hat{\beta}_1) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad Var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2-64}$$

Appendix 2.5 shows how the variance for $\hat{\beta}_2$ is obtained.

OLS estimators are BLUE

The *OLS* estimators have the least variance in the class of all linear and unbiased estimators. For this reason it is said that *OLS* estimators are the *best linear unbiased estimators (BLUE),* as illustrated in figure 2.12. This property is known as the Gauss–Markov theorem. For proof of this theorem assumptions 1-8 are used, as can be seen in appendix 2.6. This set of assumptions is known as the Gauss–Markov assumptions.

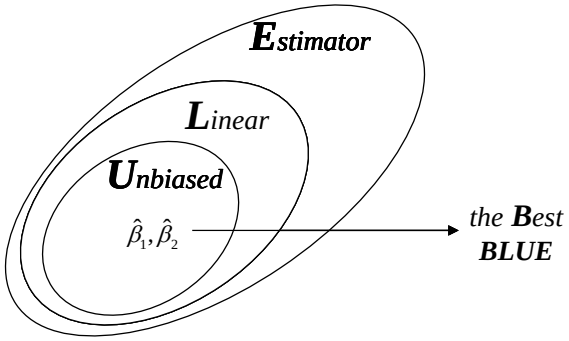


FIGURE 2.12. The *OLS* estimator is BLUE.

Estimating the disturbance variance and the variance of estimators

We do not know what the value of the disturbance variance, σ^2, is and thus we have to estimate it. But we cannot estimate it from the disturbances u_i, because they are not observable. Instead, we have to use the *OLS* residuals ($\hat{u}_i$).

The relation between disturbances and residuals is given by

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + u_i - \hat{\beta}_1 - \hat{\beta}_2 x_i \\ &= u_i - \left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i \end{aligned} \tag{2-65}$$

Hence $\hat{u}_i$ is not the same as u_i, although the difference between them- $\left(\hat{\beta}_1 - \beta_1\right) - \left(\hat{\beta}_2 - \beta_2\right) x_i$ - does have an expected value of zero. Therefore, a first estimator of σ^2 could be the residual variance:

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n} \tag{2-66}$$

However, this estimator is biased, essentially because it does not account for the two following restrictions that must be satisfied by the *OLS* residuals in the simple regression model:

$$\begin{cases} \sum_{i=1}^{n} \hat{u}_i = 0 \\ \sum_{i=1}^{n} x_i \hat{u}_i = 0 \end{cases} \tag{2-67}$$

One way to view these restrictions is the following: if we know *n*–2 of the residuals, we can get the other two residuals by using the restrictions implied by the normal equations.

Thus, there are only *n*–2 degrees of freedom in the *OLS* residuals, as opposed to *n* degrees of freedom in the disturbances. In the unbiased estimator of σ^2 shown below an adjustment is made taking into account the degrees of freedom:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \tag{2-68}$$

Under assumptions 1-8 (Gauss-Markov assumptions), and as can be seen in appendix 7, we obtain

$$E(\hat{\sigma}^2) = \sigma^2 \tag{2-69}$$

If $\hat{\sigma}^2$ is plugged into the variance formulas, we then have unbiased estimators of *var*($\hat{\beta}_1$) and *var*($\hat{\beta}_2$)

The natural estimator of σ is $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ and is called the *standard error of the regression*. The square root of the variance of $\hat{\beta}_2$ is called the *standard deviation* of $\hat{\beta}_2$, that is to say,

$$sd(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \tag{2-70}$$

Therefore, its natural estimator is called the *standard error* of $\hat{\beta}_2$:

$$se(\hat{\beta}_2) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}} \tag{2-71}$$

Note that $se(\hat{\beta}_2)$, due to the presence of the estimator $\hat{\sigma}$ in (2-71), is a random variable as is $\hat{\beta}_2$. The standard error of any estimate gives us an idea of how precise the estimator is.

Consistency of OLS and other asymptotic properties

Sometimes it is not possible to obtain an unbiased estimator. In any case *consistency* is a minimum requirement for an estimator. According to an intuitive

approach, consistency means that as $n \to \infty$, the density function of the estimator collapses to the parameter value. This property can be expressed for the estimator $\hat{\beta}_2$ as:

$$\operatorname{plim}_{n\to\infty} \hat{\beta}_2 = \beta_2 \tag{2-72}$$

where plim means probability limit. In other words, $\hat{\beta}_2$ converges in probability to β_2.

Note that the properties of unbiasedness and consistency are conceptually different. The property of unbiasedness can hold for any sample size, whereas consistency is strictly a large-sample property or an *asymptotic property*.

Under assumptions 1 through 6, the *OLS* estimators, $\hat{\beta}_1$ and $\hat{\beta}_2$, are consistent. The proof for $\hat{\beta}_2$ can be seen in appendix 2.8.

Other asymptotic properties of $\hat{\beta}_1$ and $\hat{\beta}_2$: Under the Gauss-Markov assumptions 1 through 8, $\hat{\beta}_1$ and $\hat{\beta}_2$ are *asymptotically normally distributed* and also *asymptotically efficient* within the class of consistent and asymptotically normal estimators.

OLS estimators are maximum likelihood estimators (ML) and minimum variance unbiased estimators (MVUE)

Now we are going to introduce the assumption 9 on normality of the disturbance *u*. The set of assumptions 1 through 9 is known as the *classical linear model* (*CLM*) assumptions.

Under the *CLM* assumptions, the *OLS* estimators are *also maximum likelihood estimators* (*ML*), as can be seen in appendix 2.8.

On the other hand, under *CLM* assumptions, *OLS* estimators are not only *BLUE*, but are the *minimum variance unbiased estimators (MVUE)*. This means that *OLS* estimators have the smallest variance among all unbiased, linear o nonlinear, estimators, as can be seen in figure 2.13. Therefore, we have no longer to restrict our comparison to estimators that are linear in the y_i's.

What also happens is that any linear combination of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k$ is also normally distributed, and any subset of the $\hat{\beta}_j$'s has a joint normal distribution.

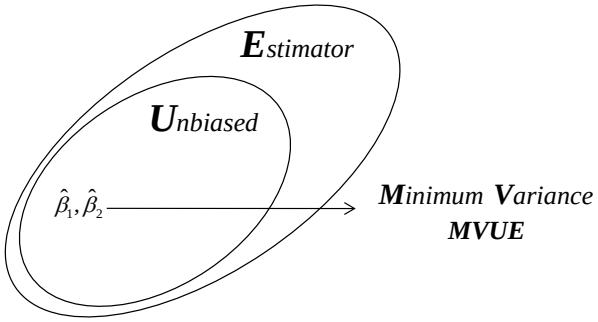


FIGURE 2.13. The *OLS* estimator is the MVUE.

In conclusion, we have seen that the *OLS* estimator has very desirable properties when the statistical basic assumptions are met.

Exercises

Exercise 2.1 The following model has been formulated to explain the annual sales (*sales*) of manufacturers of household cleaning products based as a function of a relative price index (*rpi*):

$$sales = \beta_1 + \beta_2 rpi + u$$

where the variable *sales* is expressed in a thousand million euros and *rpi* is an index obtained as the ratio between the prices of each firm and prices of the firm 1 of the sample). Thus, the value 110 in firm 2 indicates its price is 10% higher than in firm1.

Data on ten manufacturers of household cleaning products are the following:

firm	*sales*	*rpi*
1	10	100
2	8	110
3	7	130
4	6	100
5	13	80
6	6	80
7	12	90
8	7	120
9	9	120
10	15	90

a) Estimate β_1 and β_2 by *OLS*.
b) Calculate the *RSS*.
c) Calculate the coefficient of determination.
d) Check that the algebraic implications 1, 3 and 4 are fulfilled in the *OLS* estimation.

Exercise 2.2 To study the relationship between fuel consumption (*y*) and flight time (*x*) of an airline, the following model is formulated:

$$y = \beta_1 + \beta_2 x + u$$

where *y* is expressed in thousands of pounds and *x* in hours, using fractions of an hour as units of low-order decimal.

The statistics of "Flight times and fuel consumption" of an airline provides data on flight times and fuel consumption of 24 different trips made by an aircraft of the company. From these data the following statistics were drawn:

$$\sum y_i = 219.719;\ \sum x_i = 31.470;\ \sum x_i^2 = 51.075;$$

$$\sum x_i y_i = 349.486;\ \sum y_i^2 = 2396.504$$

a) Estimate β_1 and β_2 by *OLS*.
b) Decompose the variance of the variable *y* invariance explained by the regression and residual variance.
c) Calculate the coefficient of determination.
d) Estimate total consumption, in thousands of pounds, for a flight program consisting of 100 half-hour flights, 200 one hour flights and 100 two hours flights.

Exercise 2.3 An analyst formulates the following model:

$$y = \beta_1 + \beta_2 x + u$$

Using a given sample, the following results were obtained:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = 20 \qquad \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = 10 \qquad \overline{y} = 8 \quad \overline{x} = 4 \quad \hat{\beta}_2 = 3$$

Are these results consistent? Explain your answer.

Exercise 2.4 An econometrician has estimated the following model with a sample of five observations:

$$y_i = b_1 + b_2 x_i + u_i$$

Once the estimation has been made, the econometrician loses all information except what appears in the following table:

Obs.	x_i	$\hat{u}_t$
1	1	2
2	3	-3
3	4	0
4	5	¿?
5	6	¿?

With the above information the econometrician must calculate the residual variance. Do it for them.

Exercise 2.5 Given the model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad 1 = 1, 2, \ldots, n$$

the following results with a sample size of 11 were obtained:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

a) Estimate β_2 and β_1

b) Calculate the sum of square residuals.

c) Calculate the coefficient of determination.

d) Calculate the coefficient of determination under the assumption that $2F^2 = BE$

Exercise 2.6 Company A is dedicated to mounting prefabricated panels for industrial buildings. So far, the company has completed eight orders, in which the number of square meters of panels and working hours employed in the assembly are as follows:

Number of square meters (thousands)	**Number of hours**
4	7400
6	9800
2	4600
8	12200
10	14000
5	8200
3	5800

12	17000

Company A wishes to participate in a tender to mount 14000m^2 of panels in a warehouse, for which a budget is required.

In order to prepare the budget, we know the following:

a) The budget must relate exclusively to the assembly costs, since the material is already provided.
b) The cost of the working hour for Company A is 30 euros.
c) To cover the remaining costs, Company A must charge 20% on the total cost of labor employed in the assembly.

Company A is interested in participating in the tender with a budget that only covers the costs. Under these conditions, and under the assumption that the number of hours worked is a linear function of the number of square meters of panels mounted, what would be the budget provided by company A?

Exercise 2.7 Consider the following equalities:

1. $E[u] = 0$.
2. $E[\hat{u}] = 0$.
3. $\overline{u} = 0$.
4. $\overline{\hat{u}} = 0$.

In the context of the basic linear model, indicate whether each of the above equalities are true or not. Justify your answer.

Exercise 2.8 The parameters β_1 and β_2 of the following model have been estimated by *OLS*:

$$y = \beta_1 + \beta_2 x + u$$

A sample of size 3 was used and the observations for x_i were $\{1,2,3\}$. It is also known that the residual for the first observation was 0.5.

From the above information, is it possible to calculate the sum of squared residuals and obtain an estimate of σ^2? If so, carry out the corresponding calculations.

Exercise 2.9 The following data are available to estimate a relationship between *y* and *x*:

y	*x*
-2	-2
-1	0
0	1
1	0
2	1

a) Estimate the parameters α and β of the following model by *OLS*:

$$y = \alpha + \beta x + \varepsilon$$

b) Estimate var(ε_i).
c) Estimate the parameters γ and δ of the following model by *OLS*:

$$x = \gamma + \delta y + \upsilon$$

d) Are the two fitted regression lines the same? Explain the result in terms of the least-square method.

Exercise 2.10 Answer the following questions:

a) One researcher, after performing the estimation of a model by *OLS*, calculates $\sum \hat{u}_i$ and verifies that it is not equal to 0. Is this possible? Are there any conditions in which this may occur?

b) Obtain an unbiased estimator of σ^2, indicating the assumption you have to use. Explain your answer.

Exercise 2.11 In the context of a linear regression model

$$y = \beta_1 + \beta_2 x + u$$

a) Indicate whether the following equalities are true. If so explain why

$$\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n} = 0; \qquad \bar{\hat{u}} = \frac{\sum_{i=1}^{n} \hat{u}_i}{n} = 0; \qquad E\left[x_i u_i\right] = 0; \qquad E\left[u_i\right] = 0;$$

b) Establish the relationship between the following expressions:

$$E\left[u_i^2\right] = \sigma^2; \qquad \hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-k}$$

Exercise 2.12 Answer the following questions:

a) Define the probabilistic properties of *OLS* estimator under the basic assumptions of the linear regression model. Explain your answer.

b) What happens with the estimation of the linear regression model if the sample variance of the explanatory variable is null? Explain your answer.

Exercise 2.13 A researcher believes that the relationship between consumption (*cons*) and disposable income (*inc*) should be strictly proportional, and, therefore formulates the following model:

$$cons = \beta_2 inc + u$$

a) Derive the formula for estimating β_2.

b) Derive the formula for estimating σ^2.

c) In this model, is $\sum_{i=1}^{n} \hat{u}_i$ equal to 0?

Exercise 2.14 In the context of the simple linear regression model

$$y = \beta_1 + \beta_2 x + u$$

a) What assumptions must be met for the *OLS* estimators to be unbiased?

b) What assumptions are required for the estimators with variances which are the lowest within the set of linear unbiased estimators?

Exercise 2.15 In statistical terms it is often usual to make statements like the following:

"Let $x_2, \ldots x_n$, be a random sample of size n drawn from a population $N(\alpha,\sigma)$"

a) Express the previous statement with econometric language by introducing a disturbance term.

b) Derive the formula for estimating α.

c) Derive the formula for estimating σ^2.

d) In this model, is $\sum_{i=1}^{n} \hat{u}_i$ equal to 0?

Exercise 2.16 The following model relates expenditure on education (*exped*) and disposable income (*inc*):

$$exped=\beta_1+\beta_2 inc+u$$

Using the information obtained from a sample of 10 families, the following results have been obtained:

$$\overline{exped} = 7 \quad \overline{inc} = 50 \quad \sum_{i=1}^{10} inc_i^2 = 30.650 \quad \sum_{i=1}^{10} exped_i^2 = 622 \quad \sum_{i=1}^{10} inc_i' \; exped_i = 4.345$$

a) Estimate β_1 and β_2 by *OLS*.
b) Estimate the expenditure on education/ income elasticity for the sample average family.
c) Decompose the variance of the endogenous variable invariance explained by the regression and residual variance.
d) Calculate the coefficient of determination.
e) Estimate the variance of the disturbances.

Exercise 2.17 Given the population model

$$y_i=3+2x_i+u_i \qquad i= 1, 2, 3$$

where $x_i=\{1,2,3\}$:

a) Using $N(0,1)$ random number, generate 15 samples of u_1, u_2 and u_3, and obtain the corresponding values of *y*
b) Carry out the corresponding estimates of β_1 and β_2 in the model:

$$y = \beta_1 + \beta_2 x + u$$

c) Compare the sample means and variances of $\hat{\beta}_1$y $\hat{\beta}_2$ with their population expectations and variances.

Exercise 2.18 Based on the information supplied in exercise 2.17, and the 15 pairs of estimates of β_1 and β_2 obtained:

a) Calculate the residuals corresponding to each of the estimates.
b) Show why the residuals always take the form

$$\hat{u}_1 = -\hat{u}_2$$
$$\hat{u}_3 = 0$$

Exercise 2.19 The following model was formulated to explain sleeping time (*sleep*) as a function of time devoted to paid work (*paidwork*):

$$sleep = \beta_1 + \beta_2 paidwork + u$$

where *sleep* and *paidwork* are measured in minutes per day.

Using a random subsample extracted from the file *timuse03*, the following results were obtained

$$\widehat{sleep}_i = 550.17 - 0.1783 paidwork$$

$$R^2= 0.2539 \qquad n=62$$

a) Interpret the coefficient on *paidwork.*
b) What is, on average, the predicted increment in sleep if time devoted to paid work decreases in an hour per day?
c) How much of the variation in *sleep* is explained by *paidwork*?

Exercise 2.20 Quantifying happiness is not an easy task. Researchers at the Gallup World Poll went about it by surveying thousands of respondents in 155 countries, between 2006 and 2009, in order to measure two types of well-being. They asked respondents to report on the overall satisfaction with their lives, and ranked their answers using a "life evaluation" score from 1 to 10. To explain the overall satisfaction (*stsfglo*) the following model has been formulated, where observations are averages of the variables in each country:

$$stsfglo = \beta_1 + \beta_2 lifexpec + u$$

where *lifexpec* is life expectancy at birth: that is to say, number of years a newborn infant is expected to live.

Using the work file *HDR2010*, the fitted model obtained is the following:

$$\widehat{stsfglo} = -1.499 + 0.1062 lifexpec$$

$$R^2 = 0.6135 \qquad n = 144$$

a) Interpret the coefficient on *lifexpec*.
b) What would be the average overall satisfaction for a country with 80 years of life expectancy at birth?
c) What should be the life expectancy at birth to obtain a global satisfaction equal to six?

Exercise 2.21 In economics, Research and Development intensity (or simply R&D intensity) is the ratio of a company's investment in Research and Development compared to its sales.

For the estimation of a model which explains R&D intensity, it is necessary to have an appropriate database. In Spain it is possible to use the Survey of Entrepreneurial Strategies (*Encuesta sobre Estrategias Empresariales*) produced by the Ministry of Industry. This survey, on an annual basis, provides in-depth knowledge of the industrial sector's evolution over time by means of multiple data concerning business development and company decisions. This survey is also designed to generate microeconomic information that enables econometric models to be specified and tested. As far as its coverage is concerned, the reference population of this survey is companies with 10 or more workers from the manufacturing industry. The geographical area of reference is Spain, and the variables have a timescale of one year. One of the most outstanding characteristics of this survey is its high degree of representativeness.

Using the work file *rdspain*, which is a dataset consisting of 1,983 Spanish firms for 2006, the following equation is estimated to explain expenditures on research and development (*rdintens*):

$$\widehat{rdintens} = -2.639 + 0.2123 \ln(sales)$$

$$R^2 = 0.0350 \qquad n = 1983$$

where *rdintens* is expressed as a percentage of sales, and *sales* are measured in millions of euros.

a) Interpret the coefficient on ln(*sales*).
b) If *sales* increase by 50%, what is the estimated percentage point change in *rdintens*?
c) What percentage of the variation of *rdintens* is explained by *sales*? Is it large? Justify your answer.

Exercise 2.22 The following model has been formulated to explain MBA graduated salary (*salMBAgr*) as a function of tuition fees (*tuition*)

$$salMBAgr = \beta_1 + \beta_2 tuition + u$$

where *salMBApr* is the median annual salary in dollars for students enrolled in 2,010 of the 50 best American business schools and *tuition* is tuition fees including all required fees for the entire program (but excluding living expenses).

Using the data in *MBAtui10*, this model is estimated:

$$\widehat{salMBAgr}_i = 54242 + 0.4313 tuition_i$$

$$R^2=0.4275 \qquad n=50$$

a) What is the interpretation of the intercept?
b) What is the interpretation of the slope coefficient?
c) What is the predicted value of *salMBAgr* for a graduate student who paid 110000$ *tuition* fees in a 2 years *MBA*?

Exercise 2.23 Using a subsample of the Structural Survey of Wages (*Encuesta de estructura salarial*) for Spain in 2006 (*wage06sp*), the following model is estimated to explain wages:

$$\widehat{\ln(wage)} = 1.918 + 0.0527 educ$$

$$R^2=0.2445 \qquad n=50$$

where *educ* (education) is measured in years and *wage* in euros per hour.

a) What is the interpretation of the coefficient on *educ*?
b) How many more years of education are required to have a 10% higher wage?
c) Knowing that $\overline{educ} = 10.2$, calculate the wage/education elasticity. Do you consider this elasticity to be high or low?

Exercise 2.24 Using data from the Spanish economy for the period 1954-2010 (work file *consump*), the Keynesian consumption function is estimated:

$$\widehat{conspc}_t = -288 + 0.9416 incpc_t$$

$$R^2=0.994 \qquad n=57$$

where consumption (*conspc*) and disposable income (*incpc*) are expressed in constant euros per capita, taking 2008 as reference year.

a) What is the interpretation of the intercept? Comment on the sign and magnitude of the intercept.
b) Interpret the coefficient on *incpc*. What is the economic meaning of this coefficient?

c) Compare the marginal propensity to consume with the average propensity to consume at the sample mean ($\overline{conspc} = 8084$, $\overline{incpc} = 8896$). Comment on the result obtained.

d) Calculate the consumption/income elasticity for the sample mean.

Annex 2.1 Case study: Engel curve for demand of dairy products

The Engel curve shows the relationship between the various quantities of a good that a consumer is willing to purchase at varying income levels.

In a survey with 40 households, data were obtained on expenditure on dairy products and income. These data appear in table 2.6 and in work file *demand*. In order to avoid distortions due to the different size of households, both consumption and income have been expressed in terms of per capita. The data are expressed in thousands of euros per month.

There are several demand models. We will consider the following models: linear, inverse, semi-logarithmic, potential, exponential and inverse exponential. In the first three models, the regressand of the equation is the endogenous variable, whereas in the last three the regressand is the natural logarithm of the endogenous variable.

In all the models we will calculate the marginal propensity to expenditure, as well as the expenditure/income elasticity.

TABLE 2.6. Expenditure on dairy products (*dairy*), disposable income (*inc*) in terms of *per capita*. Unit: euros per month.

household	*dairy*	*inc*	*household*	*dairy*	*inc*
1	8.87	1.250	21	16.20	2.100
2	6.59	985	22	10.39	1.470
3	11.46	2.175	23	13.50	1.225
4	15.07	1.025	24	8.50	1.380
5	15.60	1.690	25	19.77	2.450
6	6.71	670	26	9.69	910
7	10.02	1.600	27	7.90	690
8	7.41	940	28	10.15	1.450
9	11.52	1.730	29	13.82	2.275
10	7.47	640	30	13.74	1.620
11	6.73	860	31	4.91	740
12	8.05	960	32	20.99	1.125
13	11.03	1.575	33	20.06	1.335
14	10.11	1.230	34	18.93	2.875
15	18.65	2.190	35	13.19	1.680
16	10.30	1.580	36	5.86	870
17	15.30	2.300	37	7.43	1.620
18	13.75	1.720	38	7.15	960
19	11.49	850	39	9.10	1.125
20	6.69	780	40	15.31	1.875

Linear model

The linear model for demand of dairy products will be the following:

$$dairy = \beta_1 + \beta_2 inc + u \tag{2-73}$$

The marginal propensity indicates the change in expenditure as income varies and it is obtained by differentiating the expenditure with respect to income in the demand equation. In the linear model the marginal propensity of the expenditure on dairy is given by

$$\frac{d\ dairy}{d\ inc} = \beta_2 \tag{2-74}$$

In other words, in the linear model the marginal propensity is constant and, therefore, it is independent of the value that takes the income. It has the disadvantage of not being adapted to describe the behavior of the consumers, especially when there are important differences in the household income. Thus, it is unreasonable that the marginal propensity of expenditure on dairy products is the same in a low-income family and a family with a high income. However, if the variation of the income is not very high in the sample, a linear model can be used to describe the demand of certain goods.

In this model the expenditure/income elasticity is the following:

$$\varepsilon_{dairy/inc}^{linear} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \beta_2 \frac{inc}{dairy} \tag{2-75}$$

Estimating the model (2-73) with the data from table 2.6, we obtain

$$\widehat{dairy} = 4.012 + 0.005288\ inc \qquad R^2 = 0.4584 \tag{2-76}$$

Inverse model

In an inverse model there is a linear relationship between the expenditure and the inverse of income. Therefore, this model is directly linear in the parameters and it is expressed in the following way:

$$dairy = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-77}$$

The sign of the coefficient β_2 will be negative if the income is correlated positively with the expenditure. It is easy to see that, when the income tends towards infinite, the expenditure tends towards a limit which is equal to β_1. In other words, β_1 represents the maximum consumption of this good.

In figure 2.14, we can see a double representation of the population function corresponding to this model. In the first one, the relationship between the dependent variable and explanatory variable has been represented. In the second one, the relationship between the regressand and regressor has been represented. The second function is linear as can be seen in the figure.

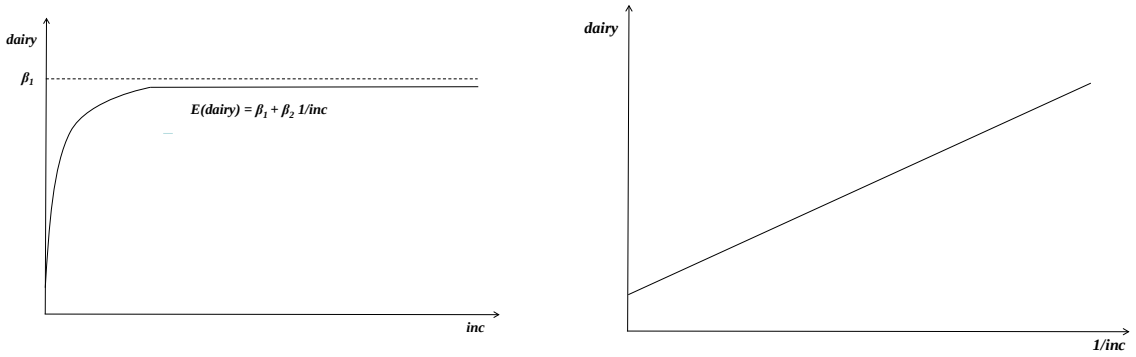


Figure 2.14. The inverse model.

In the inverse model, the marginal propensity to expenditure is given by

$$\frac{d\ dairy}{d\ inc} = -\beta_2 \frac{1}{(inc)^2} \tag{2-78}$$

According to (2-78), the marginal propensity is inversely proportional to the square of the income level.

On the other hand, the elasticity is inversely proportional to the product of expenditure and income, as can be seen in the following expression:

$$\varepsilon^{inv}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = -\beta_2 \frac{1}{inc \times dairy} \tag{2-79}$$

Estimating the model (2-77) with the data of table 2.6, we obtain

$$\hat{dairy} = 18.652 - 8702\frac{1}{inc} \qquad R^2 = 0.4281 \tag{2-80}$$

In this case the coefficient $\hat{\beta}_2$ does not have an economic meaning.

Linear-log model

This model is denominated linear-log model, because the expenditure is a linear function of the logarithm of income, that is to say,

$$dairy = \beta_1 + \beta_2 \ln(inc) + u \tag{2-81}$$

In this model the marginal propensity to expenditure is given by

$$\frac{d\ dairy}{d\ inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{inc} = \frac{d\ dairy}{d\ \ln(inc)}\frac{1}{inc} = \beta_2 \frac{1}{inc} \tag{2-82}$$

and the elasticity expenditure/income is given by

$$\varepsilon^{lin\text{-}\log}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ dairy}{d\ \ln(inc)}\frac{1}{dairy} = \beta_2 \frac{1}{dairy} \tag{2-83}$$

The marginal propensity is inversely proportional to the level of income in the linear-log model, while the elasticity is inversely proportional to the level of expenditure on dairy products.

In figure 2.15, we can see a double representation of the population function corresponding to this model.

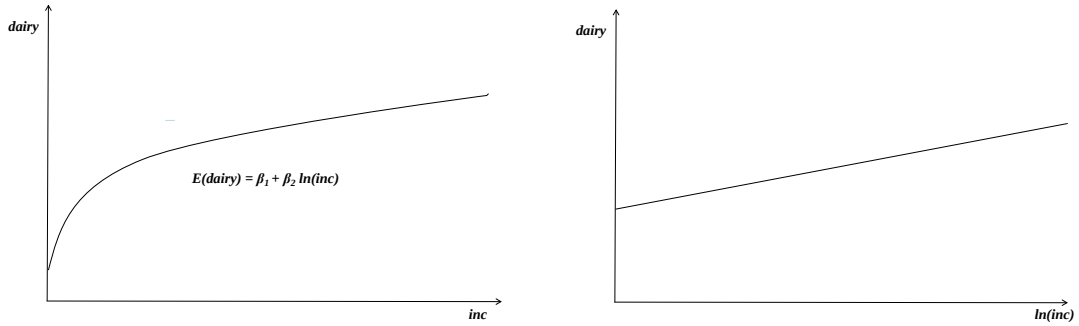


Figure 2.15. The linear-log model.

Estimating the model (2-81) with the data from table 2.6, we obtain

$$\hat{dairy} = -41.623 + 7.399 \ln(inc) \quad R^2 = 0.4567 \tag{2-84}$$

The interpretation of $\hat{\beta}_2$ is the following: if the income increases by 1%, the demand of dairy products will increase by 0.07399 euros.

Log-log model or potential model

This exponential model is defined in the following way:

$$dairy = e^{\beta_1} inc^{\beta_2} e^{u} \tag{2-85}$$

This model is not linear in the parameters, but it is linearizable by taking natural logarithms, and the following is obtained:

$$\ln(dairy) = \beta_1 + \beta_2 \ln(inc) + u \tag{2-86}$$

This model is also called log-log model, because this is the structure of the corresponding linearized model.

In this model the marginal propensity to expenditure is given by

$$\frac{d\ dairy}{d\ inc} = \beta_2 \frac{dairy}{inc} \tag{2-87}$$

In the log-log model, the elasticity is constant. Therefore, if the income increases by 1%, the expenditure will increase by β_2%, since

$$\varepsilon_{dairy/inc}^{log\text{-}log} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ \ln(inc)} = \beta_2 \tag{2-88}$$

In figure 2.16, we can see a double representation of the population function corresponding to this model.

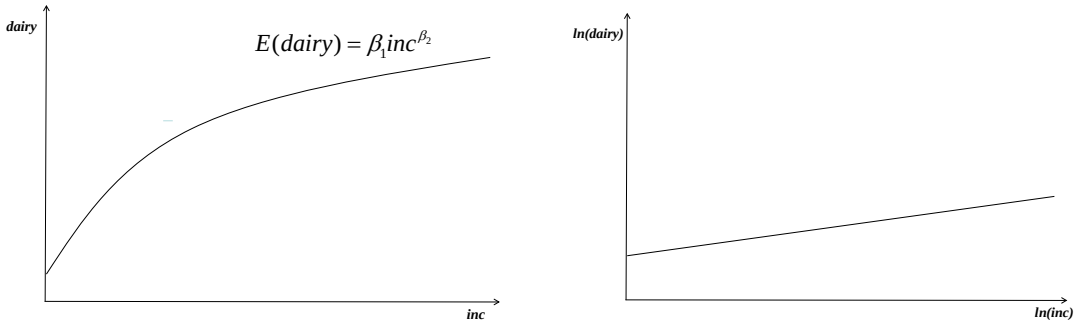


Figure 2.16. The log-log model.

Estimating the model (2-86) with the data from table 2.6, we obtain

$$\widehat{\ln(dairy)} = -2.556 + 0.6866 \ln(inc) \quad R^2 = 0.5190 \tag{2-89}$$

In this case $\hat{\beta}_2$ is the expenditure/income elasticity. Its interpretation is the following: if the income increases by 1%, the demand of dairy products will increase by 0.68%.

Log-linear or exponential model

This exponential model is defined in the following way:

$$dairy = \exp(\beta_1 + \beta_2 inc + u) \tag{2-90}$$

By taking natural logarithms on both sides of (2-90), we obtain the following model that is linear in the parameters:

$$\ln(dairy) = \beta_1 + \beta_2 inc + u \tag{2-91}$$

In this model the marginal propensity to expenditure is given by

$$\frac{d\ dairy}{d\ inc} = \beta_2 dairy \tag{2-92}$$

In the exponential model, unlike other models seen previously, the marginal propensity increases when the level of expenditure does. For this reason, this model is adequate to describe the demand of luxury products. On the other hand, the elasticity is proportional to the level of income:

$$\varepsilon^{exp}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc} inc = \beta_2 inc \tag{2-93}$$

In figure 2.17, we can see a double representation of the population function corresponding to this model.

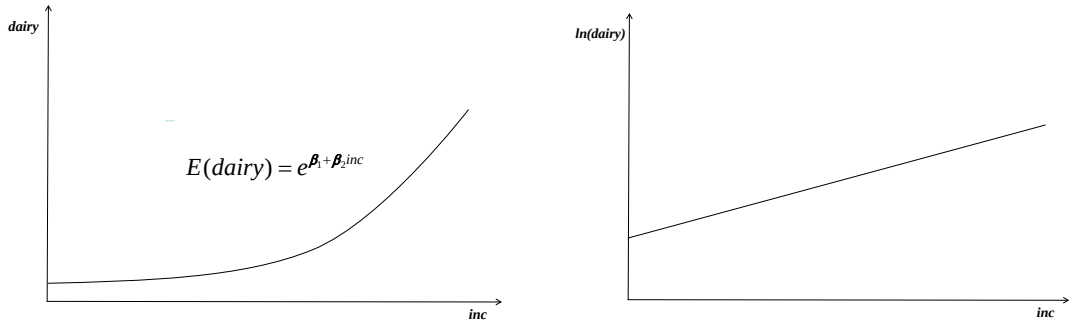


Figure 2.17. The log-linear model.

Estimating the model (2-91) with the data from table 2.6, we obtain

$$\widehat{\ln}(dairy) = 1.694 + 0.00048\,inc \qquad R^2 = 0.4978 \tag{2-94}$$

The interpretation of $\hat{\beta}_2$ is the following: if the income increases by a euro the demand of dairy products will increase by 0.048%.

Inverse exponential model

The inverse exponential model, which is a mixture of the exponential model and the inverse model, has properties that make it suitable for determining the demand for products in which there is a saturation point. This model is given by

$$dairy = \exp(\beta_1 + \beta_2 \frac{1}{inc} + u) \tag{2-95}$$

By taking natural logarithms on both sides of (2-95), we obtain the following model that is linear in the parameters:

$$\ln(dairy) = \beta_1 + \beta_2 \frac{1}{inc} + u \tag{2-96}$$

In this model the marginal propensity to expenditure is given by

$$\frac{d\ dairy}{d\ inc} = -\beta_2 \frac{dairy}{(inc)^2} \tag{2-97}$$

and the elasticity by

$$\varepsilon^{invexp}_{dairy/inc} = \frac{d\ dairy}{d\ inc}\frac{inc}{dairy} = \frac{d\ \ln(dairy)}{d\ inc} inc = -\beta_2 \frac{1}{inc} \tag{2-98}$$

Estimating the model (2-96) with the data from table 2.6, we obtain

$$\widehat{\ln}(dairy) = 3.049 - 822.02\frac{1}{inc} \qquad R^2 = 0.5040 \tag{2-99}$$

In this case, as in the inverse model, the coefficient $\hat{\beta}_2$ does not have an economic meaning.

In table 2.7, the results of the marginal propensity, the expenditure/income elasticity and R^2 in the six fitted models are shown

Table 2.7. Marginal propensity, expenditure/income elasticity and R^2 in the fitted models.

Model	*Marginal propensity*	*Elasticity*	R^2
Linear	$\hat{\beta}_2$ =0.0053	$\hat{\beta}_2 \frac{\overline{inc}}{\overline{dairy}}$ =0.6505	0.4440
Inverse	$-\hat{\beta}_2 \frac{1}{\left[\overline{inc}\right]^2}$ =0.0044	$-\hat{\beta}_2 \frac{1}{\overline{dairy} \times \overline{inc}}$ =0.5361	0.4279
Linear-log	$\hat{\beta}_2 \frac{1}{\overline{inc}}$ =0.0052	$\hat{\beta}_2 \frac{1}{\overline{dairy}}$ =0.6441	0.4566
Log-log	$\hat{\beta}_2 \frac{\overline{dairy}}{\overline{inc}}$ =0.0056	$\hat{\beta}_2$ =0.6864	0.5188
Log-linear	$\hat{\beta}_2 \times \overline{dairy}$ =0.0055	$\hat{\beta}_2 \times \overline{inc}$ =0.6783	0.4976
Inverse-log	$-\hat{\beta}_2 \frac{\overline{dairy}}{\left[\overline{inc}\right]^2}$ =0.0047	$-\hat{\beta}_2 \frac{1}{\overline{inc}}$ =0.5815	0.5038

The R^2 obtained in the first three models are not comparable with the R^2 obtained in the last three because the functional form of the regressand is different: y in the first three models and ln(y) in the last three.

Comparing the first three models the best fit is obtained by the linear-log model, if we use the R^2 as goodness of fit measure. Comparing the last three models the best fit is obtained by the log-log model. If we had used the Akaike Information Criterion (AIC), which allows the comparison of models with different functional forms for the regressand, then the-log-log model would have been the best among the six models fitted. The AIC measured will be studied in chapter 3.

Appendixes

Appendix 2.1: Two alternative forms to express $\hat{\beta}_2$

It is easy to see that

$$\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})=\sum_{i=1}^{n}(y_ix_i-\overline{x}y_i-\overline{y}x_i+\overline{yx})=\sum_{i=1}^{n}y_ix_i-\overline{x}\sum_{i=1}^{n}y_i-\overline{y}\sum_{i=1}^{n}x_i+n\overline{yx}$$

$$=\sum_{i=1}^{n}y_ix_i-n\overline{xy}-\overline{y}\sum_{i=1}^{n}x_i+n\overline{yx}=\sum_{i=1}^{n}y_ix_i-\overline{y}\sum_{i=1}^{n}x_i$$

On the other hand, we have

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}(x_i^2-2\overline{x}x_i+\overline{xx})^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{xx}$$

$$=\sum_{i=1}^{n}x_i^2-2n\overline{x}^2+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-\overline{x}\sum_{i=1}^{n}x_i$$

Therefore, (2-17) can be expressed in the following way:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Appendix 2.2. Proof: $r_{xy}^2 = R^2$

First of all, we are going to see an equivalence that will be used in the proof. By definition,

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

From the first normal equation, we have

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}$$

Subtracting the second equation from the first one:

$$\hat{y}_i - \bar{y} = \hat{\beta}_2 (x_i - \bar{x})$$

Squaring both sides

$$(\hat{y}_i - \bar{y})^2 = \hat{\beta}_2^2 (x_i - \bar{x})^2$$

and summing for all *i*, we have

$$\sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}_2^2 \sum (x_i - \bar{x})^2$$

Taking into account the previous equivalence, we have

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})\right]^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \frac{1}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = r_{xy}^2$$

Appendix 2.3. Proportional change *versus* change in logarithms

Change in logarithms is a variation rate, which is used in economics research. The relationship between proportional change and change in logarithms can be seen if we expand (2-45) by Taylor series:

$$\ln(x_1)-\ln(x_0)=\ln\left[\frac{x_1}{x_0}\right]$$

$$=\ln(1)+\left[\frac{x_1}{x_0}-1\right]\left[\frac{1}{\frac{x_1}{x_0}}\right]_{\frac{x_1}{x_0}=1}+\frac{1}{2}\left[\frac{x_1}{x_0}-1\right]^2\left[-\frac{1}{\left[\frac{x_1}{x_0}\right]}\right]_{\frac{x_1}{x_0}=1}$$

$$+\frac{1}{3\times 2}\left[\frac{x_1}{x_0}-1\right]^3\left[\frac{2}{\left[\frac{x_1}{x_0}\right]^3}\right]_{\frac{x_1}{x_0}=1}+\cdots\cdots \qquad (2\text{-}100)$$

$$=\left[\frac{x_1}{x_0}-1\right]-\frac{1}{2}\left[\frac{x_1}{x_0}-1\right]^2+\frac{1}{3}\left[\frac{x_1}{x_0}-1\right]^3+\cdots$$

$$=\frac{\Delta x_1}{x_0}-\frac{1}{2}\left[\frac{\Delta x_1}{x_0}\right]^2+\frac{1}{3}\left[\frac{\Delta x_1}{x_0}\right]^3+\cdots$$

Therefore, if we take the linear approximation in this expansion, we have

$$\Delta\ln(x)=\ln(x_1)-\ln(x_0)=\ln\left[\frac{x_1}{x_0}\right]\approx\frac{\Delta x_1}{x_0} \qquad (2\text{-}101)$$

Appendix 2.4. Proof: *OLS* estimators are linear and unbiased

We will only prove the unbiasedness of the estimator $\hat{\beta}_2$, which is the most important. In order to prove this, we need to rewrite our estimator in terms of the population parameter. The formula (2-18) can be written as

$$\hat{\beta}_2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \qquad (2\text{-}102)$$

because $\sum_{i=1}^{n}(x_i-\overline{x})\overline{y}=\overline{y}\sum_{i=1}^{n}(x_i-\overline{x})=\overline{y}\times 0=0$

Now (2-102) will be expressed in the following way:

$$\hat{\beta}_2=\sum_{i=1}^{n}c_i y_i \qquad (2\text{-}103)$$

where

$$c_i = \frac{x_i - \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2-104}$$

The c_i's have the following properties:

$$\sum_{i=1}^{n} c_i = 0 \tag{2-105}$$

$$\sum_{i=1}^{n} c_i^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2-106}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n}(x_i - \overline{x})x_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 1 \tag{2-107}$$

Now, if we substitute $y = \beta_1 + \beta_2 x + u$ (assumption 1) in (2-102), we have

$$\begin{aligned} \hat{\beta}_2 &= \sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n} c_i(\beta_1 + \beta_2 x_i + u_i) \\ &= \beta_1 \sum_{i=1}^{n} c_i + \beta_2 \sum_{i=1}^{n} c_i x_i + \sum_{i=1}^{n} c_i u_i = \beta_2 + \sum_{i=1}^{n} c_i u_i \end{aligned} \tag{2-108}$$

Since the regressors are assumed to be nonstochastic (assumption 2), the c_i are nonstochastic too. Therefore, $\hat{\beta}_2$ is an estimator that is a *linear* function of *u*'s.

Taking expectations in (2-108) and taking into account assumption 6, and implicitly assumptions 3 through 5, we obtain

$$E(\hat{\beta}_2) = \beta_2 + \sum_{i=1}^{n} c_i E(u_i) = \beta_2 \tag{2-109}$$

Therefore, $\hat{b}_2$ is an unbiased estimator of β_2

Appendix 2.5. Calculation of variance of $\hat{\beta}_2$:

$$\begin{aligned} E\left[\hat{\beta}_2 - \beta_2\right]^2 &= \left[\sum_{i=1}^{n} c_i u_i\right]^2 = \sum_{i=1}^{n} c_i^2 E(u_i^2) + \sum_{i \neq j} \sum_{i=1}^{n} c_i c_j E(u_i u_j) \\ &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{nS_X^2} \end{aligned} \tag{2-110}$$

In the above proof, to pass from the second to the third equality, we have taken into account assumptions 6 and 7.

Appendix 2.6. Proof of Gauss-Markov Theorem for the slope in simple regression

The plan for the proof is the following. First, we are going to define an arbitrary estimator $\tilde{\beta}_2$ which is linear in *y*. Second, we will impose restrictions implied by unbiasedness. Third, we will show that the variance of the arbitrary estimator must be larger than, or at least equal to, the variance of $\hat{\beta}_2$.

Let us define an arbitrary estimator $\tilde{\beta}_2$ which is linear in *y*:

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i \tag{2-111}$$

Now, we substitute y_i by its value in the population model (assumption 1):

$$\tilde{\beta}_2 = \sum_{i=1}^{n} h_i y_i = \sum_{i=1}^{n} h_i (\beta_1 + \beta_2 x_i + u_i) = \beta_1 \sum_{i=1}^{n} h_i + \beta_2 \sum_{i=1}^{n} h_i x_i + \sum_{i=1}^{n} h_i u_i \tag{2-112}$$

For the estimator $\tilde{\beta}_2$ to be unbiased, the following restrictions must be accomplished:

$$\sum_{i=1}^{n} h_i = 0 \qquad\qquad \sum_{i=1}^{n} h_i x_i = 1 \tag{2-113}$$

Therefore,

$$\tilde{\beta}_2 = \beta_2 + \sum_{i=1}^{n} h_i u_i \tag{2-114}$$

The variance of this estimator is the following:

$$E\left[\tilde{\beta}_2 - \beta_2\right]^2 = \left[\sum_{i=1}^{n} h_i u_i\right]^2 = \sigma^2 \sum_{i=1}^{n} h_i^2 =$$

$$\sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} + \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]^2 = \sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]^2 \tag{2-115}$$

$$+\sigma^2 \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]^2 + 2\sigma^2 \sum_{i=1}^{n} \left[h_i - \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right] \frac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

The third term of the last equality is 0, as shown below:

$$2\sigma^2\sum_{i=1}^{n}\left[h_i-\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$
$$=2\sigma^2\sum_{i=1}^{n}\left[h_i\frac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]-2\sigma^2\sum_{i=1}^{n}\left[\frac{(x_i-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]=2\sigma^2\times1-2\sigma^2\times1=0 \quad (2\text{-}116)$$

Therefore, taking into account (2-116) and operating, we have

$$E\left[\tilde{\beta}_2-\beta_2\right]^2=\sigma^2\sum_{i=1}^{n}\left[h_i-c_i\right]^2+\sigma^2\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \quad (2\text{-}117)$$

where $c_i=\dfrac{x_i-\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$

The second term of the last equality is the variance of $\hat{\beta}_2$, while the first term is always positive because it is a sum of squares, except that h_i=c_i, for all *i*, in which case it is equal to 0, and then $\tilde{\beta}_2=\hat{\beta}_2$. So,

$$E\left[\tilde{\beta}_2-\beta_2\right]^2\geq E\left[\hat{\beta}_2-\beta_2\right]^2 \quad (2\text{-}118)$$

Appendix 2.7. Proof: $\hat{\sigma}^2$ is an unbiased estimator of the variance of the disturbance

The population model is by definition:

$$y_i=\beta_1+\beta_2x_i+u_i \quad (2\text{-}119)$$

If we sum up both sides of (2-119) for all *i* and divide by *n*, we have

$$\overline{y}=\beta_1+\beta_2\overline{x}+\overline{u} \quad (2\text{-}120)$$

Subtracting (2-120) from (2-119), we have

$$y_i-\overline{y}=\beta_2\left(x_i-\overline{x}\right)+\left(u_i-\overline{u}\right) \quad (2\text{-}121)$$

On the other hand, $\hat{u}_i$ is by definition:

$$\hat{u}_i=y_i-\hat{\beta}_1-\hat{\beta}_2x_i \quad (2\text{-}122)$$

If we sum up both sides of (2-122) for all *i* and divide by *n*, we have

$$\overline{\hat{u}}=\overline{y}-\hat{\beta}_1-\hat{\beta}_2\overline{x} \quad (2\text{-}123)$$

Subtracting (2-123) from (2-122), and taking into account that $\overline{\hat{u}}$ =0,

$$\hat{u}_i = (y_i - \bar{y}) - \hat{\beta}_2 (x_1 - \bar{x}) \tag{2-124}$$

Substituting (2-121) in (2-124), we have

$$\begin{aligned} \hat{u}_i &= \beta_2 (x_i - \bar{x}) + (u_i - \bar{u}) - \hat{\beta}_2 (x_1 - \bar{x}) \\ &= -\left(\hat{\beta}_2 - \beta_2\right)(x_1 - \bar{x}) + (u_i - \bar{u}) \end{aligned} \tag{2-125}$$

Squaring and summing up both sides of (2-125), we have

$$\begin{aligned} \sum_{i=1}^{n} \hat{u}_i^2 = \left[\tilde{\beta}_2 - \beta_2\right]^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (u_i - \bar{u})^2 \\ -2\left[\tilde{\beta}_2 - \beta_2\right] \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u}) \end{aligned} \tag{2-126}$$

Taking expectation in (2-126), we obtain

$$\begin{aligned} E\left[\sum_{i=1}^{n} \hat{u}_i^2\right] &= \sum_{i=1}^{n} (x_i - \bar{x})^2 E\left[\tilde{\beta}_2 - \beta_2\right]^2 + E\left[\sum_{i=1}^{n} (u_i - \bar{u})^2\right] \\ &\quad -2E\left[\left(\tilde{\beta}_2 - \beta_2\right) \sum_{i=1}^{n} (x_i - \bar{x})(u_i - \bar{u})\right] \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + (n-1)\sigma^2 - 2\sigma^2 = (n-2)\sigma^2 \end{aligned} \tag{2-127}$$

To obtain the first term of the last equality of (2-127), we have used (2-64). In (2-128) and (2-129), you can find the developments used to obtain the second and the third term of the last equality of (2-127) respectively. In both cases, assumptions 7 and 8 have been taken into account.

$$\begin{aligned} E\left[\sum_{i=1}^{n} (u_i - \bar{u})^2\right] &= E\left[\sum_{i=1}^{n} u_i^2 - n\bar{u}^2\right] = E\left[\sum_{i=1}^{n} u_i^2 - n\left(\frac{\sum_{i=1}^{n} u_i}{n}\right)^2\right] \\ &= E\left[\sum_{i=1}^{n} u_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} u_i^2 + \sum_{i \neq j} u_i u_j\right)\right] = n\sigma^2 - \frac{n}{n}\sigma^2 = (n-1)\sigma^2 \end{aligned} \tag{2-128}$$

$$E\left[\left(\tilde{\beta}_2-\beta_2\right)\sum_{i=1}^{n}(x_i-\overline{x})(u_i-\overline{u})\right]=E\left[\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})u_i\sum_{i=1}^{n}(x_i-\overline{x})u_i\right]$$

$$=\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})E\left(u_i\right)\right]^2$$

$$=\frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2E\left(u_i\right)^2+\sum\sum_{i\neq j}(x_i-\overline{x})(x_i-\overline{x})E\left(u_iu_j\right)\right]=\sigma^2$$

(2-129)

According to (2-127), we have

$$E\left[\sum_{i=1}^{n}\hat{u}_i^2\right]=\left(n-2\right)\sigma^2 \tag{2-130}$$

Therefore, an unbiased estimator is given by

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{n}\hat{u}_i^2}{n-2} \tag{2-131}$$

such as

$$E\left(\hat{\sigma}^2\right)=\frac{1}{n-2}E\left(\sum_{i=1}^{n}\hat{u}_i^2\right)=\sigma^2 \tag{2-132}$$

Appendix 2.8. Consistency of the *OLS* estimator

The operator plim has the in variance property (Slutsky property). That is to say, if $\hat{\theta}$ is a consistent estimator of θ and if $g\left(\hat{\theta}\right)$ is any continuous function of $\hat{\theta}$, then

$$\operatorname*{plim}_{n\to\infty} g(\hat{\theta})=g(\theta) \tag{2-133}$$

This means is that if $\hat{\theta}$ is a consistent estimator of θ, then $1/\hat{\theta}$ and $\ln(\hat{\theta})$ are also consistent estimators of $1/\theta$ and $\ln(\theta)$ respectively. Note that these properties do not hold true for the expectation operator E; for example, if $\hat{\theta}$ is an unbiased estimator of θ [that is to say, $E(\hat{\theta})=\theta$], it is *not true* that $1/\hat{\theta}$ is an unbiased estimator of $1/\theta$; that is, $E(1/\hat{\theta})\neq 1/E(\hat{\theta})\neq 1/\theta$. This is due to the fact that the expectation operator can be only applied to *linear* functions of random variables. On the other hand, the plim operator is applicable to any continuous functions.

Under assumptions 1 through 6, the *OLS* estimators, $\hat{\beta}_1$ and $\hat{\beta}_2$, are consistent.

Now we are going to prove that $\hat{\beta}_2$ is a consistent estimator. First, $\hat{\beta}_2$ can be expressed as:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})y_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(\beta_1+\beta_2 x_i+u_i)}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$
$$= \frac{\beta_1\sum_{i=1}^{n}(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \beta_2\frac{\sum_{i=1}^{n}(x_i-\overline{x})x_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} + \frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \beta_2 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \quad (2\text{-}134)$$

In order to prove consistency, we need to take plim´s in (2-134) and apply the *Law of Large Numbers*. This law states that under general conditions, the sample moments converge to their corresponding population moments. Thus, taking plim´s in (2-134):

$$\operatorname*{plim}_{n\to\infty}\hat{\beta}_2 = \operatorname*{plim}_{n\to\infty}\left[\beta_2 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] = \beta_2 + \frac{\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2} \quad (2\text{-}135)$$

In the last equality we have divided the numerator and denominator by *n*, because if we do not do so, both summations will go to infinity when *n* goes to infinity..

If we apply the law of large numbers to the numerator and denominator of (2-135), they will converge in probability to the population moments *cov*(*x*,*u*) and *var*(*x*) respectively. Provided *var*(*x*)≠0 (assumption 4), we can use the properties of the *probability limits* to obtain

$$\operatorname{plim}\hat{\beta}_2 = \beta_2 + \frac{cov(x,u)}{var(x)} = \beta_2 \quad (2\text{-}136)$$

To reach the last equality, using assumptions 2 and 6, we obtain

$$cov(x,u) = E\left[(x-\overline{x})u\right] = (x-\overline{x})E\left[u\right] = (x-\overline{x})\times 0 = 0 \quad (2\text{-}137)$$

Therefore, $\hat{\beta}_2$ is a consistent estimator.

Appendix 2.9 Maximum likelihood estimator

Taking into account assumptions 1 through 6 the expectation of y_i is the following:

$$E(y_i) = \beta_1 + \beta_2 x_i \quad (2\text{-}138)$$

If we take into account assumptions 7, the variance of y_i is equal to

$$\operatorname{var}(y_i) = E\left[y_i - E(y_i)\right]^2 = E\left[y_i - \beta_1 + \beta_2 x_i\right]^2 = E\left[u_i\right]^2 = \sigma^2 \quad \forall i \quad (2\text{-}139)$$

According to assumption 1, y_i is a linear function of u_i, and if u_i has a normal distribution (assumption 9), then y_i will be normally and independently (assumption 8) distributed with mean $\beta_1 + \beta_2 x_i$ and variance σ^2.

Then, the joint probability density function of $y_1, y_2, \cdots, y_n$ can be expressed as a product of *n* individual density functions:

$$\begin{aligned} & f\left(y_1,\, y_2,\, \ldots, y_n \mid \beta_1 + \beta_2 x_i,\, \sigma^2\right) \\ &= f\left(y_1 \mid \beta_1 + \beta_2 x_i,\, \sigma^2\right) f\left(y_2 \mid \beta_1 + \beta_2 x_i,\, \sigma^2\right) \cdots f\left(y_n \mid \beta_1 + \beta_2 x_i,\, \sigma^2\right) \end{aligned} \tag{2-140}$$

where

$$f\left(y_i\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \tag{2-141}$$

which is the density function of a normally distributed variable with the given mean and variance.

Substituting (2-141) into (2-140)for each y_i, we obtain

$$\begin{aligned} & f\left(y_1,\, y_2, \cdots,\, y_n\right) = f\left(y_1\right) f\left(y_2\right) \cdots f\left(y_n\right) \\ &= \frac{1}{\sigma^n \left(\sqrt{2\pi}\right)^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\left[y_i - \beta_1 - \beta_2 x_i\right]^2}{\sigma^2}\right\} \end{aligned} \tag{2-142}$$

If $y_1, y_2, \cdots, y_n$ are known or given, but β_2, β_3, and σ^2 are not known, the function in (2-142) is called a likelihood function, denoted by $L(\beta_2, \beta_3, \sigma^2)$ or simply *L*. If we take natural logarithms in (2-142), we obtain

$$\begin{aligned} \ln L &= -n\ln\sigma - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i - \beta_1 - \beta_2 x_i\right)}{\sigma^2} \\ &= -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln\left(\sqrt{2\pi}\right) - \frac{1}{2}\sum_{i=1}^{n}\frac{\left(y_i - \beta_1 - \beta_2 x_i\right)}{\sigma^2} \end{aligned} \tag{2-143}$$

The *maximum likelihood* (*ML*) method, as the name suggests, consists in estimating the unknown parameters in such a manner that the probability of observing the given y_i's is as high (or maximum) as possible. Therefore, we have to find the maximum of the function (2-143). To maximize (2-143) we must differentiate with respect to β_2, β_3, and σ^2 and equal to 0. If $\tilde{\beta}_1$, $\tilde{\beta}_2$ and $\tilde{\sigma}^2$ denote the *ML* estimators, we obtain:

$$\begin{aligned} \frac{\partial \ln L}{\partial \tilde{\beta}_1} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)(-1) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\beta}_2} &= -\frac{1}{\tilde{\sigma}^2}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)(-x_i) = 0 \\ \frac{\partial \ln L}{\partial \tilde{\sigma}^2} &= -\frac{n}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}\sum\left(y - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)^2 = 0 \end{aligned} \tag{2-144}$$

If we take the first two equations of (2-144) and operate, we have

$$\sum y_i = n\tilde{\beta}_1 + \tilde{\beta}_2 \sum x_i \tag{2-145}$$

$$\sum y_i x_i = \tilde{\beta}_1 \sum x_i + \tilde{\beta}_2 \sum x_i^2 \tag{2-146}$$

As can be seen, (2-145) and (2-146) are equal to (2-13) and (2-14). That is to say, the *ML* estimators, under the *CLM* assumptions, are equal to the *OLS* estimators.

Substituting $\tilde{\beta}_1$ and $\tilde{\beta}_2$, obtained solving (2-145) and (2-146), in the third equation of (2-144), we have

$$\tilde{\sigma}^2 = \frac{1}{n}\sum\left(y_i - \tilde{\beta}_1 - \tilde{\beta}_2 x_i\right)^2 = \frac{1}{n}\sum\left(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i\right)^2 = \frac{1}{n}\sum \hat{u}_i^2 \tag{2-147}$$

The *ML* estimator for $\tilde{\sigma}^2$ is biased, since, according to (2-127),

$$E(\tilde{\sigma}^2) = \frac{1}{n}E\left[\sum_{i=1}^{n}\hat{u}_i^2\right] = \frac{n-2}{n}\sigma^2 \tag{2-148}$$

In any case, $\tilde{\sigma}^2$ is a consistent estimator because

$$\lim_{n\to 0}\frac{n-2}{n} = 1 \tag{2-149}$$

3 MULTIPLE LINEAR REGRESSION: ESTIMATION AND PROPERTIES

3.1 The multiple linear regression model

The simple linear regression model is not adequate for modeling many economic phenomena, because in order to explain an economic variable it is necessary to take into account more than one relevant factor. We will illustrate this with some examples.

In the Keynesian consumption function, disposable income is the only relevant variable:

$$cons = \beta_1 + \beta_2 inc + u \qquad (3\text{-}1)$$

However, there are other factors that may be considered relevant in consumer behavior. One of these factors could be wealth. By including this factor, we will have a model with two explanatory variables:

$$cons = \beta_1 + \beta_2 inc + \beta_3 wealth + u \qquad (3\text{-}2)$$

In the analysis of production, a potential function is often used, which can be transformed into a linear model in the parameters with an adequate specification (taking natural logs). Using a single input -labor- a model of this type would be specified as follows:

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + u \qquad (3\text{-}3)$$

The previous model is clearly insufficient for economic analysis. It would be better to use the well-known Cobb-Douglas model that considers two inputs (labor and capital):

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u \qquad (3\text{-}4)$$

According to microeconomic theory, total costs (*costot*) are expressed as a function of the quantity produced (*quantprod*). A first approximation to explain the total costs could be a model with only one regresor:

$$costot = \beta_1 + \beta_2 quantprod + u \qquad (3\text{-}5)$$

However, it is very restrictive considering that, as would be the case with the previous model, the marginal cost remains constant regardless of the quantity produced. In economic theory, a cubic function is proposed, which leads to the following econometric model:

$$costot = \beta_1 + \beta_2 quantprod + \beta_3 quantprod^2 + \beta_4 quantprod^3 + u \quad (3\text{-}6)$$

In this case, unlike the previous ones, only one explanatory variable is considered, but with three regressors.

Wages are determined by several factors. A relatively simple model could explain wages using years of education and years of experience as explanatory variables:

$$wages = \beta_1 + \beta_2 educ + \beta_3 exper + u \quad (3\text{-}7)$$

Other important factors to explain wages received can also be quantitative variables such as training and age, or qualitative variables, such as sex, industry, and so on.

Finally, in explaining the expenditure on fish relevant factors are the price of fish, the price of a substitutive commodity such as meat, and disposable income:

$$fishexp = \beta_1 + \beta_2 fishprice + \beta_3 meatprice + \beta_4 income + u \quad (3\text{-}8)$$

Thus, the above examples highlight the need for using multiple regression models. The econometric treatment of the simple regression model was made with ordinary algebra. The treatment of an econometric model with two explanatory variables by using ordinary algebra is tedious and cumbersome. Moreover, a model with three explanatory variables is virtually intractable with this tool. For this reason, the regression model will be presented using matrix algebra.

3.1.1 Population regression model and population regression function

In the model of multiple linear regression, the regressand (which can be either the endogenous variable or a transformation of the endogenous variables) is a linear function of *k* regressors corresponding to the explanatory variables -or their transformations - and of a random disturbance or error. The model also has an intercept. Designating the regressand by *y*, the regressors by x_2, x_3,..., x_k and the disturbance –or the random disturbance- by *u*, the population model of multiple linear regression is given by the following expression:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \mathrm{L} + \beta_k x_k + u \quad (3\text{-}9)$$

The parameters $\beta_1, \beta_2, \beta_3, \mathrm{L}, \beta_k$ are fixed and unknown.

On the right hand of (3-9) we can distinguish two parts: the systematic component $\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \mathrm{L} + \beta_k x_k$ and the random disturbance *u*. Calling μ_y to the systematic component, we can write:

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k \quad (3\text{-}10)$$

This equation is known as the *population regression function* (*PRF*) or *population hyperplane*. When *k*=2 the *PRF* is specifically a straight line; when *k*=3 the *PRF* is specifically a plane; finally, when *k*>3 the *PRF* is generically denominated hyperplane. This cannot to be represented in a three dimension space.

According to (3-10), μ_y is a linear function of the parameters $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$. Now, let us suppose we have a random sample of size n $\{(y_i, x_{2i}, x_{3i}, \ldots, x_{ki}): i = 1, 2, \ldots, n\}$ extracted from the population studied. If we write the population model for all observations of the sample, the following system is obtained:

$$\begin{aligned} y_1 &= \beta_1 + \beta_2 x_{21} + \beta_3 x_{31} + \ldots + \beta_k x_{k1} + u_1 \\ y_2 &= \beta_1 + \beta_2 x_{22} + \beta_3 x_{32} + \ldots + \beta_k x_{k2} + u_2 \\ &\ldots \quad\quad \ldots \quad\quad\quad \ldots \quad \ldots \\ y_n &= \beta_1 + \beta_2 x_{2n} + \beta_3 x_{3n} + \ldots + \beta_k x_{kn} + u_n \end{aligned} \qquad (3\text{-}11)$$

The previous system of equations can be expressed in a compact form by using matrix notation. Thus, we are going to denote

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{21} & x_{31} & \ldots & x_{k1} \\ 1 & x_{22} & x_{32} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{2n} & x_{3n} & \ldots & x_{kn} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \ldots \\ \beta_k \end{bmatrix} \qquad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \ldots \\ u_n \end{bmatrix}$$

The matrix **X** is called the matrix of regressors. Also included among the regressors is the regressor corresponding to the intercept. This one, which is often called *dummy* regressor, takes the value 1 for all the observations.

The model of multiple linear regression (3-11) expressed in matrix notation is the following:

$$\begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \ldots & x_{k1} \\ 1 & x_{22} & x_{32} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{2n} & x_{3n} & \ldots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \ldots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \ldots \\ u_n \end{bmatrix} \qquad (3\text{-}12)$$

If we take into account the denominations given to vectors and matrices, the model of multiple linear regression can be expressed in the following way:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad (3\text{-}13)$$

where **y** is a vector $n \times 1$, **X** is a matrix $n \times k$, $\boldsymbol{\beta}$ is a vector $k \times 1$ and **u** is a vector $n \times 1$.

3.1.2 Sample regression function

The basic idea of regression is to estimate the population parameters, $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ from a given sample.

The *sample regression function* (*SRF*) is the sample counterpart of the population regression function (*PRF*). Since the *SRF* is obtained for a given sample, a new sample will generate different estimates.

The *SRF*, which is an estimation of the *PRF*, is given by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \mathsf{L} + \hat{\beta}_k x_{ki} \qquad i = 1, 2, \mathsf{L}\ , n \text{ (3-14)}$$

The above expression allows us to calculate the *fitted value* ($\hat{y}_i$) for each y_i. In the *SRF* $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \mathsf{L}\ , \hat{\beta}_k$ are the estimators of the parameters $\beta_1, \beta_2, \beta_3, \mathsf{L}\ , \beta_k$.

We call residual to the difference between y_i and $\hat{y}_i$. That is

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \qquad \text{(3-15)}$$

In other words, the residual $\hat{u}_i$ is the difference between a sample value and its corresponding fitted value.

The system of equations (2-5) can be expressed in a compact form by using matrix notation. Thus, we are going to denote

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \cdots \\ \hat{y}_n \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \mathsf{M} \\ \hat{\beta}_k \end{bmatrix} \qquad \hat{\mathbf{u}} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \cdots \\ \hat{u}_n \end{bmatrix}$$

For all observations of the sample, the corresponding fitted model will be the following:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{(3-16)}$$

The residual vector is equal to the difference between the vector of observed values and the vector of fitted values, that is to say,

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \qquad \text{(3-17)}$$

3.2 Obtaining the *OLS* estimates, interpretation of the coefficients, and other characteristics

3.2.1 Obtaining the *OLS* estimates

Denoting *S* to the sum of the squared residuals,

$$S = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right]^2 \text{ (3-18)}$$

to apply the least squares criterion in the model of multiple linear regression, we calculate the first derivative from *S* with respect to each $\hat{\beta}_j$ in the expression (3-18):

$$\begin{aligned}
\frac{\partial S}{\partial \hat{\beta}_1} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right][-1] \\
\frac{\partial S}{\partial \hat{\beta}_2} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right][-x_{2i}] \\
\frac{\partial S}{\partial \hat{\beta}_3} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right][-x_{3i}] \qquad (3\text{-}19) \\
\mathsf{L} \quad & \mathsf{K} \quad \mathsf{L} \quad \mathsf{L} \\
\frac{\partial S}{\partial \hat{\beta}_k} &= 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right][-x_{ki}]
\end{aligned}$$

The least square estimators are obtained equaling to 0 the previous derivatives:

$$\begin{aligned}
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right] = 0 \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right] x_{2i} = 0 \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right] x_{3i} = 0 \qquad (3\text{-}20) \\
\mathsf{L} \quad & \mathsf{K} \quad \mathsf{L} \quad \mathsf{L} \\
&\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \mathsf{L} - \hat{\beta}_k x_{ki} \right] x_{ki} = 0
\end{aligned}$$

or, in matrix notation,

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \qquad (3\text{-}21)$$

The previous equations are denominated generically *hyperplane* normal equations.

In expanded matrix notation, the system of normal equations is the following:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki} \\ \sum_{i=1}^{n} x_{2i} & \sum_{i=1}^{n} x_{2i}^2 & \cdots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n} x_{ki}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{2i}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ki}y_i \end{bmatrix} \qquad (3\text{-}22)$$

Note that:

a) $\mathbf{X}'\mathbf{X}/n$ is the matrix of second order sample moments with respect to the origin, of the regressors, among which a dummy regressor (x_{1i}) associated to the intercept is included. This regressor takes the value x_{1i}=1 for all i.

b) $\mathbf{X}'\mathbf{y}/n$ is the vector of sample moments of second order, with respect to the origin, between the regressand and the regressors.

In this system there are *k* equations and *k* unknown $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k)$. This system can easily be solved using matrix algebra. In order to solve univocally the system (3-21)with respect to $\hat{\boldsymbol{\beta}}$, it must be held that the rank of the matrix $\mathbf{X}'\mathbf{X}$ is equal to *k*. If this is held, both members of (3-21) can be premultiplied by $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$:

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}$$

with which the expression of the vector of least square estimators, or more precisely, the vector of *ordinary* least square estimators (OLS), is obtained because $\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$. Therefore, the solution is the following:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \text{M} \\ \hat{\beta}_k \end{bmatrix} = \hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} \qquad (3\text{-}23)$$

Since the matrix of second derivatives, $2\mathbf{X}'\mathbf{X}$, is a positive definite matrix, the conclusion is that *S* presents a minimum in $\hat{\boldsymbol{\beta}}$.

3.2.2 Interpretation of the coefficients

A $\hat{\beta}_j$ coefficient measures the *partial* effect of the regressor x_j on *y* holding the other regressors fixed. We will see next the meaning of this expression.

The fitted model for observation *i* is given by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \ldots + \hat{\beta}_j x_{ji} + \ldots + \hat{\beta}_k x_{ki} \qquad (3\text{-}24)$$

Now, let us consider the fitted model for observation *h* in which the values of the regressors and, consequently, *y* will have changed with respect to (3-24):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_{2h} + \hat{\beta}_3 x_{3h} + \ldots + \hat{\beta}_j x_{jh} + \ldots + \hat{\beta}_k x_{kh} \qquad (3\text{-}25)$$

Subtracting (3-25) from (3-24), we have

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2 + \hat{\beta}_3 \Delta x_3 + \ldots + \hat{\beta}_j \Delta x_j + \ldots + \hat{\beta}_k \Delta x_k \qquad (3\text{-}26)$$

where $\Delta\hat{y} = \hat{y}_i - \hat{y}_h, \Delta x_2 = x_{2i} - x_{2h}, \Delta x_3 = x_{3i} - x_{3h}, \ldots \Delta x_k = x_{ki} - x_{kh}$.

The previous expression captures the variation of $\hat{y}$ due to the changes in all regressors. If only x_j changes, we will have

$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j \qquad (3\text{-}27)$$

If x_k increases in one unit, we will have

$$\Delta\hat{y} = \hat{\beta}_j \quad \text{for } \Delta x_j = 1 \qquad (3\text{-}28)$$

Consequently, the coefficient $\hat{\beta}_j$ measures the change in *y* when x_j increases in 1 unit, *holding the regressors* $x_2, x_3, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ *fixed*. It is very important to take into account this *ceteris paribus* clause when interpreting the coefficient.

This interpretation is not valid, of course, for the intercept.

EXAMPLE 3.1 Quantifying the influence of age and wage on absenteeism in the firm Buenosaires

Buenosaires is a firm devoted to manufacturing fans, having had relatively acceptable results in recent years. The managers consider that these would have been better if the absenteeism in the company were not so high. For this purpose, the following model is proposed:

$$absent = \beta_1 + \beta_2 age + \beta_3 tenure + \beta_4 wage + u$$

where *absent* is measured in days per year; *wage* in thousands of euros per year; *tenure* in years in the firm and *age* is expressed in years.

Using a sample of size 48 (file *absent*), the following equation has been estimated:

$$\widehat{absent} = \underset{(1.603)}{14.413} - \underset{(0.048)}{0.096}\, age - \underset{(0.067)}{0.078}\, tenure - \underset{(0.007)}{0.036}\, wage$$

$$R^2 = 0.694 \qquad n = 48$$

The interpretation of $\hat{\beta}_2$ is the following: holding fixed *tenure* and *wage*, if age increases by one year, worker absenteeism will be reduced by 0.096 days per year. The interpretation of $\hat{\beta}_3$ is as follows: holding fixed the *age* and *wage*, if the tenure increases by one year, worker absenteeism will be reduced by 0.078 days per year. Finally, the interpretation of $\hat{\beta}_4$ is the following: holding fixed the *age* and *tenure*, if the wage increases by 1000 euros per year, worker absenteeism will be reduced by 0.036 days per year.

EXAMPLE 3.2 Demand for hotel services

The following model is formulated to explain the demand for hotel services:

$$\ln(hostel) = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + u \qquad (3\text{-}29)$$

where *hostel* is spending on hotel services, *inc* is disposable income, both of which are expressed in euros per month. The variable *hhsize* is the number of household members.

The estimated equation with a sample of 40 households, using file *hostel*, is the following:

$$\widehat{\ln(hostel_i)} = -27.36 + 4.442 \ln(inc_i) - 0.523 hhsize_i$$

$$R^2 = 0.738 \qquad n = 40$$

As the results show, hotel services are a luxury good. Thus, the demand/income elasticity for this good is very high (4.44), which is typical of luxury goods. This means that if income increases by 1%, spending on hotel services increases by 4.44%, holding fixed the size of the household. On the other hand, if the household size increases by one member, then spending on hotel services will decrease by 52%.

EXAMPLE 3.3 A hedonic regression for cars

The hedonic model of price measurement is based on the assumption that the value of a good is derived from the value of its characteristics. Thus, the price of a car will therefore depend on the value the buyer places on both qualitative (e.g. automatic gear, power, diesel, assisted steering, air conditioning), and quantitative attributes (e.g. fuel consumption, weight, performance displacement, etc.). The data set for this exercise is file *hedcarsp* (hedonic car price for Spain) and covers years 2004 and 2005. A first model based only on quantitative attributes is the following:

$$\ln(price) = \beta_1 + \beta_2 volume + \beta_3 fueleff + u$$

where *volume* is length×width×height in m^3 and *fueleff* is the liters per 100 km/horsepower ratio expressed as a percentage.

The estimated equation with a sample of 214 observations is the following:

$$\widehat{\ln(price_i)} = 14.97 + 0.0956 volume_i - 0.1608\, fueleff_i$$

$$R^2=0.765 \quad n=214$$

The interpretation of $\hat{\beta}_2$ and $\hat{\beta}_3$ is the following. Holding fixed *fueleff*, if *volume* increases by 1 m^3, the price of a car will rise by 9.56%. Holding fixed *volume*, if the ratio liters per 100 km/horsepower increases by 1 percentage point, the price of a car price will fall by 16.08%.

EXAMPLE 3.4 Sales and advertising: the case of Lydia E. Pinkham

A model with time series data is estimated in order to measure the effect of advertising expenses, realized over different time periods, on current sales. Denoting by V_t and P_t sales and advertising expenditures, made at time *t*, the model proposed initially to explain sales, as a function of current and past advertising expenses is as follows:

$$V_t = \alpha + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 P_{t-2} + \cdots + u_t \qquad (3\text{-}30)$$

In the above expression the dots indicate that past expenditure on advertising continues to have an indefinite influence, although it is assumed that with a decreasing impact on sales. The above model is not operational given that it has an indefinite number of coefficients. Two approaches can be adopted in order to solve the problem. The first approach is to fix a priori the maximum number of periods during which advertising effects on sales are maintained. In the second approach, the coefficients behave according to some law which determines their value based on a small number of parameters, also allowing further simplification.

In the first approach the problem that arises is that, in general, there are no precise criteria or sufficient information to fix *a priori* the maximum number of periods. For this reason, we shall look at a special case of the second approach that is interesting due to the plausibility of the assumption and easy application. Specifically, we will consider the case in which the coefficients β_i decrease geometrically as we move backward in time according to the following scheme:

$$\beta_i = \beta_1 \lambda^i \quad \forall i \qquad 0 < \lambda < 1 \qquad (3\text{-}31)$$

The above transformation is called Koyck transformation, as it was this author who in 1954 introduced scheme (3-31) for the study of investment

Substituting (3-31) in (3-30), we obtain

$$V_t = \alpha + \beta_1 P_t + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \cdots + u_t \qquad (3\text{-}32)$$

The above model still has infinite terms, but only three parameters and can also be simplified. Indeed, if we express equation (3-32) for period *t*-1 and multiply both sides by λ we obtain

$$\lambda V_{t-1} = \alpha\lambda + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \beta_1 \lambda^3 P_{t-3} + \cdots + \lambda u_{t-1} \qquad (3\text{-}33)$$

Subtracting (3-33) from (3-32), and taking into account factors λ^i tend to 0 as *i* tends to infinity, the result is the following:

$$V_t = \alpha(1-\lambda) + \beta_1 P_t + \lambda V_{t-1} + u_t - \lambda u_{t-1} \qquad (3\text{-}34)$$

The model has been simplified so that it only has three regressors, although, in contrast, it has moved to a compound disturbance term. Before seeing the application of this model, we will analyze the significance of the coefficient λ and the duration of the effects of advertising expenditures on sales. The parameter λ is the decay rate of the effects of advertising expenditures on current and future sales. The cumulative effects that the advertising expenditure of one monetary unit have on sales after *m* periods are given by

$$\beta_1(1 + \lambda + \lambda^2 + \lambda^3 + \cdots + \lambda^m) \qquad (3\text{-}35)$$

To calculate the cumulative sum of effects, given in (3-35), we note that this expression is the sum of the terms of a geometric progression[2], which can be expressed as follows:

$$\frac{\beta_1(1-\lambda^m)}{1-\lambda} \tag{3-36}$$

When *m* tends to infinity, then the sum of the cumulative effects is given by

$$\frac{\beta_1}{1-\lambda} \tag{3-37}$$

An interesting point is to determine how many periods of time are required to obtain the *p*% (e.g., 50%) of the total effect. Denoting by *h* the number of periods required to obtain this percentage, we have

$$p = \frac{\text{Effect in } h \text{ periods}}{\text{Total effect}} = \frac{\dfrac{\beta_1(1-\lambda^h)}{1-\lambda}}{\dfrac{\beta_1}{1-\lambda}} = 1-\lambda^h \tag{3-38}$$

Setting *p*, *h* can be calculated according to (3-38). Solving for *h* in this expression, the following is obtained

$$h = \frac{\ln(1-p)}{\ln \lambda} \tag{3-39}$$

This model was used by Kristian S. Palda in his doctoral thesis published in 1964, entitled *The Measurement of Cumulative Advertising Effects*, to analyze the cumulative effects of advertising expenditures in the case of the company Lydia E. Pinkham. This case has been the basis for research on the effects of advertising expenditures. We will see below some features of this case:

1) The Lydia E. Pinkham Medicine Company manufactured a herbal extract diluted in an alcohol solution. This product was originally announced as an analgesic and also as a remedy for a wide variety of diseases.

2) In general, in different types of products there is often competition among different brands, as in the paradigmatic case of Coca-Cola and Pepsi-Cola. When this occurs, the behavior of the main competitors is taken into account when analyzing the effects of advertising expenditure. Lydia E. Pinkham had the advantage of having no competitors, acting as a monopolist in practice in its product line.

3) Another feature of the Lydia E. Pinkham case was that most of the distribution costs were allocated to advertising because the company had no commercial agents, with the relationship between advertising expenses and sales being very high.

4) The product was affected by different avatars. Thus, in 1914 the Food and Drug Administration (United States agency established controls for food and medicines) accused the firm of misleading advertising and so they had to change their advertising messages. Also, the Internal Revenue (IRS) threatened to apply a tax on alcohol since the alcohol content of the product was 18%. For all these reasons there were changes in the presentation and content during the period 1915-1925. In 1925 the Food and Drug Administration banned the product from being announced as medicine, having to be distributed as a tonic drink. In the period 1926-1940 spending on advertising was significantly increased and shortly after the sales of the product declined.

The estimation of the model (3-34) with data from 1907 to 1960, using file *pinkham*, is the following:

[2] Denoting by a_p, a_u and r the first term, the last term and the right respectively, the sum of the terms of a convergent geometric progression is given by

$$\frac{a_p - a_u}{1-r}$$

$$\widehat{sales}_t = 138.7 + 0.3288 advexp + 0.7593 sales_{t-1}$$

$$R^2=0.877 \quad n=53$$

The sum of the cumulative effects of advertising expenditures on sales is calculated by the formula (3-37):

$$\frac{\hat{\beta}_1}{1-\hat{\lambda}} = \frac{0.3288}{1-0.7593} = 1.3660$$

According to this result, every additional dollar spent on advertising produces an accumulated total sale of 1,366 units. Since it is important not only to determine the overall effect, but also how long the effect lasts, we will now answer the following question: how many periods of time are required to reach half of the total effects? Applying the formula (3-39) for the case of $p = 0.5$, the following result is obtained:

$$\hat{h}(0.5) = \frac{\ln(1-0.5)}{\ln(0.7593)} = 2.5172$$

3.2.3 Algebraic implications of the estimation

The algebraic implications of the estimation are derived exclusively from the application of the *OLS* method to the model of multiple linear regression:

1. *The sum of the OLS residuals is equal to 0*:

$$\sum_{i=1}^{n} \hat{u}_i = 0 \qquad (3\text{-}40)$$

From the definition of residual

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \mathrm{L} - \hat{\beta}_k x_{ki} \qquad i = 1, 2, \mathrm{L}\ , n \qquad (3\text{-}41)$$

If we add for the *n* observations, then

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \mathrm{L} - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \qquad (3\text{-}42)$$

On the other hand, the first equation of the system of normal equations (3-20) is

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \mathrm{L} - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = 0 \qquad (3\text{-}43)$$

If we compare (2-21) and (3-43), we conclude that (2-19) holds.

Note that, if (2-19) holds, it implies that

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} \hat{y}_i \qquad (3\text{-}44)$$

and, dividing (2-19) and (3-44) by *n*, we obtain

$$\bar{\hat{u}} = 0 \qquad \bar{y} = \bar{\hat{y}} \qquad (3\text{-}45)$$

2. The OLS hyperplane always goes through the point of the sample means $\left(\bar{y}, \bar{x}_2, \mathrm{L}\ , \bar{x}_k\right)$.

By dividing equation (3-43) by *n* we have:

$$\overline{y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{x}_2 + \text{L} + \hat{\beta}_k \overline{x}_k \qquad (3\text{-}46)$$

3. The sample cross product between each one of the regressors and the OLS residuals is zero

$$\sum_{i=1}^{n} x_{ji} \hat{u}_i = 0 \qquad j = 2,3,\text{L}\ ,k \quad (3\text{-}47)$$

Using the last *k* normal equations (3-20) and taking into account that by definition $\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \text{L} - \hat{\beta}_k x_{ki}$, we can see that

$$\begin{aligned} &\sum_{i=1}^{n} \hat{u}_i x_{2i} = 0 \\ &\sum_{i=1}^{n} \hat{u}_i x_{3i} = 0 \\ &\text{L} \qquad\qquad \text{L} \\ &\sum_{i=1}^{n} \hat{u}_i x_{ki} = 0 \end{aligned} \qquad (3\text{-}48)$$

4. The sample cross product between the fitted values ($\hat{y}$) and the OLS residuals is zero.

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \qquad (3\text{-}49)$$

Taking into account (2-19) and (3-48), we obtain

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \text{L} + \hat{\beta}_k x_{ki}) \hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} \hat{u}_i + \text{L} \ \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \hat{u}_i$$
$$= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 + \text{L} \ \hat{\beta}_k \times 0 = 0 \qquad (3\text{-}50)$$

3.3 Assumptions and statistical properties of the *OLS estimators*

Before studying the statistical properties of the *OLS* estimators in the multiple linear regression model, we need to formulate a set of statistical assumptions. Specifically, the set of assumptions that we will formulate are called *classical linear model* (*CLM*) *assumptions*. It is important to note that *CLM* assumptions are simple, and that the *OLS* estimators have, under these assumptions, very good properties.

3.3.1 Statistical assumptions of the *CLM* in multiple linear regression)

a) Assumption on the functional form

1) The relationship between the regressand, the regressors and the disturbance is linear in the parameters:

$$y = \beta_1 + \beta_2 x_2 + \text{L} + \beta_k x_k + u \qquad (3\text{-}51)$$

or, alternatively, for all the observations,

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u} \qquad (3\text{-}52)$$

b) Assumptions on the regressors

2) The values of $x_2, x_3 \ldots, x_k$ *are fixed in repeated sampling, or the matrix* **X** *is fixed in repeated sampling*:

This is a strong assumption in the case of the social sciences where, in general, it is not possible to experiment. An alternative assumption can be formulated as follows:

2) The regressors* $x_2, x_3, \ldots, x_k$ *are distributed independently of the random disturbance.* Formulated in another way, **X** is distributed independently of the vector of random disturbances, which implies that $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$

As we said in chapter 2, we will adopt assumption 2).

3) The matrix of regressors, **X**, *does not contain disturbances of measurement*

4) The matrix of regressors, **X**, *has rank k*:

$$\rho(\mathbf{X}) = k \qquad (3\text{-}53)$$

Recall that the matrix of regressors contains *k* columns, corresponding to the *k* regressors in the model, and *n* rows, corresponding to the number of observations. This assumption has two implications:

1. The number of observations, *n*, must be equal to or greater than the number of regressors, *k*. Intuitively, to estimate *k* parameters, we need at least *k* observations.

2. Each regressor must be linearly independent, which implies that an exact linear relationship among any subgroup of regressors cannot exist. If an independent variable is an exact linear combination of other independent variables, then there is *perfect multicollinearity*, and the model cannot be estimated.

If an approximate linear relationship exists, then estimations of the parameters can be obtained, although the reliability of such estimations would be affected. In this case, there is *non-perfect multicollinearity*.

c) Assumption on the parameters

5) The parameters $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ *are constant, or* $\boldsymbol{\beta}$ *is a constant vector*.

d) Assumptions on the disturbances

6) The disturbances have zero mean,

$$E(u_i) = 0, \quad i = 1, 2, 3, \ldots, n \quad \text{or} \quad E(\mathbf{u}) = \mathbf{0} \qquad (3\text{-}54)$$

7) The disturbances have a constant variance (homoskedasticity assumption):

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \qquad (3\text{-}55)$$

8) The disturbances with different subscripts are not correlated with each other (no autocorrelation assumption):

$$E(u_i u_j) = 0 \qquad i \neq j \qquad (3\text{-}56)$$

The formulation of *homoskedasticity* and *no autocorrelation* assumptions allows us to specify the covariance matrix of the disturbance vector:

$$E\left[\left[\mathbf{u}-E(\mathbf{u})\right]\left[\mathbf{u}-E(\mathbf{u})\right]'\right]=E\left[\left[\mathbf{u}-\mathbf{0}\right]\left[\mathbf{u}-\mathbf{0}\right]'\right]=E\left[\left[\mathbf{u}\right]\left[\mathbf{u}\right]'\right]$$

$$=E\left[\begin{bmatrix}u_1\\u_2\\\vdots\\u_n\end{bmatrix}\begin{bmatrix}u_1 & u_2 & \cdots & u_n\end{bmatrix}\right]=E\begin{bmatrix}u_1^2 & u_1u_2 & \cdots & u_1u_n\\u_2u_1 & u_2^2 & \cdots & u_2u_n\\\vdots & \vdots & \ddots & \vdots\\u_nu_1 & u_nu_2 & \cdots & u_n^2\end{bmatrix}$$

$$=\begin{bmatrix}E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n)\\E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n)\\\vdots & \vdots & \ddots & \vdots\\E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2)\end{bmatrix}=\begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\0 & \sigma^2 & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & \sigma^2\end{bmatrix}$$

(3-57)

In order to get to the last equality, it has been taken into account that the variances of each one of the elements of the vector is constant and equal to σ^2 in accordance with (3-55) and the covariances between each pair of elements is 0 in accordance with (3-56).

The previous result can be expressed in synthetic form:

$$E(\mathbf{uu}')=\sigma^2\mathbf{I} \quad (3\text{-}58)$$

The matrix given in (3-58) is denominated *scalar matrix*, since it is a scalar (σ^2, in this case) multiplied by the identity matrix.

9) The disturbance u is normally distributed

Taking into account assumptions 6 to 9, we have

$$u_i \sim NID(0,\sigma^2) \quad i=1,2,\cdots,n \qquad \text{or} \qquad \mathbf{u}\sim N(\mathbf{0},\sigma^2\mathbf{I}) \quad (3\text{-}59)$$

where *NID* stands for *normally independently distributed*.

3.3.2 Statistical properties of the OLS estimator

Under the above assumptions of the *CLM*, the *OLS* estimators possess good properties. In the proofs of this section, assumptions 3, 4 and 5 will implicitly be used.

Linearity and unbiasedness of the OLS estimator

Now, we are going to prove that the OLS estimator is linearly unbiased. First, we express $\hat{\boldsymbol{\beta}}$ as a function of the vector **u**, using assumption 1, according to (3-52):

$$\hat{\boldsymbol{\beta}}=\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y}=\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\left[\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right]=\boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u} \quad (3\text{-}60)$$

The *OLS* estimator can be expressed in this way so that the property of linearity is clearer:

$$\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}=\boldsymbol{\beta}+\mathbf{Au} \quad (3\text{-}61)$$

where $\mathbf{A} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'$ is fixed under assumption 2. Thus $\hat{\boldsymbol{\beta}}$ is a linear function of u and, consequently, it is a *linear* estimator.

Taking expectations in (3-60) and using assumption 6, we obtain

$$E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} + \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E\left[\mathbf{u}\right] = \boldsymbol{\beta} \qquad (3\text{-}62)$$

Therefore, $\hat{\boldsymbol{\beta}}$ is an *unbiased* estimator.

Variance of the OLS estimators

In order to calculate the covariance matrix of $\hat{\boldsymbol{\beta}}$ assumptions 7 and 8 are needed, in addition to the first six assumptions:

$$\begin{aligned} \text{var}(\hat{\boldsymbol{\beta}}) &= E\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]' = E\left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right]\left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}\right]' \\ &= E\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\mathbf{u}\mathbf{u}')\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} \qquad (3\text{-}63) \\ &= \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'E(\sigma^2\mathbf{I})\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} \end{aligned}$$

In the third step of the above proof it is taken into account that, according to (3-60), $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{u}$. Assumption 2 is taken into account in the fourth step. Finally, assumptions 7 and 8 are used in the last step.

Therefore, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\mathbf{X}'\mathbf{X}\right]^{-1}$ is the covariance matrix of the vector $\hat{\boldsymbol{\beta}}$. In this covariance matrix, the variance of each element $\hat{\beta}_j$ appears on the main diagonal, while the covariances between each pair of elements are outside of the main diagonal. Specifically, the variance of $\hat{\beta}_j$ (for j=2,3,…,k) is equal to σ^2 multiplied by the corresponding element of the main diagonal of $\left[\mathbf{X}'\mathbf{X}\right]^{-1}$. After operating, the variance of $\hat{\beta}_j$ can be expressed as

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{nS_j^2(1-R_j^2)} \qquad (3\text{-}64)$$

where R_j^2 is the R-squared from regressing x_j on all other x's, n is the sample size and S_j^2 is the sample variance of the regressor X.

Formula (3-64) is valid for all slope coefficients, but not for the intercept

The square root of (3-64) is called the *standard deviation of* $\hat{\beta}_j$:

$$sd(\hat{\beta}_j) = \frac{\sigma}{\sqrt{nS_j^2(1-R_j^2)}} \qquad (3\text{-}65)$$

OLS estimators are BLUE

Under assumptions 1 through 8 of the *CLM*, which are called Gauss-Markov assumptions, the OLS estimators is the *Best Linear Unbiased Estimators* (BLUE).

The Gauss Markov theorem states that the *OLS* estimator is the best estimator within the class of linear unbiased estimators. In this context, *best* means that it is an estimator with the smallest variance for a given sample size. Let us now compare the variance of an element of $\hat{\boldsymbol{\beta}}$ ($\hat{\beta}_j$), with any other estimator that is linear (so $\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$) and unbiased (so the weights, w_j, must satisfy some restrictions). The property of $\hat{\beta}_j$ being a *BLUE* estimator has the following implications when comparing its variance with the variance of $\tilde{\beta}_j$:

1) The variance of the coefficient $\tilde{\beta}_j$ is greater than, or equal to, the variance of $\hat{\beta}_j$ obtained by *OLS*:

$$\text{var}(\tilde{\beta}_j) \geq \text{var}(\hat{\beta}_j) \qquad j = 1, 2, 3, \ldots, k \qquad (3\text{-}66)$$

2) The variance of any linear combination of $\tilde{\beta}_j$´s is greater than, or equal to, the variance of the corresponding linear combination of $\hat{\beta}_j$'s.

In appendix 3.1 the proof of the theorem of Gauss-Markov can be seen.

Estimator of the disturbance variance

Taking into account the system of normal equations (3-20), if we know *n–k* of the residuals, we can get the other *k* residuals by using the restrictions imposed by that system in the residuals.

For example, the first normal equation allows us to obtain the value of $\hat{u}_n$ as a function of the remaining residuals:

$$\hat{u}_n = -\hat{u}_1 - \hat{u}_2 - \ldots - \hat{u}_{n-1}$$

Thus, there are only *n–k* degrees of freedom in the OLS residuals, as opposed to *n* degrees of freedom in the disturbances. Remember that the degree of freedom is defined as the difference between the number of observations and the number of parameters estimated.

The unbiased estimator of σ^2 is adjusted taken into account the degree of freedom:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-k} \qquad (3\text{-}67)$$

Under assumptions 1 to 8, we obtain

$$E(\hat{\sigma}^2) = \sigma^2 \qquad (3\text{-}68)$$

See appendix 3.2 for the proof.

The square root of (3-67), $\hat{\sigma}$ is called *standard error of the regression* and is an estimator of σ.

Estimators of the variances of $\hat{\boldsymbol{\beta}}$ and the slope coefficient $\hat{\beta}_j$

The estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\bar{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} \bar{\text{var}}(\hat{\beta}_1) & \bar{Cov}(\hat{\beta}_1,\hat{\beta}_2) & \cdots & \bar{Cov}(\hat{\beta}_1,\hat{\beta}_j) & \cdots & \bar{Cov}(\hat{\beta}_1,\hat{\beta}_k) \\ \bar{Cov}(\hat{\beta}_2,\hat{\beta}_1) & \bar{\text{var}}(\hat{\beta}_2) & \cdots & \bar{Cov}(\hat{\beta}_2,\hat{\beta}_j) & \cdots & \bar{Cov}(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \bar{Cov}(\hat{\beta}_j,\hat{\beta}_1) & \bar{Cov}(\hat{\beta}_j,\hat{\beta}_2) & \cdots & \bar{\text{var}}(\hat{\beta}_j) & \cdots & \bar{Cov}(\hat{\beta}_j,\hat{\beta}_k) \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ \bar{Cov}(\hat{\beta}_k,\hat{\beta}_1) & \bar{Cov}(\hat{\beta}_k,\hat{\beta}_2) & \cdots & \bar{Cov}(\hat{\beta}_k,\hat{\beta}_j) & \cdots & \bar{\text{var}}(\hat{\beta}_k) \end{bmatrix} \tag{3-69}$$

The variance of the slope coefficient $\hat{\beta}_j$, given in (3-64), is a function of the unknown parameter σ^2. When σ^2 is substituted by its estimator $\hat{\sigma}^2$, an estimator of the variance of $\hat{\beta}_j$ is obtained:

$$\bar{\text{var}}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \tag{3-70}$$

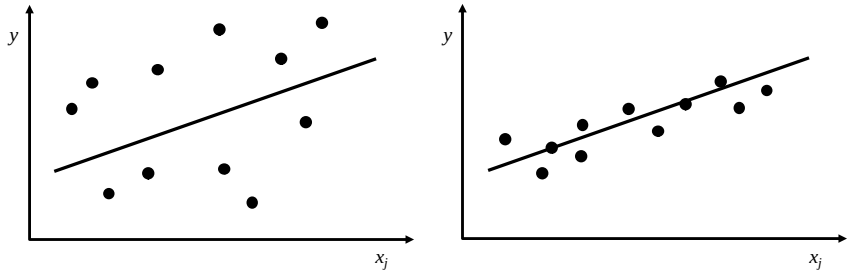
According to the previous expression, the estimator of the variance $\hat{\beta}_j$ is affected by the following factors:

a) The greater $\hat{\sigma}^2$, the greater the variance of the estimator. This is not at all surprising: more "noise" in the equation - a larger $\hat{\sigma}^2$ - makes it more difficult to estimate accurately the partial effect of any *x*'s on y. (See figure 3.1).

b) As sample size increases, the variance of the estimator is reduced.

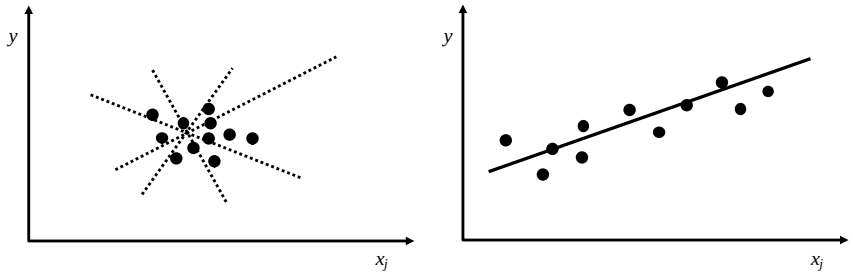
c) The smaller the sample variance of a regressor, the greater the variance of the corresponding coefficient. Everything else being equal, for estimating β_j we prefer to have as much sample variation in x_j as possible, which is illustrated in figure 3.2. As you can see, there are many hypothetical lines that could fit the data when the sample variance of x_j (S_j^2) is small, which can be seen in part a) of the figure. In any case, assumption 4 does not allow S_j^2 being equal to 0.

d) The higher R_j^2, (i.e., the higher is the correlation of regressor *j* with the rest of the regressors), the greater the variance of $\hat{\beta}_j$.



a) $\hat{s}^2$ big b) $\hat{s}^2$ small

FIGURE 3.1. Influence of $\hat{s}^2$ on the estimator of the variance.



a) S_j^2 small b) S_j^2 big

FIGURE 3.2. Influence of S_j^2 on the estimator of the variance.

The square root of (3-70) is called the *standard error of* $\hat{\beta}_j$:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{nS_j^2(1-R_j^2)}} \qquad (3\text{-}71)$$

Other properties of the OLS estimators

Under 1 through 6 *CLM* assumptions, the *OLS* estimator $\hat{\boldsymbol{\beta}}$ is *consistent*, as can be seen in appendix 3.3, *asymptotically normally distributed* and also *asymptotically efficient* within the class of the consistent and asymptotically normal estimators.

Under 1 through 9 *CLM* assumptions, the *OLS* estimator is *also* the *maximum likelihood estimator* (*ML*), as can be seen in appendix 3.4, and the *minimum variance unbiased estimator (MVUE)*. This means that the *OLS* estimator has the smallest variance among all unbiased, linear o non linear, estimators.

3.4 More on functional forms

In this section we will examine two topics on functional forms: use of natural logs in models and polynomial functions.

3.4.1 Use of logarithms in the econometric models

Some variables are often used in log form. This is the case of variables in monetary terms which are generally positive or variables with high values such as population. Using models with log transformations also has advantages, one of which is that coefficients have appealing interpretations (elasticity or semi-elasticity). Another advantage is the invariance of slopes to scale changes in the variables. Taking logs is also very useful

because it narrows the range of variables, which makes estimates less sensitive to extreme observations on the dependent or the independent variables. The *CLM* assumptions are satisfied more often in models using ln(*y*) as a regressand than in models using *y* without any transformation. Thus, the conditional distribution of *y* is frequently heteroskedastic, while ln(*y*) can be homoskedastic.

One limitation of the log transformation is that it cannot be used when the original variable takes zero or negative values. On the other hand, variables measured in years and variables that are a proportion or a percentage, are often used in level (or original) form.

3.4.2 Polynomial functions

The polynomial functions have been extensively used in econometric research. When there are only the regressors corresponding to a polynomial function we have a *polynomial model*. The general k^{th} *degree polynomial model may* be written as

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_k x^k + u \qquad (3\text{-}72)$$

Quadratic functions

An interesting case of polynomial functions is the *quadratic function*, which is a *second-degree polynomial function*. When there are only regressors corresponding to the quadratic function, we have a *quadratic model*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + u \qquad (3\text{-}73)$$

Quadratic functions are used quite often in applied economics to capture decreasing or increasing marginal effects. It is important to remark that, in such a case, β_2 does not measure the change in *y* with respect to *x* because it makes no sense to hold x^2 fixed while changing *x*. The marginal effect of *x* on *y*, which depends linearly on the value of *x*, is the following:

$$me = \frac{dy}{dx} = \beta_2 + 2\beta_3 x \qquad (3\text{-}74)$$

In a particular application this marginal effect would be evaluated at specific values of *x*. If β_2 and β_3 have opposite signs the turning point will be at

$$x^* = -\frac{\beta_2}{2\beta_3} \qquad (3\text{-}75)$$

If $\beta_2>0$ and $\beta_3<0$, then the marginal effect of *x* on *y* is positive at first, but it will be negative for values of *x* greater than x^*. If $\beta_2<0$ and $\beta_3>0$, this marginal effect is negative at first, but it will be positive for values of *x* greater than x^*.

Example 3.5 Salary and tenure

Using the data in *ceosal2* to study the type of relation between the *salary* of the Chief Executive Officers (CEOSs) in USA corporations and the number of years in the company as CEO (*ceoten*), the following model was estimated:

$$\widehat{\ln(salary)} = \underset{(0.086)}{6.246} + \underset{(0.0001)}{0.0006}\, profits + \underset{(0.0156)}{0.0440}\, ceoten - \underset{(0.00052)}{0.0012}\, ceoten^2$$

$$R^2=0.1976 \quad n=177$$

where company *profits* are in millions of dollars and *salary* is annual compensation in thousands of dollars.

The marginal effect *ceoten* on *salary* expressed in percentage is the following:

$$\widehat{me}_{salary/ceoten}\% = 4.40 - 2 \times 0.12 ceoten$$

Thus, if a CEO with 10 years in a company spends one more year in that company, their salary will increase by 2%. Equating to zero the previous expression and solving for *ceoten*, we find that the maximum effect of tenure as CEO on salary is reached by 18 years. That is, until 18 years the marginal effect of CEO tenure on the salary is positive. On the contrary, from 18 years onwards this marginal effect is negative.

Cubic functions

Another interesting case is the *cubic function*, or *third-degree polynomial function*. If in the model there are only regressors corresponding to the cubic function, we have a *cubic model*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + u \qquad (3\text{-}76)$$

Cubic models are used quite often in applied economics to capture decreasing or increasing marginal effects, particularly in the cost functions. The marginal effect (*me*) of *x* on *y*, which depends on *x* in a quadratic form, will be the following:

$$me = \frac{dy}{dx} = \beta_2 + 2\beta_3 x + 3\beta_4 x^2 \qquad (3\text{-}77)$$

The minimum of *me* will occur where

$$\frac{dme}{dx} = 2\beta_3 + 6\beta_4 x = 0 \qquad (3\text{-}78)$$

Therefore,

$$me_{\min} = \frac{-\beta_3}{3\beta_4} \qquad (3\text{-}79)$$

In a cubic model of a cost function, the restriction $\beta_3^2 < 3\beta_4\beta_2$ must be met to guarantee that the *minimum marginal cost* is positive. Other restrictions that a cost function must satisfy are as follows: β_1, β_2, and $\beta_4>0$; and $\beta_3<0$

Example 3.6 The marginal effect in a cost function

Using the data on 11 pulp mill firms (file *costfunc*) to study the cost function, the following model was estimated:

$$\widehat{cost} = \underset{(1.602)}{29.16} + \underset{(0.2167)}{2.316}\, output - \underset{(0.0081)}{0.0914}\, output^2 + \underset{(0.000086)}{0.0013}\, output^3$$

$$R^2=0.9984 \quad n=11$$

where *output* is the production of pulp in thousands of tons and *cost* is the total cost in millions of euros

The *marginal cost* is the following:

$$\widehat{marcost} = 2.316 - 2 \times 0.0914 output + 3 \times 0.0013 output^2$$

Thus, if a firm with a production of 30 thousand tons of pulp increases the pulp production by one thousand tons, the cost will increase by 0.754 million of euros. Calculating the minimum of the above expression and solving for *output*, we find that the minimum marginal cost is equal to a production of 23.222 thousand tons of pulp.

3.5 Goodness-of-fit and selection of regressors.

Once least squares have been applied, it is very useful to have some measure of the goodness of fit between the model and the data. In the event that several alternative models have been estimated, measures of the goodness of fit could be used to select the most appropriate model.

In econometric literature there are numerous measures of goodness of fit. The most popular is the coefficient of determination, which is designated by R^2 or *R*-squared, and the adjusted coefficient of determination, which is designated $\overline{R}^2$ or adjusted *R*-squared. Given that these measures have some limitations, the Akaike Information Criterion (*AIC*) and Schwarz Criterion (*SC*) will also be referred to later on.

3.5.1 Coefficient of determination

As we saw in chapter 2, the coefficient of determination is based on the following breakdown:

$$TSS = ESS + RSS \qquad (3\text{-}80)$$

where *TSS* is the *total sum of squares*, *ESS* is the *explained sum of squares* and *RSS* is the *residual sum of squares*.

Based on this breakdown, the coefficient of determination is defined as:

$$R^2 = \frac{ESS}{TSS} \qquad (3\text{-}81)$$

Alternatively, and in an equivalent manner, the coefficient of determination can be defined as

$$R^2 = 1 - \frac{RSS}{TSS} \qquad (3\text{-}82)$$

The extreme values of the coefficient of determination are: 0, when the explained variance is zero, and 1, when the residual variance is zero; that is, when the fit is perfect. Therefore,

$$0 \leq R^2 \leq 1 \qquad (3\text{-}83)$$

A small R^2 implies that the disturbance variance (σ^2) is large relative to the variance of *y*, which means that β_j is not estimated with precision. But remember that a large disturbance variance can be offset by a large sample size. Thus, if *n* is large enough, we may be able to estimate the coefficients with precision even though we have not controlled for many unobserved factors.

To interpret the coefficient of determination properly, the following caveats should be taken into account:

a) As new explanatory variables are added, the coefficient of determination increases its value or, at least, keeps the same value. This happens even though the variable (or variables) added have no relation to the endogenous variable. Thus, we can always verify that

$$R_j^2 \geq R_{j-1}^2 \qquad (3\text{-}84)$$

where R^2_{j-1} the R is squared in a model with j-1 regressors, and R^2_j is the R squared in a model with an additional regressor. That is to say, if we add variables to a given model, R^2 will never decrease, even if these variables do not have a significant influence.

b) If the model has no intercept, the coefficient of determination does not have a clear interpretation because the decomposition given (3-80) is not fulfilled. In addition, the two forms of calculation mentioned - (3-81) and (3-82) - generally lead to different results, which in some cases may fall outside the interval [0, 1].

c) The coefficient of determination cannot be used to compare models in which the functional form of the endogenous variable is different. For example, R^2 cannot be applied to compare two models in which the regressand is the original variable, y, and ln(y) respectively.

3.5.2 Adjusted *R*-Squared

To overcome one of the limitations of the R^2, we can "adjust" it in a way that takes into account the number of variables included in a given model. To see how the usual R^2 might be adjusted, it is useful to write it as

$$R^2 = 1 - \frac{RSS / n}{TSS / n} \qquad (3\text{-}85)$$

where, in the second term of the right-hand side, the residual variance is divided by the variance of the regressand.

The R^2, as it is defined in (3-85), is a *sample* measure. Now, if we want a *population* measure, we can define the *population* R^2 as

$$R^2_{POP} = 1 - \frac{\sigma_u^2}{\sigma_y^2} \qquad (3\text{-}86)$$

However, we have better estimates for these variances, σ_u^2 and σ_y^2, than the ones used in the (3-85). So, let us use unbiased estimates for these variances

$$\bar{R}^2 = 1 - \frac{SCR / (n - k)}{SCT / (n - 1)} = 1 - (1 - R^2)\frac{n - 1}{n - k} \qquad (3\text{-}87)$$

This measure is called the *adjusted R*–squared, or $\bar{R}^2$ .The primary attractiveness of $\bar{R}^2$ is that it imposes a penalty for adding additional regressors to a model. If a regressor is added to the model then *RSS* decreases, or at least is equal. On the other hand, the *degrees of freedom* of the regression *n–k* always decrease. $\bar{R}^2$ can go up or down when a new regressor is added to the model. That is to say:

$$\bar{R}^2_j \geq \bar{R}^2_{j-1} \quad \text{or} \quad \bar{R}^2_j \leq \bar{R}^2_{j-1} \qquad (3\text{-}88)$$

An interesting algebraic fact is that if we add a new regressor to a model, $\bar{R}^2$ increases if, and only if, the t statistic, which we will examine in chapter 4, on the new regressor is greater than 1 in absolute value. Thus we see immediately that $\bar{R}^2$ could be used to decide whether a certain additional regressor must be included in the model. The $\bar{R}^2$ has an upper bound that is equal to 1, but it does not strictly have a lower bound since it can take negative values.

The observations b) and c) made to the *R* squared remain valid for the adjusted *R* squared.

3.5.3 Akaike information criterion (*AIC*) and Schwarz criterion (*SC*)

These two criteria- Akaike information criterion (*AIC*) and Schwarz Criterion (*SC*) - have a very similar structure. For this reason, they will be reviewed together.

The *AIC* statistic, proposed by Akaike (1974) and based on information theory, has the following expression:

$$AIC = -\frac{2l}{n} + \frac{2k}{n} \qquad (3\text{-}89)$$

where l is the log likelihood function (assuming normally distributed disturbances) evaluated at the estimated values of the coefficients.

The *SC* statistic, proposed by Schwarz (1978), has the following expression:

$$SC = -\frac{2l}{n} + \frac{k\ln(n)}{n} \qquad (3\text{-}90)$$

The *AIC* and *SC* statistics, unlike the coefficients of determination (R^2 and $\overline{R}^2$), are better the lower their values are. It is important to remark that the *AIC* and SC statistics are not bounded unlike R^2.

a) The *AIC* and *SC* statistics penalize the introduction of new regressors. In the case of the *AIC*, as can be seen in the second term of the right hand side of (3-89), the number of regressors k appears in the numerator. Therefore, the growth of k will increase the value of *AIC* and consequently worsen the goodness of fit, if that is not offset by a sufficient growth of the log likelihood. In the case the *SC*, as can be seen in the second term of the right hand side of (3-90), the numerator is $k\ln(n)$. For $n>7$, the following happens: $k\ln(n)>2k$. Therefore, *SC* imposes a larger penalty for additional regressors than *AIC* when the sample size is greater than seven.

b) The *AIC* and *SC* statistics can be applied to statistical models without intercept.

c) The *AIC* and *SC* statistics are not relative measures as are the coefficients of determination. Therefore, their magnitude, in itself, offers no information.

d) The AIC and *SC* statistics can be applied to compare models in which endogenous variables have different functional forms. In particular, we will compare two models in which the regressands are y and $\ln(y)$. When the regressand is y, the formula (3-89) is applied in the AIC case, or (3-90) in the SC case. When the regressand is $\ln(y)$, and also when we want to carry out a comparison with another model in which the regressand is y, we must correct these statistics in the following way:

$$AIC_C = AIC + 2\overline{\ln(Y)} \qquad (3\text{-}91)$$

$$SC_C = SC + 2\overline{\ln(Y)} \qquad (3\text{-}92)$$

where AIC_C and SC_C are the corrected statistics, and *AIC* and *SC* are the statistics supplied by any econometric package such as the E-views.

Example 3.7 Selection of the best model

To analyze the determinants of expenditures on dairy the following alternative models have been considered:

1) $dairy = \beta_1 + \beta_2 inc + u$
2) $dairy = \beta_1 + \beta_2 \ln(inc) + u$
3) $dairy = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
4) $dairy = \beta_2 inc + \beta_3 punder5 + u$
5) $dairy = \beta_1 + \beta_2 inc + \beta_3 hhsize + u$
6) $\ln(dairy) = \beta_1 + \beta_2 inc + u$
7) $\ln(dairy) = \beta_1 + \beta_2 inc + \beta_3 punder5 + u$
8) $\ln(dairy) = \beta_2 inc + \beta_3 punder5 + u$

where *inc* is disposable income of household, *hhsize* is the number of household members and *punder5* is the proportion of children under five in the household.

Using a sample of 40 households (file *demand)*, and taking into account that $\overline{\ln(dairy)}$ =2.3719, the goodness of fit statistics obtained for the eight models appear in table 1. In particular, the AIC corrected for model 6) has been calculated as follows:

$$AIC_C = AIC + 2\overline{\ln(Y)} = 0.2794 + 2' \ 2.3719 = 5.0232$$

Conclusions

a) The *R*-squared can be only used to compare the following pairs of models: 1) with 2), and 3) with 5).
b) The adjusted *R*-squared can only be used to compare model 1) with 2), 3) and 5); and 6) with 7.
c) The best model out of the eight is model 7) according to *AIC* and *SC*.

TABLE 3.1. Measures of goodness of fit for eight models.

Model number	1	2	3	4	5	6	7	8
Regressand	*dairy*	*dairy*	*dairy*	*dairy*	*dairy*	ln(*dairy*)	ln(*dairy*)	ln(*dairy*)
Regressors	*intercept* *inc*	*intercept* ln(*inc)*	*intercept* *inc* *punder5*	*inc* *punder5*	*intercept* *Inc* *househsize*	*intercept* *inc*	*intercept* *inc* *punder5*	*inc* *punder5*
R-squared	0.4584	0.4567	0.5599	0.5531	0.4598	0.4978	0.5986	-0.6813
Adjusted R-squared	0.4441	0.4424	0.5361	0.5413	0.4306	0.4846	0.5769	-0.7255
Akaike information criterion	5.2374	5.2404	5.0798	5.0452	5.2847	0.2794	0.1052	1.4877
Schwarz criterion	5.3219	5.3249	5.2065	5.1296	5.4113	0.3638	0.2319	1.5721
Corrected Akaike information criterion						5.0232	4.8490	6.2314
Corrected Schwarz criterion						5.1076	4.9756	6.3159

Exercises

Exercise 3.1 Consider the linear regression model $\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$, where $\mathbf{X}$ is a matrix 50×5.

Answer the following questions, justifying your answers:

a) What are the dimensions of the vectors $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$?

b) How many equations are there in the system of normal equations $\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'y}$?

c) What conditions are needed in order to obtain $\hat{\boldsymbol{\beta}}$?

Exercise 3.2 Given the model

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

and the following data:

y	x_2	x_3
10	1	0
25	3	-1
32	4	0
43	5	1
58	7	-1
62	8	0
67	10	-1
71	10	2

a) Estimate β_1, β_2 and β_3 by *OLS*.

b) Calculate the residual sum of squares.

c) Obtain the residual variance.

d) Obtain the variance explained by the regression.

e) Obtain the variance of the endogenous variable

f) Calculate the coefficient of determination.

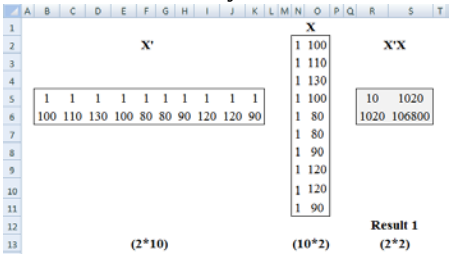
g) Obtain an unbiased estimation of σ^2.

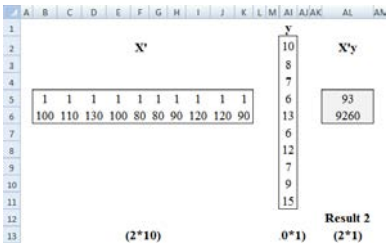
h) Estimate the variance of $\hat{\beta}_2$.

To answer these questions you can use Excel. See exhibit 3.1 as an example.

Exhibit 3.1

1) Calculation of **X'X** and **X'y**





Explanation for **X'X**

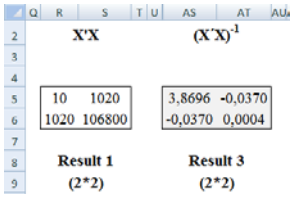
a) Enter the matrices **X'** and **X** into the Excel: B5:K6 and N2:O11

b) You can find the product **X'X** by highlighting the cells where you want to place the resulting matrix.

c) Once you have highlighted the resulting matrix, and while it is still highlighted, enter the following formula:=MMULT(B5:K6; N2:O11)

d) When the formula is entered, press the *Ctrl key* and the *Shift key* simultaneously. Then, holding these two keys, press the *Enter* key too.

2) Calculation of $\mathbf{(X'X)^{-1}}$



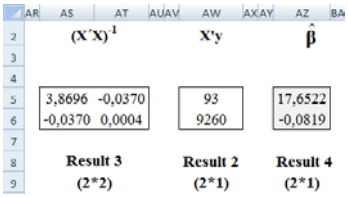
a) Enter the matrix **X'X** into the Excel**:** R5:S6

b) You can find the inverse of matrix **X'X** by highlighting the cells where you want to place the resulting matrix (R5:S6)

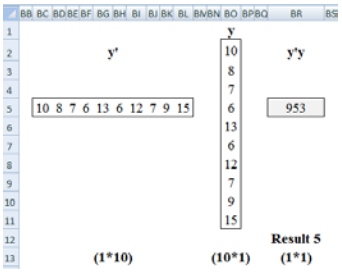
c) Once you have highlighted the resulting matrix, and while it is still highlighted, enter the following formula:=MINVERSE(R5:S6**).**

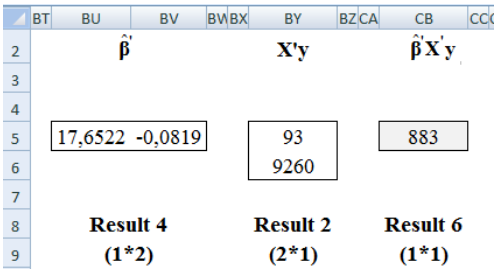
d) When the formula is entered, press the *Ctrl key* and the *Shift key* simultaneously. Then, holding these two keys, press the *Enter* key too.

3) Calculation of vector $\hat{\boldsymbol{\beta}}$



4) Calculation of $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ and σ^2





$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = R.5 - R.6 = 953 - 883 = 70$$

$$\hat{s}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-2} = \frac{70}{8} = 8.6993$$

5) Calculation of covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \hat{s}^2 (\mathbf{X}'\mathbf{X})^{-1} = 8.6993 \begin{pmatrix} 3.8696 & -0.0370 \\ -0.0370 & 0.0004 \end{pmatrix} = \begin{pmatrix} 33.6624 & -0.3215 \\ -0.3215 & 0.0032 \end{pmatrix}$$

Exercise 3.3 The following model was formulated to explain the annual sales (*sales*) of the manufacturers of household cleaning products as a function of a relative price index (*rpi*) and the advertising expenditures (*adv*):

$$sales = \beta_1 + \beta_2 rpi + \beta_3 adv + u$$

where the variable *sales* is expressed in a thousand million euros and *rpi* is a relative price index obtained as a ratio between the prices of each firm and the prices of firm 1 of the sample; *adv* is the annual expenditures on advertising and promotional campaigns and media diffusion, expressed in millions of euros.

Data on ten manufacturers of household cleaning products appear in the attached table.

firm	***sales***	***rpi***	***adv***
1	10	100	300
2	8	110	400
3	7	130	600
4	6	100	100
5	13	80	300
6	6	80	100
7	12	90	600
8	7	120	200
9	9	120	400
10	15	90	700

Using an excel spreadsheet,

a) Estimate the parameters of the proposed model
b) Estimate the covariance matrix.
c) Calculate the coefficient of determination.

Note: In exhibit 3.1 the model $sales = \beta_1 + \beta_2 rpi + u$ is estimated using excel. Instructions are also included.

Exercise 3.4 A researcher, who is developing an econometric model to explain income, formulates the following specification:

$$inc=\alpha+\beta cons+\gamma save+u \qquad [1]$$

where *inc* is the household disposable income, *cons* is the total consumption and *save* is the total savings of the household.

The researcher did not take into account that the above three magnitudes are related by the identity

$$inc=cons+save \qquad [2]$$

The equivalence between the models [1] and [2] requires that, in addition to the disappearance of the disturbance term, the model parameters [1] take the following values: α =0, β =1, and γ =1

If you estimate equation [1] with the data for a given country, can you expect, in general, that the estimates will take the values $\hat{\alpha} = 0, \hat{\beta} = 1, \hat{\gamma} = 0$?

Please justify your answer using mathematical notation.

Exercise 3.5 A researcher proposes the following econometric model to explain tourism revenue (*turtot*) in a given country:

$$turtot = \beta_1 + \beta_2 turmean + \beta_3 numtur + u$$

where *turmean* is the average expenditure per tourist and *numtur* is the total number of tourists.

a) It is obvious that *turtot*, *numtur* and *turmean* and are also linked by the relationship *turtot=turmean×numtur*. Will this somehow affect the estimation of the parameters of the proposed model?

b) Is there a model with another functional form involving tighter restrictions on the parameters? If so, indicate it.

c) What is your opinion about using the proposed model to explain the behavior of tourism revenue? Is it reasonable?

Exercise 3.6 Let us suppose you have to estimate the model

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + \beta_4 \ln(x_4) + u$$

using the following observations:

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

What problems can arise in the estimation of this model?

Exercise 3.7 Answer the following questions:

a) Explain the determination coefficient (R^2) and the adjusted determination coefficient ($\overline{R}^2$). What can you use them for? Justify your answer.

b) Given the models

$$\ln(y)=\beta_1+\beta_2\ln(x)+u \quad (1)$$
$$\ln(y)=\beta_1+\beta_2\ln(x)+\beta_3\ln(z)+u \quad (2)$$
$$\ln(y)=\beta_1+\beta_2\ln(z)+u \quad (3)$$
$$y=\beta_1+\beta_2 z+u \quad (4)$$

indicate what measure of goodness of fit is appropriate to compare the following pairs of models: (1) - (2), (1) - (3), and (1) - (4). Explain your answer.

Exercise 3.8 Let us suppose that the following model is estimated by OLS:

$$\ln(y) = \beta_1 + \beta_2 \ln(x) + \beta_3 \ln(z) + u$$

a) Can least square residuals all be positive? Explain your answer.

b) Under the assumption of no autocorrelation of disturbances, are the *OLS* residuals independent? Explain your answer

c) Assuming that the disturbances are not normally distributed, will the *OLS* estimators be unbiased? Explain your answer.

Exercise 3.9 Consider the linear regression model

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

where **y** and **u** are vectors 8×1, **X** is a matrix 8×3 and **β** is a vector 3×1. Also the following information is available:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad \hat{\mathbf{u}}'\hat{\mathbf{u}} = 22$$

Answer the following questions, by justifying your answer:

a) Indicate the sample size, the number of regressors, the number of parameters and the degrees of freedom of the residual sum of squares.

b) Derive the covariance matrix of the vector $\hat{\boldsymbol{\beta}}$, making explicit the assumptions used. Estimate the variances of the estimators.

c) Does the regression have an intercept? What implications does the answer to this question have on the meaning of R^2 in this model?

Exercise 3.10 Discuss whether the following statements are true or false:

a) In a linear regression model, the sum of the residuals is zero.

b) The coefficient of determination (R^2) is always a good measure of the model's quality.

c) The least squares estimators are biased.

Exercise 3.11 The following model is formulated to explain time spent sleeping:

$$sleep = \beta_1 + \beta_2\ totalwrk + \beta_3 leisure + u$$

where *sleep*, *totalwrk* (paid and unpaid work) and leisure (time not devoted to sleep or work) are measured in minutes per day.

The estimated equation with a sample of 1000 observations, using file *timuse03*, is the following:

$$\widehat{sleep} = 1440 - 1'\ total\_work - 1'\ leisure$$

$$R^2=1.000 \qquad n=1000$$

a) What do you think about these results?

b) What is the meaning of the estimated intercept?

Exercise 3.12 Using a subsample of the Structural Survey of Wages (*Encuesta de estructura salarial*) for Spain in 2006 (file *wage06sp*), the following model is estimated to explain wage:

$$\widehat{\ln(wage)} = 1.565 + 0.0730 educ + 0.0177 tenure + 0.0065 age$$

$$R^2=0.337 \qquad n=800$$

where *educ* (education), *tenure* (experience in the firm) and *age* are measured in years and *wage* in euros per hour.

a) What is the interpretation of coefficients on *educ*, *tenure* and *age*?

b) How many years does the age have to increase in order to have a similar effect to an increase of one year in education, holding fixed in each case the other two regressors?

c) Knowing that $\overline{educ}$ =10.2, $\overline{tenure}$ =7.2 and $\overline{age}$ =42.0, calculate the elasticities of *wage* with respect to *educ*, *tenure* and *age* for these values, holding fixed the others regressors. Do you consider these elasticities to be high or low?

Exercise 3.13 The following equation describes the price of housing in terms of house *bedrooms* (number of bedrooms), *bathrms* (number of full bathrooms) and *lotsize* (the lot size of a property in square feet):

$$price = \beta_1 + \beta_2 bedrooms + \beta_3 bathrms + \beta_4 lotsize + u$$

where *price* is the price of a house measured in dollars.

Using the data for the city of Windsor contained in file *housecan*, the following model is estimated:

$$\hat{price} = -2418 + 5827 bedrooms + 19750 bathrms + 5.411 lotsize$$

$$R^2=0.486 \quad n=546$$

a) What is the estimated increase in price for a house with one more bedroom and one more bathroom, holding *lotsize* constant?

b) What percentage of the variation in price is explained jointly by the number of bedrooms, the number of full bathrooms and the lot size?

c) Find the predicted selling price for a house of the sample with *bedrooms*=3, *bathrms*=2 and *lotsize*=3880.

d) The actual selling price of the house in *c)* was \$66,000. Find the residual for this house. Does the result suggest that the buyer underpaid or overpaid for the house?

Exercise 3.14 To examine the effects of a firm's performance on a CEO salary, the following model was formulated:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + u$$

where *roa* is the ratio profits/assets expressed as a percentage and *tenure* is the number of years as CEO (=0 if less than 6 months). Salaries are expressed in thousands of dollars, and *sales* and *profits* in millions of dollars.

The file *ceoforbes* has been used for the estimation. This file contains data on 447 CEOs of America's 500 largest corporations. (52 of the 500 firms were excluded because of missing data on one or more variables. Apple Computer was also excluded since Steve Jobs, the acting CEO of Apple in 1999, received no compensation during this period.) Company data come from Fortune magazine for 1999; CEO data come from Forbes magazine for 1999 too. The results obtained were the following:

$$\widehat{\ln(salary)} = 4.641 + 0.0054 roa + 0.2893 \ln(sales) + 0.0000564 profits + 0.0122 tenure$$

$$R^2=0.232 \quad n=447$$

a) Interpret the coefficient on the regressor *roa*

b) Interpret the coefficient on the regressor ln(*sales*). What is your opinion about the magnitude of the elasticity salary/sales?

c) Interpret the coefficient on the regressor *profits*.

d) What is the salary/profits elasticity at the sample mean ($\overline{salary}$=2028 and $\overline{profits}$=700).

Exercise 3.15 (Continuation of exercise 2.21) Using a dataset consisting of 1,983 firms surveyed in 2006 (file *rdspain*), the following equation was estimated:

$$\widehat{rdintens} = -1.8168 + 0.1482 \ln(sales) + 0.0110 exponsal$$

$$R^2 = 0.048 \quad n = 1983$$

where *rdintens* is the expenditure on research and development (R&D) as a percentage of sales, *sales* are measured in millions of euros, and *exponsal* is exports as a percentage of sales.

a) Interpret the coefficient on ln(*sales*). In particular, if *sales* increase by 100%, what is the estimated percentage point change in *rdintens*? Is this an economically large effect?
b) Interpret the coefficient on *exponsal*. Is it economically large?
c) What percentage of the variation in *rdintens* is explained by *sales* and *exponsal*?
d) What is the *rdintens*/*sales* elasticity for the sample mean ($\overline{rdintens}$ =0.732 and $\overline{sales}$ =63544960). Comment on the result.
e) What is the *rdintens*/*exponsal* elasticity for the sample mean ($\overline{rdintens}$ =0.732 and $\overline{exponsal}$ =17.657). Comment on the result.

Exercise 3.16 The following hedonic regression for cars (see example 3.3) is formulated:

$$\ln(price) = \beta_1 + \beta_2 cid + \beta_3 hpweight + \beta_4 fueleff + u$$

where *cid* is the cubic inch displacement, *hpweight* is the ratio horsepower/weight in kg expressed as percentage and *fueleff* is the ratio liters per 100 km/horsepower expressed as a percentage.

a) What are the probable signs of β_2, β_3 and β_4? Explain them.
b) Estimate the model using the file *hedcarsp* and write out the results in equation form.
c) Interpret the coefficient on the regressor *cid*.
d) Interpret the coefficient on the regressor *hpweight*.
e) To expand the model, add a regressor relative to car size, such as volume or weight. What happens if you add both of them? What is the relationship between weight and volume?

Exercise 3.17 The concept of work covers a broad spectrum of possible activities in the productive economy. An important part of work is unpaid; it does not pass through the market and therefore has no price. The most important unpaid work is housework (*houswork*) carried out mainly by women. In order to analyze the factors that influence housework, the following model is formulated:

$$houswork = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

where *educ* is the years of education attained, *hhinc* is the household income in euros per month. The variables *houswork* and *paidwork* are measured in minutes per day.

Use the data in the file *timuse03* to estimate the model. This file contains 1000 observations corresponding to a random subsample extracted from the time use survey for Spain carried out in 2002-2003.

a) Which signs do you expect for β_2, β_3, β_4 and β_5? Explain.
b) Write out the results in equation form?
c) Do you think there are relevant factors omitted in the above equation? Explain.
d) Interpret the coefficient on the regressors *educ, hhinc, age* and *paidwork.*

Exercise 3.18 (Continuation of exercise 2.20) To explain the overall satisfaction of people (*stsfglo*), the following model is formulated:

$$stsfglo = \beta_1 + \beta_2 gnipc + \beta_3 lifexpec + u$$

where *gnipc* is the gross national income per capita expressed in PPP 2008 US dollar terms and *lifexpec* is the life expectancy at birth, i.e., the number of years a newborn infant could expect to live. When a magnitude is expressed in PPP (purchasing power parity) US dollar terms, a magnitude is converted to international dollars using PPP rates. (An international dollar has the same purchasing power as a US dollar in the United States.)

Use the file *HDR2010* for the estimation of the model.

a) What are the expected signs for β_2 and β_3? Explain.
b) What would be the average overall satisfaction for a country with 80 years of life expectancy at birth and a gross national income per capita of 30000 $ expressed in PPP 2008 US dollars?
c) Interpret the coefficients on *gnipc* and *lifexpe*.
d) Given a country with a life expectancy at birth equal to 50 years, what should be the gross national income per capita to obtain a global satisfaction equal to five?

Exercise 3.19 (Continuation exercise 2.24) Due to the problems arisen in the Keynesian consumption function, Brown introduced a new regressor in the function: consumption lagged a period to reflect the persistence of consumer habits. The formulation of the model is as follows

$$conspc_t = b_1 + b_2 incpc_t + b_3 conspc_{t-1} + u_t$$

As lagged consumption is included in this model, we have to distinguish between marginal propensity to consume in the short term and long term. The short-run marginal propensity is calculated in the same way as in the Keynesian consumption function. To calculate the long-term marginal propensity it is necessary to consider equilibrium state with no changes in variables. Denoting by $conspc^e$ and $incpc^e$ consumption and income in equilibrium, and regardless of the random disturbance, the previous model in equilibrium is given by

$$conspc^e = b_1 + b_2 incpc^e + b_3 conspc^e$$

The Brown consumption function was estimated with data of the Spanish economy for the period 1954-2010 (file *consumsp*), obtaining the following results:

$$\widehat{conspc}_t = -7.156 + 0.3965 incpc_t + 0.5771 conspc_{t-1}$$

$$R^2=0.997 \qquad n=56$$

a) Interpret the coefficient on *incpc*. In the interpretation, do you have to include the clause "holding fixed the other regressor"? Justify the answer.
b) Calculate the short-term elasticity for the sample means ($\overline{conspc}$=8084, $\overline{incpc}$=8896).
c) Calculate the long-term elasticity for the sample means.
d) Discuss the difference between the values obtained for the two types of elasticity.

Exercise 3.20 To explain the influence of incentives and expenditures in advertising on sales, the following alternative models have been formulated:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u \quad (1)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(advert) + \beta_3 \ln(incent) + u \quad (2)$$

$$\ln(sales) = \beta_1 + \beta_2 advert + \beta_3 incent + u \quad (3)$$

$$sales = \beta_2 advert + \beta_3 incent + u \quad (4)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(incent) + u \quad (5)$$

$$sales = \beta_1 + \beta_2 incent + u \quad (6)$$

a) Using a sample of 18 sale areas (file *advincen*), estimate the above models:
b) In each of the following groups select the best model, indicating the criteria you have used. Justify your answer.

*b*1) (1) and (6)
*b*2) (2) and (3)
*b*3) (1) and (4)
*b*4) (2), (3) and (5)
*b*5) (1), (4) and (6)
*b*6) (1), (2), (3), (4), (5) and (6)

Appendixes

Appendix 3.1 Proof of the theorem of Gauss-Markov

To prove this theorem, the *MLC* assumptions 1 through 9 are used.

Let us now consider another estimator $\tilde{\boldsymbol{\beta}}$ which is a function of **y** (remember that $\hat{\boldsymbol{\beta}}$ is also a function of **y**), given by

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{y} \qquad (3\text{-}93)$$

where **A** is $k \times n$ arbitrary matrix, that is a function of **X** and/or other non-stochastic variables, but it is not a function of **y**. For $\tilde{\boldsymbol{\beta}}$ to be unbiased, certain conditions must be accomplished.

Taking (3-52) into account, we have

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \qquad (3\text{-}94)$$

Taking expectations on both sides of (3-94), we have

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]E(\mathbf{u}) = \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} \qquad (3\text{-}95)$$

For $\tilde{\boldsymbol{\beta}}$ to be unbiased, that is to say, $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, the following must be accomplished:

$$\mathbf{A}\mathbf{X} = \mathbf{I} \qquad (3\text{-}96)$$

Consequently,

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u} \qquad (3\text{-}97)$$

Taking into account assumptions 7 and 8, and (3-96), the $Var(\tilde{\boldsymbol{\beta}})$ is equal to

$$Var(\tilde{\boldsymbol{\beta}}) = E((\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})') = E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' + \mathbf{A}\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}'\right]\right]$$
$$= E\left[\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u}\mathbf{u}'\left[\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] + \mathbf{A}\mathbf{A}'\right] = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}'\right]$$

(3-98)

The difference between both variances is the following:

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\left[\mathbf{X}'\mathbf{X}\right]^{-1} + \mathbf{A}\mathbf{A}' - \left[\mathbf{X}'\mathbf{X}\right]^{-1}\right] = \sigma^2\mathbf{A}\mathbf{A}' \qquad (3\text{-}99)$$

The product of a matrix by its transpose is always a semi-positive definite matrix. Therefore,

$$Var(\tilde{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{A}\mathbf{A}' \geq 0 \quad (3\text{-}100)$$

The difference between the variance of an estimator $\tilde{\boldsymbol{\beta}}$ - arbitrary but linear and unbiased – and the variance of the estimator $\hat{\boldsymbol{\beta}}$ is a semi positive definite matrix. Consequently, $\hat{\boldsymbol{\beta}}$ is a Best Unbiased Linear Estimator; that is to say, it is a BLUE estimator.

Appendix 3.2 Proof: $\hat{\sigma}^2$ is an unbiased estimator of the variance of the disturbance

In order to see which is the most appropriate estimator of σ^2, we shall first analyze the properties of the sum of squared residuals. This one is precisely the numerator of the residual variance.

Taking into account (3-17) and (3-23), we are going to express the vector of residuals as a function of the regressand

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{M}\mathbf{y} \quad (3\text{-}101)$$

where **M** is an idempotent matrix.

Alternatively, the vector of residuals can be expressed as a function of the disturbance vector:

$$\begin{aligned}\hat{\mathbf{u}} &= \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{y} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{u} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\boldsymbol{\beta}\mathbf{u} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} + \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} = \left[\mathbf{I} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right]\mathbf{u} \\ &= \mathbf{M}\mathbf{u}\end{aligned} \qquad (3\text{-}102)$$

Taking into account (3-102), the sum of squared residuals (*SSR*) can be expressed in the following form:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u} \qquad (3\text{-}103)$$

Now, keeping in mind that we are looking for an unbiased estimator of σ^2, we are going to calculate the expectation of the previous expression:

$$\begin{aligned} E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] &= E\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = trE\left[\mathbf{u}'\mathbf{M}\mathbf{u}\right] = E\left[tr\mathbf{u}'\mathbf{M}\mathbf{u}\right] \\ &= E\left[tr\mathbf{M}\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}E\left[\mathbf{u}\mathbf{u}'\right] = tr\mathbf{M}\sigma^2\mathbf{I} \\ &= \sigma^2 tr\mathbf{M} = \sigma^2(n-k) \end{aligned} \qquad (3\text{-}104)$$

In deriving (3-104), we have used the property of the trace that $tr(\mathbf{AB}) = tr(\mathbf{BA})$. Taking into account that property of the trace, the value of $tr\mathbf{M}$ is obtained:

$$\begin{aligned} tr\mathbf{M} &= tr\left[\mathbf{I}_{n\times n} - \mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}'\right] = tr\mathbf{I}_{n\times n} - tr\mathbf{X}\left[\mathbf{X}'\mathbf{X}\right]^{-1}\mathbf{X}' \\ &= tr\mathbf{I}_{n\times n} - tr\mathbf{I}_{k\times k} = n-k \end{aligned}$$

According to (3-104), it holds that

$$\sigma^2 = \frac{E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right]}{n-k} \quad (3\text{-}105)$$

Keeping (3-105) in mind, an unbiased estimator of the variance will be:

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k} \qquad (3\text{-}106)$$

since, according to (3-104),

$$E(\hat{\sigma}^2) = E\left[\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}\right] = \frac{E(\hat{\mathbf{u}}'\hat{\mathbf{u}})}{n-k} = \frac{\sigma^2(n-k)}{n-k} = \sigma^2 \qquad (3\text{-}107)$$

The denominator of (3-106) is the degree of freedom corresponding to the *RSS* that appear in the numerator. This result is justified by the fact that the normal equations of the hyperplane impose *k* restrictions on the residuals. Therefore, the number of degrees of freedom of the *RSS* is equal to the number of observations (*n*) minus the number of restrictions *k*.

Appendix 3.3 Consistency of the *OLS* estimator

In appendix 2.8 we have proved the consistency of the *OLS* estimator $\hat{b}_2$ in the simple regression model. Now we are going to prove the consistency of the *OLS* vector $\hat{\boldsymbol{\beta}}$.

First, the least squares estimator $\hat{\boldsymbol{\beta}}$, given in (3-23). may be written as

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) \qquad (3\text{-}108)$$

Now, we take limits in the last factor of (3-108) and call **Q** to the result:

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{Q} \qquad (3\text{-}109)$$

If **X** is taken to be fixed in repeated samples, according to assumption 2, then (3-109) implies that $\mathbf{Q}=(1/n)\mathbf{X}'\mathbf{X}$. According to assumption 3, and because the inverse is a continuous function of the original matrix, $\mathbf{Q}^{-1}$ exists. Therefore, we can write

$$\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\operatorname{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right)$$

The last term of (3-108) can be written as

$$\frac{1}{n}\mathbf{X}'\mathbf{u} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{ji} & \cdots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{ki} & \cdots & x_{kn} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} \qquad (3\text{-}110)$$

$$= \frac{1}{n}\begin{bmatrix}\mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_i & \cdots & \mathbf{x}_n\end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i = \overline{\mathbf{x}_i u_i}$$

where $\mathbf{x}_i$ is the column vector corresponding to the i^{th} observation

Now, we are going to calculate the expectation and the variance (3-110),

$$E\left[\overline{\mathbf{x}_i u_i}\right] = \frac{1}{n}\sum_{i=1}^{n}E[\mathbf{x}_i u_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E[u_i] = \frac{1}{n}\mathbf{X}'E[\mathbf{u}] = \mathbf{0} \quad (3\text{-}111)$$

$$var\left[\overline{\mathbf{x}_i u_i}\right] = E\left[\overline{\mathbf{x}_i u_i}(\overline{\mathbf{x}_i u_i})'\right] = \frac{1}{n}\mathbf{X}'E[\mathbf{u}\mathbf{u}']\mathbf{X} = \frac{1}{n}\frac{s^2}{n}\frac{\mathbf{X}'\mathbf{X}}{n} = \frac{s^2}{n^2}\mathbf{Q} \qquad (3\text{-}112)$$

since $E[\mathbf{u}\mathbf{u}'] = s^2\mathbf{I}$, according to assumptions 7 and 8.

Taking limits in (3-112), it then follows that

$$\lim_{n\to\infty} var\left[\overline{\mathbf{x}_i u_i}\right] = \lim_{n\to\infty}\frac{s^2}{n^2}\mathbf{Q} = 0(\mathbf{Q}) = \mathbf{0} \qquad (3\text{-}113)$$

Since the expectation of $\overline{\mathbf{x}_i u_i}$ is identically zero and its variance converges to zero, $\overline{\mathbf{x}_i u_i}$ converges in mean square to zero. Convergence in mean square implies convergence in probability, and so $\operatorname{plim}(\overline{\mathbf{x}_i u_i})=0$. Therefore,

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}(\mathbf{x}_i\mathbf{u}_i) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}\left(\frac{1}{n}\mathbf{X}'\mathbf{u}\right) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot 0 = \boldsymbol{\beta} \tag{3-114}$$

Consequently, $\hat{\boldsymbol{\beta}}$ is a consistent estimator.

Appendix 3.4 Maximum likelihood estimator

The method of maximum likelihood is widely used in econometrics. This method proposes that the parameter estimators be those values for which the probability of obtaining the observations given is maximum. In the least squares estimation no prior assumption was adopted. On the contrary, the estimation by maximum likelihood requires that statistical assumptions about the various elements of the model be established beforehand. Thus, in the estimation by maximum likelihood we will adopt all the assumptions of classic linear model (*CLM*).

Therefore, in the estimation by maximum likelihood of $\boldsymbol{\beta}$ and σ^2 in the model (3-52), we take as estimators those values that maximize the probability to obtain the observations in a given sample.

Let us look at the procedure for obtaining the maximum likelihood estimators β and σ. According to the *CLM* assumptions:

$$\mathbf{u}: \quad N(\mathbf{0}, \sigma^2\mathbf{I}) \tag{3-115}$$

The expectation and variance of the distribution of **y** are given by

$$E(\mathbf{y}) = E\left[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right] = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} \tag{3-116}$$

$$\text{var}(\mathbf{y}) = E\left[\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\right] = E\left[\mathbf{u}\mathbf{u}'\right] = \sigma^2\mathbf{I} \tag{3-117}$$

Therefore,

$$\mathbf{y}: \quad N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{3-118}$$

The probability density of **y** (or likelihood function), considering **X** and **y** fixed and $\boldsymbol{\beta}$ and σ^2 variable, will be in accordance with (3-118) equal to

$$L = f(y \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left(-\left(1/2\sigma^2\right)\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right) \tag{3-119}$$

The maximum for L is reached in the same point on the $\ln(L)$ given that the logarithm function is monotonic, and thus, in order to maximize the function, we can work with $\ln(L)$ instead of L. Therefore,

$$\ln(L) = -\frac{n\ln(2\pi)}{2} - \frac{n\ln(\sigma^2)}{2} - \frac{1}{2\sigma^2} - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3-120}$$

To maximize $\ln(L)$, we differentiate it with respect to $\boldsymbol{\beta}$ and σ^2:

$$\frac{\delta\ln(L)}{\delta\boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \tag{3-121}$$

$$\frac{\delta \ln(L)}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta})}{2\sigma^4} \qquad (3\text{-}122)$$

Equating (3-121) to zero, we see that the maximum likelihood estimator of **β**, denoted by $\tilde{\boldsymbol{\beta}}$, satisfies that

$$\mathbf{X}'\mathbf{X}\tilde{\beta} = \mathbf{X}'\mathbf{y} \qquad (3\text{-}123)$$

Because we assume that $\mathbf{X}'\mathbf{X}$ is invertible,

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \mathbf{X}'\mathbf{y} \qquad (3\text{-}124)$$

Consequently, the maximum likelihood estimator of **β**, under the assumptions of the *CLM*, coincides with *OLS* estimator, that is to say,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \qquad (3\text{-}125)$$

Therefore,

$$(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} \qquad (3\text{-}126)$$

Equating (3-122) to zero and by substituting **β** by $\tilde{\boldsymbol{\beta}}$, we obtain:

$$-\frac{n}{2\tilde{\sigma}^2} + \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{2\tilde{\sigma}^4} = 0 \qquad (3\text{-}127)$$

where we have designated by $\tilde{\sigma}^2$ the maximum likelihood estimator of the variance of the random disturbances. From (3-127), it follows that

$$\tilde{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n} \qquad (3\text{-}128)$$

As we can see, the maximum likelihood estimator is not equal to the unbiased estimator that has been obtained in (3-106). In fact, if we take expectations to (3-128),

$$E\left[\tilde{\sigma}^2\right] = \frac{1}{n} E\left[\hat{\mathbf{u}}'\hat{\mathbf{u}}\right] = \frac{n-k}{n}\sigma^2 \qquad (3\text{-}129)$$

That is to say, the maximum likelihood estimator, $\tilde{\sigma}^2$, is a biased estimator, although its bias tends to zero as *n* infinity, since

$$\lim_{n\to\infty} \frac{n-k}{n} = 1 \qquad (3\text{-}130)$$

4 HYPOTHESIS TESTING IN THE MULTIPLE REGRESSION MODEL

4.1 Hypothesis testing: an overview

Before testing hypotheses in the multiple regression model, we are going to offer a general overview on hypothesis testing.

Hypothesis testing allows us to carry out inferences about population parameters using data from a sample. In order to test a hypothesis in statistics, we must perform the following steps:

1) Formulate a null hypothesis and an alternative hypothesis on population parameters.

2) Build a statistic to test the hypothesis made.

3) Define a decision rule to reject or not to reject the null hypothesis.

Next, we will examine each one of these steps.

4.1.1 Formulation of the null hypothesis and the alternative hypothesis

Before establishing how to formulate the null and alternative hypothesis, let us make the distinction between *simple* hypotheses and *composite* hypotheses. The hypotheses that are made through one or more equalities are called simple hypotheses. The hypotheses are called composite when they are formulated using the operators "inequality", "greater than" and "smaller than".

It is very important to remark that hypothesis testing is always about *population* parameters. Hypothesis testing implies making a decision, on the basis of sample data, on whether to reject that certain restrictions are satisfied by the basic assumed model. The restrictions we are going to test are known as the *null hypothesis*, denoted by H_0. Thus, null hypothesis is a statement on population parameters.

Although it is possible to make composite null hypotheses, in the context of the regression model the null hypothesis is always a simple hypothesis. That is to say, in order to formulate a null hypothesis, which shall be called H_0, we will always use the operator “equality”. Each equality implies a restriction on the parameters of the model. Let us look at a few examples of null hypotheses concerning the regression model:

a) $H_0 : \beta_1=0$

b) $H_0 : \beta_1+ \beta_2 =0$

c) $H_0 : \beta_1 = \beta_2 = 0$

d) $H_0 : \beta_2 + \beta_3 = 1$

We will also define an *alternative* hypothesis, denoted by H_1, which will be our conclusion if the experimental test indicates that H_0 is false.

Although the alternative hypotheses can be simple or composite, in the regression model we will always take a composite hypothesis as an alternative hypothesis. This hypothesis, which shall be called H_1, is formulated using the operator "inequality" in most cases. Thus, for example, given the H_0:

$$H_0 : \beta_j = 1 \qquad (4\text{-}1)$$

we can formulate the following H_1:

$$H_1 : \beta_j \neq 1 \qquad (4\text{-}2)$$

which is a "two side alternative" hypothesis.

The following hypotheses are called "one side alternative" hypotheses

$$H_1 : \beta_j < 1 \qquad (4\text{-}3)$$

$$H_1 : \beta_j > 1 \qquad (4\text{-}4)$$

4.1.2 Test statistic

A *test statistic* is a function of a random sample, and is therefore a random variable. When we compute the statistic for a given sample, we obtain an outcome of the test statistic. In order to perform a statistical test we should know the distribution of the test statistic under the null hypothesis. This distribution depends largely on the assumptions made in the model. If the specification of the model includes the assumption of normality, then the appropriate statistical distribution is the normal distribution or any of the distributions associated with it, such as the Chi-square, Student's *t*, or Snedecor's *F*.

Table 4.1 shows some distributions, which are appropriate in different situations, under the assumption of normality of the disturbances.

TABLE 4.1. Some distributions used in hypothesis testing.

	1 restriction	*1 or more restrictions*
Known σ^2	*N*	*Chi-square*
Unknown σ^2	Student's *t*	Snedecor's *F*

The statistic used for the test is built taking into account the H_0 and the sample data. In practice, as σ^2 is always unknown, we will use the distributions *t* and *F*.

4.1.3 Decision rule

We are going to look at two approaches for hypothesis testing: the classical approach and an alternative one based on *p*-values. But before seeing how to apply the

decision rule, we shall examine the types of mistakes that can be made in testing hypothesis.

Types of errors in hypothesis testing

In hypothesis testing, we *can* make two kinds of errors: *Type I error* and *Type II error*.

Type I error

We can reject H_0 when it is in fact true. This is called *Type I error*. Generally, we define the *significance level* (α) of a test as the probability of making a *Type I error*. Symbolically,

$$\alpha = \Pr(Reject\ H_0 \mid H_0) \qquad (4\text{-}5)$$

In other words, the significance level is the probability of rejecting H_0 given that H_0 is true. Hypothesis testing rules are constructed making the probability of a *Type I error* fairly small. Common values for α are 0.10, 0.05 and 0.01, although sometimes 0.001 is also used.

After we have made the decision of whether or not to reject H_0, we have either decided correctly or we have made an error. We shall never know with certainty whether an error was made. However, we can compute the *probability* of making either a *Type I error* or a *Type II error*.

Type II error

We can fail to reject H_0 when it is actually false. This is called *Type II error*.

$$\beta = \Pr(No\ reject\ H_0 \mid H_1) \qquad (4\text{-}6)$$

In words, β is the probability of not rejecting H_0 given that H_1 is true.

It is not possible to minimize both types of error simultaneously. In practice, what we do is select a low significance level.

Classical approach: Implementation of the decision rule

The classical approach implies the following steps:

a) *Choosing* α. Classical hypothesis testing requires that we initially specify a *significance level* for the test. When we specify a value for α, we are essentially quantifying our tolerance for a *Type I error*. If α=0.05, then the researcher is willing to falsely reject H_0 5% of the time.

b) *Obtaining c*, the *critical value*, using statistical tables. The value *c* is determined by α.

The critical value (*c*) for a hypothesis test is a threshold to which the value of the test statistic in a sample is compared to determine whether or not the null hypothesis is rejected.

c) *Comparing the outcome of the test* statistic, *s*, with *c*, H_0 is either rejected or not for a given α.

The rejection region (*RR*), delimited by the critical value(s), is a set of values of the test statistic for which the null hypothesis is rejected. (See figure 4.1). That is, the sample space for the test statistic is partitioned into two regions; one region (the rejection region) will lead us to reject the null hypothesis H_0, while the other will lead us not to reject the null hypothesis. Therefore, if the observed value of the test statistic *S* is in the critical region, we conclude by *rejecting* H_0; if it is not in the rejection region then we conclude by *not rejecting* H_0 or *failing to reject* H_0.

Symbolically,

$$\begin{array}{llll} \text{If} & s \geq c & \text{reject} & H_0 \\ \text{If} & s < c & \text{not reject} & H_0 \end{array} \qquad (4\text{-}7)$$

If the null hypothesis is rejected with the evidence of the sample, this is a *strong* conclusion. However, the acceptance of the null hypothesis is a *weak* conclusion because we do not know what the probability is of not rejecting the null hypothesis when it should be rejected. That is to say, we do not know the probability of making a type II error. Therefore, instead of using the expression of accepting the null hypothesis, it is more correct to say *fail to reject* the null hypothesis, or *not reject*, since what really happens is that we do not have enough empirical evidence to reject the null hypothesis.

In the process of hypothesis testing, the most subjective part is the *a priori* determination of the significance level. What criteria can be used to determine it? In general, this is an arbitrary decision, though, as we have said, the 1%, 5% and 10% levels for α are the most used in practice. Sometimes the testing is made conditional on several significance levels.

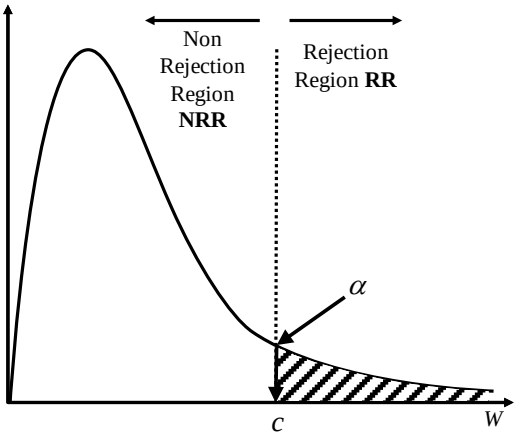


FIGURE 4.1. Hypothesis testing: classical approach.

An alternative approach: p-value

With the use of computers, hypothesis testing can be contemplated from a more rational perspective. Computer programs typically offer, together with the test statistic, a probability. This probability, which is called *p*-value (i.e., probability value), is also known as the critical or exact level of significance or the exact probability of making a

Type I error. More technically, the *p* value is defined as the lowest significance level at which a null hypothesis can be rejected.

Once the *p*-value has been determined, we know that the null hypothesis is rejected for any $\alpha \geq p$-value, while the null hypothesis is not rejected when $\alpha < p$-value. Therefore, the *p*-value is an indicator of the level of admissibility of the null hypothesis: the higher the *p*-value, the more confidence we can have in the null hypothesis. The use of the *p*-value turns hypothesis testing around. Thus, instead of fixing *a priori* the significance level, the *p*-value is calculated to allow us to determine the significance levels of those in which the null hypothesis is rejected.

In the following sections, we will see the use of *p* value in hypothesis testing put into practice.

4.2 Testing hypotheses using the *t* test

4.2.1 Test of a single parameter

The t test

Under the *CLM* assumptions 1 through 9,

$$\hat{\beta}_j \sim N\left[\beta_j, \text{var}(\hat{\beta}_j)\right] \qquad j = 1, 2, 3, \cdots, k \tag{4-8}$$

If we typify

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\text{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{sd(\hat{\beta}_j)} \sim N[0,1] \qquad j = 1, 2, 3, \cdots, k \tag{4-9}$$

The claim for normality is usually made on the basis of the Central Limit Theorem (*CLT*), but this is restrictive in some cases. That is to say, normality cannot always be assumed. In any application, whether normality of *u* can be assumed is really an empirical matter. It is often the case that using a transformation, i.e. taking logs, yields a distribution that is closer to normality, which is easy to handle from a mathematical point of view. Large samples will allow us to drop normality without affecting the results too much.

Under the *CLM* assumptions 1 through 9, we obtain a Student's *t* distribution

$$\frac{\hat{b}_j - b_j}{se(\hat{b}_j)} : \; t_{n-k} \tag{4-10}$$

where *k* is the number of unknown parameters in the population model (*k*-1 slope parameters and the intercept, β_1). The expression (4-10) is important because it allows us to test a hypothesis on β_j.

If we compare (4-10) with (4-9), we see that the Student's *t* distribution derives from the fact that the parameter σ in $sd(\hat{\beta}_j)$ has been replaced by its estimator $\hat{\sigma}$, which is a random variable. Thus, the degrees of freedom of *t* are *n*-1-*k* corresponding to the degrees of freedom used in the estimation of $\hat{\sigma}^2$.

When the degrees of freedom (*df*) in the *t* distribution are large, the *t* distribution approaches the standard normal distribution. In figure 4.2, the density function for normal and *t* distributions for different *df* are represented. As can be seen, the *t* density functions are flatter (platycurtic) and the tails are wider than normal density function, but as *df* increases, *t* density functions are closer to the normal density. In fact, what happens is that the *t* distribution takes into account that σ^2 is estimated because it is unknown. Given this uncertainty, the *t* distribution extends more than the normal one. However, as the *df* grows the *t*-distribution is nearer to the normal distribution because the uncertainty of not knowing σ^2 decreases.

Therefore, the following convergence in distribution should be kept in mind:

$$t_n \xrightarrow[n\to\infty]{} N(0,1) \qquad (4\text{-}11)$$

Thus, when the number of degrees of freedom of a Student's t tends to infinity, the *t* distribution converges towards a distribution *N*(0.1). In the context of testing a hypothesis, if the sample size grows, so will the degrees of freedom. This means that for large sizes the normal distribution can be used to test hypothesis with one unique restriction, even when you do not know the population variance. As a practical rule, when the *df* are larger than 120, we can take the critical values from the normal distribution.

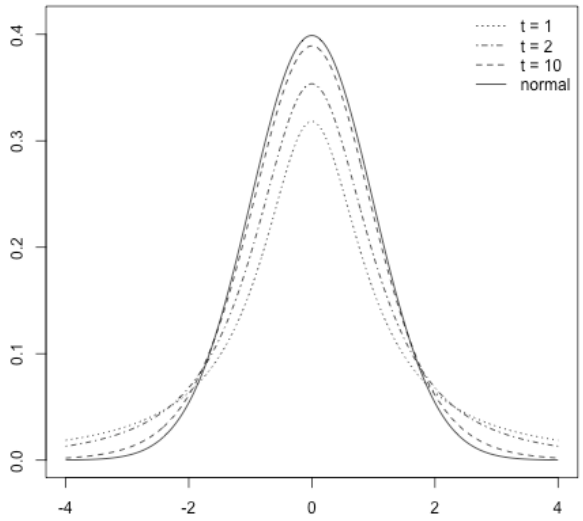


FIGURE 4.2. Density functions: normal and *t* for different degrees of freedom.

Consider the null hypothesis,

$$H_0 : \beta_j = 0$$

Since β_j measures the partial effect of x_j on *y* after controlling for all other independent variables, $H_0 : \beta_j = 0$ means that, once $x_2, x_3, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ have been accounted for, x_j has *no effect* on *y*. This is called a *significance test*. The statistic we use to test $H_0 : \beta_j = 0$, against any alternative, is called the *t statistic* or the *t ratio* of $\hat{\beta}_j$ and is expressed as

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

In order to test $H_0 : \beta_j = 0$, it is natural to look at our unbiased estimator of β_j, $\hat{\beta}_j$. In a given sample $\hat{\beta}_j$ will never be exactly zero, but a small value will indicate that the null hypothesis could be true, whereas a large value will indicate a false null hypothesis. The question is: how far is $\hat{\beta}_j$ from zero?

We must recognize that there is a sampling error in our estimate $\hat{\beta}_j$, and thus the size of $\hat{\beta}_j$ must be weighted against its sampling error. This is precisely what we do when we use $t_{\hat{\beta}_j}$, since this statistic measures how many standard errors $\hat{\beta}_j$ is away from zero. In order to determine a rule for rejecting H_0, we need to decide on the relevant *alternative hypothesis*. There are three possibilities: one-tail alternative hypotheses (right and left tail), and two-tail alternative hypothesis.

One-tail alternative hypothesis: right

First, let us consider the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j > 0$$

This is a *positive significance test*. In this case, the decision rule is the following:

Decision rule

If	$t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$	reject	H_0	(4-12)
If	$t_{\hat{\beta}_j} < t^{\alpha}_{n-k}$	not reject	H_0	

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j > 0$ at α when $t_{\hat{\beta}_j} \geq t^{\alpha}_{n-k}$ as can be seen in figure 4.3. It is very clear that to reject H_0 against $H_1 : \beta_j > 0$, we must get a positive $t_{\hat{\beta}_j}$. A negative $t_{\hat{\beta}_j}$, no matter how large, provides no evidence in favor of $H_1 : \beta_j > 0$. On the other hand, in order to obtain t^{α}_{n-k} in the *t* statistical table, we only need the significance level α and the degrees of freedom.

It is important to remark that as α decreases, t^{α}_{n-k} increases.

To a certain extent, the classical approach is somewhat arbitrary, since we need to choose α in advance, and eventually H_0 is either rejected or not.

In figure 4.4, the alternative approach is represented. As can be seen by observing the figure, the determination of the *p*-value is the inverse operation to find the value of the statistical tables for a given significance level. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, while the null hypothesis is not rejected when α<*p*-value.

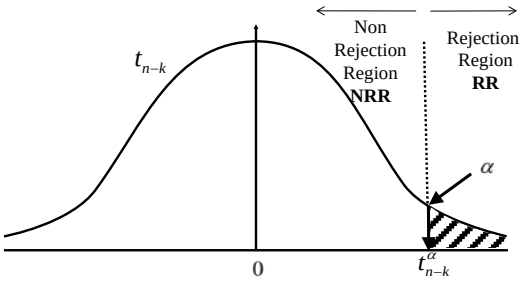


FIGURE 4.3. Rejection region using *t*: right-tail alternative hypothesis.

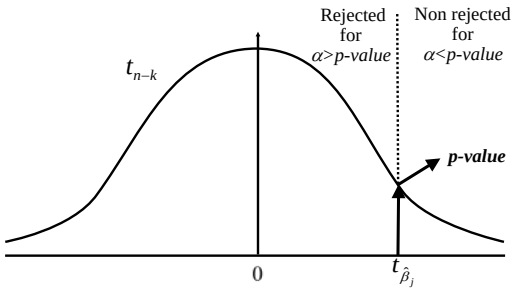


FIGURE 4.4. *p-value* using *t*: right-tail alternative hypothesis.

EXAMPLE 4.1 Is the marginal propensity to consume smaller than the average propensity to consume?

As seen in example 1.1, testing the 3rd proposition of the Keynesian consumption function in a linear model, is equivalent to testing whether the intercept is significative1y greater than 0. That is to say, in the model

$$cons = \beta_1 + \beta_2 inc + u$$

we must test whether

$$\beta_1 > 0$$

With a random sample of 42 observations, the following results have been obtained

$$\underset{}{\widehat{cons}_i} = \underset{(0.350)}{0.41} + \underset{(0.062)}{0.843} inc_i$$

The numbers in parentheses, below the estimates, are standard errors *(se*) of the estimators.

The question we pose is the following: is the third proposition of the Keynesian theory admissible? Next, we answer this question.

1) In this case, the null and alternative hypotheses are the following:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 > 0$$

2) The test statistic is:

$$t = \frac{\hat{\beta}_1 - \beta_1^0}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{0.41}{0.35} = 1.171$$

3) Decision rule

It is useful to use several significance levels. Let us begin with a significance level of 0.10 because the value of *t* is relatively small (smaller than 1.5). In this case, the degrees of freedom are 40 (42 observations minus 2 estimated parameters). If we look at the *t* statistical table (row 40 and column 0.10, or 0.20, in statistical tables with one tail, or two tails, respectively), we find $t_{40}^{0.10} = 1.303$

As *t*<1.303, we do not reject H_0 for α=0.10, and therefore we cannot reject for α=0.05 ($t_{40}^{0.05} = 1.684$) or α=0.01 ($t_{40}^{0.01} = 2.423$), as can been in figure 4.5. In this figure, the rejection region corresponds to α=0.10. Therefore, we cannot reject H_0 in favor H_1. In other words, the sample data are not consistent with Keynes's proposition 3.

In the alternative approach, as can be seen in figure 4.6, the *p*-value corresponding to a $t_{\hat{\beta}_1}$ =1.171 for a *t* with 40 *df* is equal to 0.124. For α<0.124 - for example, 0.10, 0.05 and 0.01-, H_0 is not rejected.

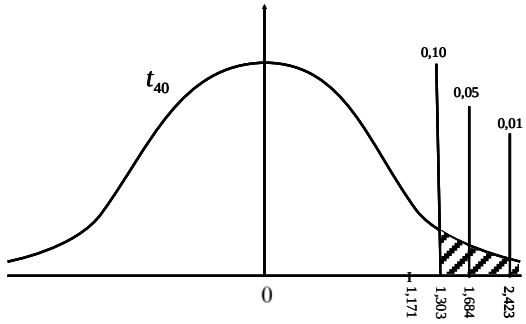


FIGURE 4.5. Example 4.1: Rejection region using *t* with a right-tail alternative hypothesis.

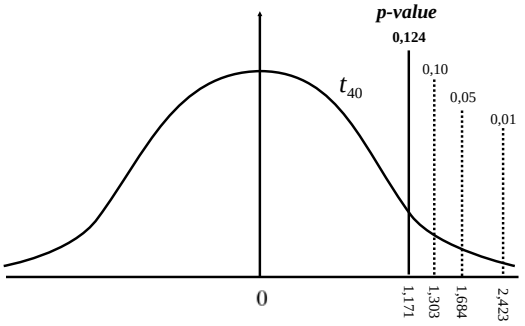


FIGURE 4.6. Example 4.1: *p*-value using *t* with right-tail alternative hypothesis.

One-tail alternative hypothesis: left

Consider now the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j < 0$$

This is a *negative significance test*.

In this case, the decision rule is the following:

Decision rule

If	$t_{\hat{\beta}_j} \leq -t^{\alpha}_{n-k}$	reject	H_0	(4-13)
If	$t_{\hat{\beta}_j} > -t^{\alpha}_{n-k}$	not reject	H_0	

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j < 0$ at a given α when $t_{\hat{\beta}_j} \leq -t^{\alpha}_{n}$, as can be seen in figure 4.7. It is very clear that to reject H_0 against $H_1 : \beta_j < 0$, we must get a negative $t_{\hat{\beta}_j}$. A positive $t_{\hat{\beta}_j}$, no matter how large it is, provides no evidence in favor of $H_1 : \beta_j < 0$.

In figure 4.8 the alternative approach is represented. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, while the null hypothesis is not rejected when α<*p*-value.

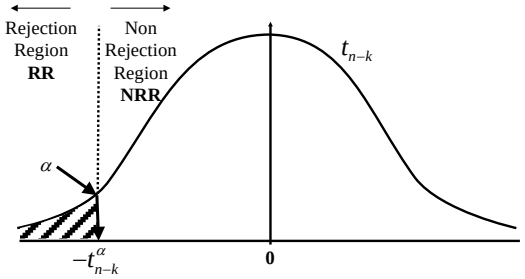


FIGURE 4.7. Rejection region using *t*: left-tail alternative hypothesis.

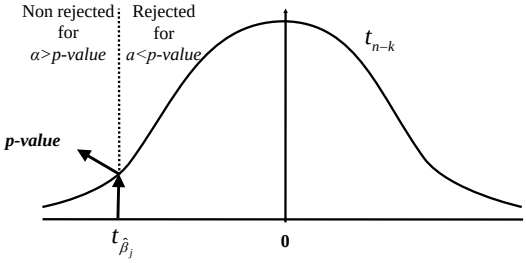


FIGURE 4.8. *p*-value using *t*: left-tail alternative hypothesis.

EXAMPLE 4.2 Has income a negative influence on infant mortality?

The following model has been used to explain the deaths of children under 5 years per 1000 live births (*deathun5*).

$$deathun5 = \beta_1 + \beta_2 gnipc + \beta_3 ilitrate + u$$

where *gnipc* is the gross national income per capita and *ilitrate* is the adult (% 15 and older) illiteracy rate in percentage.

With a sample of 130 countries (workfile *hdr2010*), the following estimation has been obtained:

$$\widehat{deathun5}_i = \underset{(5.93)}{27.91} - \underset{(0.00028)}{0.000826}\, gnipc_i + \underset{(0.183)}{2.043}\, ilitrate_i$$

The numbers in parentheses, below the estimates, are standard errors *(se*) of the estimators.

One of the questions posed by researchers is whether income has a negative influence on infant mortality. To answer this question, the following hypothesis testing is carried out:

The null and alternative hypotheses, and the test statistic, are the following:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 < 0$$
$$t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{-0.000826}{0.00028} = -2.966$$

Since the *t* value is relatively high, let us start testing with a level of 1%. For α=0.01, $t^{0.01}_{130-1-2} \approx t^{0.01}_{60} = 2.390$. Given that *t*<-2.390, as is shown in figure 4.9, we reject H_0 in favour of H_1. Therefore, the gross national income per capita has an influence that is significantly negative in mortality of children under 5.That is to say, the higher the gross national income per capita the lower the percentage of mortality of children under 5. As H_0 has been rejected for α=0.01, it will also be rejected for levels of 5% and 10%.

In the alternative approach, as can be seen in figure 4.10, the *p*-value corresponding to a $t_{\hat{\beta}_1}$ =-2.966 for a *t* with 61 *df* is equal to 0.0000. For all α>0.0000, such as 0.01, 0.05 and 0.10, H_0 is rejected.

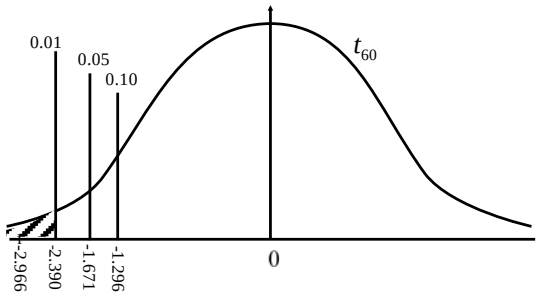


FIGURE 4.9. Example 4.2: Rejection region using *t* with a left-tail alternative hypothesis.

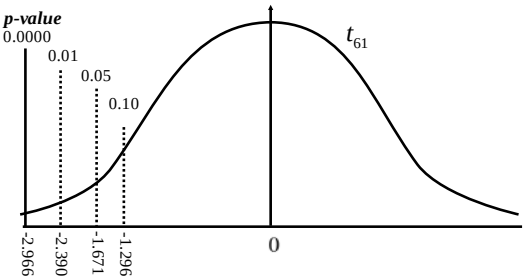


FIGURE 4.10. Example 4.2: *p*-value using *t* with a left-tail alternative hypothesis.

Two-tail alternative hypothesis

Consider now the null hypothesis

$$H_0 : \beta_j = 0$$

against the alternative hypothesis

$$H_1 : \beta_j \neq 0$$

This is the relevant alternative when the sign of β_j is not well determined by theory or common sense. When the alternative is two-sided, we are interested in the *absolute value* of the *t* statistic. This is a *significance test*.

In this case, the decision rule is the following:

Decision rule

$$\begin{array}{lll} \text{If} & \left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2} & \text{reject } H_0 \\ \text{If} & \left| t_{\hat{\beta}_j} \right| < t_{n-k}^{\alpha/2} & \text{not reject } H_0 \end{array} \qquad (4\text{-}14)$$

Therefore, we reject $H_0 : \beta_j = 0$ in favor of $H_1 : \beta_j < 0$ at α when $\left| t_{\hat{\beta}_j} \right| \geq t_{n-k}^{\alpha/2}$, as can be seen in figure 4.11. In this case, in order to reject H_0 against $H_1 : \beta_j \neq 0$, we must obtain a large enough $t_{\hat{\beta}_j}$ which is either positive or negative.

It is important to remark that as α decreases, $t_{n-k}^{\alpha/2}$ increases in absolute value.

In the alternative approach, once the *p*-value has been determined, we know that while H_0 is rejected for any level of significance of α>*p*-value, the null hypothesis is not rejected when α<*p*-value. In this case, the *p*-value is distributed between both tails in a symmetrical way, as is shown in figure 4.12.

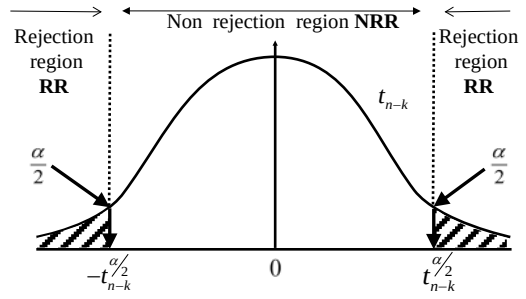


FIGURE 4.11. Rejection region using *t*: two-tail alternative hypothesis.

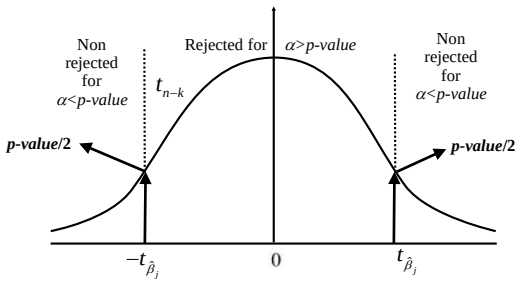


FIGURE 4.12. *p*-value using *t*: two-tail alternative hypothesis.

When a specific alternative hypothesis is not stated, it is usually considered to be two-sided hypothesis testing. If H_0 is rejected in favor of H_1 at a given α, we usually say that "x_j is *statistically significant* at the level α".

EXAMPLE 4.3 Has the rate of crime play a role in the price of houses in an area?

To explain housing prices in an American town, the following model is estimated:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

where *rooms* is the number of rooms of the house, *lowstat* is the percentage of people of "lower status" in the area and *crime* is crimes committed per capita in the area.

The output for the fitted model, using the file *hprice2* (first 55 observations), appears in table 4.2 and has been taken from E-views. The meaning of the first three columns is clear: "t-Statistic" is the outcome to perform a significance test, that is to say, it is the ratio between the "Coefficient" and the "Std error"; and "Prob" is the *p*-value to perform a two-tailed test.

In relation to this model, the researcher questions whether the rate of crime in an area plays a role in the price of houses in that area.

To answer this question, the following procedure has been carried out.

In this case, the null and alternative hypothesis and the test statistic are the following:

$$H_0 : \beta_4 = 0 \qquad t = \frac{\hat{\beta}_4}{se(\hat{\beta}_4)} = \frac{-3854}{960} = -4.016$$
$$H_1 : \beta_4 \neq 0$$

TABLE 4.2. Standard output in the regression explaining house price. *n*=55.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	-15693.61	8021.989	-1.956324	0.0559
ROOMS	6788.401	1210.720	5.606910	0.0000
LOWSTAT	-268.1636	80.70678	-3.322690	0.0017
CRIME	-3853.564	959.5618	-4.015962	0.0002

Since the *t* value is relatively high, let us by start testing with a level of 1%. For α=0.01, $t_{51}^{0.01/2} \approx t_{50}^{0.01/2} = 2.69$. (In the usual statistical tables for *t* distribution, there is no information for each *df* above 20). Given that $|t| > 2.69$, we reject H_0 in favour of H_1. Therefore, crime has a significant influence on housing prices for a significance level of 1% and, thus, of 5% and 10%.

In the alternative approach, we can perform the test with more precision. In table 4.2 we see that the *p*-value for the coefficient of crime is 0.0002. That means that the probability of the *t* statistic being greater than 4.016 is 0.0001 and the probability of *t* being smaller than -4.016 is 0.0001. That is to say, the *p*-value, as shown in Figure 4.13, is distributed in the two tails. As can be seen in this figure, H_0 is rejected for all significance levels greater than 0.0002, such as 0.01, 0.05 and 0.10.

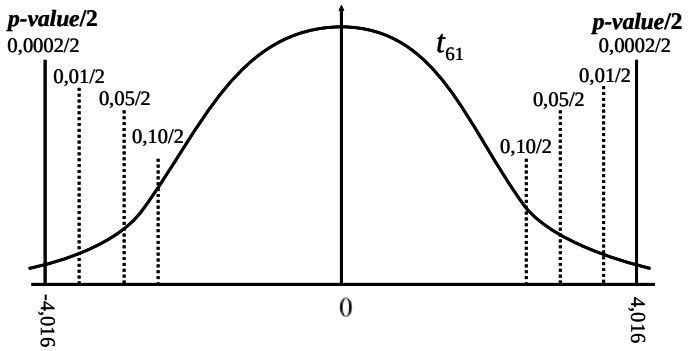


FIGURE 4.13. Example 4.3: *p*-value using *t* with a two-tail alternative hypothesis.

So far we have seen significant tests of one-tail and two-tails, in which a parameter takes the value 0 in H_0. Now we are going to look at a more general case where the parameter in H_0 takes any value:

$$H_0 : \beta_j = \beta_j^0$$

Thus, the appropriate *t* statistic is

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j^0}{se(\hat{\beta}_j)}$$

As before, $t_{\hat{\beta}_j}$ measures how many estimated standard deviations $\hat{\beta}_j$ is away from the hypothesized value of β_j^0.

EXAMPLE 4.4 Is the elasticity expenditure in fruit/income equal to 1? Is fruit a luxury good?

To answer these questions, we are going to use the following model for the expenditure in *fruit*:

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 househsize + \beta_4 punders + u$$

where *inc* is disposable income of household, *househsize* is the number of household members and *punder5* is the proportion of children under five in the household.

As the variables *fruit* and *inc* appear expressed in natural logarithms, then β_2 is the expenditure in fruit/income elasticity. Using a sample of 40 households (workfile *demand*), the results of table 4.3 have been obtained.

TABLE 4.3. Standard output in a regression explaining expenditure in fruit. *n*=40.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
C	-9.767654	3.701469	-2.638859	0.0122
LN(INC)	2.004539	0.512370	3.912286	0.0004
HOUSEHSIZE	-1.205348	0.178646	-6.747147	0.0000
PUNDER5	-0.017946	0.013022	-1.378128	0.1767

Is the expenditure in fruit/income elasticity equal to 1?

To answer this question, the following procedure has been carried out:

In this case, the null and alternative hypothesis and the test statistic are the following:

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 \neq 1 \qquad t = \frac{\hat{\beta}_2 - \beta_2^0}{se(\hat{\beta}_2)} = \frac{\hat{\beta}_2 - 1}{se(\hat{\beta}_2)} = \frac{2.005 - 1}{0.512} = 1.961$$

For α=0.10, we find that $t_{36}^{0.10/2} \approx t_{35}^{0.10/2} = 1.69$. As $|t|$>1.69, we reject H_0. For α=0.05, $t_{36}^{0.05/2} \approx t_{35}^{0.05/2} = 2.03$. As $|t|$<2.03, we do not reject H_0 for α=0.05, nor for α=0.01. Therefore, we reject that the expenditure on fruit/income elasticity is equal to 1 for α=0.10, but we cannot reject it for α=0.05, nor for α=0.01.

Is fruit a luxury good?

According to economic theory, a commodity is a luxury good when its expenditure elasticity with respect to income is higher than 1. Therefore, to answer to the second question, and taking into account that the t statistic is the same, the following procedure has been carried out:

$$H_0 : \beta_2 = 1 \qquad H_1 : \beta_2 > 1.$$

For α=0.10, we find that $t_{36}^{0.10} \approx t_{35}^{0.10} = 1.31$. As t>1.31, we reject H_0 in favour of H_1. For α=0.05, $t_{36}^{0.05} \approx t_{35}^{0.05} = 1.69$. As t>1.69, we reject H_0 in favour of H_1. For α=0.01, $t_{36}^{0.01} \approx t_{35}^{0.01} = 2.44$. As t<2.44, we do not reject H_0. Therefore, fruit is a luxury good for α=0.10 and α=0.05, but we cannot reject H_0 in favour of H_1 for α=0.01.

EXAMPLE 4.5 Is the Madrid stock exchange market efficient?

Before answering this question, we will examine some previous concepts. The *rate of return of an asset* over a period of time is defined as the percentage change in the value invested in the asset during that period of time. Let us now consider a specific asset: a share of an industrial company acquired in a Spanish stock market at the end of one year and remains until the end of next year. Those two moments of time will be denoted by *t*-1 and *t* respectively. The rate of return of this action within that year can be expressed by the following relationship:

$$RA_t = \frac{\square P_t \square D_t \square A_t}{P_{t\square 1}} \qquad (4\text{-}15)$$

where P_t: is the share price at the end of period t, D_t: are the dividends received by the share during the period t, and A_t: is the value of the rights that eventually corresponded to the share during the period t

Thus, the numerator of (4-15) summarizes the three types of capital gains that have been received for the maintenance of a share in year t; that is to say, an increase or decrease in quotation, dividends and rights on capital increase. Dividing by P_{t-1}, we obtain the rate of profit on share value at the end of the previous period. Of these three components, the most important one is the increase in quotation. Considering only that component, the yield rate of the action can be expressed by

$$RA1_t = \frac{\Delta P_t}{P_{t-1}} \qquad (4\text{-}16)$$

or, alternatively if we use a relative rate of variation, by

$$RA2_t = \Delta \ln P_t \qquad (4\text{-}17)$$

In the same way as Ra_t represents the rate of return of a particular share in either of the two expressions, we can also calculate the rate of return of all shares listed in the stock exchange. The latter rate of return, which will be denoted by RM_t, is called the market rate of return.

So far we have considered the rate of return in a year, but we can also apply expressions such as (4-16), or (4-17), to obtain daily rates of return. It is interesting to know whether the rates of return in the past are useful for predicting rates of return in the future. This question is related to the concept of market efficiency. A market is *efficient* if prices incorporate all available information, so there is no possibility of making abnormal profits by using this information.

In order to test the efficiency of a market, we define the following model, using daily rates of return defined by (4-16):

$$rmad92_t = \beta_1 + \beta_2 rmad92_{t-1} + u_t \qquad (4\text{-}18)$$

If a market is efficient, then the parameter β_2 of the previous model must be 0. Let us now compare whether the Madrid Stock Exchange is efficient as a whole.

The model (4-18) has been estimated with daily data from the Madrid Stock Exchange for 1992, using file *bolmadef*. The results obtained are the following:

$$\widehat{rmad92}_t = \underset{(0.0007)}{-0.0004} + \underset{(0.0629)}{0.1267}\, rmad92_{t-1}$$

$$R^2=0.0163 \qquad n=247$$

The results are paradoxical. On the one hand, the coefficient of determination is very low (0.0163), which means that only 1.63% of the total variance of the rate of return is explained by the previous day's rate of return. On the other hand, the coefficient corresponding to the rate of significance of the previous day is statistically significant at a level of 5% but not at a level of 1% given that the t statistic is equal to 0.1267/0.0629=2.02, which is slightly larger in absolute value than $t_{245}^{0.01} \approx t_{60}^{0.01}$=2.00. The reason for this apparent paradox is that the sample size is very high. Thus, although the impact of the explanatory variable on the endogenous variable is relatively small (as indicated by the coefficient of determination), this finding is significant (as evidenced by the statistical t) because the sample is sufficiently large.

To answer the question as to whether the Madrid Stock Exchange is an efficient market, we can say that it is not entirely efficient. However, this response should be qualified. In financial economics there is a dependency relationship of the rate of return of one day with respect to the rate corresponding to the previous day. This relationship is not very strong, although it is statistically significant in many world stock markets due to market frictions. In any case, market players cannot exploit this phenomenon, and thus the market is not inefficient, according to the above definition of the concept of efficiency.

EXAMPLE 4.6 Is the rate of return of the Madrid Stock Exchange affected by the rate of return of the Tokyo Stock Exchange?

The study of the relationship between different stock markets (NYSE, Tokyo Stock Exchange Madrid Stock Exchange, London Stock Exchange, etc.) has received much attention in recent years due to the greater freedom in the movement of capital and the use of foreign markets to reduce the risk in portfolio management. This is because the absence of perfect market integration allows diversification of risk. In any

case, there is a world trend toward a greater global integration of financial markets in general and stock markets in particular.

If markets are efficient, and we have seen in example 4.5 that they are, the *innovations* (new information) will be reflected in the different markets for a period of 24 hours.

It is important to distinguish between two types of innovations: a) *global innovations*, which is news generated around the world and has an influence on stock prices in all markets, b) *specific innovations*, which is the information generated during a 24 hour period and only affects the price of a particular market. Thus, information on the evolution of oil prices can be considered as a global innovation, while a new financial sector regulation in a country would be considered a specific innovation.

According to the above discussion, stock prices quoted at a session of a particular stock market are affected by the global innovations of a different market which had closed earlier. Thus, global innovations included in the Tokyo market will influence the market prices of Madrid on the same day. The following model shows the transmission of effects between the Tokyo Stock Exchange and the Madrid Stock Exchange in 1992:

$$rmad92_t = \beta_1 + \beta_2 rtok92_t + u_t \qquad (4\text{-}19)$$

where $rmad92_t$ is the rate of return of the Madrid Stock Exchange in period t and $rtok92_t$ is the rate of return of the Tokyo Stock Exchange in period t. The rates of return have been calculated according to (4-16).

In the working file *madtok* you can find general indices of the Madrid Stock Exchange and the Tokyo Stock Exchange during the days both exchanges were open simultaneously in 1992. That is, we eliminated observations for those days when any one of the two stock exchanges was closed. In total, the number of observations is 234, compared to the 247 and 246 days that the Madrid and Tokyo Stock Exchanges were open.

The estimation of the model (4-19) is as follows:

$$\widehat{rmad}92_t = \underset{(0.0007)}{-0.0005} + \underset{(0.0375)}{0.1244}\, rtok92_t$$

$$R^2 = 0.0452 \qquad n = 235$$

Note that the coefficient of determination is relatively low. However, for testing H_0: $\beta_2 = 0$, the statistic $t = (0.1244/0.0375) = 3.32$, which implies that we reject the hypothesis that the rate of return of the Tokyo Stock Exchange has no effect on the rate of return of the Madrid Stock Exchange, for a significance level of 0.01.

Once again we find the same apparent paradox which appeared when we analyzed the efficiency of the Madrid Stock Exchange in example 4.5 except for one difference. In the latter case, the rate of return from the previous day appeared as significant due to problems arising in the elaboration of the general index of the Madrid Stock Exchange.

Consequently, the fact that the null hypothesis is rejected implies that there is empirical evidence supporting the theory that global innovations from the Tokyo Stock Exchange are transmitted to the quotes of the Madrid Stock Exchange that day.

4.2.2 Confidence intervals

Under the *CLM*, we can easily construct a *confidence interval* (*CI*) for the population parameter, β_j. *CI* are also called interval estimates because they provide a range of likely values for β_j, and not just a point estimate.

The *CI* is built in such a way that the unknown parameter is contained within the range of the *CI* with a previously specified probability.

By using the fact that

$$\frac{\hat{b}_j - b_j}{se(\hat{b}_j)} : \; t_{n-k}$$

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \le t_{n-k}^{\alpha/2}\right] = 1-\alpha$$

Operating to put the unknown β_j alone in the middle of the interval, we have

$$\Pr\left[\hat{\beta}_j - se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2} \le \beta_j \le \hat{\beta}_j + se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha$$

Therefore, the lower and upper bounds of a (1-α) *CI* respectively are given by

$$\underline{\beta}_j = \hat{\beta}_j - se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}$$

$$\overline{\beta}_j = \hat{\beta}_j + se(\hat{\beta}_j)\times t_{n-k}^{\alpha/2}$$

If random samples were obtained over and over again with $\underline{\beta}_j$, and $\overline{\beta}_j$ computed each time, then the (unknown) population value would lie in the interval ($\underline{\beta}_j$, $\overline{\beta}_j$) for (1 – α)% of the samples. Unfortunately, for the single sample that we use to construct *CI*, we do not know whether β_j is actually contained in the interval.

Once a *CI* is constructed, it is easy to carry out two-tailed hypothesis tests. If the null hypothesis is $H_0 : \beta_j = a_j$, then H_0 is rejected against $H_1 : \beta_j \neq a_j$ at (say) the 5% significance level if, and only if, a_j is *not* in the 95% *CI*.

To illustrate this matter, in figure 4.14 we constructed confidence intervals of 90%, 95% and 99%, for the marginal propensity to consumption -β_2- corresponding to example 4.1.

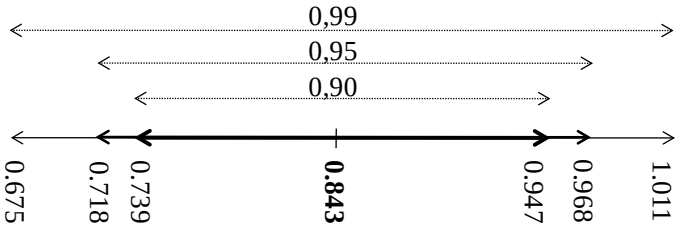


FIGURE 4.14. Confidence intervals for marginal propensity to consume in example 4.1.

4.2.3 Testing hypotheses about a single linear combination of the parameters

In many applications we are interested in testing a hypothesis involving more than one of the population parameters. We can also use the *t* statistic to test a single linear combination of the parameters, where two or more parameters are involved.

There are two different procedures to perform the test with a single linear combination of parameters. In the first, the standard error of the linear combination of parameters corresponding to the null hypothesis is calculated using information on the covariance matrix of the estimators. In the second, the model is reparameterized by introducing a new parameter derived from the null hypothesis and the reparameterized model is then estimated; testing for the new parameter indicates whether the null hypothesis is rejected or not. The following example illustrates both procedures.

EXAMPLE 4.7 Are there constant returns to scale in the chemical industry?

To examine whether there are constant returns to scale in the chemical sector, we are going to use the Cobb-Douglas production function, given by

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u \qquad (4\text{-}20)$$

In the above model parameters β_2 and β_3 are elasticities (output/labor and output/capital).

Before making inferences, remember that *returns to scale* refers to a technical property of the production function examining changes in output subsequent to a change of the same proportion in all inputs, which are labor and capital in this case. If output increases by that same proportional change then there are *constant returns to scale*. Constant returns to scale imply that if the factors *labor* and *capital* increase at a certain rate (say 10%), output will increase at the same rate (e.g., 10%). If output increases by more than that proportion, there are *increasing returns to scale*. If output increases by less than that proportional change, there are *decreasing returns to scale*. In the above model, the following occurs

- if $\beta_2+\beta_3=1$, there are *constant returns to scale*.
- if $\beta_2+\beta_3>1$, there are *increasing returns to scale*.
- if $\beta_2+\beta_3<1$, there are *decreasing returns to scale*.

Data used for this example are a sample of 27 companies of the primary metal sector (workfile *prodmet*), where *output* is gross value added, *labor* is a measure of labor input, and *capital* is the gross value of plant and equipment. Further details on construction of the data are given in Aigner, *et al.* (1977) and in Hildebrand and Liu (1957); these data were used by Greene in 1991. The results obtained in the estimation of model (4-20), using any econometric software available, appear in table 4.4.

TABLE 4.4. Standard output of the estimation of the production function: model (4-20).

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	0.602999	0.125954	4.787457	0.0001
ln(*capital*)	0.375710	0.085346	4.402204	0.0002

To answer the question posed in this example, we must test

$$H_0 : \beta_2 + \beta_3 = 1$$

against the following alternative hypothesis

$$H_1 : \beta_2 + \beta_3 \neq 1$$

According to H_0, it is stated that $\beta_2 + \beta_3 - 1 = 0$. Therefore, the *t* statistic must now be based on whether the estimated sum $\hat{\beta}_2 + \hat{\beta}_3 - 1$ is sufficiently different from 0 to reject H_0 in favor of H_1.

Two procedures will be used to test this hypothesis. In the first, the covariance matrix of the estimators is used. In the second, the model is reparameterized by introducing a new parameter.

Procedure: using covariance matrix of estimators

According to H_0, it is stated that $\beta_2 + \beta_3 - 1 = 0$. Therefore, the *t* statistic must now be based on whether the estimated sum $\hat{\beta}_2 + \hat{\beta}_3 - 1$ is sufficiently different from 0 to reject H_0 in favor of H_1. To account for the sampling error in our estimators, we standardize this sum by dividing by its standard error:

$$t_{\hat{\beta}_2+\hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{se(\hat{\beta}_2 + \hat{\beta}_3)}$$

Therefore, if $t_{\hat{\beta}_2+\hat{\beta}_3}$ is large enough, we will conclude, in a two side alternative test, that there are not *constant returns to scale*. On the other hand, if $t_{\hat{\beta}_2+\hat{\beta}_3}$ is positive and large enough, we will reject, in a one side alternative test (right), H_0 in favour of $H_1 : \beta_2 + \beta_3 > 1$. Therefore, there are *increasing returns to scale*.

On the other hand , we have

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{\overline{\text{var}}(\hat{\beta}_2 + \hat{\beta}_3)}$$

where

$$\overline{\text{var}}(\hat{\beta}_2 + \hat{\beta}_3) = \overline{\text{var}}(\hat{\beta}_2) + \overline{\text{var}}(\hat{\beta}_3) + 2 \times \overline{\text{covar}}(\hat{\beta}_2, \hat{\beta}_3)$$

Hence, to compute $se(\hat{\beta}_2 + \hat{\beta}_3)$ you need information on the estimated covariance of estimators. Many econometric software packages (such as e-views) have an option to display estimates of the covariance matrix of the estimator vector '. In this case, the covariance matrix obtained appears in table 4.5. Using this information, we have

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{0.015864 + 0.007284 - 2 \times 0.009616} = 0.0626$$

$$t_{\hat{\beta}_2 + \hat{\beta}_3} = \frac{\hat{\beta}_2 + \hat{\beta}_3 - 1}{se(\hat{\beta}_2 + \hat{\beta}_3)} = \frac{-0.02129}{0.0626} = -0.3402$$

TABLE 4.5. Covariance matrix in the production function.

	constant	ln(*labor*)	ln(*capital*)
constant	0.106786	-0.019835	0.001189
ln(*labor*))	-0.019835	0.015864	-0.009616
ln(*capital*)	0.001189	-0.009616	0.007284

Given that *t*=0.3402, it is clear that we cannot reject the existence of constant returns to scale for the usual significance levels. Given that the t statistic is negative, it makes no sense to test whether there are increasing returns to scale

Procedure: reparameterizing the model by introducing a new parameter

It is easier to perform the test if we apply the second procedure. A different model is estimated in this procedure, which directly provides the standard error of interest. Thus, let us define:

$$\theta = \beta_2 + \beta_3 - 1$$

thus, the null hypothesis that there are *constant returns to scale* is equivalent to saying that $H_0 : \theta = 0$.

From the definition of θ, we have $\beta_2 = \theta - \beta_3 + 1$. Substituting β_2 in the original equation:

$$\ln(output) = \beta_1 + (\theta - \beta_3 + 1)\ln(labor) + \beta_3 \ln(capital) + u$$

Hence,

$$\ln(output / labor) = \beta_1 + \theta \ln(labor) + \beta_3 \ln(capital / labor) + u$$

Therefore, to test whether there are constant returns to scale is equivalent to carrying out a significance test on the coefficient of ln(*labor*) in the previous model. The strategy of rewriting the model so that it contains the parameter of interest works in all cases and is usually easy to implement. If we apply this transformation to this example, we obtain the results of Table 4.6.

As can be seen we obtain the same result:

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{se(\hat{\theta})} = -0.3402$$

TABLE 4.6. Estimation output for the production function: reparameterized model.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	1.170644	0.326782	3.582339	0.0015
ln(*labor*)	-0.021290	0.062577	-0.340227	0.7366
ln(*capital/labor*)	0.375710	0.085346	4.402204	0.0002

EXAMPLE 4.8 Advertising or incentives?

The *Bush Company* is engaged in the sale and distribution of gifts imported from the Near East. The most popular item in the catalog is the *Guantanamo* bracelet, which has some relaxing properties. The sales agents receive a commission of 30% of total sales amount. In order to increase sales without expanding the sales network, the company established special incentives for those agents who exceeded a sales target during the last year.

Advertising spots were radio broadcasted in different areas to strengthen the promotion of sales. In those spots special emphasis was placed on highlighting the well-being of wearing a *Guantanamo* bracelet.

The manager of the *Bush Company* wonders whether a dollar spent on special incentives has a higher incidence on sales than a dollar spent on advertising. To answer that question, the company's econometrician suggests the following model to explain *sales*:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u$$

where *incent* are incentives to the salesmen and *advert* are expenditures in advertising. The variables *sales*, *incent* and *advert* are expressed in thousands of dollars.

Using a sample of 18 sale areas (workfile *advincen*), we have obtained the output and the covariance matrix of the coefficients that appear in table 4.7 and in table 4.8 respectively.

TABLE 4.7. Standard output of the regression for example 4.8.

Variable	*Coefficient*	*Std. Error*	*t-Statistic*	*Prob.*
constant	396.5945	3548.111	0.111776	0.9125
advert	18.63673	8.924339	2.088304	0.0542
incent	30.69686	3.604420	8.516448	0.0000

TABLE 4.8. Covariance matrix for example 4.8.

	C	*ADVERT*	*INCENT*
constant	12589095	-26674	-7101
advert	-26674	79.644	2.941
incent	-7101	2.941	12.992

In this model, the coefficient β_2 indicates the increase in sales produced by a dollar increase in spending on advertising, while β_3 indicates the increase caused by a dollar increase in the special incentives, holding fixed in both cases the other regressor.

To answer the question posed in this example, the null and the alternative hypothesis are the following:

$$H_0 : \beta_3 - \beta_2 = 0$$
$$H_1 : \beta_3 - \beta_2 > 0$$

The t statistic is built using information about the covariance matrix of the estimators:

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{se(\hat{\beta}_3 - \hat{\beta}_2)}$$

$$se(\hat{\beta}_3 - \hat{\beta}_2) = \sqrt{79.644 + 12.992 - 2 \times 2.941} = 9.3142$$

$$t_{\hat{\beta}_3 - \hat{\beta}_2} = \frac{\hat{\beta}_3 - \hat{\beta}_2}{se(\hat{\beta}_3 - \hat{\beta}_2)} = \frac{30.697 - 18.637}{9.3142} = 1.295$$

For α=0.10, we find that $t_{15}^{0.10} = 1.341$. As t<1.341, we do not reject H_0 for α=0.10, nor for α=0.05 or α=0.01. Therefore, there is no empirical evidence that a dollar spent on special incentives has a higher incidence on sales than a dollar spent on advertising.

EXAMPLE 4.9 Testing the hypothesis of homogeneity in the demand for fish

In the case study in chapter 2, models for demand for dairy products have been estimated from cross-sectional data, using disposable income as an explanatory variable. However, the price of the product

itself and, to a greater or lesser extent, the prices of other goods are determinants of the demand. The demand analysis based on cross sectional data has precisely the limitation that it is not possible to examine the effect of prices on demand because prices remain constant, since the data refer to the same point in time. To analyze the effect of prices it is necessary to use time series data or, alternatively, panel data. We will briefly examine some aspects of the theory of demand for a good and then move to the estimation of a demand function with time series data. As a postscript to this case, we will test one of the hypotheses which, under certain circumstances, a theoretical model must satisfy.

The demand for a commodity - say good *j* - can be expressed, according to an optimization process carried out by the consumer, in terms of disposable income, the price of the good and the prices of the other goods. Analytically:

$$q_j = f_j(p_1, p_2, \ldots, p_j, \ldots, p_m, di) \qquad (4\text{-}21)$$

where

- *di* is the disposable income of the consumer.
- $p_1, p_2, \ldots, p_j, \ldots p_m$ are the prices of the goods which are taken into account by consumers when they acquire the good j.

Logarithmic models are attractive in studies on demand,, because the coefficients are directly elasticities. The log model is given by

$$\ln(q_j) = \beta_1 + \beta_2 \ln(p_1) + \beta_3 \ln(p_2) + \cdots + \beta_j \ln(p_j) + \cdots + \beta_{m+1} \ln(p_m) + \beta_{m+2} \ln(R) + u \qquad (4\text{-}22)$$

It is clear to see that all β coefficients, excluding the constant term, are elasticities of different types and therefore are independent of the units of measurement for the variables. When there is no money illusion, if all prices and income grow at the same rate, the demand for a good is not affected by these changes. Thus, assuming that prices and income are multiplied by λ, if the consumer has no money illusion, the following should be satisfied

$$f_j(\lambda p_1, \lambda p_2, \ldots, \lambda p_j, \ldots, \lambda p_m, \lambda R) = f_j(p_1, p_2, \ldots, p_j, \ldots p_m, di) \qquad (4\text{-}23)$$

From a mathematical point of view, the above condition implies that the demand function must be homogeneous of degree 0. This condition is called the *restriction of homogeneity*. Applying Euler's theorem, the restriction of homogeneity in turn implies that the sum of the demand/income elasticity and of all demand/price elasticities is zero, i.e.:

$$\sum_{h=1}^{m} \varepsilon_{q_j / p_h} + \varepsilon_{q_j / R} = 0 \qquad (4\text{-}24)$$

This restriction applied to the logarithmic model (4-22) implies that

$$\beta_2 + \beta_3 + \cdots + \beta_j + \cdots + \beta_{m+1} + \beta_{m+2} = 0 \qquad (4\text{-}25)$$

In practice, when estimating a demand function, the prices of many goods are not included, but only those that are closely related, either because they are complementary or substitute goods. It is also well known that the budgetary allocation of spending is carried out in several stages.

Next, the demand for fish in Spain will be studied by using a model similar to (4-22). Let us consider that in a first assignment, the consumer distributes its income between total consumption and savings. In a second stage, the consumption expenditure by function is performed taking into account the total consumption and the relevant prices in each function. Specifically, we assume that the only relevant price in the demand for fish is the price of the good (fish) and the price of the most important substitute (meat).

Given the above considerations, the following model is formulated:

$$\ln(fish) = \beta_1 + \beta_2 \ln(fishpr) + \beta_3 \ln(meatpr) + \beta_4 \ln(cons) + u \qquad (4\text{-}26)$$

where *fish* is fish expenditure at constant prices, *fishpr* is the price of fish, *meatpr* is the price of meat and *cons* is total consumption at constant prices.

The workfile *fishdem* contains information about this series for the period 1964-1991. Prices are index numbers with 1986 as a base, and *fish* and *cons* are magnitudes at constant prices with 1986 as a base also. The results of estimating model (4-26) are as follows:

$$\widehat{\ln(fish)} = \underset{(2.30)}{7.788} - \underset{(0.133)}{0.460}\ln(fishpr) + \underset{(0.112)}{0.554}\ln(meatpr) + \underset{(0.137)}{0.322}\ln(cons)$$

As can be seen, the signs of the elasticities are correct: the elasticity of demand is negative with respect to the price of the good, while the elasticities with respect to the price of the substitute good and total consumption are positive

In model (4-26) the homogeneity restriction implies the following null hypothesis:

$$\beta_2 + \beta_3 + \beta_4 = 0 \qquad (4\text{-}27)$$

To carry out this test, we will use a similar procedure to the one used in example 4.6. Now, the parameter θ is defined as follows

$$\theta = \beta_2 + \beta_3 + \beta_4 \qquad (4\text{-}28)$$

Setting $\beta_2 = \theta - \beta_3 - \beta_4$, the following model has been estimated:

$$\ln(fish) = \beta_1 + \theta\ln(fishpr) + \beta_3\ln(meatpr / fishpr) + \beta_4\ln(cons / fishpr) + u \qquad (4\text{-}29)$$

The results obtained were the following:

$$\widehat{\ln(fish_i)} = \underset{(2.30)}{7.788} - \underset{(0.1334)}{0.4596}\ln(fishpr_i) + \underset{(0.112)}{0.554}\ln(meatpr_i) + \underset{(0.137)}{0.322}\ln(cons_i)$$

Using (4-28), testing the null hypothesis (4-27) is equivalent to testing that the coefficient of ln(*fishpr*) in (4-29) is equal to 0. Since the t statistic for this coefficient is equal to -3.44 and $t_{24}^{0.01/2}$ =2.8, we reject the hypothesis of homogeneity regarding the demand for fish.

4.2.4 Economic importance versus statistical significance

Up until now we have emphasized statistical significance. However, it is important to remember that we should pay attention to the magnitude and the sign of the estimated coefficient in addition to *t* statistics.

Statistical significance of a variable x_j is determined entirely by the size of $t_{\hat{\beta}_j}$, whereas the economic significance of a variable is related to the size (and sign) of $\hat{\beta}_j$. Too much focus on statistical significance can lead to the false conclusion that a variable is "important" for explaining *y*, even though its estimated effect is modest.

Therefore, even if a variable is statistically significant, you need to discuss the magnitude of the estimated coefficient to get an idea of its practical or economic importance.

4.3 Testing multiple linear restrictions using the *F* test.

So far, we have only considered hypotheses involving a single restriction. But frequently, we wish to test multiple hypotheses about the underlying parameters $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$.

In multiple linear restrictions, we will distinguish three types: *exclusion restrictions, model significance* and *other linear restrictions*.

4.3.1 Exclusion restrictions

Null and alternative hypotheses; unrestricted and restricted model

We begin with the leading case of testing whether a set of independent variables has no partial effect on the dependent variable, *y*. These are called *exclusion restrictions*. Thus, considering the model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u \qquad (4\text{-}30)$$

the null hypothesis in a typical example of exclusion restrictions could be the following:

$$H_0 : \beta_4 = \beta_5 = 0$$

This is an example of a set of *multiple restrictions*, because we are putting more than one restriction on the parameters in the above equation. A test of multiple restrictions is called a *joint* hypothesis test.

The alternative hypothesis can be expressed in the following way

$$H_1: H_0 \text{ is not true}$$

It is important to remark that we test the above H_0 jointly, not individually. Now, we are going to distinguish between *unrestricted* (*UR*) and *restricted* (*R*) models. The unrestricted model is the reference model or initial model. In this example the unrestricted model is the model given in (4-30). The restricted model is obtained by imposing H_0 on the original model. In the above example, the restricted model is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

By definition, the restricted model always has fewer parameters than the unrestricted one. Moreover, it is always true that

$$RSS_R \geq RSS_{UR}$$

where RSS_R is the *RSS* of the restricted model, and RSS_{UR} is the *RSS* of the unrestricted model. Remember that, because *OLS* estimates are chosen to minimize the sum of squared residuals, the *RSS* never decreases (and generally increases) when certain restrictions (such as dropping variables) are introduced into the model.

The increase in the *RSS* when the restrictions are imposed can tell us something about the likely truth of H_0. If we obtain a large increase, this is evidence against H_0, and this hypothesis will be rejected. If the increase is small, this is not evidence against H_0, and this hypothesis will not be rejected. The question is therefore whether the observed increase in the *RSS* when the restrictions are imposed is large enough, relative to the *RSS* in the unrestricted model, to warrant rejecting H_0.

The answer depends on α, but we cannot carry out the test at a chosen α until we have a statistic whose distribution is known, and is tabulated, under H_0. Thus, we need a way to combine the information in RSS_R and RSS_{UR} to obtain a test statistic with a known distribution under H_0.

Now, let us look at the general case, where the *unrestricted model* is

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (4\text{-}31)$$

Let us suppose that there are q exclusion restrictions to test. H_0 states that q of the variables have zero coefficients. Assuming that they are the last q variables, H_0 is stated as

$$H_0 : \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0 \qquad (4\text{-}32)$$

The restricted model is obtained by imposing the q restrictions of H_0 on the unrestricted model.

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{k-q} x_{k-q} + u \qquad (4\text{-}33)$$

H_1 is stated as

$$H_1: H_0 \text{ is not true} \qquad (4\text{-}34)$$

Test statistic: F ratio

The F statistic, or F ratio, is defined by

$$F = \frac{(RSS_R - RSS_{UR}) / q}{RSS_{UR} / (n-k)} \qquad (4\text{-}35)$$

where RSS_R is the *RSS* of the restricted model, and RSS_{UR} is the *RSS* of the unrestricted model and q is the number of restrictions; that is to say, the number of equalities in the null hypothesis.

In order to use the F statistic for a hypothesis testing, we have to know its sampling distribution under H_0 in order to choose the value c for a given α, and determine the rejection rule. It can be shown that, under H_0, and assuming the *CLM* assumptions hold, the F statistic is distributed as a Snedecor's F random variable with (q,n-k) *df*. We write this result as

$$F \mid H_0 : \ F_{q,n-k} \qquad (4\text{-}36)$$

A Snedecor's F with *q degrees of freedom* in the numerator and *n-k* de *degrees of freedom* in the denominator is equal to

$$F_{q,n-k} = \frac{\chi_q^2 / q}{\chi_{n-k}^2 / n-k} \qquad (4\text{-}37)$$

where x_q^2 and x_{n-k}^2 are Chi-square distributions that are independent of each other.

In (4-35) we see that the *degrees of freedom* corresponding to RSS_{UR} (df_{UR})are *n-k.* Remember that

$$\hat{\sigma}_{UR}^2 = \frac{RSS_{UR}}{n-k} \qquad (4\text{-}38)$$

On the other hand, the *degrees of freedom* corresponding to RSS_R (df_R) are n-k+q, because in the restricted model k-q parameters are estimated. The *degrees of freedom* corresponding to RSS_R-RSS_{UR} are

(n-k+q)-(n-k)=q = *numerator degrees of freedom*=df_R-df_{UR}

Thus, in the numerator of *F*, the difference in *RSS´*s is divided by *q*, which is the number of restrictions imposed when moving from the unrestricted to the restricted model. In the denominator of *F*, RSS_{UR} is divided by df_{UR}. In fact, the denominator of *F* is simply the unbiased estimator of σ^2 in the unrestricted model.

The *F* ratio must be greater than or equal to 0, since $SSR_R - SSR_{UR} \geq 0$.

It is often useful to have a form of the *F* statistic that can be computed from the R^2 of the restricted and unrestricted models.

Using the fact that $RSS_R = TSS(1-R_R^2)$ and $RSS_{UR} = TSS(1-R_{UR}^2)$, we can write (4-35) as the following

$$F = \frac{(R_{UR}^2 - R_R^2)/q}{(1-R_{UR}^2)/(n-k)} \qquad (4\text{-}39)$$

since the *SST* term is cancelled.

This is called the *R-squared* form of the *F* statistic.

Whereas the *R*-squared form of the *F* statistic is very useful for testing exclusion restrictions, it cannot be applied for testing all kinds of linear restrictions. For example, the *F* ratio (4-39) cannot be used when the model does not have intercept or when the functional form of the endogenous variable in the unrestricted model is not the same as in the restricted model.

Decision rule

The $F_{q,n-k}$ distribution is tabulated and available in statistical tables, where we look for the critical value ($F_{q,n-k}^{\alpha}$), which depends on α (the significance level), *q* (the *df* of the numerator), and *n-k*, (the *df* of the denominator). Taking into account the above, the *decision rule* is quite simple.

Decision rule

If	$F \geq F_{q,n-k}^{\alpha}$	reject	H_0
If	$F < F_{q,n-k}^{\alpha}$	not reject	H_0

(4-40)

Therefore, we reject H_0 in favor of H_1 at α when $F \geq F_{q,n-k}^{\alpha}$, as can be seen in figure 4.15. It is important to remark that as α decreases, $F_{q,n-k}^{\alpha}$ increases. If H_0 is rejected, then we say that $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ are *jointly statistically significant*, or just *jointly significant*, at the selected significance level.

This test alone does not allow us to say which of the variables has a partial effect on *y*; they may all affect *y* or only one may affect *y*. If H_0 is not rejected, then we say that $x_{k-q+1}, x_{k-q+2}, \cdots, x_k$ are jointly not statistically significant, or simply jointly not significant, which often justifies dropping them from the model. The *F* statistic is often useful for testing the exclusion of a group of variables when the variables in the group are highly correlated.

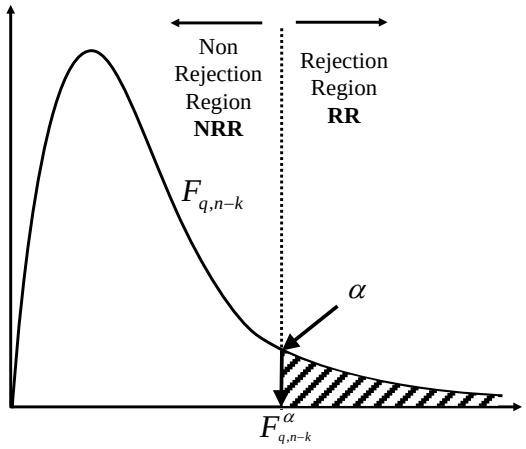


FIGURE 4.15. Rejection region and non rejection region using *F* distribution.

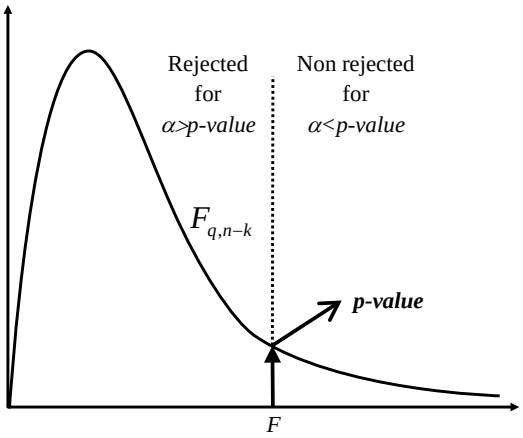


FIGURE 4.16. *p-value* using *F* distribution.

In the *F* testing context, the *p*-value is defined as

$$p\text{-}value = \Pr(F > F' \mid H_0)$$

where *F* is the actual value of the test statistic and F' denotes a Snedecor's *F* random variable with (*q*,*n*-*k*) *df*.

The *p*-value still has the same interpretation as for *t* statistics. A small *p*-value is evidence against H_0, while a large *p*-value is not evidence against H_0. Once the *p*-value has been computed, the *F* test can be carried out at any significance level. In figure 4.16 this alternative approach is represented. As can be seen by observing the figure, the determination of the *p*-value is the inverse operation to find the value in the statistical tables for a given significance level. Once the *p*-value has been determined, we know that H_0 is rejected for any level of significance of α>*p*-value, whereas the null hypothesis is not rejected when α<*p*-value.

EXAMPLE 4.10 Wage, experience, tenure and age

The following model has been built to analyze the determinant factors of wage:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + \beta_4 tenure + \beta_5 age + u$$

where *wage* is monthly earnings, *educ* is years of education, *exper is* years of work experience, *tenure* is years with current employer, and *age* is age in years.

The researcher is planning to exclude *tenure* from the model, since in many cases it is equal to experience, and also *age*, because it is highly correlated with experience. Is the exclusion of both variables acceptable?

The null and alternative hypotheses are the following:

$$H_0 : \beta_4 = \beta_5 = 0$$
$$H_1 : H_0 \text{ is not true}$$

The restricted model corresponding to this H_0 is

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 exper + u$$

Using a sample consisting of 53 observations from workfile *wage2*, we have the following estimations for the unrestricted and for the restricted models:

$$\widehat{\ln(wage_i)} = 6.476 + 0.0658 educ_i + 0.0267 exper_i - 0.0094 tenure_i - 0.0209 age_i \quad RSS = 5.954$$

$$\widehat{\ln(wage_i)} = 6.157 + 0.0457 educ_i + 0.0121 exper_i \quad RSS = 6.250$$

The *F* ratio obtained is the following:

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)} = \frac{(6.250 - 5.954)/2}{5.954/48} = 1.193$$

Given that the *F* statistic is low, let us see what happens with a significance level of 0.10. In this case the degrees of freedom for the denominator are 48 (53 observations minus 5 estimated parameters). If we look in the *F* statistical table for 2 *df* in the numerator and 45 *df* in the denominator, we find $F_{2,48}^{0.10}$; $F_{2,45}^{0.10}$ =2.42. As *F*<2.42, we do not reject H_0. If we do not reject H_0 for 0.10, we will not reject H_0 for 0.05 or 0.01, as can been in figure 4.17. Therefore, we cannot reject H_0 in favor of H_1. In other words *tenure* and *age* are not jointly significant.

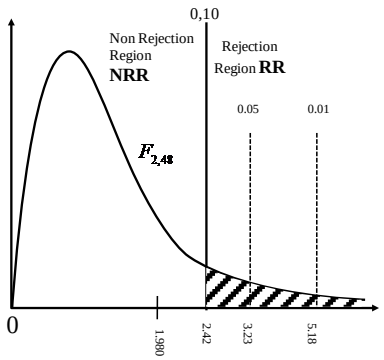


FIGURE 4.17. Example 4.10: Rejection region using *F* distribution (α values are from a $F_{2.40}$).

4.3.2 Model significance

Testing model significance, or overall significance, is a particular case of testing exclusion restrictions. Model significance means global significance of the model. One could think that the H_0 in this test is the following:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 \qquad (4\text{-}41)$$

However, this is not the adequate H_0 to test for the global significance of the model. If $\beta_2 = \beta_3 = \cdots = \beta_k = 0$, then the restricted model would be the following:

$$y = \beta_1 + u \qquad (4\text{-}42)$$

If we take expectations in (4-42), then we have

$$E(y) = \beta_1 \qquad (4\text{-}43)$$

Thus, H_0 in (4-41) states not only that the explanatory variables have no influence on the endogenous variable, but also that the mean of the endogenous variable– for example, the consumption mean- is equal to 0.

Therefore, if we want to know whether the model is globally significant, the H_0 must be the following:

$$H_0 : \beta_2 = \beta_3 = \cdots = \beta_k = 0 \qquad (4\text{-}44)$$

The corresponding restricted model given in (4-42) does not explain anything and, therefore, R_R^2 is equal to 0. Testing the H_0 given in (4-44) is very easy by using the *R-squared* form of the *F* statistic:

$$F = \frac{R^2 / k}{(1-R^2)/(n-k)} \qquad (4\text{-}45)$$

where R^2 is the R^2_{UR}, since only the unrestricted model needs to be estimated, because the R^2 of the model (4-42) – restricted model- is 0.

EXAMPLE 4.11 Salaries of CEOs

Consider the following equation to explain salaries of Chief Executive Officers (CEOs) as a function of annual firm sales, return on equity (*roe*, in percent form), and return on the firm's stock (*ros*, in percent form):

$$\ln(salary) = \beta_1 + \beta_2 \ln(sales) + \beta_3 roe + \beta_4 ros + u.$$

The question posed is whether the performance of the company (*sales*, *roe* and *ros*) is crucial to set the salaries of CEOs. To answer this question, we will carry out an overall significance test. The null and alternative hypotheses are the following:

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0$$

H_1: H_0 is not true

Table 4.9 shows an E-views complete output for *least square* (*ls*) using the filework *ceosal1*. At the bottom the "F-statistic" can be seen for overall test significance, as well as "Prob", which is the *p*-value corresponding to this statistic. In this case the *p*-value is equal to 0, that is to say, H_0 is rejected for all significance levels (See figure 4.18). Therefore, we can reject that the performance of a company has no influence on the salary of a CEO.

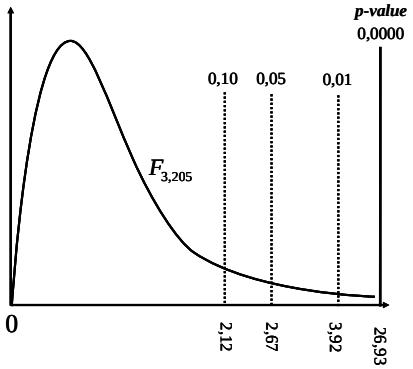


FIGURE 4.18. Example 4.11: *p*-value using *F* distribution (α values are for a $F_{3,140}$).

TABLE 4.9. Complete output from E-views in the example 4.11.

Dependent Variable: LOG(SALARY)
Method: Least Squares
Date: 04/12/12 Time: 19:39
Sample: 1 209
Included observations: 209

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.311712	0.315433	13.66919	0.0000
LOG(SALES)	0.280315	0.03532	7.936426	0.0000
ROE	0.017417	0.004092	4.255977	0.0000
ROS	0.000242	0.000542	0.446022	0.6561

R-squared	0.282685	Mean dependent var	6.950386
Adjusted R-squared	0.272188	S.D. dependent var	0.566374
S.E. of regression	0.483185	Akaike info criterion	1.402118
Sum squared resid	47.86082	Schwarz criterion	1.466086
Log likelihood	-142.5213	**F-statistic**	**26.9293**
Durbin-Watson stat	2.033496	**Prob(F-statistic)**	**0.0000**

4.3.3 Testing other linear restrictions

So far, we have tested hypotheses with exclusion restrictions using the *F* statistic. But we can also test hypotheses with linear restrictions of any kind. Thus, in the same test we can combine exclusion restrictions, restrictions that impose determined values to the parameters and restrictions on linear combination of parameters.

Therefore, let us consider the following model

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + u$$

and the null hypothesis:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_4 = 3 \\ \beta_5 = 0 \end{cases}$$

The restricted model corresponding to this null hypothesis is

$$(y - x_2 - 3x_4) = \beta_1 + \beta_3 (x_3 - x_2) + u$$

In the example 4.12, the null hypothesis consists of two restrictions: a linear combination of parameters and an exclusion restriction.

EXAMPLE 4.12 An additional restriction in the production function. (Continuation of example 4.7)

In the production function of Cobb-Douglas, we are going to test the following H_0 which has two restrictions:

$$H_0 : \begin{cases} \beta_2 + \beta_3 = 1 \\ \beta_1 = 0 \end{cases}$$

$H_1 : H_0$ is not true

In the first restriction we impose that there are constant returns to scale. In the second restriction that β_1, parameter linked to the total factor productivity is equal to 0.

Substituting the restriction of H_0 in the original model (*unrestricted model*), we have

$$\ln(output) = (1-\beta_3)\ln(labor) + \beta_3 \ln(capital) + u$$

Operating, we obtain the *restricted model*:

$$\ln(output / labor) = \beta_3 \ln(capital / labor) + u$$

In estimating the unrestricted and restricted models, we get RSS_R=3.1101 and RSS_{UR}=0.8516. Therefore, the *F ratio is*

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)} = \frac{(3.1101-0.8516)/2}{0.8516/(27-3)} = 13.551$$

There are two reasons for not using R^2 in this case. First, the restricted model has no intercept. Second, the regressand of the restricted model is different from the regressand of the unrestricted model.

Since the F value is relatively high, let us start by testing with a level of 1%. For α=0.01, $F_{2,24}^{0.01} = 5.61$. Given that F>5.61, we reject H_0 in favour of H_1. Therefore, we reject the joint hypotheses that there are constant returns to scale and that the parameter β_1 is equal to 0. If H_0 is rejected for α=0.01, it will also be rejected for levels of 5% and 10%.

4.3.4 Relation between *F* and *t* statistics

So far, we have seen how to use the F statistic to test several restrictions in the model, but it can be used to test a single restriction. In this case, we can choose between using the F statistic or the t statistic to carry out a two-tail test. The conclusions would, nevertheless, be exactly the same.

But, what is the relationship between an F with one degree of freedom in the numerator (to test a single restriction) and a t? It can be shown that

$$t_{n-k}^2 = F_{1,n-k} \qquad (4\text{-}46)$$

This fact is illustrated in figure 4.19. We observe that the tail of the F splits into the two tails of the t. Hence, the two approaches lead to exactly the same outcome, provided that the alternative hypothesis is two-sided. However, the t statistic is more flexible for testing a single hypothesis, because it can be used to test H_0 against one-tail alternatives.

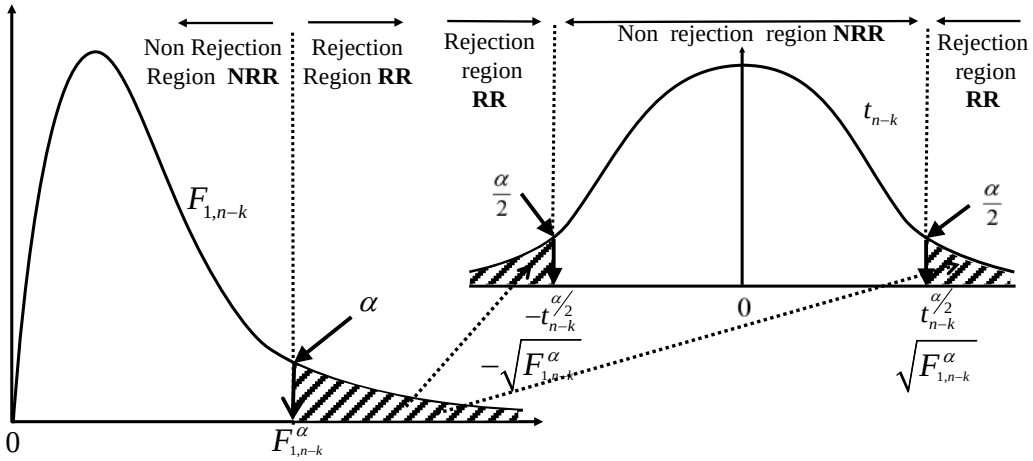


FIGURE 4.19. Relationship between a $F_{1,n-k}$ and a t_{n-k}.

Moreover, since the *t* statistics are also easier to obtain than the *F* statistics, there is no good reason for using an *F* statistic to test a hypothesis with a unique restriction.

4.4 Testing without normality

The normality of the *OLS* estimators depends crucially on the normality assumption of the disturbances. What happens if the disturbances do not have a normal distribution? We have seen that the disturbances under the Gauss-Markov assumptions, and consequently the *OLS* estimators are asymptotically normally distributed, i.e. approximately normally distributed.

If the disturbances are not normal, the *t* statistic will only have an *approximate t* distribution rather than an *exact* one. As it can be seen in the *t* student table, for a sample size of 60 observations the critical points are practically equal to the standard normal distribution.

Similarly, if the disturbances are not normal, the *F* statistic will only have an *approximate F* distribution rather than an *exact* one, when the sample size is large enough and the Gauss-Markov assumptions are fulfilled. Therefore, we can use the *F* statistic to test linear restrictions in linear models as an approximate test.

There are other asymptotic tests (the likelihood ratio, Lagrange multiplier and Wald tests) based on the likelihood functions that can be used in testing linear restriction if the disturbances are non-normally distributed. These three can also be applied when a) the restrictions are nonlinear; and b) the model is nonlinear in the parameters. For non-linear restrictions, in linear and non-linear models, the most widely used test is the Wald test.

For testing the assumptions of the model (for example, homoskedasticity and no autocorrelation) the Lagrange multiplier (*LM*) test is usually applied. In the application of the *LM* test, an *auxiliary regression* is often run. The name of auxiliary regression means that the coefficients are not of direct interest: only the R^2 is retained. In an auxiliary regression the regressand is usually the residuals (or functions of the residuals), obtained in the OLS estimation of the original model, while the regressors are often the regressors (and/or functions of them) of the original model.

4.5 Prediction

In this section two types of prediction will be examined: point and interval prediction.

4.5.1 Point prediction

Obtaining a point prediction does not pose any special problems, since it is a simple extrapolation operation in the context of descriptive methods.

Let $x_2^0, x_3^0, \cdots, x_k^0$ denote the particular values in each of the *k* regressors for prediction; these may or may not correspond to an actual data point in our sample. If we substitute these values in the multiple regression model, we have

$$y^0 = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 + u^0 = \theta^0 + u^0 \qquad (4\text{-}47)$$

Therefore, the expected, or mean, value of *y* is given by

$$E(y^0) = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \qquad (4\text{-}48)$$

The point prediction is obtained straightaway by replacing the parameters of (4-48) by the corresponding OLS estimators:

$$\hat{\theta}^0 = \hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0 \qquad (4\text{-}49)$$

To obtain (4-49) we did not need any assumption. But, if we adopt the assumptions 1 to 6, we will immediately find that that $\hat{\theta}^0$ is an unbiased predictor of θ^0:

$$E\left[\hat{\theta}^0\right] = E\left[\hat{\beta}_1 + \hat{\beta}_2 x_2^0 + \hat{\beta}_3 x_3^0 + \ldots + \hat{\beta}_k x_k^0\right] = \beta_1 + \beta_2 x_2^0 + \beta_3 x_3^0 + ... + \beta_k x_k^0 = \theta^0 \qquad (4\text{-}50)$$

On the other hand, adopting the Gauss Markov assumptions (1 to 8), it can be proved that this point predictor is the best linear unbiased estimator (BLUE).

We have a point prediction for θ^0, but, what is the point prediction for y^0? To answer this question, we have to predict u_0. As the error is not observable, the best predictor for u^0 is its expected value, which is 0. Therefore,

$$\hat{y}^0 = \hat{\theta}^0 \qquad (4\text{-}51)$$

4.5.2 Interval prediction

Point predictions made with an econometric model will in general not coincide with the observed values due to the uncertainty surrounding economic phenomena.

The first source of uncertainty is that we cannot use the population regression function because we do not know the parameters β's. Instead we have to use the sample regression function. The *confidence interval for the expected value* – i.e. for θ^0 - which will examine next, includes only this type of uncertainty.

The second source of uncertainty is that in an econometric model, in addition to the systematic part, there is a disturbance which is not observable. The *prediction interval for an individual value* – i.e. for y^0-, which will be discussed later on includes both the uncertainty arising from the estimation as well as the disturbance term.

A third source of uncertainty may come from the fact of not knowing exactly what values the explanatory variables will take for the prediction we want to make. This third source of uncertainty, which is not addressed here, complicates calculations for the construction of intervals.

Confidence interval for the expected value

If we are predicting the expected value of *y*, which is θ^0, then the prediction error $\hat{e}_1^0$ will be $\hat{e}_1^0 = \theta^0 - \hat{\theta}^0$. According to (4-50), the expected prediction error is zero. Under the assumptions of the *CLM*,

$$\frac{\hat{e}_1^0}{se(\hat{\theta}^0)} = \frac{\theta^0 - \hat{\theta}^0}{se(\hat{\theta}^0)} : \ t_{n-k}$$

Therefore, we can write that

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\theta^0 - \hat{\theta}^0}{se(\hat{\theta}^0)} \le t_{n-k}^{\alpha/2}\right] = 1-\alpha$$

Operating, we can construct a (1-α)% *confidence interval* (*CI*) for θ^0 with the following structure:

$$\Pr\left[\hat{\theta}^0 - se(\hat{\theta}^0)\times t_{n-k}^{\alpha/2} \le \theta^0 \le \hat{\theta}^0 + se(\hat{\theta}^0)\times t_{n-k}^{\alpha/2}\right] = 1-\alpha \quad (4\text{-}52)$$

To obtain a *CI* for θ^0, we need to know the standard error ($se(\hat{\theta}_0)$) for $\hat{\theta}^0$. In any case, there is an easy way to calculate it. Thus, solving (4-48) for β_1 we find that $\beta_1 = \theta^0 - \beta_2 x_2^0 - \beta_3 x_3^0 - ... - \beta_k x_k^0$. Plugging this into the equation (4-47), we obtain

$$y = \theta^0 + \beta_2(x_2 - x_2^0) + \beta_3(x_3 - x_3^0) + \ldots + \beta_k(x_k - x_k^0) + u \quad (4\text{-}53)$$

Applying OLS to (4-53), in addition to the point prediction, we obtain $se(\hat{\theta}^0)$ which is the standard error corresponding to the *intercept* in this regression. The previous method allows us to put a *CI* around the OLS estimate of *E*(*y*), for any values of the *x´*s.

Prediction interval for an individual value

We are now going to construct an interval for y^0, usually called *prediction interval for an individual value,* or for short, *prediction interval*. According to (4-47), y^0 has two components:

$$y^0 = \theta^0 + u^0 \quad (4\text{-}54)$$

The *interval for the expected value* built before is a confidence interval around θ^0 wcich is a combination of the parameters. In contrast, the interval for y^0 is random, because one of its components, u^0, is random. Therefore, the interval for y^0 is a probabilistic interval and not a confidence interval. The mechanics for obtaining it are the same, but bear in mind that now we are going to consider that the set $x_2^0, x_3^0, \ldots, x_k^0$ vis outside from of the sample used to estimate the regression.

The *prediction error* ($\hat{e}_2^0$) in using $\hat{y}^0$ to predict y^0 is

$$\hat{e}_2^0 = y^0 - \hat{y}^0 = \theta^0 + u^0 - \hat{y}^0 \quad (4\text{-}55)$$

Taking into account (4-51) and (4-50), and that $E(u^0)=0$, then the expected prediction error is zero. In finding the variance of $\hat{e}_2^0$, it must be taken into account that u^0 is uncorrelated with $\hat{y}^0$ because $x_2^0, x_3^0, \ldots, x_k^0$ is not in the sample.

Therefore, the *variance of the prediction error* (conditional on the *x´*s) is the sum of the variances:

$$Var(\hat{e}_2^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2 \quad (4\text{-}56)$$

There are two sources of variation in $\hat{e}_2^0$:

1. The sampling error in $\hat{y}^0$, which arises because we have estimated the β_j*'s*.

2. The ignorance of the unobserved factors that affect *y*, which is reflected in σ^2.

Under the *CLM* assumptions, $\hat{e}_2^0$ is also normally distributed. Using the unbiased estimator of σ^2 and taking into account that $var(\hat{y}^0) = var(\hat{\theta}^0)$, we can define the standard error (*se*) of $\hat{e}_2^0$ as

$$se(\hat{e}_2^0) = \left\{ \left[se(\hat{\theta}^0) \right]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}} \qquad (4\text{-}57)$$

Usually $\hat{\sigma}^2$ is larger than $\left[se(\hat{\theta}^0) \right]^2$. Under the assumptions of the *CLM*,

$$\frac{\hat{e}_2^0}{se(\hat{e}_2^0)} : \ t_{n-k} \qquad (4\text{-}58)$$

Therefore, we can write that

$$\Pr\left[-t_{n-k}^{\alpha/2} \le \frac{\hat{e}_2^0}{se(\hat{e}_2^0)} \le t_{n-k}^{\alpha/2} \right] = 1-\alpha \qquad (4\text{-}59)$$

Plugging in $\hat{e}_2^0 = y^0 - \hat{y}^0$ into (4-59) and rearranging it gives a (1-α)% *prediction interval* for y^0:

$$\Pr\left[\hat{y}^0 - se(\hat{e}_2^0) \times t_{n-k}^{\alpha/2} \le y^0 \le \hat{y}^0 + se(\hat{e}_2^0) \times t_{n-k}^{\alpha/2} \right] = 1-\alpha \qquad (4\text{-}60)$$

EXAMPLE 4. 13 What is the expected score in the final exam with 7 marks in the first short exam?

The following model has been estimated to compare the marks in the final exam (*finalmrk*) and in the first short exam (*shortex1*) of Econometrics:

$$\widehat{finalmrk}_i = \underset{(0.715)}{4.155} + \underset{(0.123)}{0.491} shortex1_i$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

To estimate the expected final mark for a student with $shortex1^0$=7 mark in the first short exam, the following model, according to (4-53), was estimated:

$$\widehat{finalmrk}_i = \underset{(0.497)}{7.593} + \underset{(0.123)}{0.491}(shortex1_i - 7)$$

$$\hat{\sigma} = 1.649 \quad R^2 = 0.533 \quad n = 16$$

The point prediction for $shortex1^0$=7 is $\hat{\theta}_0$ =7.593 and the lower and upper bounds of a 95% *CI* respectively are given by

$$\underline{\theta}^0 = \hat{\theta}^0 - se(\hat{\theta}^0) \times t_{14}^{0.05/2} = 7.593 - 0.497 \times 2.14 = 6.5$$

$$\overline{\theta}^0 = \hat{\theta}^0 + se(\hat{\theta}^0) \times t_{14}^{0.05/2} = 7.593 + 0.497 \times 2.14 = 8.7$$

Therefore, the student will have a 95% confidence of obtaining on average a final mark located between 6.5 and 8.7.

The point prediction could be also obtained from the first estimated equation:

$$\widehat{finalmrk} = 4.155 + 0.491 \times 7 = 7.593$$

Now, we are going to estimate a 95% probability interval for the individual value. The *se* of $\hat{e}_2^0$ is equal

$$se(\hat{e}_2^0) = \left\{ \left[se(\hat{y}^0) \right]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}} = \sqrt{0.497^2 + 1.649^2} = 1.722$$

where 1.649 is the "S. E. of regression" obtained from the E-views output directly.

The lower and upper bounds of a 95% *probability interval* respectively are given by

$$\underline{y}^0 = \hat{y}^0 - se(\hat{e}_2^0) \times t_{14}^{0.025} = 7.593 - 1.722 \times 2.14 = 3.7$$

$$\overline{y}^0 = \hat{y}^0 + se(\hat{e}_2^0) \times t_{14}^{0.025} = 7.593 + 1.722 \times 2.14 = 11.3$$

You must take into account that this *probability interval* is quite large because the size of the sample is very small.

EXAMPLE 4.14 Predicting the salary of CEOs

Using data on the most important US companies taken from Forbes (workfile *ceoforbes*), the following equation has been estimated to explain salaries (including bonuses) earned yearly (thousands of dollars) in 1999 by the CEOs of these companies:

$$\widehat{salary}_i = \underset{(104)}{1381} + \underset{(0.0013)}{0.008377}\, assets_i + \underset{(8.671)}{32.508}\, tenure_i + \underset{(0.0538)}{0.2352}\, profits_i$$

$$\hat{\sigma} = 1506 \quad R^2 = 0.2404 \quad n = 447$$

where *assets* are total assets of firm in millions of dollars, *tenure* is number of years as CEO in the company, and *profits* are in millions of dollars.

In Table 4.10 descriptive measures of explanatory variables of the model on CEOs salaries appear.

TABLE 4.10. Descriptive measures of variables of the model on CEOs salary.

	assets	*tenure*	*profits*
Mean	27054	7.8	700
Median	7811	5.0	333
Maximum	668641	60.0	22071
Minimum	718	0.0	-2669
Observations	447	447	447

The predicted salaries and the corresponding *se*($\hat{\theta}_0$) for selected values (maximum, mean, median and minimum), using a model as (4-53), appear in table 4.11.

TABLE 4.11. Predictions for selected values.

	Prediction $\hat{\theta}_0$	*Std. Error se*($\hat{\theta}_0$)
Mean values	2026	71
Median value	1688	78
Maximum values	14124	1110
Minimum values	760	195

4.5.3 Predicting *y* in a ln(*y*) model

Consider the model in logs:

$$\ln(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u \qquad (4\text{-}61)$$

Obtaining OLS estimates, we predict ln(*y*) as

$$\widehat{\ln}(y) = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \qquad (4\text{-}62)$$

Applying exponentiation to (4-62), we obtain the prediction value

$$\hat{y} = \exp(\widehat{\ln}(y)) = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k) \qquad (4\text{-}63)$$

However, this prediction is biased and inconsistent because it will systematically *underestimate* the expected value of *y*. Let us see why. If we apply exponentiation in (4-61), we have

$$y = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(u) \qquad (4\text{-}64)$$

Before taking expectation in (4-64), we must take into account that if $u \sim N(0,\sigma^2)$, then $E(\exp(u)) = \exp\left(\frac{\sigma^2}{2}\right)$. Therefore, under the *CLM* assumptions 1 through 9, we have

$$E(y) = \exp(\beta_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k) \times \exp(\sigma^2/2) \qquad (4\text{-}65)$$

Taking as a reference (4-65), the adequate predictor of *y* is

$$\hat{y} = \exp(\hat{b}_1 + \hat{b}_2 x_2 + \mathrm{L} + \hat{b}_k x_k)' \ \exp(\hat{s}^2/2) = \% \ \exp(\hat{s}^2/2) \qquad (4\text{-}66)$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2.

It is important to remark that although $\hat{y}$ is a biased predictor, it is consistent, while $\%$ is biased and inconsistent

EXAMPLE 4.15 Predicting the salary of CEOs with a log model (continuation 4.14)

Using the same data as in example 4.14, the following model was estimated:

$$\widehat{\ln(salary_i)} = \underset{(0.210)}{5.5168} + \underset{(0.0232)}{0.1885}\ln(assets_i) + \underset{(0.0032)}{0.0125}\,tenure_i + \underset{(0.0000195)}{0.00007}\,profits_i$$

$$\hat{\sigma} = 0.5499 \quad R^2 = 0.2608 \quad n = 447$$

salary and *assets* are taken in natural logs, while *profits* are in levels because some observations are negative and thus not possible to take logs.

First, we are going to calculate the inconsistent prediction, according to (4-63) for a CEO working in a corporation with *assets*=10000, *tenure*=10 years and *profits*=1000:

$$\widehat{salary}_i = \exp(\widehat{\ln(salary_i)})$$

$$= \exp(5.5168 + 0.1885\ln(10000) + 0.0125' \ 10 + 0.00007' \ 1000) = 1716$$

Using (4-66), we obtain a consistent prediction:

$$\widehat{salary} = \exp(0.5499^2/2)' \ 1716 = 1996$$

4.5.4 Forecast evaluation and dynamic prediction

In this section we will compare predictions made using an econometric model with the actual values in order to evaluate the predictive ability of the model. We will also examine the dynamic prediction in models in which lagged endogenous variables are included as regressors.

Forecast evaluation statistics

Suppose that the sample forecast is $i=n+1, n+2,\ldots, n+h$, and denote the actual and forecasted value in period *i* as y_i and $\hat{y}_i$, respectively. Now, we present some of the more common statistics used for forecast evaluation.

Mean absolute error (MAE)

The *MAE* is defined as the average of the absolute values of the errors:

$$MAE = \frac{\sum_{i=n+1}^{n+h} |\hat{y}_i - y_i|}{h} \qquad (4\text{-}67)$$

Absolute values are taken so that positive errors are compensated by the negative ones.

Mean absolute percentage error (MAPE),

$$MAPE = \frac{\sum_{i=n+1}^{n+h} \frac{|\hat{y}_i - y_i|}{y_i}}{h} \cdot 100 \qquad (4\text{-}68)$$

Root of the mean squared error (RMSE)

This statistic is defined as the square root of the mean of the squared error:

$$RMSE = \sqrt{\frac{\sum_{i=n+1}^{n+h} (\hat{y}_i - y_i)^2}{h}} \qquad (4\text{-}69)$$

As the errors are squared, the compensation between positive and negative errors are avoided. It is important to remark that the *MSE* places a greater penalty on large forecast errors than the *MAE*.

Theil Inequality Coefficient (U)

This coefficient is defined as follows:

$$U = \frac{\sqrt{\frac{\sum_{i=n+1}^{n+h} (\hat{y}_i - y_i)^2}{h}}}{\sqrt{\frac{\sum_{i=n+1}^{n+h} \hat{y}_i^2}{h}} + \sqrt{\frac{\sum_{i=n+1}^{n+h} y_i^2}{h}}} \qquad (4\text{-}70)$$

The smaller *U* is, the more accurate are the predictions. The scaling of *U* is such that it will always lie between 0 and 1. If *U*=0, then $y_i = \hat{y}_i$, for all forecasts; if *U*=1 the predictive performance is as bad as it can be. Theil's *U* statistic can be rescaled and decomposed into three proportions: bias, variance and covariance. Of course the sum of these three proportions is 1. The interpretation of these three proportions is as follows:

1) The *bias* reflects systematic errors. Whatever the value of *U*, we would hope that the bias is close to 0. A large bias suggests a systematic over or under prediction.
2) The *variance* also reflects systematic errors. The size of this proportion is an indication of the inability of the forecasts to replicate the variability of the variable to be forecasted.

3) The *covariance* measures unsystematic errors. Ideally, this should have the highest proportion of Theil inequality.

In addition of the coefficient defined in (4-70), Theil proposed other coefficients for forecast evaluation.

Dynamic prediction

Let the following model be given:

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + u_t \qquad (4\text{-}71)$$

Suppose that the sample forecast is $i=n+1,\ldots,i=n+h$, and denote the actual and forecasted value in period i as y_i and $\hat{y}_i$, respectively. The forecast for the period $n+1$ is

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1} + \hat{\beta}_3 y_n \qquad (4\text{-}72)$$

As we can see for the prediction, we use the observed value of y (y_n) because it is inside the sample used in the estimation. For the remainder of the forecast periods we use the recursively computed forecast of the lagged value of the dependent variable (dynamic prediction), that is to say,

$$\hat{y}_{n+i} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+i} + \hat{\beta}_3 \hat{y}_{n-1+i} \qquad i = 2,3,\cdots,h \qquad (4\text{-}73)$$

Thus, from period $n+2$ to $n+h$ the forecast carried out in a period is used to forecast the endogenous variable in the following period.

Exercises

Exercise 4.1 To explain the housing price in an American town, the following model is formulated:

$$price = \beta_1 + \beta_2 rooms + \beta_3 lowstat + \beta_4 crime + u$$

where *rooms* is the number of rooms in the house, *lowstat* is the percentage of people of "lower status" in the area and *crime* is crimes committed per capita in the area. Prices of houses are measured in dollars.

Using the data in *hprice2*, the following model has been estimated:

$$\widehat{price} = \underset{(8022)}{-15694} + \underset{(1211)}{6788}\,rooms - \underset{(81)}{268}\,lowstat - \underset{(960)}{3854}\,crime$$

$$R^2=0.771 \qquad n=55$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the meaning of the coefficients $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$.

b) Does the percentage of people of "lower status" have a negative influence on the price of houses in that area?

c) Does the number of rooms have a positive influence on the price of houses?

Exercise 4.2 Consider the following model:

$$\ln(fruit) = \beta_1 + \beta_2 \ln(inc) + \beta_3 hhsize + \beta_4 punder5 + u$$

where *fruit* is expenditure in fruit, *inc* is disposable income of a household, *hhsize* is the number of household members and *punder5* is the proportion of children under five in the household.

Using the data in workfile *demand*, the following model has been estimated:

$$\widehat{\ln(fruit)} = \underset{}{-} \underset{(3.701)}{9.768} + \underset{(0.512)}{2.005}\ln(inc) - \underset{(0.179)}{1.205}\,hhsize - \underset{(0.013)}{0.0179}\,punder5$$

$$R^2=0.728 \qquad n=40$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the meaning of the coefficients $\hat{\beta}_2$, $\hat{\beta}_3$ and $\hat{\beta}_4$.

b) Does the number of household members have a statistically significant effect on the expenditure in fruit?

c) Is the proportion of children under five in the household a factor that has a negative influence on the expenditure of fruit?

d) Is fruit a luxury good?

Exercise 4.3 (Continuation of exercise 2.5). Given the model

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad \mathrm{i} = 1, 2, \ldots, n$$

the following results have been obtained with a sample size of 11 observations:

$$\sum_{i=1}^{n} x_i = 0 \qquad \sum_{i=1}^{n} y_i = 0 \qquad \sum_{i=1}^{n} x_i^2 = B \qquad \sum_{i=1}^{n} y_i^2 = E \qquad \sum_{i=1}^{n} x_i y_i = F$$

(Remember that $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i}$)

a) Build a statistic to test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$.

b) Test the hypothesis of question *a)* when $EB = 2F^2$.

c) Test the hypothesis of question *a)* when $EB = F^2$.

Exercise 4.4 The following model has been formulated to explain the spending on food (*food*):

$$food = \beta_1 + \beta_2 inc + \beta_3 rpfood + u$$

where *inc* is disposable income and *rpfood* is the relative price index of food compared to other consumer products.

Taking a sample of observations for 20 successive years, the following results are obtained:

$$\widehat{food}_i = \underset{}{1.40} + \underset{(4.92)}{0.126}\,inc_i - \underset{(0.01)}{0.036}\,rpfood_i \qquad \underset{(0.07)}{}$$

$$R^2=0.996; \qquad \sum \hat{u}_t^2 = 0.196$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the null hypothesis that the coefficient of *rpfood* is less than 0.

b) Obtain a confidence interval of 95% for the marginal propensity to consume food in relation to income.

c) Test the joint significance of the model.

Exercise 4.5 The following demand function for rental housing is formulated:

$$\ln(srenhous_i)=\beta_1+\beta_2\ln(prenhous_i)+\beta_3\ln(inc_i)+\varepsilon_i$$

where *srenhous* is spending on rental housing, *prenhous* is the rental price, and *inc* is disposable income.

Using a sample of 403 observations, we obtain the following results:

$$\ln(srenhous_i) = 10 - 0.7\ln(prenhous_i) + 0.9\ln(inc_i)$$

$$R^2=0.39 \qquad \text{cov}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 1.0 & 0 & 0 \\ 0 & 0.09 & 0.085 \\ 0 & 0.085 & 0.09 \end{bmatrix}$$

a) Interpret the coefficients on ln(*prenhous*) and ln(*inc*).
b) Using a 0.01 significance level, test the null hypothesis that $\beta_2=\beta_3=0$.
c) Test the null hypothesis that $\beta_2=0$, against the alternative that $\beta_2<0$.
d) Test the null hypothesis that $\beta_3=1$ against the alternative that $\beta_3 \neq 1$.
e) Test the null hypothesis that a simultaneous increase in housing prices and income has no proportional effect on housing demand.

Exercise 4.6 The following estimated models corresponding to average cost (*ac*) functions have been obtained, using a sample of 30 firms:

$$\widehat{ac}_i = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, qty_i$$

$$R^2 = 0.838 \qquad RSS = 8090 \tag{1}$$

$$\widehat{ac}_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, qty_i + \underset{(11.61)}{26.73}\, qty_i^2 - \underset{(1.22)}{1.40}\, qty_i^3$$

$$R^2 = 0.978 \quad RSS = 1097 \tag{2}$$

where *ac* is the average cost and *qty* is the quantity produced.

(The numbers in parentheses are standard errors of estimators.)

a) Test whether the quadratic and cubic terms of the quantity produced are significant in determining the average cost.
b) Test the overall significance in the model 2.

Exercise 4.7 Using a sample of 35 observations, the following models have been estimated to explain expenditures on coffee:

$$\widehat{\ln(coffee)} = 21.32 + \underset{(0.01)}{0.11}\ln(inc) - \underset{(0.23)}{1.33}\ln(cprice) + 1.35\ln(tprice) \tag{1}$$

$$R^2 = 0.905 \qquad RSS = 254$$

$$\widehat{\ln(coffee)} = 19.9 + \underset{(0.02)}{0.14}\ln(inc) - \underset{(0.21)}{1.42}\ln(cprice) \tag{2}$$

$$RSS = 529$$

where *inc* is disposable income, *cprice* is coffee price and *tprice* is tea price.

(The numbers in parentheses are standard errors of estimators.)

a) Test the overall significance of model (1)
b) The standard error of ln(*tprice*) is missing in model (1), can you calculate it?
c) Test whether the price of tea is statistically significant.
d) How would you test the assumption that the price elasticity of coffee is equal but opposite to the price elasticity of tea? Detail the procedure.

Exercise 4.8 The following model has been formulated to analyse the determinants of air quality (*airqual*) in 30 Standard Metropolitan Statistical Areas (SMSA) of California:

$$airqual = \beta_1 + \beta_2 popln + \beta_3 medincm + \beta_4 poverty + \beta_5 fueoil + \beta_6 valadd + u$$

where *airqual* is weight in μg/m^3 of suspended particular matter, *popln* is population in thousands, *medincm* is medium per capita income in dollars, *poverty* is the percentage of families with income less than poverty levels, *fueloil* is thousands of barrels of fuel oil consumed in industrial manufacturing, and *valadd* is value added by industrial manufactures in 1972 in thousands of dollars.

Using the data in workfile *airqualy*, the above model has been estimated:

$$\widehat{airqual}_i = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$
$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i$$
$$R^2 = 0.415 \qquad n = 30$$

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the coefficients on *medincm*, *poverty* and *valadd*
b) Are the slope coefficients individually significant at 10%?
c) Test the joint significance of *fueloil* and *valadd*, knowing that

$$\widehat{airqual}_i = \underset{(10.41)}{97.67} + \underset{(0.020)}{0.0566}\, popln_i - \underset{(0.0039)}{0.0102}\, medincm_i - \underset{(0.0078)}{0.0174}\, poverty_i$$
$$R^2 = 0.339 \qquad n = 30$$

d) If you omit the variable *poverty* in the first model, the following results are obtained:

$$\widehat{airqual}_i = \underset{(10.02)}{82.98}_i + \underset{(0.031)}{0.0523}\, popln_i - \underset{(0.0055)}{0.0097}\, medincm_i$$
$$- \underset{(0.0017)}{0.00063}\, fueoil_i - \underset{(0.0028)}{0.00037} valadd_i$$
$$R^2 = 0.218 \qquad n = 30$$

Are the slope coefficients individually significant at 10% in the new model? Do you consider these results to be reasonable in comparison with those obtained in part *b*).
Comparing the R^2 of the two estimated models, what is the role played by *poverty* in determining air quality?

e) If you regress *airqual* using as regressors only the intercept and *poverty*, you will obtain that R^2=0.037. Do you consider this value to be reasonable taking into account the results obtained in part *d*)?

Exercise 4.9 With a sample of 39 observations, the following production functions by *OLS* was estimated:

$$\bar{output}_t = \hat{a}\, labor_t^{1.30} capital_t^{0.32} \exp(0.0055 trend_t) \qquad R^2 = 0.9945$$

$$\bar{output}_t = \hat{b}\, labor_t^{1.41} capital_t^{0.47} \qquad R^2 = 0.9937$$

$$\bar{output}_i = \hat{g} \exp(0.0055 trend_t) \qquad R^2 = 0.9549$$

a) Test the joint significance of *labor* and *capital*.
b) Test the significance of the coefficient of the variable *trend*.
c) Identify the statistical assumptions under which the test carried out in the two previous sections are correct. A further question: Specify the population model of the first of the three previous specifications.

Exercise 4.10 A researcher has developed the following model:

$$y = \beta_1 + \beta_2\ x_2 + \beta_3 x_3 + u$$

Using a sample of 43 observations, the following results were obtained:

$$\hat{y}_i = -\ 0.06 + 1.44\ x_{1i} - 0.48\ x_{2i}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1011 & -0.0007 & -0.0005 \\ & 0.0231 & -0.0162 \\ & & 0.0122 \end{bmatrix}$$

$$\sum y_i^2 = 444 \qquad \sum \hat{y}_i^2 = 424.92$$

a) Test that the intercept is less than 0.
b) Test that β_2=2.
c) Test the null hypothesis that $\beta_2+3\beta_3=0$.

Exercise 4.11 Given the function of production

$$q = ak^{\alpha} l^{\beta} \exp(u)$$

and using data from the Spanish economy over the past 20 years, the following results were obtained:

$$\bar{\ln}(q_i) = 0.15 + 0.73 \ln(k_i) + 0.47 \ln(l_i)$$

$$\left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4129 & -95 & -266 \\ -95 & 3 & 5 \\ -266 & 5 & 19 \end{bmatrix} \qquad RSS = 0.017$$

a) Test the individual significance of the coefficients on *k* and *l*.
b) Test whether the parameter α is significantly different from 1.
c) Test whether there are increasing returns to scale.

Exercise 4.12 Let the following multiple regression model be:

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + u$$

With a sample of 33 observations, this model is estimated by *OLS*, obtaining the following results:

$$\hat{y}_i = 12.7 + 14.2x_{1i} + 2.1x_{2i}$$

$$\hat{\sigma}^2 \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \begin{bmatrix} 4.1 & -0.95 & -0.266 \\ -0.95 & 3.8 & 0.5 \\ -0.266 & 0.5 & 1.9 \end{bmatrix}$$

a) Test the null hypothesis $\alpha_0 = \alpha_1$.

b) Test whether $\alpha_1 / \alpha_2 = 7$.

c) Are the coefficients α_0, α_1, y α_2 individually significant?

Exercise 4.13 Using a sample of 30 companies, the following cost functions have been estimated:

a) $\hat{cost}_i = \underset{(11.97)}{172.46} + \underset{(3.70)}{35.72}\, x_i \qquad R^2 = 0.838 \quad \bar{R}^2 = 0.829 \quad RSS = 8090$

b) $\hat{cost}_i = \underset{(29.44)}{310.07} - \underset{(33.81)}{85.39}\, x_i + \underset{(11.61)}{26.73}\, x_i^2 - \underset{(1.22)}{1.40}\, x_i^3 \quad R^2 = 0.978 \quad \bar{R}^2 = 0.974 \quad RSS = 1097$

where *cost* is the average cost and *x* is the quantity produced.

(The numbers in parentheses are standard errors of estimators.)

a) Which of the two models would you choose? What would be the criteria?

b) Test whether the quadratic and cubic terms of the quantity produced are significant in determining the average cost.

c) Test the overall significance of the model *b).*

Exercise 4.14 A researcher formulates the following model:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

Using a sample of 13 observations the following results are obtained:

$$\hat{y}_i = 1.00 - 1.82x_{2i} + 0.36x_{3i} \qquad (1)$$

$$R^2 = 0.50 \qquad n = 13$$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.25 & -0.01 & 0.04 \\ -0.01 & 0.16 & -0.15 \\ 0.04 & -0.15 & 0.81 \end{bmatrix}$$

a) Test the null hypothesis that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 < 0$.

b) Test the null hypothesis that $\beta_2 + \beta_3 = -1$ against the alternative hypothesis that $\beta_2 + \beta_3 \neq -1$, with a significance level of 5%.

c) Is the whole model significant?

d) Assuming that the variables in the estimated model are measured in natural logarithms, what is the interpretation of the coefficient for x_3?

Exercise 4.15 With a sample of 50 automotive companies the following production functions were estimated taking the gross value added of the automobile production (*gva*) as the endogenous variable and labor input (*labor*) and capital input (*capital*) as explanatory variables.

1) $$\widehat{\ln(gva_i)} = \underset{}{3.87} + \underset{(0.11)}{0.80}\ln(labor_i) + \underset{(0.24)}{1.24}\ln(capital_i), \quad RSS = 254 \quad R^2 = 0.75 \quad \bar{R}^2 = 0.72$$

2) $$\widehat{\ln(gva_i)} = 19.9 + 1.04\ln(capital_i), \quad RSS = 529 \quad R^2 = 0.84, \bar{R}^2 = 0.81$$

3) $$\widehat{\ln(gva / labor_i)} = 15.2 + 0.87\ln(capital_i / labor_i), \quad RSS = 380$$

(The numbers in parentheses are standard errors of estimators.)

a) Test the joint significance of both factors in the production function.
b) Test whether labor has a significant positive influence on the gross value added of automobile production.
c) Test the hypothesis of constant returns to scale. Explain your answer.

Exercise 4.16 With a sample of 35 annual observations two demand functions of Rioja wine have been estimated. The endogenous variable is spending on Rioja reserve wine (*wine*) and the explanatory variables are disposable income (*inc*), the average price of a bottle of Rioja reserve wine (*pwinrioj*) and the average price of a bottle of Ribera Duero reserve wine (*pwinduer*). The results are as follows:

$$\widehat{\ln(vino_i)} = \underset{}{21.32} + \underset{(0.01)}{0.11}\ln(renta_i) - \underset{(0.23)}{1.33}\ln(pvinrioj_i) + \underset{(0.233)}{1.35}\ln(pvinduer_i)$$

$$R^2 = 0.905 \quad RSS = 254$$

$$\widehat{\ln(vino_i)} = 19.9 + \underset{(0.02)}{0.14}\ln(renta_i) - \underset{(0.21)}{1.42}\ln(pvinrioj_i)$$

$$RSS = 529$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the joint significance of the first model.
b) Test whether the price of wine from Ribera del Duero has a significant influence, using two statistics that do not use the same information. Show that both procedures are equivalent.
c) How would you test the hypothesis that the price elasticity of Rioja wine is the same but with an opposite sign to the price elasticity of Ribera del Duero wine? Detail the procedure to follow.

Exercise 4.17 To analyze the demand for Ceylon tea (*teceil*) the following econometric model is formulated:

$$\ln(teceil) = \beta_1 + \beta_2\ln(inc) + \beta_3\ln(pteceil) + \beta_4\ln(pteind) + \beta_5\ln(pcobras) + u$$

where *inc* is the disposable income, *pteceil* the price of tea in Ceylon, *pteind* is the price of tea in India and *pcobras* is the price of Brazilian coffee.

With a sample of 22 observations the following estimates were made:

$$\widehat{\ln(teceil_i)} = \underset{}{2.83} + \underset{(0.17)}{0.25}\ln(inc_i) - \underset{(0.98)}{1.48}\ln(pteceil_i) + \underset{(0.69)}{1.18}\ln(pteind_i) + \underset{(0.16)}{0.19}\ln(pcofbras_i)$$

$$RSS=0.4277$$

$$\ln(\overline{teceil}_i \acute{} \; pteceil) = 0.74 + \underset{(0.16)}{0.26}\ln(inc_i) + \underset{(0.15)}{0.20}\ln(pcofbras_i)$$

$$RSS=0.6788$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the significance of disposable income.

b) Test the hypothesis that $\beta_3 = -1$ y $\beta_4 = 0$, and explain the procedure applied.

c) If instead of having information on *RSS*, only R^2 was known for each model, how would you proceed to test the hypothesis of section *b)*?

Exercise 4.18 The following fitted models are obtained to explain the deaths of children under 5 years per 1000 live births (*deathu5*) using a sample of 64 countries.

1) $\widehat{deathun5}_i = 263.64 - \underset{(0.0019)}{0.0056}\,inc_i + \underset{(0.21)}{2.23}\,fertrate_i$; $R^2 = 0.7077$

2) $\widehat{deathun5}_i = 168.31 - \underset{(0.0018)}{0.0055}\,inc_i + \underset{(0.25)}{1.76}\,femilrat_i + 12.87\,fertrate_i, R^2 = 0.7474$

where *inc* is income per capita, *femiltrat* is the female illiteracy rate, and *fertrate* is the fertility rate

(The numbers in parentheses are standard errors of the estimators.)

a) Test the joint significance of income, illiteracy and fertility rates.

b) Test the significance of the fertility rate.

c) Which of the two models would you choose? Explain your answer.

Exercise 4.19 Using a sample of 32 annual observations, the following estimations were obtained to explain the car sales (*car*) of a particular brand:

$$\dot{car}_i = \underset{(6.48)}{104.8} - \underset{(3.19)}{6.64}\,pcar_i + \underset{(0.16)}{2.98}\,adv_i$$

$$\sum \hat{u}_i^2 = 1805.2; \qquad \sum (car_i - \overline{car})^2 = 13581.4$$

where *pcar* is the price of cars and *adv* are spending on advertising.

(The numbers in parentheses are standard errors of the estimators.)

a) Are price and advertising expenditures significant together? Explain your answer.

b) Can you accept that prices have a negative influence on sales? Explain your answer.

c) Describe in detail how you would test the hypothesis that the impact of advertising expenditures on sales is greater than minus 0.4 times the impact of the price.

Exercise 4.20 In a study of the production costs (*cost*) of 62 coal mines, the following results are obtained:

$$\dot{cost}_i = \underset{(3.4)}{2.20} - \underset{(0.005)}{0.104}\,dmec_i + \underset{(2.2)}{3.48}\,geodif_i + \underset{(0.15)}{0.104}\,absent_i$$

$$\sum \left[cp_i - \overline{cp} \right]^2 = 109.6 \qquad \sum \hat{u}_i^2 = 18.48$$

where *dmec* is the degree of mechanization, *geodif* is a measurement of geological difficulties and *absent* is the percentage of absenteeism.

a) Test the significance of each of the model coefficients.
b) Test the overall significance of the model.

Exercise 4.21 With fifteen observations, the following estimation was obtained:

$$\hat{y}_i = 8.04 - \underset{(1.00)}{2.46}\, x_{i2} + \underset{(0.60)}{0.23}\, x_{i3}$$

$$\overline{R}^2 = 0.30$$

where the values between parentheses are standard deviations and the coefficient of determination is the adjusted one.

a) Is the coefficient of the variable x_2 significant?
b) Is the coefficient of the variable x_3 significant?
c) Discuss the joint significance of the model.

Exercise 4.22 Consider the following econometric specification:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

With a sample of 26 observations, the following estimations were obtained:

1) $\hat{y}_i = 2 + \underset{(1.9)}{3.5}\, x_{1i} - \underset{(2.2)}{0.7}\, x_{2i} - \underset{(1.5)}{2}\, x_{3i} + u_i$ $\qquad R^2$=0.982

2) $\hat{y}_i = 1.5 + \underset{(2.7)}{3}\,(x_{1i} + x_{2i}) - \underset{(2.4)}{0.6}\, x_{3i} + u_i$ $\qquad R^2$= 0.876

(The *t* statistics are between brackets)

a) Show that the following expressions for the *F*-statistic are equivalent:

$$F = \frac{\left(RSS_R - RSS_{UR}\right)/r}{RSS_{UR}/(n-k)} \qquad F = \frac{\left(R^2_{UR} - R^2_R\right)/q}{(1-R^2_{UR})/(n-k)}$$

b) Test the null hypothesis $\beta_2 = \beta_3$.

Exercise 4.23 In the estimation of the Brown model in exercise 3.19, using the workfile *consumsp*, we obtained the following results:

$$\bar{c}onspc_t = -\underset{(84.88)}{7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

R^2=0.997 $\quad$ *RSS*=1891320 $\quad$ *n*=56

Two additional estimations are now obtained:

$$\overline{consp}c_t - conspc_{t-1} = -\underset{(84.43)}{98.13} + \underset{(0.0803)}{0.2757}(incpc_t - conspc_{t-1})$$

R^2=0.1792 $\quad$ *RSS*=2199474 $\quad$ *n*=56

$$\overline{consp}c_t - incpc_{t-1} = -\underset{(84.88)}{7.156} - \underset{(0.0090)}{0.0264}\, incpc + \underset{(0.0903)}{0.5771}(conspc_{t-1} - incpc_t)$$

R^2=0.6570 $\quad$ *RSS*=1891320 $\quad$ *n*=56

(The numbers in parentheses are standard errors of the estimators.)

a) Test the significance of each of the coefficients for the first model.
b) Test that the coefficient on *incpc* in the first model is smaller than 0.5.
c) Test the overall significance of the first model.

d) Is it admissible that $\beta_2 + \beta_3 = 1$?

e) Show that by operating in the third model you can reach the same coefficients as in the first model.

Exercise 4.24 The following model was formulated to analyze the determinants of the median base salary in $ for graduating classes of 2010 from the best American business schools (*salMBAgr*):

$$salMBAgr = \beta_1 + \beta_2 tuition + \beta_3 salMBApr + u$$

where *tuition* is tuition fees including all required fees for the entire program (but excluding living expenses) and *salMBApr* is the median annual salary in $ for incoming classes in 2010.

Using the data in *MBAtui10*, the previous model has been estimated:

$$\widehat{salMBAgr}_i = \underset{(5415)}{42489} + \underset{(0.0628)}{0.1881} tuition_i + \underset{(0.1015)}{0.5992} salMBApr_i$$

$$R^2=0.703 \quad n=39$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the above model are individually significant at 1% and at 5%?

b) Test the overall significance of the model.

c) What is the predicted value of *salMBAgr* for a graduate student who paid 100000$ *tuition* fees in a two-year MBA master and previously had a *salMBApr* equal to 70000$? How many years of work does the student require to offset tuition expenses? To answer this question, suppose that the discount rate equals the expected rate of salary increase and that the student received no wage income during the two master courses.

d) If we added the regressor *rank2010* (the rank of each business school in 2010), the following results were obtained:

$$\widehat{salMBAgr}_i = \underset{(8520)}{61320} + \underset{(0.0626)}{0.1229} tuition_i + \underset{(0.1055)}{0.4662} salMBApr_i$$

$$\underset{(85.13)}{-232.06} rank2010_i$$

$$R^2=0.755 \quad n=39$$

Which of the regressors included in this model are individually significant at 5%?

What is the interpretation of the coefficient on *rank2010*?

e) The variable *rank2010* is based on three components: *gradpoll* is a rank based on surveys of MBA grads and contributes 45 percent to final ranking; *corppoll* is a rank based on surveys of MBA recruiters and contributes 45 percent to final ranking; and *intellec* is a rank based on a review of faculty research published over a five-year period in 20 top academic journals and faculty books reviewed in *The New York Times*, *The Wall Street Journal*, and *Bloomberg Businessweek* over the same period; this last rank contributes 10 percent to the final ranking. In the following estimated model *rank2010* has been substituted for its three components:

$$\widehat{salMBAgr}_i = \underset{(10700)}{79904} + \underset{(0.0696)}{0.0305}\,tuition_i + \underset{(0.107)}{0.3751}\,salMBApr_i$$
$$\underset{(94.54)}{-303.82}\,gradpoll_i - \underset{(61.26)}{33.829}\,corppoll_i - \underset{(64.09)}{113.36}\,intellec_i$$
$$R^2=0.797 \quad n=39$$

What is the weight in percentage of each one of these three components in determining the *salMBAgr*? Compare the results with the contribution of each in defining *rank2010*.

f) Are *gradpoll*, *corppoll* and *intellec* jointly significant at 5%? Are they individually significant at 5%?

Exercise 4.25 (Continuation of exercise 3.12). The population model corresponding to this exercise is:

$$\ln(wage) = \beta_1 + \beta_2 educ + \beta_3 tenure + \beta_4 age + u$$

Using workfile *wage06sp*, the previous model was estimated:

$$\widehat{\ln(wage)}_i = \underset{(0.073)}{1.565} + \underset{(0.0035)}{0.0448}\,educ_i + \underset{(0.0019)}{0.0177}\,tenure_i + \underset{(0.0016)}{0.0065}\,age_i$$
$$R^2=0.337 \quad n=800$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the overall significance of the model.

b) Is *tenure* statistically significant at 10%? Is *age* positively significant at 10%?

c) Is it admissible that the coefficient of *educ* is equal to that of *tenure*? Is it admissible that the coefficient of *educ* is triple to that of *tenure*? To answer these questions you have the following additional information:

$$\widehat{\ln(wage)}_i = \underset{(0.073)}{1.565} + \underset{(0.0042)}{0.0271}\,educ_i + \underset{(0.0019)}{0.0177}(educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

$$\widehat{\ln(wage)}_i = \underset{(0.073)}{1.565} - \underset{(0.0071)}{0.0082}\,educ_i + \underset{(0.0019)}{0.0177}(3\times educ + tenure)_i + \underset{(0.0016)}{0.0065}\,age_i$$

Can you calculate the R^2 in the two equations in part *c*)? Please do it.

Exercise 4.26 (Continuation of exercise 3.13). Let us take the population model of this exercise as the reference model. In the estimated model, using workfile *housecan*, the standard errors of the coefficients appear between brackets:

$$\widehat{price}_i = \underset{(3379)}{-2418} + \underset{(1207)}{5827}\,bedrooms_i + \underset{(1785)}{19750}\,bathrms_i + \underset{(0.388)}{5.411}\,lotsize_i$$
$$R^2=0.486 \quad n=546$$

a) Test the overall significance of this model.

b) Test the null hypothesis that an additional bathroom has the same influence on housing prices than four additional bedrooms. Alternatively, test that an additional bathroom has more influence on housing prices than four additional bedrooms. (Additional information: $\text{var}(\hat{\beta}_2)$ =1455813; $\text{var}(\hat{\beta}_3)$ =3186523; and $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =-764846).

c) If we add the regressor *stories* (number of stories excluding the basement) to the model, the following results have been obtained:

$$\widehat{price}_i = \underset{(3603)}{-4010} + \underset{(1215)}{2825}\, bedrooms_i + \underset{(1734)}{17105}\, bathrms_i + \underset{(0.369)}{5.429}\, lotsize_i + \underset{(1008)}{7635}\, stories_i$$

$$R^2=0.536 \quad n=546$$

What do you think about the sign and magnitude of the coefficient on *stories*? Do you find it surprising? What is the interpretation of this coefficient? Test whether the number of stories has a significant influence on housing prices.

d) Repeat the tests in part *b*) with the model estimated in part *c*). (Additional information: $\text{var}(\hat{\beta}_2)$ =1475758; $\text{var}(\hat{\beta}_3)$ =3008262; and $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =-554381).

Exercise 4.27 (Continuation of exercise 3.14). Let us take the population model of this exercise as the reference model. Using workfile *ceoforbes*, the estimated model was the following:

$$\widehat{\ln(salary)}_i = \underset{(0.377)}{4.641} + \underset{(0.0033)}{0.0054}\, roa_i + \underset{(0.0425)}{0.2893}\ln(sales_i) + \underset{(0.0000220)}{0.0000564}\, profits_i + \underset{(0.0032)}{0.0122}\, tenure_i$$

$$R^2=0.232 \quad n=447$$

(The numbers in parentheses are standard errors of the estimators.)

a) Does *roa* have a significant effect on salary? Does *roa* have a significant positive effect on salary? Carry out both tests at the 10% and 5% significance level.

b) If *roa* increases by 20 points, by what percentage is *salary* predicted to increase?

c) Test the null hypothesis that the elasticity *salary/sales* is equal to 0.4.

d) If we add the regressor *age*, the following results are obtained:

$$\widehat{\ln(salary)}_i = \underset{(0.442)}{4.159} + \underset{(0.0033)}{0.0055}\, roa_i + \underset{(0.0423)}{0.2903}\ln(sales_i) + \underset{(0.0000220)}{0.0000539}\, profits + \underset{(0.0035)}{0.00924}\, tenure_i + \underset{(0.0043)}{0.00880}\, age_i$$

$$R^2=0.240 \quad n=447$$

Are the estimated coefficients very different from the estimates in the reference model? What about the coefficient on *tenure*? Explain it.

e) Does *age* have a significant effect on the salary of a CEO?

f) Is it admissible that the coefficient of *age* is equal to the coefficient of *tenure*? (Additional information: $\text{var}(\hat{\beta}_5)$ =1.24E-05; $\text{var}(\hat{\beta}_6)$ =1.82E-05; and $\text{var}(\hat{\beta}_5, \hat{\beta}_6)$ =-6.09E-06).

Exercise 4.28 (Continuation of exercise 3.15). Let us take the population model of this exercise as the reference model. Using workfile *rdspain*, the estimated model was the following:

$$\widehat{rdintens}_i = \underset{(0.428)}{-1.8168} + \underset{(0.0278)}{0.1482}\ln(sales_i) + \underset{(0.0021)}{0.0110}\, exponsal_i$$

$$R^2=0.048 \quad n=1983$$

(The numbers in parentheses are standard errors of the estimators.)

a) Is the *sales* variable individually significant at 1%?

b) Test the null hypothesis that the coefficient on *sales* is equal to 0.2?
c) Test the overall significance of the reference model.
d) If we add the regressor ln(*workers*), the following results are obtained:

$$\widehat{rdintens} = \underset{(0.750)}{0.480} - \underset{(0.0687)}{0.08585}\ln(sales) + \underset{(0.0021)}{0.01049}\,exponsal + \underset{(0.09198)}{0.3422}\ln(workers)$$

$$R^2=0.055 \quad n=1983$$

Is *sales* individually significant at 1% in the new estimated model?
e) Test the null hypothesis that the coefficient on ln(*workers*) is greater than 0.5?

Exercise 4.29 (Continuation of exercise 3.16). Let us take the population model of this exercise as the reference model. Using workfile *hedcarsp*, the corresponding fitted model is the following:

$$\widehat{\ln(price)}_i = \underset{(0.154)}{14.42} + \underset{(0.0000438)}{0.000581}cid_i + \underset{(0.0079)}{0.003823}hpweight_i - \underset{(0.0122)}{0.07854}\,fueleff_i$$

$$R^2=0.830 \quad n=214$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 1%?
b) Add the variable *volume* to the reference model. Does *volume* have a statistically significant effect on ln(*price*)? Does *volume* have a statistically significant positive effect on ln(*price*)?
c) Is it admissible that the coefficient of *volume* estimated in part *b*) is equal but is the opposite of the coefficient of *fueloff*?
d) Add the variables *length*, *width* and *height* to the model estimated in part *b*). Taking into account that *volume*=*length*×*width*×*height*, is there perfect multicollinearity in the new model? Why? Why not? Estimate the new model if it is possible.
e) Add the variable ln(*volume*) to the reference model. Test the null hypothesis that the *price*/*volume* elasticity is equal to 1?
f) What happens if you add the regressors ln(*length*), ln(*width*) and ln(*height*) to the model estimated in part *e*)?

Exercise 4.30 (Continuation of exercise 3.17). Let us take the population model of this exercise as the reference model. Using workfile *timuse03*, the corresponding fitted model is the following:

$$\widehat{houswork}_i = \underset{(23.27)}{141.9} + \underset{(1.621)}{3.850}\,educ_i - \underset{(0.00539)}{0.00917}\,hhinc_i + \underset{(0.311)}{1.767}\,age_i - \underset{(0.0229)}{0.2289}\,paidwork_i$$

$$R^2=0.1440 \quad n=1000$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 5% and at 1%?
b) Estimate a model in which you could test directly whether one additional year of education has the same effect on time devoted to house work as two additional years of age. What is your conclusion?
c) Test the joint significance of *educ* and *hhnc*.

d) Run a regression in which you add the variable *childup3* (number of children up to three years) to the reference model. In the new model, which of the regressors are individually significant at 5% and at 1%?

e) In the model formulated in *d*), what is the most influential variable? Why?

Exercise 4.31 (Continuation of exercise 3.18). Let us take the population model of this exercise as the reference model. Using workfile *hdr2010*, the corresponding fitted model is the following:

$$\widehat{stsfglo}_i = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i$$

$$R^2=0.642 \quad n=144$$

(The numbers in parentheses are standard errors of the estimators.)

a) Which of the regressors included in the reference model are individually significant at 1%?

b) Run a regression by adding the variables *popnosan* (population in percentage without access to improved sanitation services) and *gnirank* (rank in *gni*) to the reference model. Which of the regressors included in the new model are individually significant at 1%? Interpret the coefficients on *popnosan* and *gnirank.*

c) Are *popnosan* and *gnirank* jointly significant?

d) Test the overall significance of the model formulated in *b*).

Exercise 4.32 Using a sample of 42 observations, the following model has been estimated:

$$\hat{y}_t = -670.591 + 1.008 x_t$$

For observation 43, it is known that the value of *x* is 1571.9.

a) Calculate the point predictor for observation 43.

b) Knowing that the variance of the prediction error $\hat{e}_2^{43} = y^{43} - \hat{y}^{43}$ is equal to $(24.9048)^2$, calculate a 90% probability interval for the individual value.

Exercise 4.33 Besides the estimation presented in exercise 4.23, the following estimation on the Brown consumption function is also available:

$$\widehat{conspc}_t = \underset{(64.35)}{12729} + \underset{(0.0857)}{0.3965}(incpc_t - 13500) + \underset{(0.0903)}{0.5771}(conspc_{t-1} - 12793.6)$$

$$R^2=0.997 \quad RSS=1891320 \quad n=56$$

(The numbers in parentheses are standard errors of the estimators.)

a) Obtain the point predictor for consumption per capita in 2011, knowing that $incpc_{2011}$=13500 and $conspc_{2010}$=12793.6.

b) Obtain a 95%confidence interval for the expected value of consumption per capita in 2011.

c) Obtain a 95% prediction interval for the individual value of consumption per capita in 2011.

Exercise 4.34 (Continuation of exercise 4.30) Answer the following questions:

a) Using the first estimation in exercise 4.30, obtain a prediction for *houswork* (minutes devoted to house-work per day), when you plug in the

equation *educ*=10 (years), *hhinc*=1200 (euros per month), *age*=50 (years) and *paidwork*=400 (minutes per day).

b) Run a regression, using workfile *timuse03*, which allows you to calculate a 95% CI with the characteristics used in part *a).*

c) Obtain a 95% prediction interval for the individual value of *houswork* with the characteristics used in parts *a).*

Exercise 4.35 (Continuation of exercise 4.29) Answer the following questions:

a) Plug in the first equation of the exercise 4.29 of *cid*=2000 (cubic inch displacement), *hpweight*=10 (ratio horsepower/weight in kg expressed as percentage), and *fueleff*=6 (minutes per day) Obtain the point predictor of consumption per capita in 2011, knowing that $incpc_{2011}$=12793.6 and $conspc_{2010}$=13500.

b) Obtain a consistent estimate of *price* with the characteristics used in parts *a).*

c) Run a regression that allows you to calculate a 95% CI with the characteristics used in part a).

d) Obtain a 95% prediction interval for the individual value of the consumption per capita 2011.

5 MULTIPLE REGRESSION ANALYSIS WITH QUALITATIVE INFORMATION

5.1 Introducing qualitative information in econometric models.

Up until now, the variables that we have used in explaining the endogenous variable have a quantitative nature. However, there are other variables of a qualitative nature that can be important when explaining the behavior of the endogenous variable, such as sex, race, religion, nationality, geographical region etc. For example, holding all other factors constant, female workers are found to earn less than their male counterparts. This pattern may result from gender discrimination, but whatever the reason, qualitative variables such as gender seem to influence the regressand and clearly should be included in many cases among the explanatory variables, or the regressors. Qualitative factors often (although not always) come in the form of binary information, i.e. a person is male or female, is either married or not, etc. When qualitative factors come in the form of dichotomous information, the relevant information can be captured by defining a binary variable or a zero-one variable. In econometrics, binary variables used as regressors are commonly called *dummy* variables. In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero.

In the case of gender, we can define

$$female = \begin{cases} 1 & \text{if the person is a female} \\ 0 & \text{if the person is a male} \end{cases}$$

But of course we can also define

$$male = \begin{cases} 1 & \text{if the person is a male} \\ 0 & \text{if the person is a female} \end{cases}$$

Nevertheless, it is important to remark that both variables, male and female, contain the same information. Using zero-one variables for capturing qualitative information is an arbitrary decision, but with this election the parameters have a natural interpretation.

5.2 A single dummy independent variable

Let us see how we incorporate dichotomous information into regression models. Consider the simple model of hourly *wage* determination as a function of the years of education (*educ*):

$$wage = \beta_1 + \beta_2 educ + u \qquad (5\text{-}1)$$

To measure gender wage discrimination, we introduce a dummy variable for gender as an independent variable in the model defined above,

$$wage = \beta_1 + \delta_1 female + \beta_2 educ + u \qquad (5\text{-}2)$$

The *attribute* gender has two *categories*: *male and female*. The *female* category has been included in the model, while the *male* category, which was omitted, is the *reference category*. Model 1 is shown in Figure 5.1, taking $\delta_1<0$. The interpretation of δ_1 is the following: δ_1 is the difference in hourly wage between females and males, given the same amount of education (and the same error disturbance *u*). Thus, the coefficient δ_1 determines whether there is discrimination against women or not. If $\delta_1<0$ then, for the same level of other factors (education, in this case), women earn less than men on average. Assuming that the disturbance mean is zero, if we take expectation for both categories we obtain:

$$\begin{aligned} \mu_{wage|\,female} &= E(wage \mid female = 1, educ) = \beta_1 + \delta_1 + \beta_2 educ \\ \mu_{wage|male} &= E(wage \mid female = 0, educ) = \beta_1 + \beta_2 educ \end{aligned} \qquad (5\text{-}3)$$

As can be seen in (5-3), the intercept is β_1 for males, and $\beta_1+\delta_1$ for females. Graphically, as can be seen in Figure 5.1, there is a shift of the intercept, but the lines for men and women are parallel.

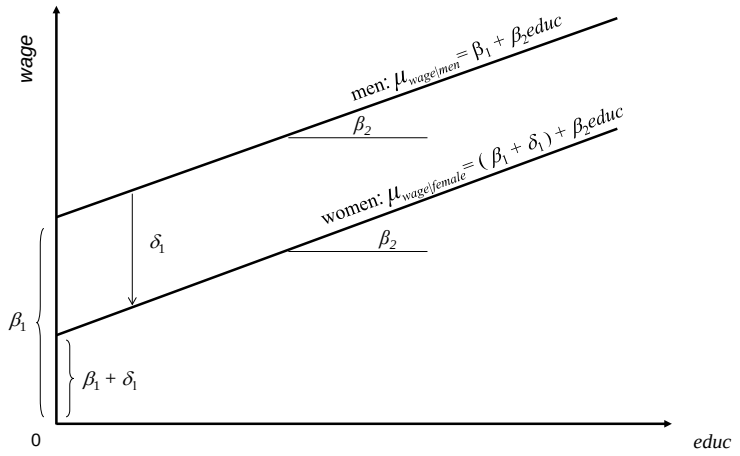


FIGURE 5.1. Same slope, different intercept.

In (5-2) we have included a dummy variable for *female* but not for *male*, because if we had included both dummies this would have been redundant. In fact, all we need is two intercepts, one for females and another one for males. As we have seen, if we introduce the *female* dummy variable, we have an intercept for each gender. Introducing two dummy variables would cause perfect multicollinearity given that *female*+*male*=1, which means that *male* is an exact linear function of *female* and of the intercept. Including dummy variables for both genders plus the intercept is the simplest example of the so-called dummy variable trap, as we shall show later on.

If we use *male* instead of *female*, the wage equation would be the following:

$$wage = \alpha_1 + \gamma_1 male + \beta_2 educ + u \qquad (5\text{-}4)$$

Nothing has changed with the new equation, except the interpretation of α_1 and γ_1: α_1 is the intercept for women, which is now the *reference category*, and $\alpha_1+\gamma_1$ is the intercept for men. This implies the following relationship between the coefficients:

$$\alpha_1=\beta_1+\delta_1 \text{ and } \alpha_1+\gamma_1=\beta_1 \Rightarrow \gamma_1=-\delta_1$$

In any application, it does not matter how we choose the reference category, since this only affects the interpretation of the coefficients associated to the dummy variables, but it is important to keep track of which category is the reference category. Choosing a reference category is usually a matter of convenience. It would also be possible to drop the intercept and to include a dummy variable for each category. The equation would then be

$$wage = \mu_1 male + \nu_1 female + \beta_2 educ + u \qquad (5\text{-}5)$$

where the intercept is μ_1 for men and ν_1 for women.

Hypothesis testing is performed as usual. In model (5-2), the null hypothesis of no difference between men and women is $H_0 : \delta_1 = 0$, while the alternative hypothesis that there is discrimination against women is $H_1 : \delta_1 < 0$. Therefore, in this case, we must apply a one sided (left) *t* test.

A common specification in applied work has the dependent variable as the logarithm transformation ln(*y*) in models of this type. For example:

$$\ln(wage) = \beta_1 + \delta_1 female + \beta_2 educ + u \qquad (5\text{-}6)$$

Let us see the interpretation of the coefficient of the dummy variable in a log model. In model (5-6), taking *u*=0, the wage for a female and for a male is as follows:

$$\ln(wage_F) = \beta_1 + \delta_1 + \beta_2 educ \qquad (5\text{-}7)$$

$$\ln(wage_M) = \beta_1 + \beta_2 educ \qquad (5\text{-}8)$$

Given the same amount of education, if we subtract (5-7) from (5-8), we have

$$\ln(wage_F) - \ln(wage_M) = \delta_1 \qquad (5\text{-}9)$$

Taking antilogs in (5-9) and subtracting 1 from both sides of (5-9), we get

$$\frac{wage_F}{wage_M} - 1 = e^{\delta_1} - 1 \qquad (5\text{-}10)$$

That is to say

$$\frac{wage_F - wage_M}{wage_M} = e^{\delta_1} - 1 \qquad (5\text{-}11)$$

According to (5-11), the proportional change between the female wage and the male wage, for the same amount of education, is equal to $e^{\delta_1} - 1$. Therefore, the exact percentage change in hourly wage between men and women is $100 \times (e^{\delta_1} - 1)$. As an approximation to this change, $100 \times \delta_1$ can be used. However, if the magnitude of the percentage is high, then this approximation is not so accurate.

EXAMPLE 5.1 Is there wage discrimination against women in Spain?

Using data from the *wage structure survey* of Spain for 2002 (file *wage02sp*), model (5-6) has been estimated and the following results were obtained:

$$\widehat{\ln(wage)} = \underset{(0.026)}{1.731} - \underset{(0.022)}{0.307}\, female + \underset{(0.0025)}{0.0548}\, educ$$

$$RSS=393 \quad R^2=0.243 \quad n=2000$$

where *wage* is hourly wage in euros, *female* is a dummy variable that takes the value 1 if it is a woman, and *educ* are the years of education. (The numbers in parentheses are standard errors of the estimators.)

To answer the question posed above, we need to test $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 < 0$. Given that the *t* statistic is equal to -14.27, we reject the null hypothesis for α=0.01. That is to say, there is a negative discrimination in Spain against women in the year 2002. In fact, the percentage difference in hourly wage between men and women is $100 \times (e^{0.307} - 1) = 35.9\%$, given the same years of education.

EXAMPLE 5.2 Analysis of the relation between market capitalization and book value: the role of ibex35

A researcher wants to study the relationship between market capitalization and book value in shares quoted on the continuous market of the Madrid stock exchange. In this market some stocks quoted are included in the ibex35, a selective index. The researcher also wants to know whether the stocks included in the ibex35 have, on average, a higher capitalization.. With this purpose in mind, the researcher formulates the following model:

$$\ln(marketcap) = \beta_1 + \delta_1 ibex35 + \beta_2 \ln(bookvalue) + u \qquad (5\text{-}12)$$

where

- ***marktval*** is the capitalization value of a company, which is calculated by multiplying the price of the stock by the number of stocks issued.
- *bookval* is the book value of a company, also referred to as the net worth of the company. The book value is calculated as the difference between a company's assets and its liabilities.
- *ibex35* is a dummy variable that takes the value 1 if the corporation is included in the selective Ibex 35.

Using the 92 stocks quoted on 15th November 2011 which supply information on book value (file *bolmad11*), the following results were obtained:

$$\widehat{\ln(marketcap)} = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.690}\, ibex35 + \underset{(0.037)}{0.675} \ln(bookvalue)$$

$$RSS=35.672 \quad R^2=0.893 \quad n=92$$

The *marketcap*/*bookvalue* elasticity is equal to 0.690; that is to say, if the book value increases by 1%, then the market capitalization of the quoted stocks will increase by 0.675%.

To test whether the stocks included in *ibex35* have on average a higher capitalization implies testing $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 > 0$. Given that the *t* statistic is (0.690/0.179)=3.85, we reject the null hypothesis for the usual levels of significance. On the other hand, we see that the stocks included in *ibex35* are quoted 99.4% higher than the stocks not included. The percentage is obtained as follows: $100 \times (e^{0.690} - 1) = 99.4\%$.

In the case of β_2, we can test $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$. Given that the *t* statistic is (0.675/0.037)=18, we reject the null hypothesis for the usual levels of significance.

EXAMPLE 5.3 Do people living in urban areas spend more on fish than people living in rural areas?

To see whether people living in urban areas spend more on fish than people living in rural areas, the following model is proposed:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + u \qquad (5\text{-}13)$$

where *fish* is expenditure on fish, *urban* is a dummy variable which takes the value 1 if the person lives in an urban area and *inc* is disposable income.

Using a sample of size 40 (file *demand*), model (5-13) was estimated:

$$\widehat{\ln(fish)} = - \underset{(0.511)}{6.375} + \underset{(0.055)}{0.140}\, urban + \underset{(0.070)}{1.313} \ln(inc)$$

RSS=1.131 R^2=0.904 n=40

According to these results, people living in urban areas spend 14% more on fish than people living in rural areas. If we test $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 > 0$, we find that the t statistic is (0.140/0.055)=2.55. Given that $t_{37}^{0.01} \approx t_{35}^{0.01}$ =2.44, we reject the null hypothesis in favor of the alternative for the usual levels of significance. That is to say, there is empirical evidence that people living in urban areas spend more on fish than people living in rural areas.

5.3 Multiple categories for an attribute

In the previous section we have seen an attribute (gender) that has two categories (male and female). Now we are going to consider attributes with more than two categories. In particular, we will examine an attribute with three categories

To measure the impact of firm size on wage, we can use a dummy variable. Let us suppose that firms are classified in three groups according to their size: small (up to 49 workers), medium (from 50 to 199 workers) and large (more than 199 workers). With this information, we can construct three dummy variables:

$$small = \begin{cases} 1 & \text{up to 49 workers} \\ 0 & \text{in other case} \end{cases}$$

$$medium = \begin{cases} 1 & \text{from 50 to 199 workers} \\ 0 & \text{in other case} \end{cases}$$

$$large = \begin{cases} 1 & \text{more than 199 workers} \\ 0 & \text{in other case} \end{cases}$$

If we want to explain hourly wages by introducing the firm size in the model, we must omit one of the categories. In the following model, the omitted category is *small* firms:

$$wage = \beta_1 + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (5\text{-}14)$$

The interpretation of the θ_j coefficients is the following: θ_1 (θ_2) is the difference in hourly wage between medium (large) firms and small firms, given the same amount of education (and the same error term u).

Let us see what happens if we also include the category *small* in (5-14). We would have the model:

$$wage = \beta_1 + \theta_0 small + \theta_1 medium + \theta_2 large + \beta_2 educ + u \quad (5\text{-}15)$$

Now, let us consider that we have a sample of six observations: the observations 1 and 2 correspond to small firms; 3 and 4 to medium ones; and 5 and 6 to large ones. In this case the matrix of regressors **X** would have the following configuration:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & educ_1 \\ 1 & 1 & 0 & 0 & educ_2 \\ 1 & 0 & 1 & 0 & educ_3 \\ 1 & 0 & 1 & 0 & educ_4 \\ 1 & 0 & 0 & 1 & educ_5 \\ 1 & 0 & 0 & 1 & educ_6 \end{bmatrix}$$

As can be seen in matrix **X**, column 1 of this matrix is equal to the sum of columns 2, 3 and 4. Therefore, there is perfect multicollinearity due to the so-called *dummy variable trap*. Generalizing, if an attribute has *g* categories, we need to include only *g*−1 dummy variables in the model along with the intercept. The intercept for the reference category is the overall intercept in the model, and the dummy variable coefficient for a particular group represents the estimated difference in intercepts between that category and the reference category. If we include *g* dummy variables along with an intercept, we will fall into the dummy variable trap. An alternative is to include *g* dummy variables and to exclude an overall intercept. In the case we are examining, the model would be the following:

$$wage = \theta_0 small + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (5\text{-}16)$$

This solution is not advisable for two reasons. With this configuration of the model it is more difficult to test differences with respect to a reference category. Second, this solution only works in the case of a model with only one unique attribute.

EXAMPLE 5.4 Does firm size influence wage determination?

Using the sample of example 5.1 (file *wage02sp*), model (5-14), taking log for *wage*, was estimated:

$$\widehat{\ln(wage)} = \underset{(0.027)}{1.566} + \underset{(0.025)}{0.281}\, medium + \underset{(0.024)}{0.162}\, large + \underset{(0.0025)}{0.0480}\, educ$$

$$RSS=406 \quad R^2=0.218 \quad n=2000$$

To answer the question above, we will not perform an *individual* test on θ_1 or θ_2. Instead we must *jointly* test whether the size of firms has a significant influence on wage. That is to say, we must test whether medium and large firms together have a significant influence on the determination of wage. In this case, the null and the alternative hypothesis, taking (5-14) as the unrestricted model, will be the following:

$$H_0 : \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ is not true}$$

The restricted model in this case is the following:

$$\ln(wage) = \beta_1 + \beta_2 educ + u \qquad (5\text{-}17)$$

The estimation of this model is the following:

$$\widehat{\ln(wage)} = \underset{(0.026)}{1.657} + \underset{(0.0026)}{0.0525}\, educ$$

$$RSS=433 \quad R^2=0.166 \quad n=2000$$

Therefore, the *F* statistic is

$$F = \frac{[RSS_R - RSS_{UR}]/q}{RSS_{UR}/(n-k)} = \frac{[433-406]/2}{406/(2000-4)} = 66.4$$

So, according to the value of the *F* statistic, we can conclude that the size of the firm has a significant influence on wage determination for the usual levels of significance.

Example 5.5 In the case of Lydia E. Pinkham, are the time dummy variables introduced significant individually or jointly?

In example 3.4, we considered the case of Lydia E. Pinkham in which *sales* of a herbal extract from this company (expressed in thousands of dollars) were explained in terms of advertising expenditures in thousands of dollars (*advexp*) and last year's sales ($sales_{t-1}$). However, in addition to these two variables, the author included three time dummy variables: *d*1, *d*2 and *d*3. These dummy variables encompass the various situations which took place in the company. Thus, *d*1 takes 1 in the period 1907-1914 and 0 in the remaining periods, *d*2 takes 1 in the period 1915-1925 and 0 in other periods, and finally, *d*3 takes 1 in the period 1926 - 1940 and 0 in the remaining periods. Thus, the reference category is the period 1941-1960. The final formulation of the model was therefore the following:

$$sales_t=\beta_1+\beta_2 advexp_t+\beta_3 sales_{t-1}+\beta_4 d1_t+\beta_5 d2_t+\beta_6 d3_t+u_t \qquad (5\text{-}18)$$

The results obtained in the regression, using file *pinkham*, were the following:

$$\widehat{sales}_t = \underset{(96.3)}{254.6} + \underset{(0.136)}{0.5345}\, advexp_t + \underset{(0.0814)}{0.6073}\, sales_{t-1} - \underset{(89)}{133.35}\, d1_t + \underset{(67)}{216.84}\, d2_t - \underset{(67)}{202.50}\, d3_t$$

$$R^2=0.929 \qquad n=53$$

To test whether the dummy variables individually have a significant effect on sales, the null and alternative hypotheses are:

$$\begin{cases} H_0 : \gamma_i = 0 \\ H_1 : \gamma_i \neq 0 \end{cases} \qquad i = 1,2,3$$

The corresponding *t* statistics are the following:

$$t_{\hat{\gamma}_1} = \frac{-133.35}{89} = -1.50 \qquad t_{\hat{\gamma}_2} = \frac{216.84}{67} = 3.22 \qquad t_{\hat{\gamma}_3} = \frac{-202.50}{67} = -3.02$$

As can be seen, the regressor *d*1 is not significant for any of the usual levels of significance, whereas on the contrary the regressors *d*2 and *d*3 are significant for any of the usual levels.

The interpretation of the coefficient of the regressor *d*2, for example, is as follows: holding fixed the advertising spending and given the previous year's sales, sales for one year of the period 1915-1920 are $ 2.684 higher than for a year of the period 1941-1960.

To test jointly the effect of the time dummy variables, the null and alternative hypotheses are

$$\begin{cases} H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0 \\ H_1 : H_0 \text{ is not true} \end{cases}$$

and the corresponding test statistic is

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k)} = \frac{(0.9290-0.8770)/3}{(1-0.9290)/(53-6)} = 11.47$$

For any of the usual significance levels the null hypothesis is rejected. Therefore, the time dummy variables have a significant effect on sales

5.4 Several attributes

Now we will consider the possibility of taking into account two attributes to explain the determination of wage: gender and length of workday (part-time and full-time). Let *partime* be a dummy variable that takes value 1 when the type of contract is part-time and 0 if it is full-time. In the following model, we introduce two dummy variables: *female* and *partime*:

$$wage = \beta_1 + \delta_1 female + \phi_1 partime + \beta_2 educ + u \qquad (5\text{-}19)$$

In this model, ϕ_1 is the difference in hourly wage between those who work part-time, given gender and the same amount of education (and also the same disturbance term *u*).

Each of these two attributes has a reference category, which is the omitted category. In this case, male is the reference category for gender and full-time for type of contract. If we take expectations for the four categories involved, we obtain:

$$\mu_{wage|\,female,\,partime} = E\left[wage \mid female,\, partime, educ\right] = \beta_1 + \delta_1 + \phi_1 + \beta_2 educ$$
$$\mu_{wage|\,female,\,fulltime} = E\left[wage \mid female,\, fulltime, educ\right] = \beta_1 + \delta_1 + \beta_2 educ$$
$$\mu_{wage|male,\,partime} = E\left[wage \mid male,\, partime, educ\right] = \beta_1 + \phi_1 + \beta_2 educ$$
$$\mu_{wage|male,\,fulltime} = E\left[wage \mid male,\, fulltime, educ\right] = \beta_1 + \beta_2 educ$$

(5-20)

The overall intercept in the equation reflects the effect of both reference categories, male and full-time, and so full-time male is the reference category. From (5-20), you can see the intercept for each combination of categories.

EXAMPLE 5.6 The influence of gender and length of the workday on wage determination

Model (5-19), taking log for wage, was estimated by using data from the *wage structure survey* of Spain for 2006 (file *wage06sp*):

$$\widehat{\ln(wage)} = \underset{(0.026)}{2.006} - \underset{(0.021)}{0.233}\, female - \underset{(0.027)}{0.087}\, partime + \underset{(0.0023)}{0.0531}\, educ$$

$$RSS=365 \quad R^2=0.235 \quad n=2000$$

According to the values of the coefficients and corresponding standard errors, it is clear that each one of the two dummy variables, *female* and *partime*, are statistically significant for the usual levels of significance.

EXAMPLE 5.7 Trying to explain the absence from work in the company Buenosaires

Buenosaires is a firm devoted to the manufacturing of fans, having had relatively acceptable results in recent years. The managers consider that these would have been better if absenteeism in the company were not so high. In order to analyze the factors determining absenteeism, the following model is proposed:

$$absent = \beta_1 + \delta_1 bluecoll + \phi_1 male + \beta_2 age + \beta_3 tenure + \beta_4 wage + u \qquad (5\text{-}21)$$

where *bluecoll* is a dummy indicating that the person is a manual worker (the reference category is white collar) and *tenure* is a continuous variable reflecting the years worked in the company.

Using a sample of size 48 (file *absent*), the following equation has been estimated:

$$\widehat{absent} = \underset{(1.640)}{12.444} + \underset{(0.669)}{0.968}\, bluecoll + \underset{(0.712)}{2.049}\, male - \underset{(0.047)}{0.037}\, age - \underset{(0.065)}{0.151}\, tenure - \underset{(0.007)}{0.044}\, wage$$

$$RSS=161.95 \quad R^2=0.760 \quad n=48$$

Next, we will look at whether *bluecoll* is significant. Testing $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 \neq 0$, the *t* statistic is (0.968/0.669)=1.45. As $t_{40}^{0.10/2}$=1.68, we fail to reject the null hypothesis for α=0.10. And so there is no empirical evidence to state that absenteeism amongst blue collar workers is different from white collar workers. But if we test $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 > 0$, as $t_{40}^{0.10}$=1.30 for α=0.10, then we cannot reject that absenteeism amongst blue collar workers is greater than amongst white collar workers.

On the contrary, in the case of the *male* dummy, testing $H_0 : \varphi_1 = 0$ against $H_1 : \varphi_1 \neq 0$, given that the *t* statistic is (2.049/0.712)=2.88 and $t_{40}^{0.01/2}$=2.70, we reject that absenteeism is equal in men and women for the usual levels of significance.

EXAMPLE 5.8 Size of firm and gender in determining wage

In order to know whether the size of the firm and gender jointly are two relevant factors in determining wage, the following model is formulated:

$$\ln(wage) = \beta_1 + \delta_1\, female + \theta_1 medium + \theta_2 large + \beta_2 educ + u \qquad (5\text{-}22)$$

In this case, we must perform a joint test where the null and the alternative hypotheses are

$$H_0 : \delta_1 = \theta_1 = \theta_2 = 0$$
$$H_1 : H_0 \text{ is not true}$$

In this case, the restricted model is model (5-17) which was estimated in example 5.4 (file *wage02sp*). The estimation of the unrestricted model is the following:

$$\widehat{\ln(wage)} = \underset{(0.026)}{1.639} - \underset{(0.021)}{0.327}\, female + \underset{(0.023)}{0.308}\, medium + \underset{(0.023)}{0.168}\, large + \underset{(0.0024)}{0.0499}\, educ$$

$$RSS=361 \quad R^2=0.305 \quad n=2000$$

The F statistic is

$$F = \frac{[RSS_R - RSS_{UR}]/q}{RSS_{UR}/(n-k)} = \frac{[433-361]/3}{361/(2000-5)} = 133$$

Therefore, according to the value of F, we can conclude that the size of the firm and gender jointly have a significant influence in wage determination.

5.5 Interactions involving dummy variables.

5.5.1 Interactions between two dummy variables

To allow for the possibility of an interaction between gender and length of the workday on wage determination, we can add an interaction term between *female* and *partime* in model (5-19), with the model to estimate being the following:

$$wage = \beta_1 + \delta_1 female + \phi_1 partime + \varphi_1 female \times partime + \beta_2 educ + u \quad (5\text{-}23)$$

This allows working time to depend on gender and vice versa.

EXAMPLE 5.9 Is the interaction between females and part-time work significant?

Model (5-23), taking log for wage, was estimated by using data from the *wage structure survey* of Spain for 2006 (file *wage06sp*):

$$\widehat{\ln(wage)} = \underset{(0.026)}{2.007} - \underset{(0.022)}{0.259}\, female - \underset{(0.047)}{0.198}\, partime + \underset{(0.058)}{0.167}\, female' \ partime + \underset{(0.0024)}{0.0538}\, educ$$

$$RSS=363 \quad R^2=0.238 \quad n=2000$$

To answer the question posed, we have to test $H_0 : \varphi_1 = 0$ against $H_0 : \varphi_1 \neq 0$. Given that the t statistic is (0.167/0.058)=2.89 and taking into account that $t_{60}^{0.01/2}$ =2.66, we reject the null hypothesis in favor of the alternative hypothesis. Therefore, there is empirical evidence that the interaction between females and part-time work is statistically significant.

EXAMPLE 5.10 Do small firms discriminate against women more or less than larger firms?

To answer this question, we formulate the following model:

$$\ln(wage) = \beta_1 + \delta_1 female + \theta_1 medium + \theta_2 large + \varphi_1 female \times medium + \varphi_2 female \times large + \beta_2 educ + u \quad (5\text{-}24)$$

Using the sample of example 5.1 (file *wage02sp*), model (5-24) was estimated:

$$\widehat{\ln(wage)} = \underset{(0.027)}{1.624} - \underset{(0.034)}{0.262}\, female + \underset{(0.028)}{0.361}\, medium + \underset{(0.027)}{0.179}\, large - \underset{(0.050)}{0.159}\, female' \ medium - \underset{(0.051)}{0.043}\, female' \ large + \underset{(0.0024)}{0.0497}\, educ$$

$$RSS=359 \quad R^2=0.308 \quad n=2000$$

If in (5-24) the parameters φ_1 and φ_2 are equal to 0, this will imply that in the equation for wage determination, there will be non interaction between gender and firm size. Thus to answer the above question, we take (5-24) as the *unrestricted* model. The null and the alternative hypothesis will be the following:

$$H_0 : \varphi_1 = \varphi_2 = 0$$
$$H_1 : H_0 \text{ is not true}$$

In this case, the restricted model is therefore model (5-22) estimated in example 5.7. The *F* statistic takes the value

$$F = \frac{[RSS_R - RSS_{UR}]/q}{RSS_{UR}/(n-k)} = \frac{[361-359]/2}{359/(2000-7)} = 5.55$$

For α=0.01, we find that $F_{2,1993}^{0.01}$; $F_{2,60}^{0.01} = 4.98$. As F>5.61, we reject H_0 in favor of H_1. As H_0 has been rejected for α=0.01, it will also be rejected for levels of 5% and 10%. Therefore, the usual levels of significance, the interaction between gender and firm size is relevant for wage determination.

5.5.2 Interactions between a dummy variable and a quantitative variable

So far, in the examples for wage determination a dummy variable has been used to shift the intercept or to study its interaction with another dummy variable, while keeping the slope of *educ* constant. However, one can also use dummy variables to shift the slopes by letting them interact with any continuous explanatory variables. For example, in the following model the *female* dummy variable interacts with the continuous variable *educ*:

$$wage = \beta_1 + \beta_2 educ + \delta_1 female \times educ + u \qquad (5\text{-}25)$$

As can be seen in figure 5.2, the intercept is the same for men and women in this model, but the slope is greater in men than in women because δ_1 is negative.

In model (5-25), the returns to an extra year of education depend upon the gender of the individual. In fact,

$$\frac{\partial wage}{\partial educ} = \begin{cases} \beta_2 + \delta_1 & \text{for women} \\ \beta_2 & \text{for men} \end{cases} \qquad (5\text{-}26)$$

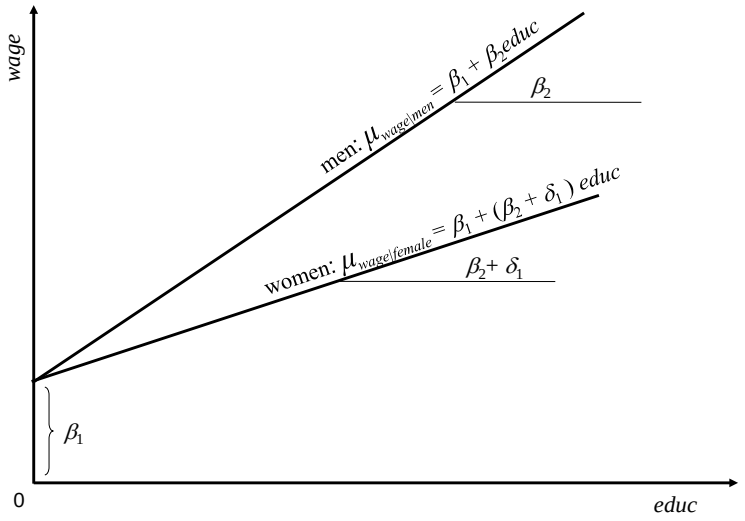


FIGURE 5.2. Different slope, same intercept.

EXAMPLE 5.11 Is the return to education for males greater than for females?

Using the sample of example 5.1 (file *wage02sp*), model (5-25) was estimated by taking log for wage*:*

$$\widehat{\ln(wage)} = \underset{(0.025)}{1.640} + \underset{(0.0026)}{0.0632}\,educ - \underset{(0.0021)}{0.0274}\,educ' \quad female$$

$$RSS=400 \quad R^2=0.229 \quad n=2000$$

In this case, we need to test $H_0 : \delta_1 = 0$ against $H_1 : \delta_1 < 0$. Given that the *t* statistic is (-0.0274/0.0021) =-12.81, we reject the null hypothesis in favor of the alternative hypothesis for any level of significance. That is to say, there is empirical evidence that the return for an additional year of education is greater for men than for women.

5.6 Testing structural changes

So far we have tested hypotheses in which one parameter, or a subset of parameters of the model, is different for two groups (women and men, for example). But sometimes we wish to test the null hypothesis that two groups have the same population regression function, against the alternative that it is not the same. In other words, we want to test whether the same equation is valid for the two groups. There are two procedures for this: using dummy variables and running separate regressions through the Chow test.

5.6.1 Using dummy variables

In this procedure, testing for differences across groups consists in performing a joint significance test of the dummy variable, which distinguishes between the two groups and its interactions with all other independent variables. We therefore estimate the model with (*unrestricted model*) and without (*restricted model*) the dummy variable and all the interactions.

From the estimation of both equations we form the *F* statistic, either through the *RSS* or from the R^2. In the following model for the determination of wages, the intercept and the slope are different for males and females:

$$wage = \beta_1 + \delta_1\, female + \beta_2 educ + \delta_2\, female \times educ + u \qquad (5\text{-}27)$$

The population regression function corresponding to this model is represented in figure 5.3. As can be seen, if *female*=1, we obtain

$$wage = (\beta_1 + \delta_1) + (\beta_2 + \delta_2) educ + u \qquad (5\text{-}28)$$

For women the intercept is $\beta_1 + \delta_1$, and the slope $\beta_2 + \delta_2$. For *female*=0, we obtain equation (5-1). In this case, for men the intercept is β_1, and the slope β_2. Therefore, δ_1 measures the difference in intercepts between men and women and, δ_2 measures the difference in the return to education between males and females. Figure 5.3 shows a lower intercept and a lower slope for women than for men. This means that women earn less than men at all levels of education, and the gap increases as *educ* gets larger; that is to say, an additional year of education shows a lower return for women than for men.

Estimating (5-27) is equivalent to estimating two wage equations separately, one for men and another for women. The only difference is that (5-27) imposes the same variance across the two groups, whereas separate regressions do not. This set-up is ideal, as we will see later on, for testing the equality of slopes, equality of intercepts, and equality of both intercepts and slopes across groups.

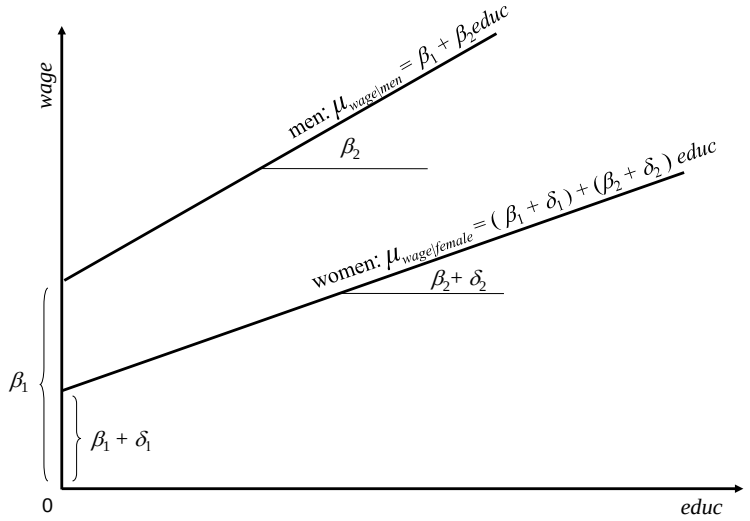


FIGURE 5.3. Different slope, different intercept.

EXAMPLE 5.12 Is the wage equation valid for both men and women?

If parameters δ_1 and δ_2 are equal to 0 in model (5-27), this will imply that the equation for wage determination is the same for men and women. In order to answer the question posed, we take (5-27), as the *unrestricted* model but express *wage* in logs. The null and the alternative hypothesis will be the following:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ is not true}$$

Therefore, the restricted model is model (5-17). Using the same sample as in example 5.1 (file *wage02sp*), we have obtained the following estimation of models (5-27) and (5-17):

$$\widehat{\ln(wage)} = \underset{(0.030)}{1.739} - \underset{(0.0546)}{0.3319}\, female + \underset{(0.0030)}{0.0539}\, educ - \underset{(0.0054)}{0.0027}\, educ' \quad female$$

$$RSS=393 \qquad R^2=0.243 \qquad n=2000$$

$$\widehat{\ln(wage)} = \underset{(0.026)}{1.657} + \underset{(0.0026)}{0.0525}\, educ$$

$$RSS=433 \qquad R^2=0.166 \qquad n=2000$$

The F statistic takes the value

$$F = \frac{[RSS_R - RSS_{UR}]/q}{RSS_{UR}/(n-k)} = \frac{[433-393]/2}{393/(2000-4)} = 102$$

It is clear that for any level of significance, the equations for men and women are different.

When we tested in example 5.1 whether there was discrimination in Spain against women ($H_0 : \delta_1 = 0$ against $H_1 : \delta_1 < 0$), it was assumed that the slope of *educ* (model (5-6)) is the same for men and women. Now it is also possible to use model (5-27) to test the same null hypothesis, but assuming a different slope. Given that the *t* statistic is (-0.3319/0.0546)=-6.06, we reject the null hypothesis by using this more general model than the one in example 5.1.

In example 5.11 it was tested whether the coefficient δ_2 in model (5-25), taking log for wage, was 0, assuming that the intercept is the same for males and females. Now, if we take (5-27), taking log for wage, as the unrestricted model, we can test the same null hypothesis, but assuming that the intercept is different for males and females. Given that the *t* statistic is (0.0027/0.0054)=0.49, we cannot reject the null hypothesis which states that there is no interaction between gender and education.

EXAMPLE 5.13 Would urban consumers have the same pattern of behavior as rural consumers regarding expenditure on fish?

To answer this question, we formulate the following model which is taken as the *unrestricted* model:

$$\ln(fish) = \beta_1 + \delta_1 urban + \beta_2 \ln(inc) + \delta_2 \ln(inc) \times urban + u \qquad (5\text{-}29)$$

The null and the alternative hypothesis will be the following:

$$H_0 : \delta_1 = \delta_2 = 0$$
$$H_1 : H_0 \text{ is not true}$$

The restricted model corresponding to this H_0 is

$$\ln(fish) = \beta_1 + \beta_2 \ln(inc) + u \qquad (5\text{-}30)$$

Using the sample of example 5.3 (file *demand*), models (5-29) and (5-30) were estimated:

$$\widehat{\ln(fish)} = \underset{(0.627)}{-6.551} + \underset{(1.095)}{0.678}\, urban + \underset{(0.087)}{1.337} \ln(inc) - \underset{(0.152)}{0.075} \ln(inc) \times urban$$

$$RSS=1.123 \qquad R^2=0.904 \qquad n=40$$

$$\widehat{\ln(fish)} = \underset{(0.542)}{-6.224} + \underset{(0.075)}{1.302} \ln(inc)$$

$$RSS=1.325 \qquad R^2=0.887 \qquad n=40$$

The F statistic takes the value

$$F = \frac{[RSS_R - RSS_{UR}]/q}{RSS_{UR}/(n-k)} = \frac{[1.325-1.123]/2}{1.123/(40-4)} = 3.24$$

If we look up in the F table for 2 *df* in the numerator and 35 *df* in the denominator for α=0.10, we find $F_{2,36}^{0.10}$; $F_{2,35}^{0.10} = 2.46$. As F>2.46 we reject H_0 . However, as $F_{2,36}^{0.05}$; $F_{2,35}^{0.05} = 3.27$, we fail to reject H_0 in favour of H_1 for α=0.05 and, therefore, for α=0.01. Conclusion: there is no strong evidence that families living in rural areas have a different pattern of fish consumption than families living in rural areas.

Example 5.14 Has the productive structure of Spanish regions changed?

The question to be answered is specifically the following: Did the productive structure of Spanish regions change between 1995 and 2008? The problem posed is a problem of structural stability. To specify the model to be taken as a reference in the test, let us define the dummy *y2008*, which takes the value 1 if the year is 2008 and 0 if the year is 1995.

The reference model is a Cobb-Douglas model, which introduces additional parameters to collect the structural changes that may have occurred. Its expression is:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + \gamma_2 y2008 + \alpha_2 y2008 \times \ln(k) + \beta_2 y2008 \times \ln(l) + u \qquad (5\text{-}31)$$

It is easily seen, according to the definition of the dummy *y2008*, that the elasticities production/capital are different in the periods 1995 and 2008. Specifically, they take the following values:

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(Q)}{\partial \ln(K)} = \alpha_1 + \alpha_2$$

In the case that α_2 is equal to 0, then the elasticity of production/capital is the same in both periods.

Similarly, the production/labor elasticities for the two periods are given by

$$\varepsilon_{Q/K(1995)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 \qquad \varepsilon_{Q/K(2008)} = \frac{\partial \ln(L)}{\partial \ln(K)} = \beta_1 + \beta_2$$

The intercept in the Cobb-Douglas is a parameter that measures efficiency. In model (5-31), the possibility that the efficiency parameter (*PEF*) is different in the two periods is considered. Thus

$$PEF(1995) = \gamma_1 \qquad PEF(2008) = \gamma_1 + \gamma_2$$

If the parameters α_1, β_1 and γ_1 are zero in model (5-31), the production function is the same in both periods. Therefore, in testing structural stability of the production function, the null and alternative hypotheses are:

$$H_0 : \gamma_2 = \alpha_2 = \beta_2$$
$$H_1 : H_0 \text{ is not true} \qquad (5\text{-}32)$$

Under the null hypothesis, the restrictions given in (5-32) lead to the following restricted model:

$$\ln(q) = \gamma_1 + \alpha_1 \ln(k) + \beta_1 \ln(l) + u \qquad (5\text{-}33)$$

The file *prodsp* contains information for each of the Spanish regions in 1995 and 2008 on gross value added in millions of euros (*gdp*), occupation in thousands of jobs (labor), and productive capital in millions of euros (*captot*). You can also find the dummy y2008 in that file.

The results of the unrestricted regression model (5-31) are shown below. It is evident that we cannot reject the null hypothesis that each of the coefficients α_1, β_1 and γ_1, taken individually, are 0, since none of the *t* statistics reaches 0.1 in absolute value.

$$\widehat{\ln(gva)} = \underset{(0.916)}{0.0559} + \underset{(0.185)}{0.6743}\ln(captot) + \underset{(0.185)}{0.3291}\ln(labor)$$
$$- \underset{(2.32)}{0.1088}\, y20108 + \underset{(0.419)}{0.0154}\, y2008' \ln(captot) - \underset{(0.418)}{0.0094}\, y2008' \ln(labor)$$
$$R^2=0.99394 \quad n=34$$

The results of the restricted model (5-33) are the following:

$$\widehat{\ln(gva)} = -\underset{(0.200)}{0.0690} + \underset{(0.036)}{0.6959}\ln(captot) + \underset{(0.042)}{0.311}\ln(labor)$$
$$R^2=0.99392 \quad n=34$$

As can be seen, the R^2 of the two models are virtually identical because they differ only from the fifth decimal. It is not surprising, therefore, that the *F* statistic for testing the null hypothesis (5-32) takes a value close to 0:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k)} = \frac{(0.99394 - 0.99392)/3}{(1-0.99394)/(34-6)} = 0.0308$$

Thus, the alternative hypothesis that there is structural change in the productive economy of the Spanish regions between 1995 and 2008 is rejected for any significance level.

5.6.2 Using separate regressions: The Chow test

This test was introduced by the econometrician Chow (1960). He considered the problem of testing the equality of two sets of regression coefficients. In the Chow test, the *restricted* model is the same as in the case of using dummy variables to distinguish between groups. The *unrestricted* model, instead of distinguishing the behaviour of the two groups by using dummy variables, consists simply of separate regressions. Thus, in the wage determination example, the *unrestricted* model consists of two equations:

$$\begin{aligned} female: &\quad wage = \beta_{11} + \beta_{21} educ + u \\ male: &\quad wage = \beta_{12} + \beta_{22} educ + u \end{aligned} \qquad (5\text{-}34)$$

If we estimate both equations by *OLS*, we can show that the *RSS* of the unrestricted model, RSS_{UR}, is equal to the sum of the *RSS* obtained from the estimates for women, RSS_1, and for men, RSS_2. That is to say,

$$RSS_{UR}=RSS_1+RSS_2$$

The null hypothesis states that the parameters of the two equations in (5-34) are equal. Therefore

$$H_0 : \begin{cases} \beta_{11} = \beta_{12} \\ \beta_{21} = \beta_{22} \end{cases}$$
$$H_1 : \text{No } H_0$$

By applying the null hypothesis to model (5-34), you get model (5-17), which is the restricted model. The estimation of this model for the whole sample is usually called

the *pooled* (*P*) regression. Thus, we will consider that the RSS_R and RSS_P are equivalent expressions.

Therefore, the *F* statistic will be the following:

$$F = \frac{\left[RSS_P - \left(RSS_1 + RSS_2\right)\right]/k}{\left[RSS_1 + RSS_2\right]/\left[n-2k\right]} \qquad (5\text{-}35)$$

It is important to remark that, under the null hypothesis, the error variances for the groups must be equal. Note that we have *k* restrictions: the slope coefficients (interactions) plus the intercept. Note also that in the unrestricted model we estimate two different intercepts and two different slope coefficients, and so the *df* of the model are $n-2k$.

One important limitation of the Chow test is that under the null hypothesis there are no differences at all between the groups. In most cases, it is more interesting to allow partial differences between both groups as we have done using dummy variables.

The Chow test can be generalized to more than two groups in a natural way. From a practical point of view, to run separate regressions for each group to perform the test is probably easier than using dummy variables.

In the case of three groups, the *F* statistic in the Chow test will be the following:

$$F = \frac{\left[RSS_P - (RSS_1 + RSS_2 + RSS_3)\right]/2\times k}{(RSS_1 + RSS_2 + RSS_3)/(n-3k)} \qquad (5\text{-}36)$$

Note that, as a general rule, the number of the *df* of the numerator is equal to the (number of groups-1)$\times k$, while the number of the *df* of the denominator is equal to *n* minus (number of groups)$\times k$.

EXAMPLE 5.15 Another way to approach the question of wage determination by gender

Using the same sample as in example 5.1 (file *wage02sp*), we have obtained the estimation of the equations in (5-34), taking log for wage, for men and women, which taken together gives the estimation of the *unrestricted* model:

Female equation $\quad \widehat{\ln(wage)} = \underset{(0.042)}{1.407} + \underset{(0.0041)}{0.0566}\,educ$

RSS=104 $\quad R^2$=0.236 $\quad n$=617

Male equation $\quad \widehat{\ln(wage)} = \underset{(0.031)}{1.739} + \underset{(0.0032)}{0.0539}\,educ$

RSS=289 $\quad R^2$=0.175 $\quad n$=1383

The *restricted* model, estimated in example 5.4, has the same configuration as the equations in (5-34) but in this case refers to the whole sample. Therefore, it is the pooled regression corresponding to the restricted model. The *F* statistic takes the value

$$F = \frac{\left[RSS_P - (RSS_F + RSS_M)\right]/k}{RSS_F + RSS_M)/(n-2k)} = \frac{\left[433-(104+289)\right]/2}{(104+289)/(2000-2\times 2)} = 102$$

The *F* statistic must be, and is, the same as in example 5.12. The conclusions are therefore the same.

EXAMPLE 5.16 Is the model of wage determination the same for different firm sizes?

In other examples the intercept, or the slope on education, was different for three different firm sizes (small, medium and large). Now we shall consider a completely different equation for each firm size. Therefore, the unrestricted model will be composed by three equations:

$$\begin{aligned} &samall: \ln(wage) = \beta_{11} + \delta_{11}\, female + \beta_{21} edu + u \\ &medium: \ln(wage) = \beta_{12} + \delta_{12}\, female + \beta_{22} edu + u \\ &large: \ln(wage) = \beta_{13} + \delta_{13}\, female + \beta_{23} edu + u \end{aligned} \qquad (5\text{-}37)$$

The null and the alternative hypothesis will be the following:

$$H_0: \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} \\ \delta_{11} = \delta_{12} = \delta_{13} \\ \beta_{21} = \beta_{22} = \beta_{23} \end{cases}$$
$$H_1: \text{No } H_0$$

Given this null hypothesis, the restricted model is model (5-2).

The estimations of the three equations of (5-37), by using file *wage02sp*, are the following:

small $\quad \widehat{\ln(wage)} = \underset{(0.034)}{1.706} - \underset{(0.031)}{0.249}\, female + \underset{(0.0038)}{0.0396}\, educ$

RSS=121 $\quad R^2$=0.160 $\quad n$=801

medium $\quad \widehat{\ln(wage)} = \underset{(0.051)}{1.934} - \underset{(0.039)}{0.422}\, female + \underset{(0.0046)}{0.0548}\, educ$

RSS =123 $\quad R^2$=0.302 $\quad n$=590

large $\quad \widehat{\ln(wage)} = \underset{(0.046)}{1.749} - \underset{(0.039)}{0.303}\, female + \underset{(0.0044)}{0.0554}\, educ$

RSS =114 $\quad R^2$=0.273 $\quad n$=609

The pooled regression has been estimated in example 5.1. The F statistic takes the value

$$F = \frac{\left[RSS_P - (RSS_S + RSS_M + RSS_L)\right] / 2 \times k}{(RSS_S + RSS_M + RSS_L) / (n - 3k)}$$
$$= \frac{\left[393 - (121 + 123 + 114)\right] / 6}{(121 + 123 + 114) / (2000 - 3 \times 3)} = 32.4$$

For any level of significance, we reject that the equations for wage determination are the same for different firm sizes.

EXAMPLE 5.17 Is the Pinkham model valid for the four periods?

In example 5.5, we introduced time dummy variables and we tested whether the intercept was different for each period. Now, we are going to test whether the whole model is valid for the four periods considered. Therefore, the unrestricted model will be composed by four equations:

$$\begin{aligned} &1907\text{-}1914 \quad sales_t = \beta_{11} + \beta_{21} advexp_t + \beta_{31}\, sales_{t-1} + u_t \\ &1915\text{-}1925 \quad sales_t = \beta_{12} + \beta_{22} advexp_t + \beta_{32}\, sales_{t-1} + u_t \\ &1926\text{-}1940 \quad sales_t = \beta_{13} + \beta_{23} advexp_t + \beta_{33}\, sales_{t-1} + u_t \\ &1941\text{-}1960 \quad sales_t = \beta_{14} + \beta_{24} advexp_t + \beta_{34}\, sales_{t-1} + u_t \end{aligned} \qquad (5\text{-}38)$$

The null and the alternative hypothesis will be the following:

$$H_0: \begin{cases} \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} \\ \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} \\ \beta_{31} = \beta_{32} = \beta_{33} = \beta_{34} \end{cases}$$
$$H_1: \text{No } H_0$$

Given this null hypothesis, the restricted model is the following model:

$$sales_t = \beta_1 + \beta_2 advexp_t + \beta_3\, sales_{t-1} + u_t \qquad (5\text{-}39)$$

The estimations of the four equations of (5-38) are the following:

1907-1914 $\quad \widehat{sales}_t = \underset{(603)}{64.84} + \underset{(1.025)}{0.9149}\, advexp + \underset{(0.425)}{0.4630}\, sales_{t-1} \quad SSR = 36017 \quad n = 7$

1915-1925 $\widehat{sales}_t = \underset{(190)}{221.5} + \underset{(0.557)}{0.1279}\, advexp + \underset{(0.300)}{0.9319}\, sales_{t-1}$ $SSR = 400605$ $n = 11$

1926-1940 $\widehat{sales}_t = \underset{(112)}{446.8} + \underset{(0.115)}{0.4638}\, advexp + \underset{(0.0827)}{0.4445}\, sales_{t-1}$ $SSR = 201614$ $n = 15$

1941-1960 $\widehat{sales}_t = -\underset{(134)}{182.4} + \underset{(0.241)}{1.6753}\, advexp + \underset{(0.111)}{0.3042}\, sales_{t-1}$ $SSR = 187332$ $n = 20$

The pooled regression, estimated in example 3.4, is the following:

$$\widehat{sales}_t = \underset{(95.7)}{138.7} + \underset{(0.156)}{0.3288}\, advexp + \underset{(0.0915)}{0.7593}\, sales_{t-1} \qquad SSR = 2527215 \quad n = 53$$

The F statistic takes the value

$$F = \frac{\left[SSR_P - (SSR_1 + SSR_2 + SSR_3 + SSR_4)\right] / 3 \times k}{(SSR_1 + SSR_2 + SSR_3 + SSR_4) / (n - 4k)}$$

$$= \frac{\left[2527215 - (36017 + 400605 + 201614 + 187332)\right] / 9}{(36017 + 400605 + 201614 + 187332) / (53 - 4 \times 3)} = 9.16$$

For any level of significance, we reject that the model (5-39) is the same for the four periods considered.

Exercises

Exercise 5.1 Answer the following questions for a model with explanatory dummy variables:

a) What is the interpretation of the dummy coefficients?

b) Why are not included in the model so many dummy variables as categories there are?

Exercise 5.2 Using a sample of 560 families, the following estimations of demand for rental are obtained:

$$\hat{q}_i = \underset{(0.11)}{4.17} - \underset{(0.017)}{0.247}\, p_i + \underset{(0.026)}{0.960}\, y_i$$

$$R^2 = 0.371 \qquad n = 560$$

$$\hat{q}_i = \underset{(0.13)}{5.27} - \underset{(0.030)}{0.221}\, p_i + \underset{(0.031)}{0.920}\, y_i + \underset{(0.120)}{0.341}\, d_i y_i$$

$$R^2 = 0.380$$

where q_i is the log of expenditure on rental housing of the i^{th} family, p_i is the logarithm of rent per m^2 in the living area of the i^{th} family, y_i is the log of household disposable income of the i^{th} family and d_i is a dummy variable that takes value one if the family lives in an urban area and zero in a rural area.

(The numbers in parentheses are standard errors of the estimators.)

a) Test the hypothesis that the elasticity of expenditure on rental housing with respect to income is 1, in the first fitted model.

b) Test whether the interaction between the dummy variable and income is significant. Is there a significant difference in the housing expenditure elasticity between urban and rural areas? Justify your answer.

Exercise 5.3 In a linear regression model with dummy variables, answer the following questions:

a) The meaning and interpretation of the coefficients of dummy variables in models with endogenous variable in logs.

b) Express how a model is affected when a dummy variable is introduced in a multiplicative way with respect to a quantitative variable.

Exercise 5.4 In the context of a multiple linear regression model,

a) What is a dummy variable? Give an example of an econometric model with dummy variables. Interpret the coefficients.

b) When is there perfect multicollinearity in a model with dummy variables?

Exercise 5.5 The following estimation is obtained using data for workers of a company:

$$\widehat{wage}_i = 500 + 50tenure_i + 200college_i + 100male_i$$

where *wage* is the wage in euros per month, *tenure* is the number of years in the company, *college* is a dummy variable that takes value 1 if the worker is graduated from college and 0 otherwise and *male* is a dummy variable which takes value 1 if the worker is male and 0 otherwise.

a) What is the predicted wage for a male worker with six years of *tenure* and *college* education?

b) Assuming that all working women have college education and none of the male workers do, write a hypothetical matrix of regressors (**X**) for six observations. In this case, would you have any problem in the estimation of this model? Explain it.

c) Formulate a new model that allows to establish whether there are wage differentials between workers with primary, secondary and college education.

Exercise 5.6 Consider the following linear regression model:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + u_i \tag{1}$$

where y is the monthly salary of a teacher, x is the number of years of teaching experience y d_1 y d_2 are two dummy variables taking the following values:

$$d_{1i} = \begin{cases} 1 & \text{if the teacher is male} \\ 0 & \text{otherwise} \end{cases} \qquad d_{2i} = \begin{cases} 1 & \text{if the teacher is white} \\ 0 & \text{otherwise} \end{cases}$$

a) What is the reference category in the model?

b) Interpret γ_1 and γ_2. What is the expected salary for each of the possible categories?

c) To improve the explanatory power of the model, the following alternative specification was considered:

$$y_i = \alpha + \beta x_i + \gamma_1 d_{1i} + \gamma_2 d_{2i} + \gamma_3 (d_{1i} d_{2i}) + u_i \tag{2}$$

d) What is the meaning of the term $(d_{1i} d_{2i})$? Interpret γ_3.

e) What is the expected salary for each of the possible categories in model (2)?

Exercise 5.7 Using a sample of 36 observations, the following results are obtained:

$$\hat{y}_t = \underset{(0.12)}{1.10} - \underset{(0.34)}{0.96} x_{t1} - \underset{(3.35)}{4.56} x_{t2} + \underset{(0.07)}{0.34} x_{t3}$$

$$\sum_{t=1}^{n} \left(\hat{y}_t - \bar{y} \right)^2 = 109.24 \quad \sum_{t=1}^{n} \hat{u}_t^2 = 20.22$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test the individual significance of the coefficient associated with x_2.
b) Calculate the coefficient of determination, R^2, and explain its meaning.
c) Test the joint significance of the model.
d) Two additional regressions, with the same specification, were made for the two categories *A* and *B* included in the sample (n_1=21 y n_2=15). In these estimates the following *RSS* were obtained: 11.09 y 2.17, respectively. Test if the behavior of the endogenous variable is the same in the two categories.

Exercise 5.8 To explain the time devoted to sport (*sport*), the following model was formulated

$$sport = \beta_1 + \delta_1 female + \gamma_1 smoker + \beta_2 age + u \qquad (1)$$

where sport is the minutes spent on sports a day, on average; *female* and *smoker* are dummy variables taking the value 1 if the person is a woman or smoker of at least five cigarettes per day, respectively.

a) Interpret the meaning of δ_1, γ_1 and β_2.
b) What is the expected time spent on sports activities for all possible categories?
c) To improve the explanatory power of the model, the following alternative specification was considered:

$$\begin{aligned} depor = {} & \beta_1 + \delta_1 mujer + \gamma_1 fumador + \varphi_1 mujer' \ fumador \\ & + \delta_2 mujer' \ edad + \gamma_2 fumador' \ edad + \beta_2 edad + u \end{aligned} \qquad (2)$$

In model (2), what is the meaning of φ_1? What is the meaning of δ_2 and γ_2?
d) What are the possible marginal effects of *sport* with respect to *age* in the model (2)? Describe them.

Exercise 5.9 Using information for Spanish regions in 1995 and 2000, several production functions were estimated.

For the whole of the two periods, the following results were obtained:

$$\widehat{\ln(q)} = 5.72 + 0.26\ln(k) + 0.75\ln(l) - 1.14 f + 0.11 f' \ln(k) - 0.05 f' \ln(l) \qquad (1)$$

$$R^2 = 0.9594 \quad \bar{R}^2 = 0.9510 \quad RSS = 0.9380 \quad n = 34$$

$$\widehat{\ln(q)} = 3.91 + 0.45\ln(k) + 0.60(l) \qquad (2)$$

$$R^2 = 0.9567 \quad \bar{R}^2 = 0.9525 \quad RSS = 1.0007$$

Moreover, the following models were estimated separately for each of the years:

1995
$$\widehat{\ln(q)} = 5.72 + 0.26\ln(k) + 0.75l \qquad (3)$$

$$R^2 = 0.9527 \quad \bar{R}^2 = 0.9459 \quad RSS = 0.6052$$

2000
$$\widehat{\ln(q)} = 4.58 + 0.37\ln(k) + 0.70l \qquad (4)$$

$$R^2 = 0.9629 \quad \bar{R}^2 = 0.9555 \quad RSS = 0.3331$$

where q is output, k is capital, l is labor and f is a dummy variable that takes the value 1 for 1995 data and 0 for 2000.

a) Test whether there is structural change between 1995 and 2000.
b) Compare the results of estimations (3) and (4) with estimation (1).
c) Test the overall significance of model (1).

Exercise 5.10 With a sample of 300 service sector firms, the following *cost* function was estimated:

$$\widehat{cost}_i = \underset{(0.025)}{0.847 + 0.899\, qty_i} \qquad RSS = 901.074 \qquad n = 300$$

where qty_i is the quantity produced.

The 300 firms are distributed in three big areas (100 in each one). The following results were obtained:

$$\text{Area 1: } \widehat{cost}_i = \underset{(0.038)}{1.053 + 0.876\, qty_i} \qquad \hat{s}^2 = 0.457$$

$$\text{Area 2: } \widehat{cost}_i = \underset{(0.096)}{3.279 + 0.835\, qty_i} \qquad \hat{s}^2 = 3.154$$

$$\text{Area 3: } \widehat{cost}_i = \underset{(0.10)}{5.279 + 0.984\, qty_i} \qquad \hat{s}^2 = 4.255$$

a) Calculate an unbiased estimation of σ^2 in the cost function for the sample of 300 firms.
b) Is the same cost function valid for the three areas?

Exercise 5.11 To study spending on magazines (*mag*), the following models have been formulated:

$$\ln(mag) = \beta_1 + \beta_2 \ln(inc) + \beta_3 age + \beta_4 male + u \qquad (1)$$

$$\ln(mag) = \beta_1 + \beta_2 \ln(inc) + \beta_3 age + \beta_4 male + \beta_5 prim + \beta_6 sec + u \qquad (2)$$

where *inc* is disposable income, *age* is age in years, *male* is a dummy variable that takes the value 1 if he is male, *prim* and *sec* are dummy variables that take the value 1 when the individual has reached, at most, primary and secondary level respectively.

With a sample of 100 observations, the following results have been obtained

$$\widehat{\ln(mag)}_i = \underset{(0.124)}{1.27} + \underset{(0040)}{0.756 \ln(inc_i)} + \underset{(0.001)}{0.031 age_i} - \underset{(0.022)}{0.017 male_i}$$

$$RSS=1.1575 \qquad R^2=0.9286$$

$$\widehat{\ln(mag_i)} = \underset{(0.020)}{1.26} + \underset{(0.007)}{0.811 \ln(inc_i)} + \underset{(0.0002)}{0.030\, age_i} + \underset{(0.003)}{0.003 male_i} - \underset{(0.004)}{0.250\, prim_i} + \underset{(0.005)}{0.108 sec_i}$$

$$RSS=0.0306 \qquad R^2=0.9981$$

a) Is education a relevant factor to explain spending on magazines? What is the reference category for education?
b) In the first model, is spending on magazines higher for men than for women? Justify your answer.
c) Interpret the coefficient on the *male* variable in the second model. Is spending on magazines higher for men than for women? Compare with the result obtained in section *a)*.

Exercise 5.12 Let *fruit* be the expenditure on fruit expressed in euros over a year carried out by a household and let r_1, r_2, r_3, and r_4 be dichotomous variables which reflect the four regions of a country.

a) If you regress *fruit* only on r_1, r_2, r_3, and r_4 without an intercept, what is the interpretation of the coefficients?
b) If you regress *fruit* only on r_1, r_2, r_3, and r_4 with an intercept, what would happen? Why?
c) If you regress *fruit* only on r_2, r_3, and r_4 without an intercept, what is the interpretation of the coefficients?
d) If you regress *fruit* only on r_1- r_2, r_2, r_4-r_3, and r_4 without an intercept, what is the interpretation of the coefficients?

Exercise 5.13 Consider the following model

$$wage = \beta_1 + \delta_1 female + \beta_2 educ + u$$

Now, we are going to consider three possibilities of defining the *female* dummy variable.

1) $female = \begin{cases} 1 & \text{for female} \\ 0 & \text{for male} \end{cases}$ 2) $female = \begin{cases} 2 & \text{for female} \\ 1 & \text{for male} \end{cases}$ 3) $female = \begin{cases} 2 & \text{for female} \\ 0 & \text{for male} \end{cases}$

a) Interpret the dummy variable coefficient for each definition.
b) Is one dummy variable definition preferable to another? Justify the answer.

Exercise 5.14 In the following regression model:

$$wage = \beta_1 + \delta_1 female + u$$

where *female* is a dummy variable, taking value 1 for female and value 0 for a male.

Prove that applying the *OLS* formulas for simple regression you obtain that

$$\hat{\beta}_1 = \overline{wage_M}$$

$$\hat{\delta}_1 = \overline{wage_F} - \overline{wage_M}$$

where *F* indicates female and *M* male.

In order to obtain a solution, consider that in the sample there are n_1 females and n_2 males: the total sample is $n= n_1+n_2$.

Exercise 5.15 The data of this exercise were obtained from a controlled marketing experiment in stores in Paris on coffee expenditure, as reported in A. C. Bemmaor and D. Mouchoux, "*Measuring the Short-Term Effect of In-Store Promotion and Retail Advertising on Brand Sales: A Factorial Experiment'*, Journal of Marketing Research, 28 (1991), 202–14. In this experiment, the following model has been formulated to explain the quantity sold of coffee per week:

$$\ln(coffqty) = \beta_1 + \delta_1 advert + \beta_2 \ln(coffpric) + \delta_2 advert \times \ln(coffpric) + u$$

where *coffpric* takes three values: 1, for the usual price, 0.95 and 0.85; *advert* is a dummy variable that takes value 1 if there is advertising in this week and 0 if there is not. The experiment lasted for 18 weeks. The original model and three other models were estimated, using file *coffee2*:

1)
$$\widehat{\ln(coffqty_i)} = \underset{(0.04)}{5.85} + \underset{(0.099)}{0.2565}\, advert_i - \underset{(0.450)}{3.9760}\ln(coffpric_i) - \underset{(0.883)}{1.069}\, advert_i' \ln(coffpric_i)$$
$$R^2 = 0.9468 \quad n = 18$$

2)
$$\widehat{\ln(coffqty_i)} = \underset{(0.04)}{5.83} + \underset{(0.057)}{0.3559}\, advert_i - \underset{(0.393)}{4.2539}\ln(coffpric_i)$$
$$R^2 = 0.9412 \quad n = 18$$

3)
$$\widehat{\ln(coffqty_i)} = \underset{(0.04)}{5.88} - \underset{(0.513)}{3.6939}\ln(coffpric_i) - \underset{(0.582)}{2.9575}\, advert_i' \ln(coffpric_i)$$
$$R^2 = 0.9214 \quad n = 18$$

4)
$$\widehat{\ln(coffqty_i)} = \underset{(0.07)}{5.89} - \underset{(0.674)}{5.1727}\ln(coffpric)$$
$$R^2 = 0.7863 \quad n = 18$$

a) In model (2), what is the interpretation of the coefficient on *advert*?

b) In model (3), what is the interpretation of the coefficient on $advert \times \ln(coffpric$?

c) In model (2), does the coefficient on *advert* have a significant positive effect at 5% and at 1%?

d) Is model (4) valid for weeks with advertising and for weeks without advertising?

e) In model (1), is the intercept the same for weeks with advertising and for weeks without advertising?

f) In model (3), is the coffee demand/price elasticity different for weeks with advertising and for weeks without advertising?

g) In model (4), is the coffee demand/price elasticity smaller than -4?

Exercise 5.16 (Continuation of exercise 4.39). Using file *timuse03*, the following models have been estimated:

$$\widehat{housework}_i = \underset{(23)}{132} + \underset{(1.497)}{2.787}\, educ_i + \underset{(0.308)}{1.847}\, age_i - \underset{(0.023)}{0.2337}\, paidwork_i \qquad (1)$$
$$R^2 = 0.142 \quad n = 1000$$

$$\widehat{housework}_i = -\underset{(22.29)}{3.02} + \underset{(1.356)}{3.641}\, educ_i + \underset{(0.279)}{1.775}\, age_i - \underset{(0.021)}{0.1568}\, paidwork_i + \underset{(2.16)}{32.11}\, female_i \qquad (2)$$
$$R^2 = 0.298 \quad n = 1000$$

$$\widehat{housework}_i = -\underset{(35.18)}{8.04} + \underset{(2.352)}{4.847}\, educ_i + \underset{(0.502)}{1.333}\, age_i - \underset{(0.032)}{0.0871}\, paidwork_i + \underset{(8.15)}{32.75}\, female_i$$
$$- \underset{(0.546)}{0.1650}\, educ_i \times female_i + \underset{(0.112)}{0.1019}\, age_i \times female_i - \underset{(0.009)}{0.02625}\, paidwork_i \times female_i \qquad (3)$$
$$R^2 = 0.306 \quad n = 1000$$

a) In model (1), is there a statistically significant tradeoff between time devoted to paid work and time devoted to housework?

b) All other factors being equal and taking as a reference model (2), is there evidence that women devote more time to housework than men?

c) Compare the R^2 of models (1) and (2). What is your conclusion?

d) In model (3), what is the marginal effect of time devoted to housework with respect to time devoted to paid work?
e) Is interaction between *paidwork* and gender significant?
f) Are the interactions between gender and the quantitative variables of the model jointly significant?

Exercise 5.17 ***Using data from*** *Bolsa de Madrid* (***Madrid Stock Exchange***) on November 19, 2011 (file *bolmad11*), the following models have been estimated:

$$\widehat{\ln(marktval_i)} = \underset{(0.243)}{1.784} + \underset{(0.179)}{0.6998}\, ibex35_i + \underset{(0.0369)}{0.6749}\ln(bookval_i) \qquad (1)$$

$$RSS=35.69 \quad R^2=0.8931 \quad n=92$$

$$\widehat{\ln(marktval_i)} = \underset{(0.275)}{1.828} + \underset{(0.778)}{0.4236}\, ibex35_i + \underset{(0.0423)}{0.6678}\ln(bookval_i) + \underset{(0.088)}{0.0310}\, ibex35_i' \ln(bookval_i) \qquad (2)$$

$$RSS=35.622 \quad R^2=0.8933 \quad n=92$$

$$\widehat{\ln(marktval_i)} = \underset{(0.310)}{2.323} + \underset{(0.785)}{0.1987}\, ibex35_i + \underset{(0.0405)}{0.6688}\ln(bookval_i) + \underset{(0.089)}{0.0369}\, ibex35_i' \ln(bookval_i) - \underset{(0.236)}{0.6613}\, services_i - \underset{(0.221)}{0.6698}\, consump_i - \underset{(0.263)}{0.1931}\, energy_i - \underset{(0.207)}{0.3895}\, industry_i - \underset{(0.324)}{0.7020}\, itt_i \qquad (3)$$

$$RSS=30.781 \quad R^2=0.9078 \quad n=92$$

$$\widehat{\ln(marktval_i)} = \underset{(0.234)}{1.366} + \underset{(0.0305)}{0.7658}\ln(bookval_i) \qquad (4)$$

$$RSS=41.625 \quad R^2=0.8753 \quad n=92$$

For *finance*=1 $$\widehat{\ln(markval_i)} = \underset{(0.560)}{0.558} + \underset{(0.0702)}{0.9346}\ln(bookval_i) \qquad (5)$$

$$RSS=2.7241 \quad R^2=0.9415 \quad n=13$$

where

- ***marktval*** is the capitalization value of a company.
- *bookval* is the book value of a company.
- *ibex35* is a dummy variable that takes the value 1 if the corporation is included in the selective Ibex 35.
- *services*, *consumption*, *energy*, *industry* and *itc* (information technology and communication) are dummy variables. Each of them takes the value 1 if the corporation is classified in this sector in Bolsa de Madrid. The category of reference is *finance*.

a) In model (1), what is interpretation of the coefficient on *ibex35*?
b) In model (1), is the *marktval/bookval* elasticity equal to 1?
c) In model (2), is the elasticity *marktval/*bookval the same for all corporations included in the sample?
d) Is model (4) valid both for corporations included in ibex 35 and for corporations excluded?
e) In model (3), what is interpretation of the coefficient on *consump*?
f) Is the coefficient on *consump* significatively negative?

g) Is the introduction of dummy variables for different sectors statistically justifiable?

h) Is the *marktval/bookval* elasticity for the financial sector equal to 1?

Exercise 5.18 (Continuation óf exercise 4.37). Using file *rdspain*, the equations which appear in the attached table have been estimated.

The following variables appear in the table:

- *rdintens* is expenditure on research and development (R&D) as a percentage of sales,
- *sales* are measured in millions of euros,
- *exponsal* is exports as a percentage of sales;
- *medtech* and *hightech* are two dummy variables which reflects if the firm belongs to a medium or a high technology sector. The reference category corresponds to the firms with *low* technology,
- *workers* is the number of workers.

	(1) *rdintens*	(2) *rdintens*	(3) *rdintens*	(4) *rdintens* for *hightech*=1	(5) *rdintens* for *medtech*=1	(6) *rdintens* for *lowtech*=1
exponsal	0.0136	0.0101	0.00968	0.00584	0.0116	0.00977
	(0.00195)	(0.00193)	(0.00189)	(0.00792)	(0.00300)	(0.00169)
workers	0.000433	0.000392	0.000394	0.00196	0.0000563	0.000393
	(0.0000740)	(0.0000725)	(0.000208)	(0.000338)	(0.0000815)	(0.000121)
hightech		1.448	0.976			
		(0.141)	(0.151)			
medtech		0.361	0.472			
		(0.109)	(0.112)			
hightech× *workers*			0.00153			
			(0.000271)			
medtech× *workers*			-0.000326			
			(0.000222)			
intercept	0.394	0.137	0.143	1.211	0.577	0.142
	(0.0598)	(0.0691)	(0.0722)	(0.313)	(0.103)	(0.0443)
n	1983	1983	1983	296	616	1071
R^2	0.0507	0.0986	0.138	0.113	0.0278	0.0459
RSS	9282.7	8815.0	8425.3	4409.0	2483.6	1527.5
F	52.90	54.06	52.90	18.71	8.776	25.72
df_n	2	4	6	2	2	2
df_d	1980	1978	1976	293	613	1068

Standard errors in parentheses

a) In model (2), all other factors being equal, is there evidence that expenditure on research and development (expressed as a percentage of sales) in high technology firms is greater than in low technology firms? How strong is the evidence?

b) In model (2), all other factors being equal, is there evidence that *rdintens* in medium technology firms is equal to low technology firms? How strong is the evidence?

c) Taking as reference model (2), if you had to test the hypothesis that *rdintens* in high technology firms is equal to medium technology firms, formulate a model that allows you to test this hypothesis without using information on covariance matrix of the estimators

d) Is the influence of workers on *rdintens* associated with the level of technology in the firms?

e) Is the model (1) valid for all firms regardless of their technological level?

Exercise 5.19 To explain the overall satisfaction of people (*stsfglo*), the following model were estimated using data from the file *hdr2010*:

$$\widehat{stsfglo}_i = \underset{(0.584)}{-0.375} + \underset{(0.00000617)}{0.0000207}\, gnipc_i + \underset{(0.009)}{0.0858}\, lifexpec_i \qquad (1)$$

$$R^2 = 0.642 \quad n = 144$$

$$\widehat{stsfglo}_i = \underset{(0.897)}{2.911} + \underset{(0.00000572)}{0.0000381}\, gnipc_i + \underset{(0.18)}{1.215}\, lifexpec_i + \underset{(0.179)}{1.215}\, dlatam_i - \underset{(0.259)}{0.7901}\, dafrica_i \qquad (2)$$

$$R^2 = 0.748 \quad n = 144$$

$$\widehat{stsfglo}_i = \underset{(1.014)}{1.701} + \underset{(0.000006)}{0.0000327}\, gnipc_i + \underset{(0.0147)}{0.0527}\, lifexpec_i + \underset{(0.177)}{1.166}\, dlatam_i - \underset{(1.712)}{3.096}\, dafrica_i + \underset{(0.0000456)}{0.0000673}\, gnipc_i \times dafrica_i - \underset{(0.0295)}{0.0699}\, lifexpec_i \times dafrica_i \qquad (3)$$

$$R^2 = 0.760 \quad n = 144$$

where

- *gnipc* is gross national income per capita expressed in PPP 2008 US dollar terms,
- *lifexpec* is life expectancy at birth, i.e., number of years a newborn infant could be expected to live,
- *dafrica* is a dummy variable that takes value 1 if the country is in Africa,
- *dlatam* is a dummy variable that takes value 1 if the country is in Latin America.

a) In model (2), what is the interpretation of the coefficients on *dlatam* and *dafrica*?

b) In model (2), do *dlatam* and *dafrica* individually have a significant positive influence on global satisfaction?

c) In model (2), do *dlatam* and *dafrica* have a joint influence on global satisfaction?

d) Is the influence of life expectancy on global satisfaction smaller in Africa than in other regions of the world?

e) Is the influence of the variable *gnipc* greater in Africa than in other regions of the world at 10%?

f) Are the interactions of people living in Africa and the variables *gnipc* and *lifexpec* jointly significant?

Exercise 5.20 The equations which appear in the attached table have been estimated using data from the file *timuse03*. This file contains 1000 observations corresponding to a random subsample extracted from the time use survey for Spain carried out in 2002-2003.

The following variables appear in the table:

- *educ* is years of education attained,
- *sleep*, *paidwork* and *unpaidwrk* are measured in minutes per day,
- *female*, *workday* (Monday to Friday), *spaniard* and *houswife* are dummy variables.

a) In model (1), is there a statistically significant tradeoff between time devoted to paid work and time devoted to sleep?
b) In model (1), is the coefficient on *unpaidwk* statistically significant?
c) In model (1), is there evidence that women sleep more than men?
d) In model (2), are *workday* and *spaniard* individually significant? Are they jointly significant?
e) Is the coefficient on *housewife* statistically significant?
f) Are the interactions between *female* and *educ*, *paidwork* and *unpaidwk* jointly significant?

	(1) Sleep	(2) Sleep	(3) Sleep	(4) Sleep	(5) Sleep	(6) sleep
educ	-4.669 (0.916)	-4.787 (0.912)	-4.805 (0.912)	-4.754 (0.913)	-4.782 (0.917)	-4.792 (0.917)
persinc	0.0238 (0.00587)	0.0207 (0.00600)	0.0195 (0.00607)	0.0210 (0.00601)	0.0208 (0.00601)	0.0208 (0.00601)
age	0.854 (0.174)	0.879 (0.174)	0.895 (0.174)	0.884 (0.174)	0.879 (0.174)	0.891 (0.302)
paidwork	-0.258 (0.0150)	-0.247 (0.0159)	-0.246 (0.0159)	-0.248 (0.0160)	-0.246 (0.0210)	-0.247 (0.0159)
unpaidwk	-0.205 (0.0184)	-0.198 (0.0184)	-0.188 (0.0196)	-0.224 (0.0365)	-0.198 (0.0185)	-0.198 (0.0184)
female	4.161 (1.465)	3.588 (1.467)	3.981 (1.493)	2.485 (1.975)	3.638 (1.691)	3.727 (3.287)
workday		-19.31 (7.168)	-19.46 (7.165)	-19.47 (7.171)	-19.30 (7.173)	-19.30 (7.172)
spaniard		-47.50 (19.99)	-46.88 (19.98)	-47.90 (20.00)	-47.63 (20.10)	-47.51 (20.00)
houswife			-14.71 (10.42)			
unpaidwk ×*female*				0.00607 (0.00726)		
paidwork ×***female***					-0.000324 **(0.00540)**	
age× *female*						-0.00308 (0.0652)
intercept	588.9 (13.62)	648.3 (24.34)	646.6 (24.36)	651.9 (24.73)	648.2 (24.39)	647.8 (26.40)
N	1000	1000	1000	1000	1000	1000
R^2	0.316	0.325	0.326	0.325	0.325	0.325
RSS	9913901.3	9789312.3	9769648.2	9782424.0	9789276.9	9789290.3
F	76.58	59.62	53.27	53.06	52.95	52.95
df_n	6	8	9	9	9	9
df_d	993	991	990	990	990	990

Standard errors in parentheses

Exercise 5.21 To study infant mortality in the world, the following models have been estimated using data from the file *hdr2010*:

$$\widehat{deathinf}_i = \underset{(4.58)}{93.02} - \underset{(0.0002)}{0.00037}\, gnipc_i - \underset{(0.1866)}{0.6046}\, physicn_i - \underset{(0.003)}{0.003}\, contrcep_i \quad (1)$$

$$RSS=40285 \quad R^2=0.6598 \quad n=108$$

$$\widehat{deathinf}_i = \underset{(5.96)}{78.55} - \underset{(0.0002)}{0.00042}\, gnipc - \underset{(0.1879)}{0.3809}\, physicn_i - \underset{(0.1042)}{0.6989}\, contrcep_i + \underset{(5.05)}{17.92}\, dafrica \quad (2)$$

$$RSS=35893 \quad R^2=0.6851 \quad n=108$$

$$\widehat{deathinf}_i = \underset{(6.76)}{72.58} - \underset{(0.0002)}{0.00044}\, gnipc - \underset{(0.1879)}{0.3994}\, physicn_i - \underset{(0.1234)}{0.5857}\, contrcep_i + \underset{(5.05)}{17.92}\, dafrica - \underset{(0.000826)}{0.0000914}\, gnipc' \; dafrica - \underset{(2.2351)}{2.0013}\, physicn' \; dafrica - \underset{(0.2716)}{0.2172}\, contrcep_i' \; dafrica \quad (3)$$

$$RSS=34309 \quad R^2=0.7109 \quad n=108$$

where

- *deatinf* is number of ***infant deaths (***one year or younger***)*** per 1000 live births in 2008,
- *gnipc* is gross national income per capita expressed in *PPP* 2008 US dollar terms,
- *physicn* are physicians per 10,000 people in the period 2000-2009,
- *contrcep* is the contraceptive prevalence rate using any method, expressed as % of married women aged 15–49 for the period 1990-2008,
- *dafrica* is a dummy variable that takes value 1 if the country is in Africa.
  - *a)* In model (1), what is interpretation of the coefficients on *gnipc*, *physicn* and *contrcep*?
  - *b)* In model (2), what is the interpretation of the coefficient on *dafrica*?
  - *c)* In model (2), all other factors being equal, do the countries of Africa have a greater infant mortality than the countries of other regions of the world?
  - *d)* What is the marginal effect of variable *gnipc* on infant mortality in model (3)?
  - *e)* Is the slope corresponding to the regressor *contrcep* significantly greater for the countries of Africa?
  - *f)* Are the slopes corresponding to the regressors *gnipc*, *physicn* and *contrcep* jointly different for the countries of Africa?
  - *g)* Is the model (1) valid for all countries of the world?

Exercise 5.22 Using a random subsample of 2000 observations extracted from the time use surveys for Spain carried out in the periods 2002-2003 and 2009-2010 (file *timus309*), the following models have been estimated to explain time spent watching television:

$$watchtv = \underset{(9.46)}{114} - \underset{(0.620)}{3.523}\, educ + \underset{(0.130)}{1.330}\, age - \underset{(0.0102)}{0.1111}\, paidwork \quad (1)$$

$$R^2 = 0.169 \quad n = 2000$$

$$watchtv = \underset{(9.915)}{127} - \underset{(0.615)}{3.653}\, educ + \underset{(0.129)}{1.291}\, age - \underset{(0.010)}{0.120}\, paidwork - \underset{(4.903)}{25.146}\, female + \underset{(5.247)}{17.137}\, y2009 \quad (2)$$

$$R^2 = 0.184 \quad n = 2000$$

$$watchtv = \underset{(10.01)}{123} - \underset{(0.615)}{3.583}\,educ + \underset{(0.129)}{1.302}\,age - \underset{(0.012)}{0.105}\,paidwork - \underset{(4.899)}{24.869}\,female$$
$$+ \underset{(6.115)}{24.536}\,y2009 - \underset{(0.021)}{0.050}\,y2009 \times paidwork \qquad R^2 = 0.186 \quad n = 2000 \tag{3}$$

where

- *educ* is years of education attained,
- *watchtv* and *paidwork* are measured in minutes per day.
- *female* is a dummy variable that takes value 1 if the interviewee is a female
- *y2009* is a dummy variable that takes value 1 if the survey was carried out in 2008-2009

a) In model (1), what is interpretation of the coefficient on *educ*?

b) In model (1), is there a statistically significant tradeoff between time devoted to work and time devoted to watching television?

c) All other factors being equal and taking as reference model (2), is there evidence that men watch television more than women? How strong is the evidence?

d) In model (2), what is the estimated difference in watching television between females surveyed in 2008-2009 and males surveyed in 2002-2003? Is this difference statistically significant?

e) In model (3), what is the marginal effect of time devoted to paid work on time devoted to watching television?

f) Is there a significant interaction between the year of the survey and time devoted to paid work?

Exercise 5.23 Using the file *consumsp*, the following models were estimated to analyze if the entry of Spain into the European community in 1986 had any impact on the behavior of Spanish consumers:

$$\bar{consp}c_t = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\,incpc_t + \underset{(0.0903)}{0.5771}\,conspc_{t-1} \tag{1}$$
$$R^2=0.9967 \qquad RSS=1891320 \qquad n=56$$

$$\bar{consp}c_t = \underset{(108)}{-102.4} + \underset{(0.0879)}{0.3573}\,incpc_t + \underset{(0.0901)}{0.5992}\,conspc_{t-1} + \underset{(92.56)}{148.60}\,y1986_t \tag{2}$$
$$R^2=0.9968 \qquad RSS=1802007 \qquad n=56$$

$$\bar{consp}c_t = \underset{(114)}{79.17} + \underset{(0.1100)}{0.5181}\,incpc_t + \underset{(0.1199)}{0.4186}\,conspc_{t-1} + \underset{(456.3)}{819.82}\,y1986_t$$
$$- \underset{(0.2338)}{0.5403}\,incpc_t \times y1986_t + \underset{(0.2182)}{0.5424}\,conspc_{t-1} \times y1986_t \tag{3}$$
$$R^2=0.9972 \qquad RSS=1600714 \qquad n=56$$

$$\bar{consp}c_t = \underset{(118)}{117.03} + \underset{(0.0968)}{0.3697}\,incpc_t + \underset{(0.1051)}{0.5823}\,conspc_{t-1} + \underset{(348)}{41.62}\,y1986_t$$
$$+ \underset{(0.0326)}{0.0104}\,incpc_t \times y1986_t \tag{4}$$
$$R^2=0.9968 \qquad RSS=1798423 \qquad n=56$$

$$\bar{consp}c_t = \underset{(114)}{120.1} + \underset{(0.0854)}{0.3750}\,incpc_t + \underset{(0.0890)}{0.5758}\,conspc_{t-1} + \underset{(0.0087)}{0.0141}\,incpc_t \times y1986_t \tag{5}$$
$$R^2=0.9968 \qquad RSS=1798927 \qquad n=56$$

(The numbers in parentheses are standard errors of the estimators.)

a) Test in model (5) whether the marginal propensity to consume in the short term was reduced in 1986 and beyond.

b) Are the interactions between *y1986* and the quantitative variables of the model jointly significant?

c) Test whether there was a structural change in the consumption function in 1986.

d) Test whether the coefficient on $conspc_{t-1}$ changed in 1986 and beyond.

e) Was there a gap between consumption before 1986, with respect to 1986 and beyond?

6 RELAXING THE ASSUMPTIONS IN THE LINEAR CLASSICAL MODEL

6.1 Relaxing the assumptions in the linear classical model: an overview

In chapters 2 and 3, single and multiple linear regression models were formulated, including the set of statistical assumptions called the classical linear model (*CLM*) assumptions. Now, let us examine the problems posed by the failure of each one of the *CLM* assumptions and alternative methods for estimating the linear model.

Assumption on the functional form

Assumption 1 postulates the following population model:

$$y = \beta_1 + \beta_2 x_1 + \ldots + \beta_k x_k + u \qquad (6\text{-}1)$$

This assumption specifies what the endogenous variable is and its functional form, as well as what the explanatory variables are and their functional forms. It also states that the model is linear on the parameters

If we estimate a different population model, a misspecification error is made. The consequences of such errors will be discussed in section 6.2.

Assumptions on the regressors

The assumptions 2, 3 and 4 were made on the regressors. In the multiple linear regression, assumption 2 postulated that the values $x_2, x_3, \ldots, x_k$ are fixed in repeated samples, that is to say, the regressors are non-stochastic. This is a reasonable assumption when the regressors are obtained from experiments. However, it is less admissible for variables obtained by observation in a passive way, as in the case of income in the consumption function.

When the regressors are stochastic, the statistical relationship between the regressors and the random disturbance is crucial in building an econometric model. For this reason, an alternative assumption was formulated as 2*: the regressors $x_2, x_3, \ldots, x_k$ are distributed independently of the random disturbance. When we assume this alternative assumption, the inference, conditional on the matrix of regressors, leads to results that are virtually coincident with the case where the matrix **X** is fixed. In other words, in the case of independence between the regressors and the random disturbance, the ordinary least squares method is still the optimal method for estimating the vector of coefficients.

In assumption 3 it was postulated that the matrix of regressors **X** contains no measurement errors. If there are measurement errors, a very serious econometric problem will arise with a complex solution.

Assumption 4 states that there is no exact linear relationship between the regressors, or, in other words, it establishes that there is no perfect multicollinearity in the model. This assumption is necessary to calculate the *OLS* estimators. Perfect multicollinearity is not used in practice. Instead, there is often an approximately linear relationship between the regressors. In this case the estimators obtained will not be accurate, although they still retain the property of being *BLUE* estimators. In other words, the relationship between the regressors makes it difficult to quantify the effect that each one has on the regressand. This is due to the fact that the variances of the estimators are high. When an approximately linear relationship between the regressors exists, multicollinearity is not perfect. Section 6.3 will be devoted to examining the detection of non-perfect multicollinearity, along with some possible solutions

Assumptions on the parameters

In assumption 5 it was assumed that the parameters are not random. The real world suggests that this coefficient constancy is not reasonable. In models using time series data, there are often changes in patterns of behavior over time, which would naturally involve changes in the regression coefficients. In any case, section 5.6 examines the test of structural change which determines whether there has been any change in the parameters over time.

Assumptions on the random disturbance term

In assumption 6 it is assumed that $E(\mathbf{u})=\mathbf{0}$. This assumption is not empirically testable in the general case of models with intercept.

Before moving on to other assumptions on the random disturbance u_i, it should be noted that this is an unobservable variable. Information on u_i is obtained indirectly through the residuals, which will be used for testing the behavior of the disturbances. However, the use of residuals to perform tests on disturbances poses some problems. When the *CLM* assumptions are fulfilled, the random disturbances are neither autocorrelated nor homoskedastic, whereas the residuals are heteroskedastic and autocorrelated under these assumptions. These circumstances are important in the design of statistical tests on heteroskedasticity and no autocorrelation.

If assumptions 7 of homoscedasticity and/or 8 of no autocorrelation are not fulfilled, the least squares estimators are still linear and unbiased but they are not the best.

The assumptions of homoskedasticity and no autocorrelation formulated in chapter 3, respectively, may be formulated together indicating that the covariance matrix of random disturbances is a scalar matrix, i.e.:

$$E(\mathbf{uu}') = \sigma^2\mathbf{I} \qquad (6\text{-}2)$$

When one or both assumptions indicated are not fulfilled, then the covariance matrix will be less restrictive. Thus, we will consider the following covariance matrix of the disturbances:

$$E(\mathbf{uu}') = \sigma^2\mathbf{\Omega} \qquad (6\text{-}3)$$

where the only restriction imposed on $\mathbf{\Omega}$ is that it is a positive definite matrix

When the covariance matrix is a non-scalar matrix such as (6-3), then one can obtain linear, unbiased and best estimators by applying the method of generalized least squares (*GLS*). The expression of these estimators is as follows:

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y} \qquad (6\text{-}4)$$

In practice, formula (6-4) is not directly applied. Instead a two-step process that leads to exactly the same results is applied.

In section 6.5, we will examine the tests to determine whether there is heteroskedasticity, as well as the particularization of the *GLS* method in this case. Section 6.6 will present testing methods and the appropriate treatment of autocorrelation.

Assumption 9 of normality postulated in the *CLM* allows us to make statistical inferences with known distributions. If the normality assumption is not adequate, then the tests will only be approximately valid. In section 6.4, a normality test of the disturbances is used to determine whether this assumption is acceptable.

6.2 Misspecification

Misspecification occurs when we estimate a different model from the population model. The problem in social sciences, and in particular in economics, is that we do not usually know the population model.

Bearing in mind this observation, we shall consider three types of misspecification:

- Inclusion of irrelevant variables.
- Exclusion of relevant variables.
- Incorrect functional form.

6.2.1 Consequences of misspecification

We will examine the consequences of each type of misspecification on the *OLS* estimators

Inclusion of an irrelevant variable

Let us consider, for example, that the population model is the following:

$$y = \beta_1 + \beta_2 x_2 + u \qquad (6\text{-}5)$$

Consequently, the *population regression function* (*PRF*) is given by

$$\mu_y = \beta_1 + \beta_2 x_2 \qquad (6\text{-}6)$$

Now let us suppose that the *sample regression function* (*SRF*) estimated is the following

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} \qquad (6\text{-}7)$$

This is the case of *inclusion of an irrelevant variable*: specifically, in (6-7) we have introduced the irrelevant variable x_3. What are the effects of including an irrelevant variable in the *OLS* estimators?

It can be shown that the estimators corresponding to (6-7) are unbiased, that is to say,

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_2) = \beta_2 \qquad E(\hat{\beta}_3) = 0$$

However, the variances of these estimators will be greater than those obtained by estimating (6-5) in which x_3 is (correctly) omitted.

This result can be extended to the case of including one or more irrelevant variables. In this case *OLS* estimators are unbiased, but with variances greater than when the irrelevant variables are not included in the estimated model.

Exclusion of a relevant variable

Let us consider, for example, that the population model is the following:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \qquad (6\text{-}8)$$

The *PRF* is therefore given by:

$$\mu_y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 \quad (6\text{-}9)$$

Now let us suppose that the *SRF* we estimate, due to ignorance or data unavailability, is the following

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} \quad (6\text{-}10)$$

This is a case of *exclusion of a relevant variable*: in (6-10) we have omitted the relevant variable x_3. Is $\hat{\beta}_2$, obtained by applying *OLS* in (6-10), an unbiased estimator of β_2?

As appendix 6.1 shows, the estimator $\hat{\beta}_2$ is biased. The bias is

$$Bias(\hat{\beta}_2) = \beta_3 \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2} \qquad (6\text{-}11)$$

The bias is null if, according to (6-11), the covariance between x_2 and x_3 is 0. It is important to remark that the ratio

$$\frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}$$

is just the *OLS* slope ($\hat{\delta}_2$) coefficient from regression of x_3 on x_2. That is to say,

$$\hat{x}_2 = \hat{\delta}_1 + \hat{\delta}_2 \hat{x}_2 = \hat{\delta}_1 + \frac{\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)x_{3i}}{\sum_{i=1}^{n}(x_{2i} - \overline{x}_2)^2}\hat{x}_2 \qquad (6\text{-}12)$$

Thus, according to (6-72) - in appendix 6.1-, and (6-12), we can write that

$$E(\tilde{\beta}_2) = \beta_2 + \beta_3 \hat{\delta}_2 \qquad (6\text{-}13)$$

Therefore, the bias is equal to $\beta_3\hat{\delta}_2$. In table 6.1, there is a summary of the sign of the bias in $\tilde{\beta}_2$ when x_3 is omitted in estimating equation. It must be taken into account that the sign of $\hat{\delta}_2$ is the same as the sign of the sample correlation between x_2 and x_3.

TABLE.1. Summary of bias in $\tilde{\beta}_2$ when x_3 is omitted in estimating equation.

	$Corr(x_2,x_3)>0$	$Corr(x_2,x_3)<0$
$\beta_3>0$	Positive bias	Negative bias
$\beta_3<0$	Negative bias	Positive bias

Incorrect functional form

If we use a functional form different from the true population model, then the *OLS* estimators will be biased.

In conclusion, if there is exclusion of relevant variables or/and an incorrect functional form has been used, then the *OLS* estimators will be biased and also inconsistent. Therefore, the conventional inference procedures will be invalidated in these two cases.

6.2.2 Specification tests: the RESET test

To test whether irrelevant variables are included in the model we can apply the exclusion restriction tests, which we have examined in chapter 4.

To test the exclusion of relevant variables or the use of an incorrect functional form, we can apply the RESET (Regression Equation Specification Error Test) test. This test is a general test for specification errors proposed by Ramsey (1969). In order to explain it, consider that the *initial* model is the following:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + u \qquad (6\text{-}14)$$

Now, we introduce an *augmented* model in which two new variables (z_1 and z_2) appear:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 z_1 + \alpha_2 z_2 + u \qquad (6\text{-}15)$$

Taking into account the specification of the two models, the null and alternative hypotheses will be the following:

$$H_0 : \alpha_1 = \alpha_2 = 0$$
$$H_1 : H_0 \text{ is not true} \qquad (6\text{-}16)$$

The crucial question in building the test is to determine the *z* variables or regressors to be introduced. In the case of exclusion of relevant variables, the *z* variables will be the omitted regressors which may be new variables or also squares and powers of previous variables. The test to be applied would be similar to the exclusion tests, but with the roles reversed: the restricted model is now the *initial* model, while the unrestricted model corresponds to the *augmented* model.

In testing for incorrect functional form, consider, for example, that (6-14) is specified instead of the true relationship:

$$\ln(y) = \beta_1 + \beta_2 \ln(x_2) + \beta_3 \ln(x_3) + u \qquad (6\text{-}17)$$

In model (6-17), there is a multiplicative relationship between the regressors. Ramsey took into account that a Taylor series approximation of the multiplicative relationship would yield an expression involving powers and cross-products of the explanatory variables. For this reason, he suggests including, in the augmented model, powers of the predicted values of the dependent variable (which are, of course, linear combinations of power and cross-product terms of the explanatory variables):

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha_1 \hat{y}^2 + \alpha_2 \hat{y}^3 + u \qquad (6\text{-}18)$$

where the $\hat{y}$ ´s are the *OLS* fitted values corresponding to the model (6-14). The superscripts indicate the powers to which these predictions are raised. The first power is not included since it is perfectly collinear with the rest of the regressors of the initial model.

The steps involved in the RESET test are as follows:

Step 1. The *initial* model is estimated and the *fitted values,* $\hat{y}_i$, are calculated.

Step 2. The *augmented* model, which can include one or more powers of $\hat{y}_i$, is estimated.

Step 3. Taking the R^2_{init} corresponding to the initial model and the R^2_{augm} corresponding to the augmented model, the *F* statistic is calculated:

$$F = \frac{(R^2_{augm} - R^2_{init}) / r}{(1 - R^2_{augm}) / (n - h)} \qquad (6\text{-}19)$$

where *r* is the number of new parameters added to the initial model, and *h* is the number of parameters of the augmented model, including the intercept.

Under the null hypothesis, this statistic is distributed as follows:

$$F \mid H_0 : \; F_{r,n-h} \qquad (6\text{-}20)$$

Step 4. For a significance level α, and designating by $F^{\alpha}_{r,n-h}$ the corresponding value in the *F* table, the decision to make is the following:

$$\text{If} \quad F \geq F^{\alpha}_{r,n-h} \quad \text{reject} \quad H_0$$
$$\text{If} \quad F < F^{\alpha}_{r,n-h} \quad \text{not reject} \quad H_0$$

Therefore, high values of the statistic lead to the rejection of the initial model.

In RESET test we test the null hypothesis against an alternative hypothesis that does not indicate what the correct specification should be. This test is therefore a misspecification test which may indicate that there is some form of misspecification but does not give any indication of what the correct specification should be.

EXAMPLE 6.1 Misspecification in a model for determination of wages

Using a subsample of data from the *wage structure survey* of Spain for 2006 (file *wage06sp*), the following model is estimated:

$$\widehat{wage}_i = \underset{(1.55)}{4.679} + \underset{(0.146)}{0.681}educ_i + \underset{(0.071)}{0.293}tenure_i$$
$$R^2=0.249 \quad n=150$$

where *educ* (education) and *tenure* (experience in the firm) are measured in years and *wage* in euros per hour.

Considering that we may have a problem of incorrect functional form, an augmented model is estimated. In this augmented model - besides *educ*, *tenure*, and the intercept - $\widehat{wage}_i^{\,2}$ and $\widehat{wage}_i^{\,3}$ from the initial model are included as regressors. The *F* statistic calculated using the R^2_{init} and R^2_{augm}, according to (6-19), is equal to 4.18. Given that $F^{0.05}_{2,145}$; $F^{0.05}_{2,60} = 3.15$, we reject that, for the levels α=0.05 and α=0.10, the linear form is adequate to explain wage determination. On the contrary, given that $F^{0.01}_{2,145}$; $F^{0.01}_{2,60} = 4.98$ H_0 is not rejected for α=0.01.

6.3 Multicollinearity

6.3.1 Introduction

Perfect multicollinearity is not usually seen in practice, unless the model is wrongly designed as we saw in chapter 5. Instead, an approximately linear relationship between the regressors often exists. In this case, the estimators obtained will generally not be very accurate, despite still being *BLUE*. In other words, the relationship between regressors makes it difficult to quantify accurately the effect each one has on the regressand. This is due to the fact that the variances of the estimators are high. When there is an approximately linear relationship between the regressors, then it is said that there is *not perfect multicollinearity*. The multicollinearity problem arises because there is insufficient information to get an accurate estimation of model parameters.

To analyze the problem of multicollinearity, we will examine the variance of an estimator. In the multiple linear regression model, the estimator of the variance of any slope coefficient - for example, $\hat{\beta}_j$ - is equal, as we saw in (3-68), to

$$\widehat{\text{var}}(\hat{b}_j) = \frac{\hat{s}^2}{nS_j^2(1- R_j^2)} \qquad (6\text{-}21)$$

where $\hat{s}^2$ is the unbiased estimator of σ^2, *n* is the sample size, S_j^2 is the sample variance of the regressor x_j, and R_j^2 is the *R*-squared obtained from regressing x_j on all other *x*'s.

The last of these four factors which determines the value of the variance of $\hat{\beta}_j$, (1- R_j^2), is precisely an indicator of multicollinearity. Multicollinearity arises in estimating β_j when R_j^2 is "close" to one, but there is no absolute number that we can quote to conclude that multicollinearity is really a problem for the precision of the estimators. Although the problem of multicollinearity cannot be clearly defined, it is true that, for estimating β_j, the lower the correlation between x_j and the other independent variables the better. If R_j^2 is equal to 1, then we would have perfect multicollinearity and it is not possible to obtain the estimators of the coefficients. In any case, when one or more R_j^2 are close to 1, multicollinearity is a serious problem. In this case, when making inferences with the model, the following problems arise:

a) The variances of the estimators are very large.

b) The estimated coefficients will be very sensitive to small changes in the data.

6.3.2 Detection

Multicollinearity is a problem of the *sample*, because it is associated with the specific configuration of the sample of the *x*'s. For this reason, there are no statistical tests. (Remember that statistical tests only work with *population* parameters). Instead, many practical rules were developed attempting to determine to what extent multicollinearity seriously affects the inference made with a model. These rules are not always reliable, and in some cases are questionable. In any case, we are going to look at some measures that are very useful to detect the degree of multicollinearity: the *variance inflation factor* (*VIF*) and the *tolerance*, and the *condition number* and the *coefficient variance decomposition*.

Variance inflation factor (VIF) and tolerance

In order to explain the meaning of these measures, let us suppose there is *no* linear relationship between x_j and the other explanatory variables in the model, that is to say, the regressor x_j is *orthogonal* to the remaining regressors. In this case, R_j^2 will be zero and the variance of $\hat{\beta}_j$ will be

$$\widehat{\text{var}}(b_j^*) = \frac{\hat{s}^2}{nS_j^2} \qquad (6\text{-}22)$$

Dividing (6-21) by (6-22), we obtain the variance inflation factor (*VIF*) as

$$VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2} \qquad (6\text{-}23)$$

The *VIF* statistic calculated according to (6-23) is sometimes called "centered *VIF*" to be distinguished from the "uncentered *VIF*" which is interesting in models without intercept. The E-views programme supplies both statistics.

Tolerance, which is the inverse of *VIF*, is defined as

$$Tolerance(\hat{\beta}_j) = \frac{1}{VIF} = 1 - R_j^2 \qquad (6\text{-}24)$$

Thus, $VIF(\hat{\beta}_j)$ is the ratio between the estimated variance and the one that there would have been if x_j was uncorrelated with the other regressors in the model. In other words, the *VIF* shows the extent to which the variance of the estimator is "inflated" as a result of non-orthogonallity of the regressors. It is readily seen that the higher the *VIF* (or the lower the tolerance index), the higher the variance of $\hat{\beta}_j$.

The procedure is to choose each one of the regressors at a time as the dependent variable and to regress them against a constant and the remaining explanatory variables. We would then get *k* values for the *VIF*'s. If any of them is high, then multicollinearity is detected. Unfortunately, however, there is no theoretical indicator to determine whether the *VIF* is "high." Also, there is no theory that tells us what to do if multicollinearity is found.

The variance inflation factor (*VIF*) and the tolerance are both widely used measures of the degree of multicollinearity. Unfortunately, several rules of thumb – most commonly the rule of 10 – associated with the *VIF*– are regarded by many practitioners as a sign of severe or serious multicollinearity (this rule appears in both scholarly articles and advanced statistical textbooks), but this rule has no scientific justification

The problem with the *VIF* (or the tolerance) is that it does not provide any information that could be used to treat the problem.

EXAMPLE 6.2 Analyzing multicollinearity in the case of labor absenteeism

In example 3.1 a model was formulated and estimated, using file *absent*, to explain absenteeism from work as a function of the variables *age*, *tenure* and *wage*.

Table 6.2 provides information on the tolerance and the *VIF* of each regressor. According to these statistics, multicollinearity does not appear to affect the *wage* but there is a certain degree of multicollinearity in the variables *age* and *tenure*. In any case, the problem of multicollinearity in this model does not appear to be serious because all *VIF* are below 5.

TABLE 6.2. Tolerance and *VIF*.

	Collinearity statistics	
	Tolerance	*VIF*
age	0.2346	4.2634
tenure	0.2104	4.7532
wage	0.7891	1.2673

Condition number and coefficient variance decomposition

This method, developed by Belsey *et al*. (1982), is based on the variance decomposition of each regression coefficient as a function of the eigenvalues λ_h of the matrix **X'X** and the corresponding elements of the associate eigenvectors. We will not discuss eigenvalues and eigenvectors here, because they are beyond the scope of this book, but in any case we will see their application.

The *condition number* is a standard measure of ill-conditioning in a matrix. It indicates the potential sensitivity of the computed inverse matrix to small changes in the original matrix (**X'X** in the case of the regression). Multicollinearity reveals its presence

by one or more eigenvalues of **X'X** being "small". The closer a matrix is to singularity the smaller the eigenvalues. The condition number (κ) is defined as the square root of the largest eigenvalue ($\lambda_{\max}$) divided by the smallest eigenvalue ($\lambda_{\min}$):

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{6-25}$$

When there is no multicollinearity at all, then all the eigenvalues and the condition number will be equal to one. As multicollinearity increases, eigenvalues will be both greater and smaller than 1 (eigenvalues close to zero indicate a multicollinearity problem), and the condition number will increase. An informal rule of thumb is that if the condition number is greater than 15, multicollinearity is a concern; if it is greater than 30 multicollinearity is a very serious concern.

The variance of $\hat{\beta}_j$ can be decomposed into the contributions from each one of the eigenvalues and can be expressed in the following way:

$$\text{var}(\hat{\beta}_j) = \sigma^2 \sum_h \frac{u_{jh}^2}{\lambda_h} \tag{6-26}$$

Thus, the proportion of the contribution of eigenvalue λ_h in the variance of $\hat{\beta}_j$ is equal to

$$\phi_{jh} = \frac{\frac{u_{jh}^2}{\lambda_h}}{\sum_{h=0}^{k} \frac{u_{jh}^2}{\lambda_h}} \tag{6-27}$$

High values of ϕ_{jh} indicate that, as a consequence of multicollinearity, there is an inflation of the variance. Given that eigenvalues close to zero indicate a multicollinearity problem, it is important to pay special attention to the contribution of the smallest eigenvalues. The contributions corresponding to the smallest eigenvalue may give a clue of the regressors which are involved in the multicollinearity problem.

EXAMPLE 6.3 ***Analyzing the multicollinearity of factors determining time devoted to housework***

In order to analyze the factors that influence time devoted to housework, the following model was formulated in exercise 3.17, using file *timuse03*:

$$houswork = \beta_1 + \beta_2 educ + \beta_3 hhinc + \beta_4 age + \beta_5 paidwork + u$$

where *educ* is the years of education attained, and *hhinc* is the household income in euros per month. The variables *houswork* and *paidwork* are measured in minutes per day.

Table 6.3 provides information on eigenvalues, sorted from the smallest to the largest, and the variance decomposition proportions for each eigenvalue are calculated according to (6-27). The condition number is equal to

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} = \sqrt{\frac{542.14}{7.06E-06}} = 8782$$

The condition number is very big, which would indicate a large amount of multicollinearity.

As can be seen in table 6.3[3], the greater proportions associated with the smallest eigenvalue, which is the main cause of multicollinearity in this model, correspond to the regressors *educ* and *age*. These two regressors are inversely correlated. The greatest proportions associated with the second smallest eigenvalue correspond to the regressors *educ* and the household income, which are positively correlated.

TABLE 6.3. Eigenvalues and variance decomposition proportions.

Eigenvalues	7.03E-06	0.000498	0.025701	1.861396	542.1400

Variance decomposition proportions

	Associated Eigenvalue				
Variable	1	2	3	4	5
C	0.999995	4.72E-06	8.36E-09	1.23E-13	1.90E-15
EDUC	0.295742	0.704216	4.22E-05	2.32E-09	3.72E-11
HHINC	0.064857	0.385022	0.209016	0.100193	0.240913
AGE	0.651909	0.084285	0.263805	5.85E-07	1.86E-08
PAIDWORK	0.015405	0.031823	0.007178	0.945516	7.80E-05

6.3.3 Solutions

In principle, the problem of multicollinearity is related to deficiencies in the sample. The non-experimental design of the sample is often responsible for these deficiencies. Let us look at some of the solutions to solve the problem of multicollinearity.

Elimination of variables

Multicollinearity can be mitigated if the regressors most affected by multicollinearity are removed. The problem with this solution is that the estimators of the new model would be biased if the original model was correct. On this issue the following reflection should be made. In any case, the researcher is interested in obtaining an unbiased estimator (or at least with very small bias) with a reduced variance. The mean square error (*MSE*) includes both factors. Thus, for the estimator $\hat{\beta}_j$, the *MSE* is defined as follows:

$$MSE(\hat{\beta}_j) = \left[bias(\hat{\beta}_j) \right]^2 + var(\hat{\beta}_j) \qquad (6\text{-}28)$$

If a regressor is eliminated from the model, the estimator of a regressor that is maintained (for example, $\hat{\beta}_j$) will be biased. Nevertheless, its *MSE* can be lower than that of the original model, because the omission of a variable can sufficiently reduce the variance of the estimator. In sum, although the elimination of a variable is not a desirable practice in principle, under certain circumstances it can be justified when it contributes to decreasing the *MSE*.

[3] In table 6.3, the eigenvalues are ordered from the lowest to the highest as the associated eigenvalues in the variance decomposition proportions. It is important to remark that in E-views eigenvalues are ordered from the highest to the lowest. However, in this package the condition number is defined differently than usual in the econometrics manuals which we have followed.

Increasing the sample size

Given that some degree of multicollinearity is a problem particularly when the variances of the estimators increase significantly, the solutions should aim to reduce these variances. A solution for increasing the variability of the regressors across the sample consists in introducing additional observations. However, this is not always feasible, since the data used in empirical analysis generally come from different data sources given the researcher only collects information on rare occasions.

Furthermore, when dealing with experimental designs, the variability of the regressors can be directly increased without increasing the size of the sample.

Using outside sample information

Another possibility is the use of outside sample information, either by setting constraints on the parameters of the model, or by using estimates from other studies.

Establishing restrictions on the parameters of the model reduces the number of parameters to be estimated and therefore alleviates the possible shortcomings of the sample information. In any case, these restrictions must be inspired by the theoretical model itself, or at least have an economic meaning.

In general, a disadvantage of this approach is that the meaning attributed to the estimator obtained in cross sectional data is very different from that obtained with time series data, in the case when both types of data are jointly used. Sometimes these estimators can be truly "foreign" or outside the object of study.

Using ratios

If instead of the regressand and the regressors of the original model, we use ratios with respect to the most affected regressor by collinearity, the correlations among the regressors of the model may decrease. One such solution is very attractive for the simplicity of implementation. However, the transformations of the original variables of the model using ratios can cause other problems. Assuming the original model fulfills the *CLM* assumptions, this transformation implicitly modifies the properties of the model, and therefore the disturbances of the transformed model will no longer be homoskedastic but heteroskedastic.

6.4 Normality test

The F and t significance tests built in chapter 4 are based on the normality assumption of the disturbances. But it is not usual to perform a normality test, given that a sufficiently large sample -e.g. 50 or more observations - is not often available. However, normality tests have recently been receiving a growing interest in both theoretical and applied studies.

Let us examine one test for verifying the assumptions of normality of disturbances in an econometric model. This test was proposed by Bera and Jarque, and is based on the statistics of skewness and kurtosis of the residuals.

The skewness statistic is the standardized third-order moment, applied to the residuals, and its expression is the following:

$$\gamma_{1(\hat{u})} = \frac{\sum \hat{u}_i^3 / n}{\left[\sum \hat{u}_i^2 / n\right]^{3/2}} \qquad (6\text{-}29)$$

In a symmetric distribution, as is the case of the normal distribution, the coefficient of skewness is 0.

The kurtosis statistic is the standardized fourth-order moment, applied to residuals, and its expression is the following:

$$\gamma_{2(\hat{u})} = \frac{\sum \hat{u}_i^4 / n}{\left[\sum \hat{u}_i^2 / n\right]^{2}} \qquad (6\text{-}30)$$

In a standard normal distribution, i.e. in an $N(0.1)$, the coefficient of kurtosis is equal to 3.

The Bera and Jarque statistic *(BJ)* is given by:

$$BJ = \left[\frac{n}{6}\left(\gamma_{1(\hat{u})}\right)^2 + \frac{n}{24}\left(\gamma_{2(\hat{u})} - 3\right)^2\right] \qquad (6\text{-}31)$$

In a theoretical normal distribution, the above expression will be equal to 0, as the coefficient of skewness and kurtosis respectively take the values 0 and 3. The statistic *BJ* will take higher values as the coefficient of asymmetry is far from 0 and the coefficient of kurtosis is far from 3. Under the null hypothesis of normality, the statistic *BJ* has the following distribution

$$BJ \xrightarrow[n\to\infty]{} \chi_2^2 \qquad (6\text{-}32)$$

The indication $n \to \infty$ means that *BJ* is an asymptotic test, i.e. valid when the sample is sufficiently large.

EXAMPLE 6.4 Is the hypothesis of normality acceptable in the model to analyze the efficiency of the Madrid Stock Exchange?

In example 4.5, using file *bolmadef*, we analyzed the market efficiency of the Madrid Stock Exchange in 1992, using a model that relates the daily rate of return on the rate of the previous day. Now we will test the normality assumption on the disturbances of this model. Given the low proportion of the variance explained with this model (see example 4.5), the test of normality of the disturbances is roughly equivalent to test the normality of the endogenous variable.

Table 6.4 shows the coefficients of skewness, kurtosis and the Bera and Jarque statistic, applied to the residuals. The asymmetry coefficient (-0.04) is not far from the value 0 corresponding to a distribution $N(0.1)$. On the other hand, the coefficient of kurtosis (4.43) is slightly different from 3, which is the value in the normal distribution. In this case, we reject the assumption of normality for the usual levels of significance, as the Bera and Jarque statistic takes the value of 21.02, which is larger than $C_2^{2(0.01)} = 9.21$.

TABLE 6.4. Normality test in the model on the Madrid Stock Exchange.

skewness coefficient	*kurtosis coefficient*	*Bera and Jarque statistic*
-0.0421	4.4268	21.0232

The fact that the normality assumption is rejected may seem paradoxical, since the values of kurtosis and especially of skewness do not differ substantially from the values taken by these coefficients in a normal distribution. However, the discrepancies are significant enough because they are supported by a large sample size (247 observations). If *n* (the size of the sample) had been 60 rather than 247, the *BJ* statistic, calculated according to (6-31) and using the same coefficient of skewness and kurtosis, takes the

value of 5.11, which is smaller than $\chi_2^{2(0.01)}$ = 9.21. To put it another way, with the same coefficients, but with a smaller sample, there is not enough empirical evidence to reject the null hypothesis of normality. Note that this is due to the fact that the *BJ* statistic increases proportionally to the size of the sample, but the degrees of freedom (2) remain unchanged.

6.5 Heteroskedasticity

The homoskedasticity assumption (assumption 7 of the *CLM*) states that the disturbances have a constant variance, that is to say:

$$var(u_i) = \sigma^2 \qquad i = 1, 2, \cdots n \qquad (6\text{-}33)$$

Assuming that there is only one independent variable, the homoskedasticity assumption means that the variability around of the regression line is the same for any value of *x*. In other words, variability does not increase or decrease when *x* varies, as shown in figure 2.7, part a) of chapter 2. In figure 6.1, a scatter plot is shown corresponding to a model in which disturbances are homoskedastic.

If the homoskedasticity assumption is not satisfied, then there is heteroskedasticity, or disturbances are heteroskedastic. In figure 2.7, part b) a model with heteroskedastic disturbances was represented: the dispersion increases with increasing values of *x*. Figure 6.2 shows the scatter diagram corresponding to a model in which the dispersion grows when *x* grows.

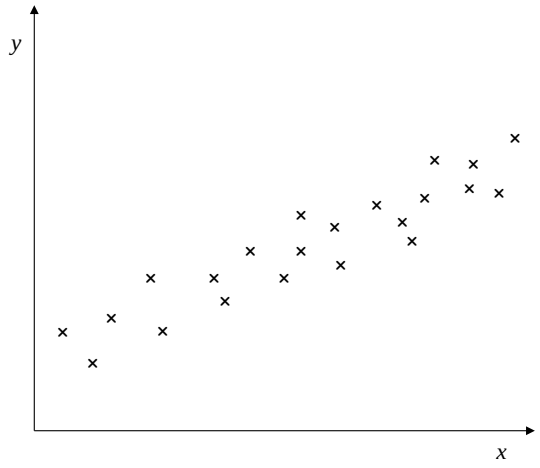


FIGURE 6.1. Scatter diagram corresponding to a model with homoskedastic disturbances.

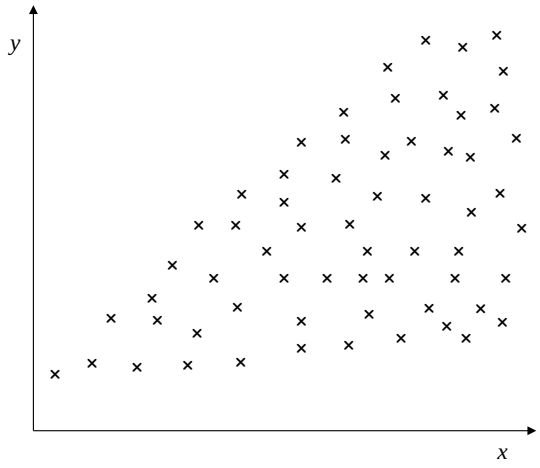


FIGURE 6.2. Scatter diagram corresponding to a model with heteroskedastic disturbances.

6.5.1 Causes of heteroskedasticity

In models estimated with cross sectional data (for example, demand studies based on surveys of household budgets) there are often problems of heteroskedasticity. However, heteroskedasticity can also occur in models estimated with time series.

Let us now consider some factors that can cause disturbances to be heteroskedastic:

a) Influence of the size of an explanatory variable in the size of the disturbance. Let us examine this factor using an example. Consider a model in which spending on hotels is a linear function of disposable income. If you have a representative sample of the population of a country, the great variability of the income received by families can

be seen. Logically, low income families are unlikely to spend large amounts on hotels, and in this case we can expect that the oscillations in the expenditure of one family to another are not significant. In contrast, in high-income families a greater variability in this type of expenditure can be expected. Indeed, high-income families may choose between spending a substantial part of their income on hotels or spending virtually nothing. The scatter diagram in figure 6.2 may be adequate to represent what happens in a model to explain the demand for a luxury good such as spending on hotels.

b) The presence of outliers can cause heteroskedasticity. An outlier is an observation generated apparently by a different population to that generating the remaining sample observations. When the sample size is small, the inclusion or exclusion of such an observation can substantially alter the results of regression analysis and cause heteroskedasticity.

c) Data transformation. As we saw in a previous section, one of the solutions to solve the problem of multicollinearity consisted in transforming the model taking ratios with respect to a variable (say x_{ji}), i.e. dividing both sides of the model by x_{ji}. Therefore, the disturbance will now be u_i/x_{ji}, instead of u_i. Assuming that u_i fulfills the homoskedasticity assumption, the disturbances of the transformed model (u_i/x_{ji}) will no longer be homoskedastic but heteroskedastic.

6.5.2 Consequences of heteroskedasticity

When there is heteroskedasticity, the *OLS* method is not the most appropriate because the estimators obtained are not the *best*, i.e. the estimators are not *BLUE*.

Moreover, the *OLS* estimators obtained when there is heteroskedasticity, in addition to not being *BLUE*, have the following problem. The covariance matrix of the estimators obtained by applying the usual formula is not valid when there is heteroskedasticity (and/or autocorrelation). Consequently, the *t* and *F* statistics based on the estimated covariance matrix can lead to erroneous inferences.

6.5.3 Heteroskedasticity tests

We are going to examine two heteroskedasticity tests: Breusch-Pagan-Godfrey and White. Both of them are asymptotic and have the form of a Lagrange multiplier (*LM*) test.

Breusch-Pagan-Godfrey (BPG) test

Breusch and Pagan (1979) developed a test for heteroskedasticity and Godfrey (1978) developed another one. Because they are similar, they are usually known as Breusch–Pagan–Godfrey (*BPG*) heteroskedasticity tests.

The *BPG* test is an asymptotic test, that is to say, it is only valid for large samples. The null and alternative hypotheses of this test can be formulated as follows:

$$\begin{aligned} &H_0 : E\left(u_i^2\right) = \sigma^2 \quad \forall i \\ &H_1 : \sigma_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} \end{aligned} \tag{6-34}$$

where the z_i's can be some or all of the x_i's of the model.

Taking into account the above H_1, H_0 can be expressed as

$$H_0 : \alpha_2 = \alpha_3 = \cdots \alpha_m = 0 \qquad (6\text{-}35)$$

The steps involved in this test are as follows:

Step 1. The original model is estimated and the *OLS* residuals are calculated.

Step 2. The following auxiliary regression is estimated, taking as the regressand the square of the residuals ($\hat{u}_i^2$) obtained in estimating the original model, since we know neither σ_i^2 nor u_i^2:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 z_{2i} + \alpha_3 z_{3i} + \cdots + \alpha_m z_{mi} + \varepsilon_i \qquad (6\text{-}36)$$

The auxiliary regression should have an intercept, although the original model is estimated without it. In accordance with expression (6-36), in the auxiliary regression there are *m* regressors in addition to the intercept.

Step 3. Designating by R_{ar}^2 the coefficient of determination of the auxiliary regression, the statistic nR_{ar}^2 is calculated.

Under the null hypothesis, this statistic (*BPG*) is distributed as follows:

$$BPG = nR_{ar}^2 \xrightarrow[n\to\infty]{} \chi_m^2 \qquad (6\text{-}37)$$

Step 4 For a significance level α, and designating by $\chi_m^{2(\alpha)}$ the corresponding value in χ^2 table, the decision to make is the following:

If $BPG > \chi_m^{2(\alpha)}$ H_0 is rejected

If $BPG \leq \chi_m^{2(\alpha)}$ H_0 is not rejected

In this test, high values of the statistic correspond to a situation of heteroskedasticity, that is to say, to the rejection of the null hypothesis.

EXAMPLE 6.5 Application of the Breusch-Pagan-Godfrey test

This test will be applied to a sub-sample of 10 observations, which have been used for estimating hotel expenditures (*hostel*) as a function of disposable income (*inc*). The data appear in table 6.5.

TABLE 6.5. *Hostel* and *inc* data.

i	*hostel*	*inc*
1	17	500
2	24	700
3	7	250
4	17	430
5	31	810
6	3	200
7	8	300
8	42	760
9	30	650
10	9	320

Step 1. Applying *OLS* to the model,

$$hostel = b_1 + b_2 inc + u$$

using data from table 6.5, the following estimated model is obtained:

$$\widehat{hostel}_i = \underset{(3.48)}{-7.427} + \underset{(0.0065)}{0.0533} inc_i$$

The residuals corresponding to this fitted model appear in table 6.6.

TABLE 6.6. Residuals of the regression of *hostel* on *inc*.

i	1	2	3	4	5	6	7	8	9	10
$\hat{u}_i$	-2.226	-5.888	1.100	1.505	-4.751	-0.234	-0.565	8.913	2.777	-0.631

Step 2. The auxiliary regression which must be estimated is the following:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 inc_i + \eta_i$$

Applying *OLS*, the following results are obtained:

$$\hat{u}_i^2 = -23.93 + 0.0799 inc \qquad R^2=0.5045$$

Step 3. Using the value of R^2, the *BPG* statistics is:

$$BPG = nR_{ar}^2 = 10(0.56) = 5.05.$$

Step 4. Given that $\chi_1^{2(0.01)} = 3.84$, the null hypothesis of homoskedasticity is rejected for a significance level of 5%, because $BPG > 3.84$, but not for the significance level of 1%.

Note that the validity of this test is asymptotic. However, the sample used in this example is very small.

White test

In the White test the hypothetical variables determining the heteroskedasticity are not specified. This test is a non-constructive test because it gives no indication of the heteroskedasticity scheme when the null hypothesis is rejected

The White test is based on the fact that the standard errors are asymptotically valid if we substitute the homoskedasticity assumption for the weaker assumption that the squared disturbance u^2 is uncorrelated with all the regressors, their squares, and their cross products. Taking this into account, White proposed to carry out the auxiliary regression of $\hat{u}_i^2$, since u_i^2 is unknown, on the factors mentioned above. If the coefficients of the auxiliary regression are jointly non-significant, then we can admit that the disturbances are homoskedastic. According to the assumption adopted, the White test is an asymptotic test.

The application of the White test can pose problems in models with many regressors. For example, if the original model has five independent variables, the White auxiliary regression would involve 16 regressors (unless some are redundant), which implies that the estimation is done with a loss of 16 degrees of freedom. For this reason, when the model has many regressors a *simplified* version of the White test is often applied. In the simplified version, the cross products are omitted from the auxiliary regression.

The steps involved in the *complete* version of the White test are as follows:

Step 1. The original model is estimated and the *OLS* residuals are calculated.

Step 2. The following auxiliary regression is estimated, taking as the regressand the square of the residuals obtained in the previous step:

$$\hat{u}_i^2 = \alpha_1 + \alpha_2\psi_{2i} + \alpha_3\psi_{3i} + \cdots + \alpha_m\psi_{mi} + \varepsilon_i \qquad (6\text{-}38)$$

In the above auxiliary regression, the regressors ψ_{ji} are the regressors of the original model, their squared values, and the crossproduct(s) of the regressors.

In any case, it is necessary to eliminate any redundancies that occur (i.e. regressors that appear repeatedly). For example, the intercept (which is 1 for all observations) and the square of the intercept cannot appear simultaneously as regressors, since they are identical. The simultaneous introduction of these two regressors will lead to perfect multicollinearity.

The auxiliary regression should have an intercept, even if the original model is estimated without it. In accordance with expression (6-38), in the auxiliary regression there are *m* regressors as well as the intercept.

Step 3. Designating by R_{ar}^2 the coefficient of determination of the auxiliary regression, the statistic nR_{ar}^2 is calculated.

Under the null hypothesis, this statistic (*W*) is distributed as follows:

$$W = nR_{ar}^2 \xrightarrow[n\to\infty]{} \chi_m^2 \qquad (6\text{-}39)$$

This statistic is used to test the overall significance of model (6-38).

Step 4. It is similar to step 4 in Breusch-Pagan-Godfrey test.

EXAMPLE 6.6 Application of the White test

This test is going to be applied to data from table 6.5.

Step 1. This step is the same as in the Breusch-Pagan-Godfrey test.

Step 2. Since there are two regressors in the original model (the intercept and *inc*), the regressors of the auxiliary regression will be

$$\psi_{1i} = 1 \qquad \forall i$$
$$\psi_{2i} = 1 \times inc_i$$
$$\psi_{3i} = inc_i^2$$

Consequently, the model to be estimated is

$$\hat{u}_i^2 = \alpha_1 + \alpha_2 inc_i + \alpha_3 inc_i^2 + \eta_i$$

By applying OLS to the data from table 6.5, we obtain the following

$$\hat{u}_i^2 = 14.29 - 0.10 inc_i + 0.00018 inc_i^2 \qquad R^2=0.56$$

Step 3. By using the $R^{2,}$ we obtain the *W* statistic:

$$W=nR^2 = 10(0.56) = 5.60.$$

The number of degrees of freedom is two.

Step 4. Given that $\chi_2^{2(0.10)} = 4.61$, the null hypothesis of homoskedasticity is rejected for a 10% significance level because $W=nR^2 > 4.61$, but not for significance levels of 5% and 1%.

Note that the validity of this test is asymptotic too.

EXAMPLE 6.7 Heteroskedasticity tests in models explaining the market value of the Spanish banks

To explain the market value (*marktval*) of Spanish banks as a function of their book value (*bookval*) two models were formulated: one linear (example 2.8) and another one doubly logarithmic (example 2.10).

Heteroskedasticity in the linear model

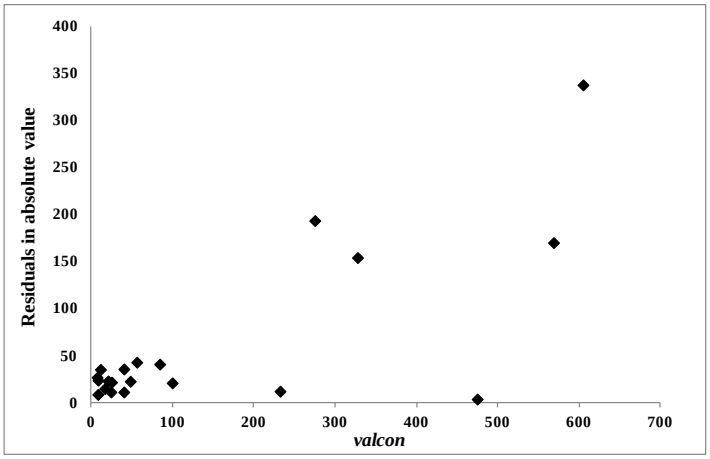
The linear model is given by

$$marktval=\beta_1+\beta_2 bookval+u$$

Using data from 20 banks and insurance companies (filework *bolmad95*), the following results were obtained:

$$\widehat{marktval} = \underset{(30.85)}{29.42} + \underset{(0.127)}{1.219}\, bookval$$

In graphic 6.1, the scatter plot between the residuals in absolute value (ordinate) and the variable *bookval* (in abscissa) is represented. This graphic shows that the absolute values of the residuals, which are indicative of the spread of this series, grow with increasing values of the variable *bookval*. In other words, this graph provides an indication but not a formal proof of the existence of heteroskedasticity of the disturbances associated with the variable *bookval*.



GRAPHIC 6.1. Scatter plot between the residuals in absolute value and the variable *bookval* in the linear model.

The *BPG* statistic takes the following value:

$$BPG= nR_{ra}^2 = 20\times 0.5220=10.44$$

As $\chi_1^{2(0.01)}$ =6.64<10.44, the null hypothesis of homoskedasticity is rejected for a significance level of 1%, and therefore for α=0.05 and for α=0.10.

Now we will apply the White test. In this case, the auxiliary regression includes as regressors the intercept, the variable *bookval*, and the square of this variable. The White statistic takes the following value:

$$W= nR_{ra}^2 = 20\times 0.6017=12.03$$

As $\chi_2^{2(0.01)}$ =9.21=<12.03, the null hypothesis of homoskedasticity is rejected for a significance level of 1%.

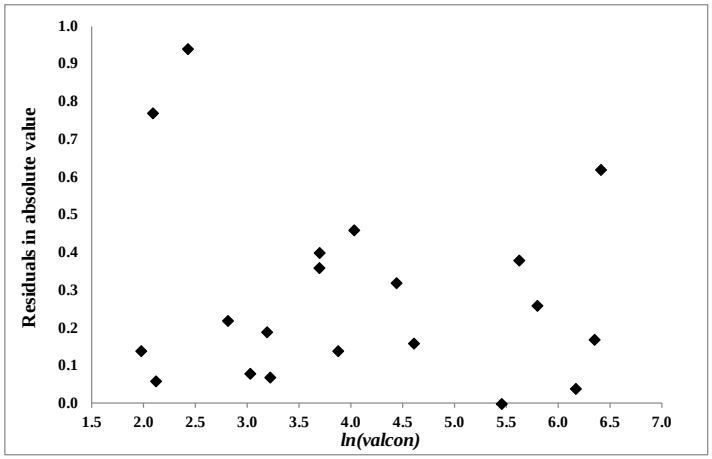
Therefore, both tests are conclusive in rejecting the null hypothesis for the usual levels of significance.

Heteroskedasticity in the log-log model

The estimated log-log model with the same sample was as follows:

$$\widehat{\ln(marktval)} = \underset{(0.265)}{0.676} + \underset{(0.062)}{0.9384}\ln(bookval)$$

In graphic 6.2 the scatter plot between the residuals in absolute value (ordinate), corresponding to this estimated model, and the variable ln(*bookval*) (in abscissa) is represented. As shown, the two largest residuals correspond to two banks with small market value. Even disregarding these two cases, apparently there is no relationship between the residuals and the explanatory variable of the model.



GRAPHIC 6.2. Scatter plot between the residuals in absolute value and the variable *bookval* in the log-log model.

The results of the two tests of heteroskedasticity applied are shown in table 6.7.

TABLE 6.7. Tests of heteroskedasticity on the log-log model to explain the market value of Spanish banks.

Test	*Statistic*	*Table values*
Breusch-Pagan	$BP = nR_{ra}^2 = 1.05$	$\chi_2^{2(0.10)} = 4.61$
White	$W = nR_{ra}^2 = 2.64$	$\chi_2^{2(0.10)} = 4.61$

Both tests carried out indicate that the null hypothesis of homoskedasticity cannot be rejected against the alternative hypothesis that the variance of the disturbances is associated with the explanatory variable of the model.

An important conclusion is that, if an econometric model is estimated with cross sectional data, it is easy to find observations with very different size. These problems of scale can cause heteroskedasticity in the disturbances but can often be solved by using log-log models.

EXAMPLE 6.8 Is there heteroskedasticity in demand of hostel services?

In general, heteroskedasticity in the disturbances does not usually appear in demand for food commodities. By contrast, heteroskedasticity is usually much more frequent in demand for luxury goods, because in the demand for these goods there is a large disparity in the behavior of high income households, while in households with low incomes such disparity is very unlikely.

In view of these considerations, the specification for analyzing the demand for hostel services is the following:

$$\ln(hostel) = b_1 + b_2 \ln(inc) + b_3 secstud + b_4 terstud + b_5 hhsize + u \qquad (6\text{-}40)$$

where *inc* is disposable income of a household, *hhsize* is the number of household members, and *secstud* and *terstud* are two dummies that take the value one if individuals have completed secondary and tertiary studies respectively.

The results obtained, using file *hostel*, are the following (file *hostel*):

$$\widehat{\ln(hostel)}_i = \underset{(2.26)}{-16.37} + \underset{(0.324)}{2.732} \ln(inc)_i + \underset{(0.258)}{1.398} secstud_i + \underset{(0.333)}{2.972} terstud_i - \underset{(0.088)}{0.444} hhsize_i$$

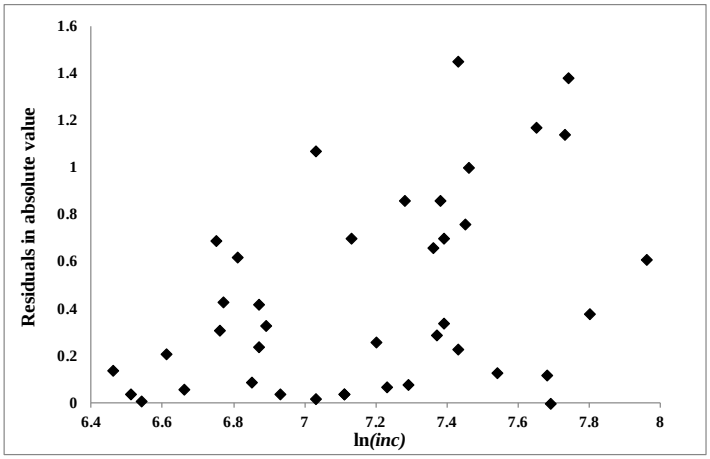
$$R^2 = 0.921 \qquad n = 40$$

Note that hostel services are a luxury good, as the elasticity of demand/income for this good is very high (2.73). This means that if income increases by 1%, spending on hostel services will increase, on average, by 2.73%. As can be seen, families where the main breadwinner has secondary studies (*secstud*) or, especially, higher education (*terstud*), spend more on hostel services than if the main breadwinner only

has primary education. However, spending on hostel services will decrease as household size (*hhsize*) increases.

Graphic 6.3 shows the scatter plot between the residuals in absolute value and the variable ln(*inc*). Income (or a transformation of it) is the main candidate, if not the only one, to explain the hypothetical heteroskedasticity in the disturbances. As shown in the graphic, the dispersion of residuals is smaller for low incomes than for middle or upper incomes.

We will now apply the two tests of heteroskedasticity that have been discussed in this section.



GRAPHIC 6.3. Scatter plot between the residuals in absolute value and the variable *ln(inc)* in the hostel model.

The results of the two tests of heteroskedasticity applied are shown in table 6.8

TABLE 6.8. Tests of heteroskedasticity in the model of demand for hostel services.

Test	*Statistic*	*Table values*
Breusch-Pagan-Godfrey	$BPG = nR_{ra}^2 = 7.83$	$\chi_2^{2(0.05)} = 5.99$
White	$W = nR_{ra}^2 = 12.24$	$\chi_2^{2(0.01)} = 9.21$

In the *BPG* test we reject the null hypothesis of homoskedasticity for a significance level of α=0.05, but not for α=0.01.

Since there are many dummy variables in the model, including cross products in the auxiliary regression, this can lead to serious problems of multicollinearity. For this reason, in the auxiliary regression cross products are not included. Not surprisingly, among the regressors of the auxiliary regression squares of *secstud* and *terstud* are not included because they are dummies. Given the value obtained in the White statistic, we reject the null hypothesis of homoskedasticity for a significance level of α=0.01. Therefore, the White test is more conclusive in rejecting the homoskedasticity assumption.

6.5.4 Estimation of heteroskedasticity-consistent covariance

When there is heteroskedasticity and we apply *OLS*, we cannot make correct inferences by using the covariance matrix associated to the *OLS* estimates, because this matrix is not a consistent estimator of the covariance matrix of the coefficients. Consequently, the *t* and *F* statistics based on that estimated covariance matrix can lead to erroneous inferences.

Therefore, in the case that there is heteroskedasticity and *OLS* have been applied, a consistent estimate of the covariance matrix should be looked for to make inferences. White derived a consistent estimator of the covariance matrix under heteroskedasticity.

However, it is important to note that this estimator does not work well if the sample is small, given that it is an asymptotic approximation.

Most econometric packages allow standard errors to be calculated by the White procedure. By using these consistent standard deviations, adequate tests can be made under the heteroskedasticity assumption.

EXAMPLE 6.9 Heteroskedasticity consistent standard errors in the models explaining the market value of Spanish banks (Continuation of example 6.7)

In the following estimated equation of the linear model, using file *bolmad95*, standard deviations of the estimates are calculated by the White procedure and therefore they are consistent under heteroskedasticity:

$$\underset{}{\widehat{marktval}} = \underset{(18.67)}{29.42} + \underset{(0.249)}{1.219}\, bookval$$

As can be seen, the standard error of the *bookval* coefficient goes from 0.127 in the usual procedure to 0.249 in the White procedure. However, the *p*-value remains very low (0.0001). Accordingly, the significance of the variable *bookval* for all usual levels is still maintained. By contrast, the intercept, which has no special meaning in the model, now has a standard error (18.67), which is lower than that obtained with the usual procedure (30.85).

If we apply the White procedure to the log-log model, the following results are obtained:

$$\underset{}{\widehat{\ln(marktval)}} = \underset{(0.3218)}{0.676} + \underset{(0.0698)}{0.9384}\ln(bookval)$$

In this case, the standard error of ln(*bookval*) coefficient is practically the same in the two procedures.

From the above results, the following conclusions can be obtained. In determining the market value of Spanish banks, disturbances of the linear model are strongly heteroskedastic. Therefore, when using a consistent estimate, the standard deviation is almost doubled compared to the standard one. By contrast, in the log-log model, which is not affected by heteroskedasticity, there is little difference between the standard errors obtained with both procedures.

6.5.5 The treatment of the heteroskedasticity

In order to estimate a model with heteroskedastic disturbances it is necessary to know or, if it is unknown, to estimate the pattern of heteroskedasticity. Thus, suppose that the standard deviation of the disturbances follows this scheme:

$$\sigma_i = f\left(x_{ji}\right) \qquad (6\text{-}41)$$

As indicated in epigraph 6.1, the method *GLS* allows *BLUE* estimators to be obtained when disturbances are heteroskedastic. If we know scheme (6-41), the application of *GLS* is performed in two stages. In the first stage, the original model is transformed by dividing both sides by the standard deviation. Therefore, according to (6-41), the transformed model is given by

$$\frac{y_i}{f\left(x_{ji}\right)} = \beta_1 \frac{1}{f\left(x_{ji}\right)} + \beta_2 \frac{x_{1i}}{f\left(x_{ji}\right)} + \beta_3 \frac{x_{2i}}{f\left(x_{ji}\right)} + \ldots + \beta_k \frac{x_{ki}}{f\left(x_{ki}\right)} + \frac{u_i}{f\left(x_{ji}\right)} \qquad (6\text{-}42)$$

It is easily seen that the disturbances of the previous model, $(u_i/f(x_{ji}))$, are homoskedastic. Therefore, in the second stage *OLS* is applied to the transformed model, thus obtaining *BLUE* estimators. When we divide each observation by $f(x_{ji})$, we are weighting by the inverse of the value taken by this function. For this reason the above

procedure is often called *weighted least squares* (*WLS*). In this case, the weighting factor is $1/f(x_{ji})$.

If the function $f(x_{ji})$ is not known, it is necessary to estimate it. In that case, the estimation method will not be exactly the *GLS* method because the application of this method involves the knowledge of the covariance matrix, or, at least, knowledge of a matrix that is proportional to it. If we estimate the covariance matrix, in addition to the parameters, it is said that *feasible GLS* is applied. In the case of heteroskedastic disturbances, the particularization of the feasible *GLS* method is called *WLS* (*weighted least squares*) in two stages. In the first the function $f(x_{ij})$ stage is estimated, whereas in the second stage *OLS* is applied to the model transformed using the $f(x_{ji})$ estimates.

To see how to apply the *WLS* method in two stages, let us consider the following relationship, which simply defines the variance of the disturbances, in the case of heteroskedasticity,

$$E\left(u_i^2\right) = \sigma_i^2 \tag{6-43}$$

Therefore, the squared disturbance can be made equal, as in the regression model, to its expectation plus a random variable. That is to say:

$$u_i^2 = \sigma_i^2 + \varepsilon_i \tag{6-44}$$

As the disturbances are not observable, one can establish a relationship analogous to the above using residuals instead of disturbances. Therefore,

$$\hat{u}_i^2 = \sigma_i^2 + \eta_{2i} \tag{6-45}$$

It should be noted that the above relationship does not have exactly the same properties as (6-44) because the residuals are correlated and heteroskedastic, even if the disturbances fulfill the *CLM* assumptions. However, in large samples they will have the same properties.

If we use the residuals as the regressand instead of the squared residuals, we must take the absolute values, since the standard deviation takes only positive values. Taking into account (6-45), the following relationship can be established:

$$\left|\hat{u}_i\right| = \sigma_i^2 + \eta_{2i} = f\left(x_{ij}\right) + \eta_{2i} \tag{6-46}$$

Since the function $f(x_{ij})$ is generally unknown, different functions are often tried. Here there are some of the most common:

$$\begin{aligned}
\left|\hat{u}_i\right| &= \alpha_1 + \alpha_2 x_{ji} + \eta_{2i} \\
\left|\hat{u}_i\right| &= \alpha_1 + \alpha_2 \sqrt{x_{ji}} + \eta_{2i} \\
\left|\hat{u}_i\right| &= \alpha_1 + \alpha_2 \frac{1}{x_{ji}} + \eta_{2i} \\
\left|\hat{u}_i\right| &= \alpha_1 + \alpha_2 \ln(x_{ji}) + \eta_{2i}
\end{aligned} \tag{6-47}$$

The functional form with the best fit (a higher coefficient of determination or a smaller *AIC* statistic) is selected. For the transformation two circumstances are contemplated, depending on the significance of the intercept. If this coefficient is

statistically significant, the model is transformed by dividing by the fitted values of the selected equation. If it is not statistically significant, the model is transformed by dividing by the regressor corresponding to the selected equation. Thus, if the selected equation were the second one of (6-47), with the intercept not being significant, the transformed model would be as follows:

$$\frac{y_i}{\sqrt{x_{ji}}} = \beta_1 \frac{1}{\sqrt{x_{ji}}} + \beta_2 \frac{x_{2i}}{\sqrt{x_{ji}}} + \beta_3 \frac{x_{3i}}{\sqrt{x_{ji}}} + \ldots + \beta_k \frac{x_{ki}}{\sqrt{x_{ji}}} + \frac{u_i}{\sqrt{x_{ji}}} \tag{6-48}$$

Note that if the intercept is not significant, the estimated parameters are not involved in the transformation of the model, but they are if the intercept is significant. As the estimators in models (6-47) are biased, although consistent, it is not convenient to transform the models by applying the fitted values, $|\hat{u}_i|$-obtained by using $\hat{\alpha}_0$ and $\hat{\alpha}_1$- except when the significance of the intercept is very high (e.g., exceeding 1%).

EXAMPLE 6.10 Application of weighted least squares in the demand of hotel services (Continuation of example 6.8)

Since the two tests applied to the model to explain the cost of hotel services indicate that the disturbances are heteroskedastic, we apply the weighted least squares method to estimate the model (6-40).

First, we estimate the four models (6-47), using as the regressand the residuals $|\hat{u}_i|$ -in absolute value- obtained in the estimation of model (6-40) by *OLS*. The results are presented below:

$$\widehat{|u_i|} = \underset{(0.143)}{0.0239} + \underset{(2.73)}{0.0003}\,inc \qquad R^2 = 0.1638$$

$$\widehat{|u_i|} = \underset{(-1.34)}{-0.4198} + \underset{(2.82)}{0.0235}\sqrt{inc} \qquad R^2 = 0.1733$$

$$\widehat{|u_i|} = \underset{(5.39)}{0.8857} - \underset{(-2.87)}{532.1}\frac{1}{inc} \qquad R^2 = 0.1780$$

$$\widehat{|u_i|} = \underset{(-2.46)}{-2.7033} + \underset{(2.88)}{0.4389}\ln(inc) \qquad R^2 = 0.1788$$

In the above results, the *t*-statistic appears below each coefficient.

The functional form in which ln(*inc*) appears as a regressor is selected because it corresponds to the highest R^2 obtained. Since the coefficient of the independent term is not statistically significant at 1%, following the recommendation, *WLS* are applied taking 1/ln(*inc*) as the weighting variable. In estimating *WLS*, the following results were obtained:

$$\widehat{\ln(hostel)}_i = \underset{(2.15)}{-16.21} + \underset{(0.309)}{2.709}\ln(inc)_i + \underset{(0.247)}{1.401}\,secstud_i + \underset{(0.326)}{2.982}\,terstud_i - \underset{(0.085)}{0.445}\,hhsize_i$$

$$R^2=0.914 \qquad n=40$$

Compared to the *OLS* estimates of example 6.5, it can be seen that the differences are very small, which is indicative of the robustness of the model.

6.6 Autocorrelation

No autocorrelation, or *no serial correlation* assumption (assumption 8 of the *CLM*) states that disturbances with different subscripts are not correlated with each other:

$$E(u_i u_j) = 0 \qquad i \neq j \tag{6-49}$$

That is, the disturbances corresponding to different periods of time, or to different individuals, are not correlated with each other. Figure 6.3 shows a plot corresponding to disturbances which are not autocorrelated. The *x* axis is time. As can be seen, disturbances

are randomly distributed above and below the line 0 (theoretical mean of *u*). In the figure, each disturbance is linked by a line to the disturbance of the following period: in total this line crosses the line 0 on 13 occasions.

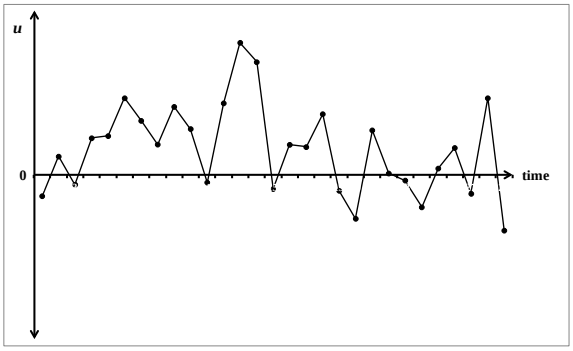


FIGURE 6.3. Plot of non-autocorrelated disturbances.

The transgression of the no autocorrelation assumption occurs quite frequently in models using time series data. It should be noted also that autocorrelation can be positive as well as negative. Positive autocorrelation is characterized by leaving a trail over time, because the value of each disturbance is near the value of the disturbance which precedes it. Positive autocorrelation occurs, by far, much more frequently in practice than the negative one. Figure 6.4 shows a plot corresponding to disturbances which are positively autocorrelated. As can be seen, the line which links successive disturbances crosses the line 0 only 4times.

By contrast, disturbances affected by negative autocorrelation present a saw tooth configuration, since each disturbance often takes the opposite sign of the disturbance which precedes it. In figure 6.5, the plot corresponds to disturbances which are negatively autocorrelated. Now the line 0 is crossed 21 times by the line which links successive disturbances.

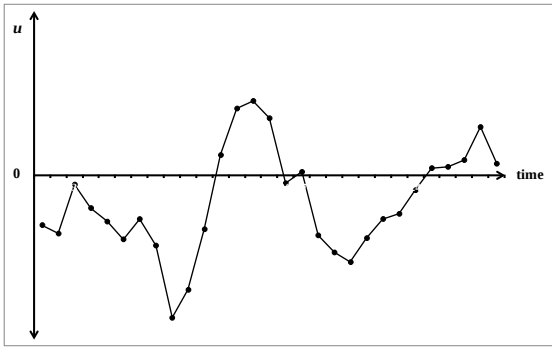


FIGURE 6.4. Plot of positive autocorrelated disturbances.

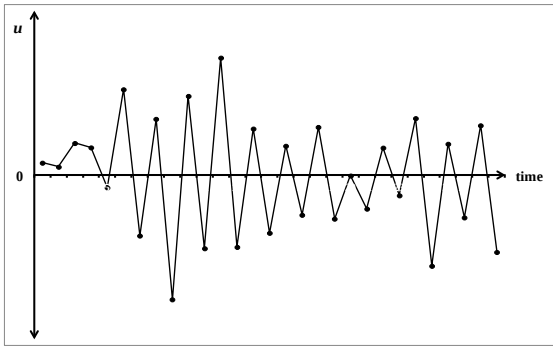


FIGURE 6.5. Plot of negative autocorrelated disturbances.

6.6.1 Causes of autocorrelation

There are several reasons for the presence of autocorrelation in a model, some of which are as follows:

a) Specification bias. That is, it can be caused by using an incorrect functional form or the omission of a relevant variable.

Let us suppose the correct functional form for determining *wage* as a function of years of experience (*exp*) is as follows:

$$wage = \beta_1 + \beta_2 exp + \beta_3 exp^2 + u$$

Instead of this model, the following one is fitted:

$$wage = \beta_1 + \beta_2 exp + v$$

In the second model, the disturbance has a systematic component ($v = \beta_3 exp^2 + u$). In figure 6.5, a scatter diagram (generated for the first model) and the fitted function of the second model are represented. As can be seen, for the low values of *exp* the fitted model overestimates wages; for intermediate values of *exp* wages are underestimated; finally, for high values the fitted model again overestimates wages. This example illustrates a case in which the use of an uncorrected functional form provokes positive autocorrelation.

On the other hand, the omission of a relevant variable in the model could induce positive autocorrelation if that variable has, for example, a cyclical behavior.

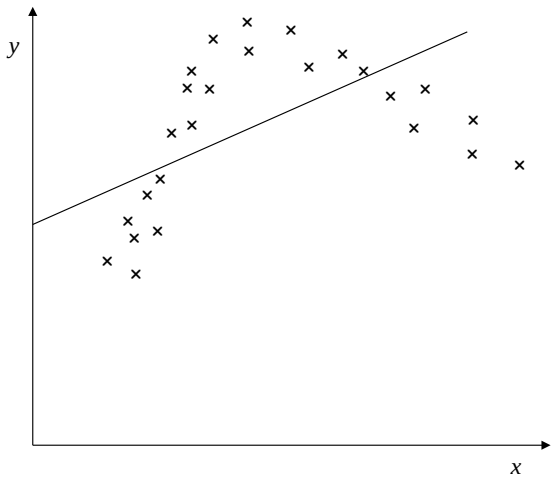


FIGURE 6.6. Autocorrelated disturbances due to a specification bias.

b) Inertia. The disturbance term in a regression equation reflects the influence of those variables affecting the dependent variable that have not been included in the regression equation. To be precise, inertia or the persisting effects of excluded variables of the model –and included in *u*- is probably the most frequent cause of positive autocorrelation. As is well known, macroeconomic time series -such as *GDP*, production, employment and price indexes- tend to move together: during expansion periods these series tend to increase in parallel, while in times of contraction they tend to decrease also in a parallel form. For this reason, in regressions involving time series data, successive observations of the disturbance are likely to be dependent on the previous ones. Thus, this cyclical behavior can produce autocorrelation in the disturbances.

c) Data Transformation. As an example let us consider the following model to explain consumption as a function of income:

$$cons_t = b_1 + b_2 inc_t + u_t \qquad (6\text{-}50)$$

For the observation *t*-1, we can write

$$cons_{t-1} = b_1 + b_2 inc_{t-1} + u_{t-1} \qquad (6\text{-}51)$$

If we subtract (6-51) from (6-50), we obtain

$$\Delta cons_t = b_2 \Delta inc_t + \Delta u_t \qquad (6\text{-}52)$$

where $\Delta cons_t = cons_t - cons_{t-1}$, $\Delta inc_t = inc_t - inc_{t-1}$ and $v_t = \Delta u_t = u_t - u_{t-1}$.

The equation (6-50) is known as a *level form* equation, while the equation (6-52) is known as the *first difference form* equation. Both of them are used in empirical analysis. If disturbance in (6-50) is not autocorrelated, the disturbance in (6-52), which is equal to $v_t = u_t - u_{t-1}$, will be autocorrelated, because v_t and v_{t-1} have a common element (u_{t-1}). In any case it should be noted the model (6-52), as specified, poses other econometric problems which will not be addressed here.

6.6.2 Consequences of autocorrelation

The consequences of autocorrelation for *OLS* are somewhat similar to those of heteroskedasticity. Thus, if the disturbances are autocorrelated, then the *OLS* estimator is not *BLUE* because one can find an alternative unbiased estimator with smaller variance. In addition to not being *BLUE*, the estimator obtained by *OLS* under the assumption of autocorrelation presents the problem that the estimation of the covariance matrix of the estimators calculated by the *OLS* usual formulas is biased. Consequently, the *t* and *F* statistics based on this covariance matrix can lead to erroneous inferences.

6.6.3 Autocorrelation tests

In order to test autocorrelation, a scheme of autocorrelation of disturbances in the alternative hypothesis must be defined. We will examine three of the best known tests. In two of them (the Durbin and Watson test and Durbin's *h* test) the alternative hypothesis is a first-order autoregressive scheme, while the third one, called the Breusch–Godfrey test, is a general test of autocorrelation applicable to higher-order autoregressive schemes.

Durbin and Watson test

The econometricians Durbin and Watson proposed the *d* test in 1950. *DW* is also used to refer to this statistic.

Durbin and Watson proposed the following scheme for the disturbances u_i:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \rightarrow NID(0, \sigma^2) \qquad (6\text{-}53)$$

The proposed scheme for u_t is a first-order autoregressive scheme, since the disturbances appear as regressand and also as regressor lagged a period. In the terminology of time series analysis, the scheme (6-53) is called *AR*(1), that is to say, an autoregressive process of order 1. The coefficient of this scheme is ρ, which is required to be less than 1 in absolute value so that the disturbances do not have an explosive character, when *n* grows indefinitely. The variable ε_t is a random variable with a normal and independent distribution (which means *NID*) with mean 0 and variance σ^2. Consequently, the variable ε_t fulfills the same assumptions as u_t in the *CLM* assumptions. The variables with these properties are often called white noise variables.

According to the sign of ρ being positive or negative, the autocorrelation will be positive or negative. On the other hand, almost always one-tailed test is performed, namely the alternative hypothesis is taken as either positive autocorrelation or negative autocorrelation.

The problem of constructing an autocorrelation test is that the disturbances are not observable. The test must therefore be based on the residuals obtained from the *OLS* estimation. This raises problems, since, under the null hypothesis that disturbances are not autocorrelated, residuals are autocorrelated. In the construction of their test, Durbin and Watson took these factors into account.

Let us now apply this test. Taking as a reference the scheme defined in (6-53), Durbin and Watson formulate the following null and alternative hypothesis of positive autocorrelation

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned} \qquad (6\text{-}54)$$

Thus, $u_t=\varepsilon_t$ is verified under the null hypothesis, i.e. the model fulfills the *CLM* assumptions.

The statistic used by Durbin and Watson for testing hypotheses (6-54) is the *d* or *DW* statistic, defined as follows:

$$d = DW = \frac{\sum_{t=2}^{n}\left(\hat{u}_t - \hat{u}_{t-1}\right)}{\sum_{t=1}^{n}\hat{u}_t^2} \qquad (6\text{-}55)$$

The statistical distribution of *d*, which is symmetrical with a mean equal to 2, is very complicated, since it depends on the particular form of the matrix of regressor **X**, the sample size (*n*) and the number of regressors (*k*) excluding the intercept.

However, for different levels of significance, Durbin and Watson obtained two values (d_L and d_U) for each value of *n* and *k*. The rules to test positive autocorrelation are:

$$\begin{aligned} &\text{If } d < d_L \text{ , there is positive autocorrelation.} \\ &\text{If } d_L \le d \le d_U \text{ , the test is not conclusive.} \\ &\text{If } d > d_U \text{ , there is not positive autocorrelation.} \end{aligned} \qquad (6\text{-}56)$$

As can be seen, there are values where the test is not conclusive. This is due to the effect that the particular configuration of the matrix **X** has on the distribution of *d*.

If you want to test negative autocorrelation, the alternative hypothesis is the following:

$$H_1 : \rho < 0 \qquad (6\text{-}57)$$

In order to apply the negative autocorrelation test, it is taken into account that the statistic *d* has a symmetrical distribution ranging between 0 and 4. The rules, therefore, are the following:

Si $d > 4 - d_L$, there is negative autocorrelation.

Si $4 - d_U \leq d \leq 4 - d_L$, the test is not conclusive.

Si $d < 4 - d_U$, there is not positive autocorrelation.

(6-58)

The Durbin and Watson test is not applicable if there are lagged endogenous variables as regressors.

To be applied to quarterly data, Wallis considered a fourth-order autoregressive scheme:

$$u_t = \rho_4 u_{t-4} + \varepsilon_i \qquad |\rho_4| < 1 \qquad \varepsilon_t \rightarrow NID(0, \sigma^2) \qquad (6\text{-}59)$$

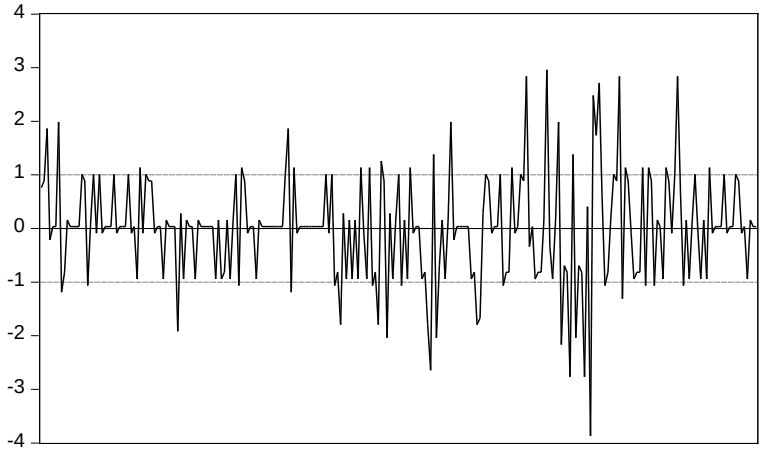
The above scheme is similar to (6-53), the difference being that the disturbance of the right hand side is lagged four periods. The Wallis statistic is similar to (6-55), but takes into account that the residuals are lagged four periods. This author designed *ad hoc* tables for testing models in which disturbances follow scheme (6-59).

EXAMPLE 6.11 Autocorrelation in the model to determine the efficiency of the Madrid Stock Exchange

In example 4.5, a model was formulated to determine the efficiency of the Madrid stock exchange. Graphic 6.4 shows the standardized residuals[4] corresponding to the estimation of this model, using file *bolmadef*. The *DW* statistic is equal to 2.04. (The *DW* statistic appears in the output of any econometric package). As the *DW* table does not have values for a sample size of 247, we use the corresponding values to *n*=200 and *k'*= 1. (In the nomenclature of this test, *k*' is used for the total number of regressors excluding the intercept). As the sample size is large we use a significance level of 1%. Upper and lower tabulated values, which correspond to the above specification, are as follows:

d_L=1.664; d_U=1.684

Since DW=2.04>d_U, we do not reject the null hypothesis that the disturbances are not autocorrelated for a significance level of α=0.01, i.e. of 1%, versus the alternative hypothesis of positive autocorrelation according to the scheme (6-53).



GRAPHIC 6.4. Standardized residuals in the estimation of the model to determine the efficiency of the Madrid Stock Exchange.

EXAMPLE 6.12 Autocorrelation in the model for the demand for fish

[4] Standardized residuals are equal to residuals divided by $\hat{\sigma}$.

In example 4.9 we estimated model (4-44), using file *fishdem*, to explain the demand for fish in Spain. The graphic 6.5 shows the standardized residuals obtained in the estimation of this model. This graph does not show that there is a significant autocorrelation scheme. In this regard, it should be noted that, over a total of 28 observations, the line joining the points of the residuals crosses the axis 0 11 times, which indicates a degree of randomness of the distribution of the residuals.

The value of the *DW* statistic for testing the scheme (6-53) is 1.202. For n=28 and k'=3, and for a significance level of 1%, we get the following tabulated values:

$$d_L=0.969 \qquad d_U=1.415$$

Since $d_L<1.202<d_U$, there is not enough evidence to accept the null hypothesis, or to reject it.

GRAPHIC 6.5. Standardized residuals in the model on the demand for fish.

Durbin's h test

Durbin (1970) proposed a statistic, called *h*, to test the hypothesis (6-54) in the case that one or more lagged endogenous variables appear as explanatory variables. The expression of the *h* statistic is the following:

$$h=\hat{r}\sqrt{\frac{n}{1-n\widehat{\mathrm{var}}(\hat{b}_j)}} \qquad (6\text{-}60)$$

where $\hat{r}$ is the correlation coefficient between $\hat{u}_i$ and $\hat{u}_{i-1}$, n is the sample size, and $\widehat{\mathrm{var}}(\hat{b}_j)$ is the variance corresponding to the coefficient of the lagged endogenous variable.

The statistic $\hat{r}$ can be estimated using the following approximation, $d \;;\; 2(1-\hat{r})$. If the regressand appears with different time lags as regressors, the variance corresponding to the regressor with the lowest lag is selected.

Under assumptions (6-54), the *h* statistic has the following distribution:

$$h \xrightarrow[n\to\infty]{a} N(0,1) \qquad (6\text{-}61)$$

The critical region is therefore in the tails of the standard normal distribution: the tail on the right for positive autocorrelation and the tail on the left for negative autocorrelation.

The statistic (6-60) cannot be calculated if $n\widehat{\mathrm{var}}(\hat{b}_j)\geq 1$. In this case, Durbin proposed an alternative procedure to estimate an auxiliary regression: the residuals are

taken as the regressand, the regressors are the same as those of the original model and the residuals also lagged a period. This procedure is a particular case of the Breusch–Godfrey test, which we will see next.

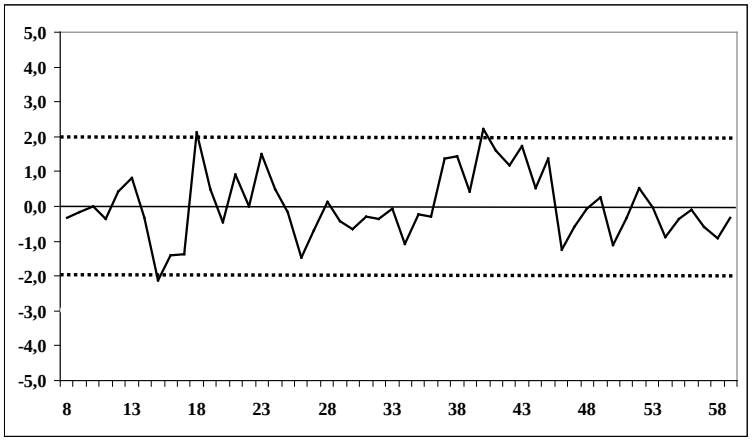
EXAMPLE 6.13 Autocorrelation in the case of Lydia E. Pinkham

In example 5.5 with the case of Lydia E. Pinkham, a model to explain the sales of a herbal extract was estimated using file *pinkham*. Graphic 6.6 shows the graph of standardized residuals corresponding to this model. As can be seen, it appears that the residuals are not distributed in a random way. Note, for example, that from 1936 the residuals take positive values for 8 consecutive years.

The adequate test for autocorrelation in this model is Durbin's *h* statistic, as there is a lagged endogenous variable $sales_{t-1}$ in this model. The *h* statistic is:

$$h = \hat{r}\sqrt{\frac{n}{1-n\widehat{\mathrm{var}}\left(\hat{b}_j\right)}} = \left(\frac{d}{2}\right)\sqrt{\frac{n}{1-n\widehat{\mathrm{var}}\left(\hat{b}_j\right)}} = \left(\frac{1.2012}{2}\right)\sqrt{\frac{53}{1-53\times 0.0814^2}} = 3.61$$

Given this value of *h*, the null hypothesis of no autocorrelation is rejected for α=0.01 or, even, for α=0.001, according to the table of the normal distribution.



GRAPHIC 6.6. Standardized residuals in the estimation of the model of the Lydia E. Pinkham case.

Breusch–Godfrey (BG) test

The Breusch–Godfrey (1978) test is a general test of autocorrelation applicable to higher-order autoregressive schemes, and it can be used when there are stochastic regressors such as the lagged regressand. This is an asymptotic test which is also known as the *LM* (Lagrange multipliers) general test for autocorrelation.

In the *BG* test, it is assumed that the disturbances u_t follow a *p*th-order autoregressive model *AR*(*p*):

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t \qquad |\rho| < 1 \qquad \varepsilon_t \to NID(0, \sigma^2) \tag{6-62}$$

This is simply the extension of the *AR*(1) scheme of the Durbin and Watson test.

The null hypothesis and the alternative hypotheses to be tested are:

$$H_0 : \rho_1 = \rho_2 = \cdots = \rho_p = 0$$

$$H_1 : H_0 \text{ is not true}$$

The *BG* test involves the following steps:

Step 1. The original model is estimated and the *OLS* residuals ($\hat{u}_i$) are calculated.

Step 2. An auxiliary regression is estimated, in which the residuals ($\hat{u}_i$) are taken as the regressand and the regressors of the original model and the residuals lagged 1, 2, ... and *p* periods are taken as regressors:

$$\hat{u}_t = \alpha_1 + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + \gamma_1 \hat{u}_{t-1} + \cdots + \gamma_1 \hat{u}_{t-p} + \varepsilon_i \qquad (6\text{-}63)$$

The auxiliary regression should have an intercept, even if the original model is estimated without it. In accordance with expression (6-63), in the auxiliary regression there are *k*+*p* regressors in addition to the intercept.

Step 3. Designating by R^2_{ar} the coefficient of determination of the auxiliary regression, the statistic nR^2_{ar} is calculated.

Under the null hypothesis, the *BG* statistic is distributed as follows:

$$BG = nR^2_{ar} \xrightarrow[n\to\infty]{} \chi^2_{k+p} \qquad (6\text{-}64)$$

The *BG* statistic is used to test the overall significance of the model (6-63). For this purpose, the *F* statistic can also be used. However, in this case it has only asymptotic validity, in the same way as with the *BG* statistic.

Step 4 For a significance level α, and designating by $\chi^{2(\alpha)}_{k+p}$ the corresponding value in χ^2 table, the decision to make is the following:

If $BG > \chi^{2(\alpha)}_{k+p}$ H_0 is rejected

If $BG \leq \chi^{2(\alpha)}_{k+p}$ H_0 is not rejected

As a particular case the *BG* test can be applied to quarterly data using a *AR*(4) scheme.

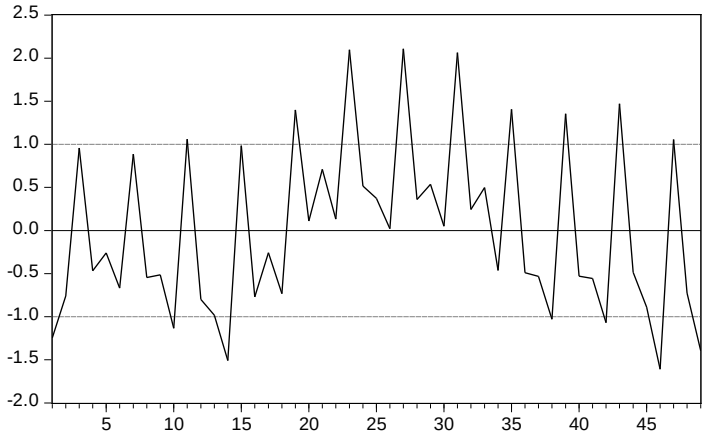
EXAMPLE 6.14 Autocorrelation in a model to explain the expenditures of residents abroad

To explain the expenditures of residents abroad (*turimp*), the following model was estimated by using quarterly data for the Spanish economy (file *qnatacsp*):

$$\widehat{\ln(turimp_t)} = \underset{(3.43)}{-17.31} + \underset{(0.276)}{2.0155}\ln(gdp_t)$$

R^2=0.531 *DW*=2.055 *n*=49

where *gdp* is gross domestic product.



GRAPHIC 6.7. Standardized residuals in the estimation of the model explaining the expenditures of residents abroad.

Graphic 6.7 shows the standardized residuals corresponding to this model. As can be seen, it appears that the residuals are not distributed in a random way because, for example, there are peaks every 4 quarters, indicating that the autocorrelation has a scheme *AR*(4).

The *BG* statistic, calculated for a *AR*(4) scheme, is equal to nR_{ar}^2 =36.35. Given this value of *BG*, the null hypothesis of no autocorrelation is rejected for α=0.01, since $\chi_5^{2(\alpha)}$ =15.09. In the auxiliary regression, in which $\hat{u}_{t-1}, \hat{u}_{t-2}, \hat{u}_{t-3}$ and $\hat{u}_{t-4}$ have been used as regressors, $\hat{u}_{t-4}$ is the only significant regressor.

6.6.4 HAC standard errors

As an extension of White's heteroskedasticity-consistent standard errors that we have seen in section 6.5.2, Newey and West proposed a method known as *HAC* (heteroskedasticity and autocorrelation consistent) standard errors that allows *OLS* standard errors to be corrected not only in situations of autocorrelation, but also in the case of heteroskedasticity. Remember that the White method was designed specifically for heteroskedasticity. It is important to point out that the Newey and West procedure is, strictly speaking, valid in large samples and may not be appropriate in small ones. Note that a sample of 50 observations is a reasonably large sample.

EXAMPLE 6.15 HAC standard errors in the case of Lydia E. Pinkham (Continuation of example 6.13)

Given the existence of autocorrelation in the model for the case of Lydia E. Pinkham, we have calculated the standard errors according to the Newey and West procedure. These standard errors allow us to make hypothesis tests on parameters correctly. The available sample is 53 observations. In table 6.9 you can find the statistics *t* obtained by the conventional procedure and the procedure *HAC*, and the ratio between them. The *t* obtained by the procedure *HAC* are slightly lower than those obtained by the conventional method, except the *advexp* coefficient whose *t* is surprisingly much higher when the procedure HAC is applied. In any case, the same conclusions are obtained for the two methods for significance levels of 0.1, 0.05 and 0.01 in the significance test of each parameter.

TABLE 6.9.The *t* statistics, conventional and HAC, in the case of Lydia E. Pinkham.

regressor	*t* conventional	*t* HAC	ratio
intercept	2.644007	1.779151	1.49
advexp	3.928965	5.723763	0.69
sales(-1)	7.45915	6.9457	1.07
*d*1	-1.499025	-1.502571	1.00
*d*2	3.225871	2.274312	1.42
*d*3	-3.019932	-2.658912	1.14

6.6.5 Autocorrelation treatment

In order to estimate an econometric model where the disturbances follow the *AR*(1) scheme, we first consider the case that the value of ρ is known. Although this is more an academic assumption which would not happen in reality, it is convenient to adopt this assumption initially for presentation purposes. Let us suppose the following linear regression model:

$$y_t = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \mathrm{L} + \beta_k x_{kt} + u_t \qquad (6\text{-}65)$$

If we lag a period in (6-65) and multiply both sides by ρ both, we obtain

$$\rho y_{t-1} = \rho\beta_1 + \rho\beta_2 x_{2,t-1} + \rho\beta_3 x_{3,t-1} + \mathrm{L} + \rho\beta_k x_{k,t-1} + \rho u_{t-1} \qquad (6\text{-}66)$$

Subtracting (6-66) from (6-65), we have:

$$y_t - \rho y_{t-1} = \beta_1(1-\rho) + \beta_2\left(x_{2t} - \rho x_{2,t-1}\right) + \mathrm{L} + \beta_k\left(x_{kt} - \rho x_{k,t-1}\right) + \left(u_t - \rho u_{t-1}\right) \qquad (6\text{-}67)$$

As can be seen, according to the scheme given in (6-53), the disturbance term of (6-67) fulfills the *CLM* assumptions.

Model (6-67) can be estimated directly by least squares if you know the value of ρ. The estimator obtained is close to the *GLS* method if the sample is large enough. The *GLS* method needs to strictly transform the observations 2 through *n* according to (6-67) scheme, but also to transform the first observation in the following way:

$$y_t\sqrt{1-\rho^2} = \beta_1\sqrt{1-\rho^2} + \beta_2\sqrt{1-\rho^2}\,x_{2t} + \mathrm{L} + \beta_k\sqrt{1-\rho^2}\,x_{kt} + \varepsilon_t \qquad (6\text{-}68)$$

When we estimate ρ together with the other model parameters, then the method is called *feasible GLS*.

In general, in the application of feasible *GLS* the transformation of the first observation according to (6-68) is ignored. Feasible *GLS* methods for estimating a model in which the disturbances follow a *AR*(1) scheme can be grouped into three blocks: a) two-step methods, b) iterative methods, and c) scanning methods.

Here we present two methods for block a), called direct method and Durbin two stages method.

In the first stage of these two methods, ρ is estimated. In the direct method, ρ is easily estimated from the *DW* statistic, using this approximate ratio DW ; $2(1-\hat{r})$. In the method of Durbin in two stages, we estimate the following regression model in which the explanatory variables are the regressors of the original model, the regressors lagged one period and the endogenous variable lagged one period:

$$y_t = \alpha_1 + \alpha_{2,0}x_{2t} + \alpha_{2,1}x_{2,t-1} + \ldots + \alpha_{k0}x_{kt} + \alpha_{k1}x_{k,t-1} + \rho y_{t-1} + \upsilon_t \qquad (6\text{-}69)$$

The coefficient of the lagged endogenous variable is precisely the parameter ρ. In the first stage, the model (6-69) is estimated by *OLS*, taking from it the estimate of ρ. In the second stage, applicable to both methods, the model is transformed with the estimation of ρ calculated in the first stage as follows:

$$y_t - \hat{\rho}y_{t-1} = \beta_1(1-\hat{\rho}) + \beta_2\left(x_{2t} - \hat{\rho}x_{2,t-1}\right) + \ldots + \beta_k\left(x_{kt} - \hat{\rho}x_{k,t-1}\right) + \xi_t \qquad (6\text{-}70)$$

Applying *OLS* to the transformed model we obtain the parameter estimates. An exposition of iterative and scanning methods can be seen in Uriel, E.; Contreras, D.; Moltó, M. L. and Peiró, A. (1990): *Econometría. El modelo lineal*. Editorial AC. Madrid.

Exercises

Exercise 6.1 Let us consider that the population model is the following**:**

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad (1)$$

Instead, the following model is estimated:

$$\hat{y}_i = \hat{\beta}_2 x_{2i} \qquad (2)$$

Is $\hat{\beta}_2$, obtained by applying *OLS* in (2), an unbiased estimator of β_3?

Exercise 6.2 Let us consider that the population model is the following:

$$y_i = \beta_2 x_i + u_i \quad (1)$$

Instead, the following model is estimated:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} \quad (2)$$

Is $\hat{\beta}_2$, obtained by applying *OLS* in (2), an unbiased estimator of β_2?

Exercise 6.3 Let the following models be:

$$imp = b_1 + b_2 gdp + b_3 rpimp + u \qquad (1)$$

$$\ln(imp) = b_1 + b_2 \ln(gdp) + b_3 \ln(rpimp) + u \qquad (2)$$

where *imp* is the import of goods, *gdp* is gross domestic product at market prices, and *rpimp* are the relative prices imports/gdp. The magnitudes *imp* and *gdp* are expressed in millions of pesetas.

a) Using a sample of the period 1971-1977 for Spain (file *imports*p), estimate models (1) and (2).

b) Interpret coefficients β_2 and β_3 in both models.

c) Apply the RESET procedure to model (1).
d) Apply the RESET procedure to model (2).
e) Choose the most adequate specification using the *p*-values obtained in sections *c)* and *d)*.

Exercise 6.4 Consider the following model of food demand

$$food = \beta_1 + \beta_2 rp + \beta_3 inc + u$$

where *food* is spending on food, *rp* are the relative prices and *inc* is disposable income.

Researcher A omitted variable *inc*, obtaining the following estimation:

$$\widehat{food}_i = \underset{(11.85)}{89.97} + \underset{(0.118)}{0.107}\ rp_i$$

Researcher B, who is more careful, got the following estimation:

$$\widehat{food}_i = \underset{(5.84)}{92.05} - \underset{(0.067)}{0.142}\, rp_i + \underset{(0.031)}{0.236}\ inc_i$$

(The numbers in parentheses are standard errors of estimators.)

Throughout the discussion between researcher A and researcher B about which of the two estimated models is most appropriate, researcher A tries to justify his oversight on account of the omission being due to a problem of multicollinearity.

a) In favor of which researcher would you be in view of the results obtained? Explain your choice.
b) Obtain analytically the bias of the estimator of β_2 in the estimation carried out by researcher A.

Exercise 6.5 The following production function is formulated:

$$\ln(output) = \beta_1 + \beta_2 \ln(labor) + \beta_3 \ln(capital) + u$$

where *output* is the amount of output produced, *labor* is the amount of labor, capital is the amount of capital.

The following data correspond to 9 companies:

$output_i$	230	140	180	270	300	240	230	350	120
$labor_i$	30	10	20	40	50	20	30	60	40
$capital_i$	160	50	100	200	240	190	160	300	150

A researcher estimates the model mistaking only 8 observations, and obtains the following results:

$$\widehat{output}_i = \underset{(1.956)}{97.259} + \underset{(0.124)}{0.970}\, labor_i + \underset{(0.027)}{0.650}\, capital_i$$

$$R^2 = 0.999 \qquad F=3422$$

The numbers in parentheses are the standard errors of the estimators and the *F* statistic corresponds to the test of the whole model.

When he realizes his mistake, he estimates the model with all observations (n=9), obtaining in this case the following results:

$$\widehat{output}_i = \underset{(32.046)}{75.479} - \underset{(1.742)}{1.970}\, labor_i + \underset{(0.376)}{1.272}\, capital_i$$

$$R^2 = 0.824 \qquad F= 14.056$$

His confusion is great when comparing the two estimates, and he cannot understand why the results become very different when using one more observation. Can we find any reason that could justify these differences?

Exercise 6.6 Suppose in the model

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

the *R*-squared obtained from regressing x_1 on x_2, which will be called $R^2_{1/2}$, is zero.

Run the following regressions:

$$y = \lambda_0 + \lambda_1 x_1 + u$$

$$y = \gamma_0 + \gamma_1 x_2 + u$$

a) Will $\hat{\lambda}_1$ be equal to $\hat{\beta}_1$ and $\hat{\gamma}_1$ be equal to $\hat{\beta}_2$?

b) Will $\hat{\beta}_0$ be equal to $\hat{\lambda}_0$ or $\hat{\beta}_0$ be equal to $\hat{\gamma}_0$?

c) Will var($\hat{\lambda}_1$) be equal to var($\hat{\beta}_1$) and var($\hat{\gamma}_1$) be equal to var($\hat{\beta}_2$)?

Exercise 6.7 An analyst wants to estimate the following model using the observations of the attached table:

$$y_i = e^{\beta_1} x_{2i}^{\beta_2} x_{3i}^{\beta_3} x_{4i}^{\beta_4} e^{u_i}$$

x_2	x_3	x_4
3	12	4
2	10	5
4	4	1
3	9	3
2	6	3
5	5	1

What problems can occur in the estimation of this model with these data?

Exercise 6.8 In exercise 4.8, using the file *airqualy*, the following model was estimated:

$$\bar{a}irqual_i = \underset{(10.19)}{97.35} + \underset{(0.0311)}{0.0956}\, popln_i - \underset{(0.0055)}{0.0170}\, medincm_i - \underset{(0.0089)}{0.0254}\, poverty_i$$

$$- \underset{(0.0017)}{0.0031}\, fueoil_i - \underset{(0.0025)}{0.0011} valadd_i$$

$$R^2 = 0.415 \qquad n = 30$$

a) Calculate the statistic *VIF* for each coefficient.

b) What is your conclusion?

Exercise 6.9 To examine the effects of firm performance on CEO salary, the following model is formulated:

$$\ln(salary) = \beta_1 + \beta_2 roa + \beta_3 \ln(sales) + \beta_4 profits + \beta_5 tenure + \beta_6 age + u$$

where *roa* is the ratio profits/assets expressed as a percentage, *tenure* is the number of years as CEO (=0 if less than six months), and *age* is age in years. Salaries are expressed in thousands of dollars, and *sales* and *profits* in millions of dollars.

a) Using the full sample (447 observations) of the file *ceoforbes*, estimate the model by *OLS.*

b) Apply the normality test to the residuals.
c) Using the first 60 observations, estimate the model by *OLS.* Compare the coefficients and the R^2 of this estimation with that obtained in section *a)*. What is your conclusion?
d) Apply the normality test to the residuals obtained in section *c)*. What is your conclusion comparing this result with that obtained in section *b)*?

Exercise 6.10 Let the following model be

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

where

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

Apply generalized least squares to estimate β_2 in model [1].

Exercise 6.11 Let the following model be

$$y_i = \beta x_i + u_i \qquad [1]$$

where

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

a) Estimate β in model [1] using generalized least squares.
b) Calculate the variance of the estimator of β.

Exercise 6.12 Let the model be

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad [1]$$

where the variance of the disturbances is equal to

$$\sigma_i^2 = \sigma^2 x_i, \quad x_i > 0, \quad \forall i$$

1) Applying *OLS* to the model [1] and taking into account the Gauss-Markov assumptions, the variance of the estimator according to (2-16) is

$$\frac{\sigma^2}{\sum (x_i - \overline{x})^2} \qquad [2]$$

2) Applying *OLS* to the model [1] and considering that $\sigma_i^2 = \sigma^2 x_i$ and the remaining Gauss-Markov assumptions, the variance of the estimator is therefore equal to

$$\frac{\sigma^2 \sum (x_i - \overline{x})^2 x_i}{(\sum (x_i - \overline{x})^2)^2} \qquad [3]$$

3) Applying *GLS* to model [1] and considering that $\sigma_i^2 = \sigma^2 x_i$ and the remaining Gauss-Markov assumptions, the variance of the estimator is therefore equal to

$$\frac{\sigma^2}{\sum \frac{(x_i - \overline{x})^2}{x_i}} \qquad [4]$$

a) Are the variances [2] and [3] correct?

b) Show that [4] is less than or equal to [3]. (*Hint*: Apply the Cauchy-Schwarz inequality which says that $\left(\sum w_i z_i\right)^2 \leq \left(\sum w_i^2\right)\left(\sum z_i^2\right)$ is true)

Exercise 6.13 Let the following model be

$$hostel = \alpha_1 + \alpha_2 inc + u$$

where *hostel* is the spending on hotels and *inc* the yearly disposable income

The following information on 9 families was obtained:

family	*hostel*	*inc*
1	13	300
2	3	200
3	38	700
4	47	900
5	14	400
6	18	500
7	25	800
8	1	100
9	21	600

Hostel and income variables are expressed in thousands of pesetas.

a) Estimate the model by *OLS.*
b) Apply the White heteroskedasticity test.
c) Apply the Breusch-Pagan-Godfrey heteroskedasticity test.
d) Do you think it is appropriate to use the above heteroskedasticity tests in this case?

Exercise 6.14 With reference to the model seen in exercise 4.5, we assume now that

$$\text{var}(\varepsilon_i) = \sigma^2 \ln(y_i)$$

a) Are, in this case, the *OLS* estimators unbiased?
b) Are the *OLS* estimators efficient?
c) Could you suggest an estimator better than *OLS*?

Exercise 6.15 Indicate and explain which of the following statements are true when there is heteroskedasticity:

a) The *OLS* estimators are no longer *BLUE*.
b) The *OLS* estimators $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ are inconsistent.
c) The conventional *t* and *F* tests are not valid.

Exercise 6.16 In exercise 3.19, using the file *consumsp*, the Brown model was estimated for the Spanish economy in the period 1954-2010. The results obtained were the following:

$$\widehat{conspc}_t = \underset{(84.88)}{-7.156} + \underset{(0.0857)}{0.3965}\, incpc_t + \underset{(0.0903)}{0.5771}\, conspc_{t-1}$$

$$R^2=0.997 \qquad RSS=1891320 \qquad n=56$$

Using the residuals of the above fitted model, the following regression was obtained:

$$(\hat{u}_t^2) = 141568 + 89.71 incpc_t - 149.2 conspc_{t-1}$$
$$- 0.183 incpc_t^2 - 0.221 conspc_{t-1}^2 + 0.406 incpc_t \acute{} conspc_{t-1}$$
$$R^2=0.285$$

a) Is there heteroskedasticity in the consumption function?

b) The following estimation, with White heteroskedasticity-consistent standard errors, is obtained:

$$\bar{conspc}_t = \underset{(66.92)}{?} + \underset{(0.0669)}{?}\, incpc_t + \underset{(0.0741)}{?}\, conspc_{t-1}$$

Can you fill the blanks above? Please do so.

Explain the difference between the White heteroskedasticity- consistent standard errors and the usual standard errors of the initial equation.

c) Test whether the coefficient on *incpc* is equal to 0.5. What standard errors are you going to use in the inference process? Why?

Exercise 6.17 Assume the following specification:

$$c_i = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i$$
$$\sigma_i^2 = \sigma^2 h_i^2$$

Would it be appropriate to eliminate the heteroskedasticity to perform the following transformation?

$$\frac{c_i}{h_i} = \gamma_1 + \gamma_2 h_i + \gamma_3 m_i + u_i \text{ ?}$$

Explain your answer.

Exercise 6.18 Let the following model be

$$y = \beta_1 + \beta_2 x + u$$

and we have the following information:

y_i	x_i	$\hat{u}_i$
2	-3	1.37
3	-2	-0.42
7	-1	0.79
6	0	-3.00
15	1	3.21
8	2	-6.58
22	3	4.63

a) Apply the White heteroskedasticity test.

b) Apply the Breusch-Pagan-Godfrey heteroskedasticity test.

c) Why is the significance obtained with both tests so different?

Exercise 6.19 Answer the following questions

a) Explain in detail what is the problem of heteroskedasticity in the linear regression model.

b) Illustrate briefly the problem of heteroskedasticity with an example.

c) Propose solutions to the heteroskedasticity problem.

Exercise 6.20 Using a sample corresponding to 17 regions, the following estimations were obtained:

$$\hat{y}_i = -309.8 + 0.76z_i + 3.05h_i \qquad R^2 = 0.989$$

$$\hat{u}_i^2 = -1737.2 - 17.8z_i + 0.09z_i^2 + 0.65z_i h_i + 10.6h_i - 0.31h_i^2 \qquad R^2 = 0.705$$

where *y* is the expenditure on education, *z* is GDP and *h* is the number of inhabitants.

a) Is there a problem of heteroskedasticity? Detail the procedure followed in testing.

b) Assuming that the presence of heteroskedasticity is detected in the regression model, what solution would you take to test the significance of the explanatory variables of the model? Explain your answer.

Exercise 6.21 Using data from Spanish economy for the period 1971-1997 (file *importsp*), the following model was estimated to explain the Spanish imports (*imp*):

$$\widehat{\ln}(imp_t) = \underset{(2.81)}{-26.58} + \underset{(0.162)}{2.4336}\ln(gdp_t) - \underset{(0.021)}{0.4494}\ln(rpimp_t)$$

$$R^2=0.997 \quad n=27$$

where *gdp* is the gross domestic product at market prices, and *rpimp* are the relative prices imports/gdp. The variables *imp* and *gdp* are expressed in millions of pesetas.

a) Set up and estimate the auxiliary regression to perform the Breusch-Pagan-Godfrey heteroskedasticity test.

b) Apply the Breusch-Pagan-Godfrey heteroskedasticity test using the auxiliary regression run in section *a)*.

c) Set up the auxiliary regression to perform the *complete* White heteroskedasticity test.

d) Apply the *complete* White heteroskedasticity test using the auxiliary regression run in section *c)*.

e) Set up the auxiliary regression to perform the *simplified* White heteroskedasticity test.

f) Apply the *simplified* White heteroskedasticity test using the auxiliary regression run in section *e)*.

g) Compare the results of the test carried out in sections *b*), *d*) and *f*).

Exercise 6.22 Using data from file *tradocde*, the following model has been estimated to explain the imports (*impor*) in OECD countries:

$$\widehat{\ln}(impor_i) = \underset{(6.67)}{18.01} + \underset{(0.658)}{1.6425}\ln(gdp_i) - \underset{(0.636)}{0.5151}\ln(popul_i)$$

$$R^2=0.614 \qquad n=34$$

where *gdp* is gross domestic product at market prices, and *popul* is the population of each country.

a) What is the interpretation of the coefficient on ln(*gdp*)?

b) Set up the auxiliary regression to perform the White heteroskedasticity test.

c) Apply the White heteroskedasticity test using the auxiliary regression run in section *b)*.

d) Test whether the *import*/*gdp* elasticity is greater than 1. To make this test, do you need to use the White heteroskedasticity-robust standard errors?

Exercise 6.23 Explain in detail what the appropriate autocorrelation test would be in each situation:

a) When the model has no lagged endogenous variables and the observations are annual.

b) When the model has lagged endogenous variables and the observations are annual.

c) When the model has no lagged endogenous variables and the observations are quarterly.

Exercise 6.24 Two alternative models were used to estimate the average cost of annual car production of a particular brand in the period 1980-1999:

$$c = \alpha + \beta p + u \qquad R^2 = 0.848; \quad \bar{R}^2 = 0.812; \quad d = DW = 0.51$$

$$c = \alpha + \beta p + \gamma p^2 + u \qquad R^2 = 0.852; \quad \bar{R}^2 = 0.811; \quad d = DW = 2.11$$

a) When comparing the two estimations, indicate if you detect any econometric problem. Explain it.

b) Depending on your answer to the previous section, which of the two models would you choose?

Exercise 6.25 In the period 1950-1980, the following production is estimated

$$\ln(o_t) = \underset{(0.24)}{-3.94} + \underset{(0.083)}{1.45}\ln(l_t) + \underset{(0.048)}{0.38}\ln(k_t)$$

$$R^2 = 0.994 \qquad DW = 0.858 \qquad \hat{\rho} = 0.559$$

where o is output, l is labor, and k is capital.

(The numbers in parentheses are standard errors of the estimators.)

a) Test whether there is autocorrelation.

b) If the model had a lagged endogenous variable as an explanatory variable, indicate how you would test whether there is autocorrelation.

Exercise 6.26 Using 38 annual observations, the following demand function for a product was estimated:

$$d_i = 2.47 + \underset{(0.39)}{0.35}\, p_i + \underset{(0.06)}{0.9}\, d_{i-1} \qquad R^2 = 0.98 \qquad DW = 1.82$$

where d is the quantity demanded, and p is the price.

(The numbers in parentheses are standard errors of the estimators.

a) Is there a problem of autocorrelation? Explain your answer.

b) List the conditions under which it would be appropriate to use the Durbin Watson statistic.

Exercise 6.27 The following model of housing demand with annual observations for the period 1960-1994 is estimated:

$$\widehat{\ln(rent_t)} = \underset{(0.15)}{-0.39} + \underset{(0.05)}{0.31}\ln(inc_t) - \underset{(0.02)}{0.67}\ln(price_t) + \underset{(0.04)}{0.70}\ln(rent_{t-1})$$

$$R^2 = 0.999 \qquad DW = 0.52$$

where v is spending on rent, r is disposable income, p is the price of housing

(The numbers in parentheses are standard deviations of the estimators).

a) Test whether there is autocorrelation.
b) Taking into account the conclusions reached in section *a)*, how would you carry out the significance tests for each one of the coefficients? Explain your answer.

Exercise 6.28 Answer the following questions:

a) In a model to explain the sales, the estimation is carried out using quarterly data. Explain how you can reasonably test whether there is autocorrelation.
b) Describe in detail, introducing assumptions that you consider appropriate, how you would estimate the model when the null hypothesis of no autocorrelation is rejected.

Exercise 6.29 In the estimation of the Keynesian consumption function for the French economy, the following results were obtained:

$$\hat{cons}_t = \underset{(0.73)}{1.22} + \underset{(79.39)}{0.854}\, inc_t$$

$$R^2 = 0.983 \quad \text{DW}=0.4205 \qquad n=30$$

(The numbers in parentheses are the *t* statistics of the estimators).

A researcher believes the focus should be placed on the saving function, rather than on the consumption function, proposing the following model:

$$saving_t = a_1 + a_2 inc_t + v_t$$

where

$$saving_t = inc_t - cons_t$$

a) Obtain the estimates of α_1 and α_2.
b) Estimate the variances of $\hat{\alpha}_1$ and $\hat{\alpha}_2$.
c) Calculate the DW statistic of the saving model.
d) Calculate the R^2 of the saving model.

Exercise 6.30 Let the model be

$$\begin{aligned} y_t &= \beta x_t + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t; \quad E\left[\varepsilon_t^2\right] = \sigma^2 \qquad \forall i \end{aligned} \qquad [1]$$

a) If model [1] is transformed by taking differences first, under what circumstances is the transformed model preferable to model [1]?
b) Is it appropriate to use the R^2 to compare model [1] and the transformed model? Explain your answer.

Exercise 6.31 Let the model be:

$$y_t = \beta_1 + \beta_2 x_t + u_t \qquad [1]$$

The following sample of observations is disposable for the variables *x* and *y*:

y_i	6	3	1	1	1	4	6	16	25	36	49	64
x_i	-4	-3	-2	-1	1	2	3	4	5	6	7	8

a) Estimate the model [1] by OLS and calculate the corresponding adjusted determination coefficient.
b) Calculate the Durbin-Watson statistic for the estimations made in *a*).

c) In view of the Durbin and Watson test and the representation of the fitted line and residuals, is it appropriate to reformulate model [1]? Justify your answer and, if it is yes, estimate the alternative model that you consider the most appropriate for the data.

Exercise 6.32 Let the model be:

$$y_t = \beta_1 + \beta_2 x_t + u_t$$

$$u_t = r\,u_{t-1} + e_t; \qquad e_t : \; NI\left(0, s^2\right)$$

The following additional information is also disposable:

$$\rho = 0.5$$

y_i	22	26	32	31	40	46	46	50
x_i	4	6	10	12	13	16	20	22

a) Estimate the model by *OLS*.

b) Estimate the model by *GLS* without transforming the first observation.

c) Which of the two estimators of β_2 is more efficient?

Exercise 6.33 In a study on product demand, the following results were obtained:

$$\hat{y}_t = \underset{(7.17)}{2.30} + \underset{(0.05)}{0.86}\, x_t$$

$$R^2 = 0.9687 \quad \text{DW=3.4} \quad n = 15$$

(The numbers in parentheses are standard errors of the estimators.)

Furthermore, the following additional information about the residual regressions is disposable:

1. $\left|\hat{u}_t\right| = \underset{(0.210)}{0.167} + \underset{(0.180)}{0.127}\, x_t$

2. $\left|\hat{u}_t\right| = \underset{(0.098)}{0.231} + \underset{(0.095)}{0.218}\, x_t^{1/2}$

a) Detect whether there is autocorrelation.

b) Detect whether there is heteroskedasticity.

c) What would be the most appropriate procedure to solve the potential problem of heteroskedasticity?

Exercise 6.34 Using a sample of the period 1971-1997 (file *importsp*), the following model was estimated, using *HAC* standard errors, to explain the imports of goods in Spain (*imp*):

$$\widehat{\ln}(imp_t) = -\underset{(3.65)}{26.58} + \underset{(0.210)}{2.434}\ln(gdp_t) - \underset{(0.023)}{0.4494}\ln(rpimp_{t-1})$$

$$R^2 = 0.997 \quad \text{DW=0.73} \quad n = 27$$

where *gdp* is gross domestic product at market prices, and *rpimp* are the relative prices import/gdp. Both magnitudes are expressed in millions of pesetas.

(The numbers in parentheses are standard errors of the estimators.)

a) Interpret the coefficient on *rpimp*.

b) Is there autocorrelation in this model?

c) Test whether the *imp/gdp* elasticity plus four times the *imp/rpimp* elasticity is equal to zero. (Additional information: $\text{var}(\hat{\beta}_2)$ =0.044247; $\text{var}(\hat{\beta}_3)$ =0.000540; and $\text{var}(\hat{\beta}_2, \hat{\beta}_3)$ =0.004464).

d) Test the overall significance of this model.

Exercise 6.35 Using a sample for the period 1954-2009 (file *electsp*), the following model was estimated to explain the electricity consumption in Spain (*conselec*):

$$\widehat{\ln(conselec_t)} = \underset{}{-} \underset{(0.46)}{9.98} + \underset{(0.035)}{1.469}\ln(gdp_t)$$

$$R^2 = 0.9805 \quad \text{DW}=0.18 \quad n = 37 \qquad (1)$$

where *gdp* is gross domestic product at 1986 market prices. The variable *conselec* is expressed in a thousand tonnes of oil equivalent (*ktoe*) and *gdp* is expressed in millions of pesetas.

(The numbers in parentheses are standard errors of the estimators.)

a) Test whether there is autocorrelation applying the Durbin-Watson statistic.

b) Test whether there is autocorrelation applying the Breusch-Godfrey statistic for a *AR*(2) scheme.

c) The following model is also estimated:

$$\widehat{\log(conselec_t)} = \underset{}{-} \underset{(0.75)}{0.917} + \underset{(0.107)}{0.164}\log(gdp_t) + \underset{(0.072)}{0.871}\log(conselec_{t-1})$$

$$R^2 = 0.997 \quad \text{DW}=0.93 \quad n = 36 \qquad (2)$$

Test whether there is autocorrelation applying the procedure you consider appropriate.

d) Test whether the *conselec/gdp* elasticity in an equilibrium situation ($\ln(conselec^e) = b_1 + b_2 \ln(gdp^e) + b_3 \ln(conselec^e)$) is greater than 1, using an adequate procedure.

Exercise 6.36 The Phillips curve represents the relationship between the rate of inflation (*inf*) and the unemployment rate (*unemp*). While it has been observed that there is a stable short run tradeoff between unemployment and inflation, this has not been observed in the long run.

The following model reflects the Phillips curve:

$$inf = \beta_1 + \beta_2 unempl + u$$

Using a sample for the Spanish economy in the period 1970-2010 (file *phillipsp*), the following results were obtained:

$$\widehat{inf}_t = \underset{(1.79)}{12.59} - \underset{(0.120)}{0.3712}\,unempl_t$$

$$R^2=0.198; \qquad DW=0.219; \qquad n=41$$

(The numbers in parentheses are standard deviations of the estimators).

a) Interpret the coefficient on *unempl*.

b) Test whether there is first order autocorrelation using Durbin and Watson test.

c) Using the disposable information so far, can you test the significance of the coefficient on *unempl* adequately?

d) Using the *HAC* standard errors, test the significance of the coefficient on *unempl*.

Exercise 6.37 It is important to remark that the Phillips curve is a relative relationship. Inflation is considered low or high relative to the expected rate of inflation and unemployment is considered low or high relative to the so-called natural rate of unemployment. In the *augmented* Phillips curve this is taken into account:

$$inf_t - inf^e_{t/t-1} = \beta_2(unempl_t - \lambda_0) + u_t$$

where λ_0 is the natural rate of unemployment and $inf^e_{t/t-1}$ is the expected rate of inflation for *t* formed in *t*-1. If we consider that the expected inflation for *t* is equal to the inflation in *t*-1 ($inf^e_{t/t-1} = inf_{t-1}$) and $\beta_1 = -\beta_2\lambda_0$, the augmented Phillips curve can be written as:

$$inf_t - inf_{t-1} = \beta_1 + \beta_2 unempl_t + u_t$$

a) Using file *phillipsp*, estimate the above model.
b) Interpret the coefficient on *unempl*.
c) Test whether there is second order autocorrelation.
d) Test whether the natural rate of unemployment is greater than 10.

Appendix 6.1

First we are going to express the $\hat{\beta}_2$ taking into account that *y* is generated by the model (6-8):

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(y_i-\bar{y})}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} = \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)y_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}$$

$$= \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)(\beta_1+\beta_2 x_{1i}+\beta_3 x_{2i}+u_i)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}$$

$$= \beta_2 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{1i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)u_i}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_2)^2} \qquad (6\text{-}71)$$

If we take expectations on both sides of (6-71), we have

$$E(\tilde{\beta}_2) = \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)^2} + \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)E(u_i \mid x_2, x_3)}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)^2}$$
$$= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)x_{2i}}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_2)^2} \qquad (6\text{-}72)$$