3 Multiple linear regression: estimation and properties

Ezequiel Uriel
Universidad de Valencia
Version: 09-2013

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:

$$cons = \beta_1 + \beta_2 inc + u \tag{3-1}$$

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:

$$cons = \beta_1 + \beta_2 inc + \beta_3 wealth + u \tag{3-2}$$

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:

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

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:

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

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:

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

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:

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

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:

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

The parameters $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ are fixed and unknown.

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

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

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

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

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

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

The matrix $\mathbf{X}$ 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:

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

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

3.1.2 Sample regression function

The basic idea of regression is to estimate the population parameters, $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ 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

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \tag{3-14}$$

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

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

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \tag{3-15}$$

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

The system of equations (3-14) can be expressed in a compact form by using matrix notation. Thus, we are going to denote

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

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

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-16}$$

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

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{3-17}$$

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,

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

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):

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-1]$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{2i}]$$

$$\frac{\partial S}{\partial \hat{\beta}_3} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{3i}] \qquad (3\text{-}19)$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$\frac{\partial S}{\partial \hat{\beta}_k} = 2\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right][-x_{ki}]$$

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

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right] = 0$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]x_{2i} = 0$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]x_{3i} = 0 \qquad (3\text{-}20)$$

$$\cdots \qquad \cdots \qquad \cdots \qquad \cdots$$

$$\sum_{i=1}^{n}\left[y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki} \right]x_{ki} = 0$$

or, in matrix notation,

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

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) $\mathbf{X'X}/n$ 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) $\mathbf{X'y}/n$ 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 $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \cdots, \hat{\beta}_k)$. 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 $\mathbf{X'X}$ is equal to k. If this is held, both members of (3-21) can be premultiplied by $[\mathbf{X'X}]^{-1}$:

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

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 $[\mathbf{X'X}]^{-1}\mathbf{X'X} = \mathbf{I}$. Therefore, the solution is the following:

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \hat{\boldsymbol{\beta}} = [\mathbf{X'X}]^{-1}\mathbf{X'y} \tag{3-23}$$

Since the matrix of second derivatives, $2\mathbf{X'X}$, 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

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \cdots + \hat{\beta}_j x_{ji} + \cdots + \hat{\beta}_k x_{ki} \tag{3-24}$$

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):

$$\hat{y}_h = \hat{\beta}_1 + \hat{\beta}_2 x_{2h} + \hat{\beta}_3 x_{3h} + \cdots + \hat{\beta}_j x_{jh} + \cdots + \hat{\beta}_k x_{kh} \tag{3-25}$$

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

$$\Delta\hat{y} = \hat{\beta}_2 \Delta x_2 + \hat{\beta}_3 \Delta x_3 + \cdots + \hat{\beta}_j \Delta x_j + \cdots + \hat{\beta}_k \Delta x_k \tag{3-26}$$

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

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

$$\Delta\hat{y} = \hat{\beta}_j \Delta x_j \tag{3-27}$$

If x_k increases in one unit, we will have

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

Consequently, the coefficient $\hat{\beta}_j$ measures the change in y when x_j increases in 1 unit, *holding the regressors* $x_2, x_3, \cdots, x_{j-1}, x_{j+1}, \cdots, x_k$ *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:

$$V_t = \alpha + \beta_1 P_t + \beta_2 P_{t-1} + \beta_3 P_{t-2} + \cdots + u_t \tag{3-30}$$

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:

$$\beta_i = \beta_1 \lambda^i \quad \forall i \qquad 0 < \lambda < 1 \tag{3-31}$$

The above transformation is called Koyck transformation, as it was this author who in 1954 introduced scheme $(3\text{-}31)$ for the study of investment

Substituting (3-31) in (3-30), we obtain

$$V_t = \alpha + \beta_1 P_t + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \cdots + u_t \tag{3-32}$$

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

$$\lambda V_{t-1} = \alpha\lambda + \beta_1 \lambda P_{t-1} + \beta_1 \lambda^2 P_{t-2} + \beta_1 \lambda^3 P_{t-3} + \cdots + \lambda u_{t-1} \tag{3-33}$$

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:

$$V_t = \alpha(1-\lambda) + \beta_1 P_t + \lambda V_{t-1} + u_t - \lambda u_{t-1} \tag{3-34}$$

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

$$\beta_1(1 + \lambda + \lambda^2 + \lambda^3 + \cdots + \lambda^m) \tag{3-35}$$

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[1], which can be expressed as follows:

---

[1] 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

$$\frac{\beta_1(1-\lambda^m)}{1-\lambda} \tag{3-36}$$

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 $(3\text{-}34)$ with data from 1907 to 1960, using file *pinkham*, is the following:

$$\widehat{sales_t} = 138.7 + 0.3288\,advexp + 0.7593\,sales_{t-1}$$

$$R^2 = 0.877 \quad n = 53$$

The sum of the cumulative effects of advertising expenditures on sales is calculated by the formula (3-37):

$$\frac{a_p - a_u}{1-r}$$

$$\frac{\hat{\beta_1}}{1-\hat{\lambda}} = \frac{0.3288}{1-0.7593} = 1.3660$$

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

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki} \qquad i = 1, 2, \cdots, n \qquad (3\text{-}41)$$

If we add for the *n* observations, then

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} \qquad (3\text{-}42)$$

On the other hand, the first equation of the system of normal equations (3-20) is

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_1 - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \cdots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = 0 \qquad (3\text{-}43)$$

If we compare (3-42) and (3-43), we conclude that (3-40) holds.

Note that, if (3-40) holds, it implies that

$$\sum_{i=1}^{n} y = \sum_{i=1}^{n} \hat{y}_i \qquad (3\text{-}44)$$

and, dividing (3-40) and (3-44) by *n,* we obtain

$$\overline{\hat{u}} = 0 \qquad \overline{y} = \overline{\hat{y}} \qquad (3\text{-}45)$$

2. *The OLS hyperplane always goes through the point of the sample means* $(\overline{y}, \overline{x}_2, \cdots, \overline{x}_k)$.

By dividing equation (3-43) by *n* we have:

$$\overline{y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{x}_2 + \cdots + \hat{\beta}_k \overline{x}_k \qquad (3\text{-}46)$$

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

$$\sum_{i=1}^{n} x_{ji}\hat{u}_i = 0 \qquad\qquad j = 2,3,\cdots,k \tag{3-47}$$

Using the last k normal equations (3-20) and taking into account that by definition $\hat{u}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{2i} - \hat{\beta}_3 x_{3i} - \cdots - \hat{\beta}_k x_{ki}$, we can see that

$$\sum_{i=1}^{n} \hat{u}_i x_{2i} = 0$$

$$\sum_{i=1}^{n} \hat{u}_i x_{3i} = 0 \tag{3-48}$$

$$\cdots \qquad\qquad \cdots$$

$$\sum_{i=1}^{n} \hat{u}_i x_{ki} = 0$$

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

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0 \tag{3-49}$$

Taking into account (3-40) and (3-48), we obtain

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} (\hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki})\hat{u}_i = \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}\hat{u}_i + \cdots \hat{\beta}_k \sum_{i=1}^{n} x_{ki}\hat{u}_i \tag{3-50}$$
$$= \hat{\beta}_1 \times 0 + \hat{\beta}_2 \times 0 + \cdots \hat{\beta}_k \times 0 = 0$$

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:

$$y = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u \tag{3-51}$$

or, alternatively, for all the observations,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{3-52}$$

b) Assumptions on the regressors

2) The values of $x_2, x_3 \cdots, x_k$ are fixed in repeated sampling, or the matrix $\mathbf{X}$ 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 $x_2, x_3, \cdots, x_k$ are distributed independently of the random disturbance. Formulated in another way, **X** is distributed independently of the vector of random disturbances, which implies that $E(\mathbf{X'u}) = \mathbf{0}$

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 $\beta_1, \beta_2, \beta_3, \cdots, \beta_k$ are constant, or $\boldsymbol{\beta}$ is a constant vector.

d) Assumptions on the disturbances

6) The disturbances have zero mean,

$$E(u_i) = 0, \quad i = 1, 2, 3, \ldots, n \ \text{ or } \ E(\mathbf{u}) = \mathbf{0} \qquad (3\text{-}54)$$

7) The disturbances have a constant variance (homoskedasticity assumption):

$$var(u_i) = \sigma^2 \quad i = 1, 2, \cdots n \qquad (3\text{-}55)$$

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:

$$E\Big[[\mathbf{u}-E(\mathbf{u})][\mathbf{u}-E(\mathbf{u})]'\Big] = E\Big[[\mathbf{u}-\mathbf{0}][\mathbf{u}-\mathbf{0}]'\Big] = E\Big[[\mathbf{u}][\mathbf{u}]'\Big]$$

$$= E\left[\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}\right] = E\begin{bmatrix} u_1^2 & u_1 u_2 & \cdots & u_1 u_n \\ u_2 u_1 & u_2^2 & \cdots & u_2 u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_n u_1 & u_n u_2 & \cdots & u_n^2 \end{bmatrix} \quad \text{(3-57)}$$

$$= \begin{bmatrix} E(u_1^2) & E(u_1 u_2) & \cdots & E(u_1 u_n) \\ E(u_2 u_1) & E(u_2^2) & \cdots & E(u_2 u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_n u_1) & E(u_n u_2) & \cdots & E(u_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

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:

$$E(\mathbf{u}\mathbf{u}') = \sigma^2 \mathbf{I} \quad \text{(3-58)}$$

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

$$u_i \sim NID(0, \sigma^2) \quad i = 1, 2, \cdots, n \qquad \text{or} \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad \text{(3-59)}$$

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):

$$\hat{\boldsymbol{\beta}} = \big[\mathbf{X}'\mathbf{X}\big]^{-1}\mathbf{X}'\mathbf{y} = \big[\mathbf{X}'\mathbf{X}\big]^{-1}\mathbf{X}'\big[\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\big] = \boldsymbol{\beta} + \big[\mathbf{X}'\mathbf{X}\big]^{-1}\mathbf{X}'\mathbf{u} \quad \text{(3-60)}$$

The *OLS* estimator can be expressed in this way so that the property of linearity is clearer:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \big[\mathbf{X}'\mathbf{X}\big]^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + \mathbf{A}\mathbf{u} \quad \text{(3-61)}$$

where $\mathbf{A} = \big[\mathbf{X}'\mathbf{X}\big]^{-1}\mathbf{X}'$ 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

$$E\left[\hat{\boldsymbol{\beta}}\right] = \boldsymbol{\beta} + \left[\mathbf{X'X}\right]^{-1} \mathbf{X'}E\left[\mathbf{u}\right] = \boldsymbol{\beta} \tag{3-62}$$

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:

$$\text{var}(\hat{\boldsymbol{\beta}}) = E\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]\left[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})\right]' = E\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right]\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right]'$$

$$= E\left[\left[\mathbf{X'X}\right]^{-1} \mathbf{X'uu'X}\left[\mathbf{X'X}\right]^{-1}\right] = \left[\mathbf{X'X}\right]^{-1} \mathbf{X'}E(\mathbf{uu'})\mathbf{X}\left[\mathbf{X'X}\right]^{-1} \tag{3-63}$$

$$= \left[\mathbf{X'X}\right]^{-1} \mathbf{X'}E(\sigma^2\mathbf{I})\mathbf{X}\left[\mathbf{X'X}\right]^{-1} = \sigma^2\left[\mathbf{X'X}\right]^{-1}$$

In the third step of the above proof it is taken into account that, according to (3-60), $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left[\mathbf{X'X}\right]^{-1}\mathbf{X'u}$. Assumption 2 is taken into account in the fourth step. Finally, assumptions 7 and 8 are used in the last step.

Therefore, $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\mathbf{X'X}\right]^{-1}$ 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 $j=2,3,\ldots,k$) is equal to σ^2 multiplied by the corresponding element of the main diagonal of $\left[\mathbf{X'X}\right]^{-1}$. After operating, the variance of $\hat{\beta}_j$ can be expressed as

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{nS_j^2(1-R_j^2)} \tag{3-64}$$

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* $\hat{\beta}_j$:

$$sd(\hat{\beta}_j) = \frac{\sigma}{\sqrt{nS_j^2(1-R_j^2)}} \tag{3-65}$$

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

14

$\tilde{\beta}_j = \sum_{i=1}^{n} w_{ij} y_i$) 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:

$$\hat{u}_n = -\hat{u}_1 - \hat{u}_2 - \cdots - \hat{u}_{n-1}$$

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

$$\widehat{Var(\hat{\boldsymbol{\beta}})} = \hat{\sigma}^2\left[\mathbf{X'X}\right]^{-1} = \begin{bmatrix} \widehat{\text{var}(\hat{\beta}_1)} & \widehat{Cov(\hat{\beta}_1,\hat{\beta}_2)} & \cdots & \widehat{Cov(\hat{\beta}_1,\hat{\beta}_j)} & \cdots & \widehat{Cov(\hat{\beta}_1,\hat{\beta}_k)} \\ \widehat{Cov(\hat{\beta}_2,\hat{\beta}_1)} & \widehat{\text{var}(\hat{\beta}_2)} & \cdots & \widehat{Cov(\hat{\beta}_2,\hat{\beta}_j)} & \cdots & \widehat{Cov(\hat{\beta}_2,\hat{\beta}_k)} \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \widehat{Cov(\hat{\beta}_j,\hat{\beta}_1)} & \widehat{Cov(\hat{\beta}_j,\hat{\beta}_2)} & \cdots & \widehat{\text{var}(\hat{\beta}_j)} & \cdots & \widehat{Cov(\hat{\beta}_j,\hat{\beta}_k)} \\ \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ \widehat{Cov(\hat{\beta}_k,\hat{\beta}_1)} & \widehat{Cov(\hat{\beta}_k,\hat{\beta}_2)} & \cdots & \widehat{Cov(\hat{\beta}_k,\hat{\beta}_j)} & \cdots & \widehat{\text{var}(\hat{\beta}_k)} \end{bmatrix} \quad (3\text{-}69)$$

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:

$$\widehat{\text{var}(\hat{\beta}_j)} = \frac{\hat{\sigma}^2}{nS_j^2(1-R_j^2)} \quad (3\text{-}70)$$

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 $\hat{\beta}_j$.
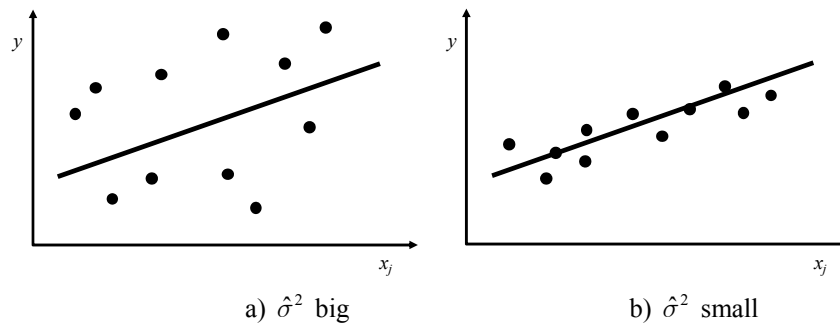
a) $\hat{\sigma}^2$ big   b) $\hat{\sigma}^2$ small

FIGURE 3.1. Influence of $\hat{\sigma}^2$ on the estimator of the variance.

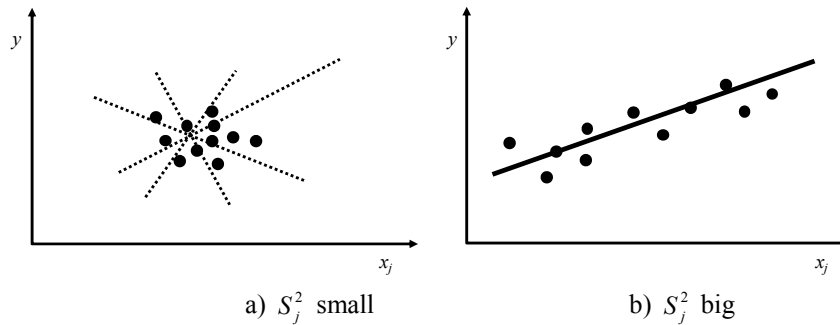

a) S_j^2 small   b) S_j^2 big

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 $\hat{\beta}_j$* :

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{nS_j^2(1-R_j^2)}} \tag{3-71}$$

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

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + \cdots + \beta_k x^k + u \tag{3-72}$$

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*:

$$y = \beta_1 + \beta_2 x + \beta_3 x^2 + u \tag{3-73}$$

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:

$$me = \frac{dy}{dx} = \beta_2 + 2\beta_3 x \tag{3-74}$$

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

$$x^* = -\frac{\beta_2}{2\beta_3} \tag{3-75}$$

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:

$$\widehat{\ln(salary)} = 6.246 + \underset{(0.086)}{0.0006} profits + \underset{(0.0001)}{0.0440} ceoten - \underset{(0.0156)}{0.0012} ceoten^2$$
$$\underset{(0.00052)}{}$$

$$R^2 = 0.1976 \quad n = 177$$

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 $\bar{R}^2$ 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_{j-1}^2 the R is squared in a model with j-1 regressors, and R_j^2 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

$$R^2 = 1 - \frac{RSS/n}{TSS/n} \tag{3-85}$$

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

$$R^2_{POP} = 1 - \frac{\sigma_u^2}{\sigma_y^2} \tag{3-86}$$

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

$$\bar{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-k} \tag{3-87}$$

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:

$$\bar{R}^2_j \geq \bar{R}^2_{j-1} \quad \text{or} \quad \bar{R}^2_j \leq \bar{R}^2_{j-1} \tag{3-88}$$

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:

$$AIC = -\frac{2l}{n} + \frac{2k}{n} \tag{3-89}$$

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:

$$SC = -\frac{2l}{n} + \frac{k\ln(n)}{n} \tag{3-90}$$

The *AIC* and *SC* statistics, unlike the coefficients of determination (R^2 and $\bar{R}^2$), 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:

$$AIC_C = AIC + 2\overline{\ln(Y)} \tag{3-91}$$

$$SC_C = SC + 2\overline{\ln(Y)} \tag{3-92}$$

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:

$$AIC_C = AIC + 2\overline{\ln(Y)} = 0.2794 + 2 \times 2.3719 = 5.0232$$

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 $y = X\beta + u$, where X is a matrix 50×5.

Answer the following questions, justifying your answers:

      a) What are the dimensions of the vectors y, β, u ?

      b) How many equations are there in the system of normal equations $X'X\hat{\beta} = X'y$?

      c) What conditions are needed in order to obtain $\hat{\beta}$?

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**

	X
	1 100
	1 110
	1 130
	1 100
	1 80
	1 80
	1 90
	1 120
	1 120
	1 90

X'
1	1	1	1	1	1	1	1	1	1
100	110	130	100	80	80	90	120	120	90

X'X
10	1020
1020	106800

(2*10)   (10*2)   (2*2)   Result 1

X'
1	1	1	1	1	1	1	1	1	1
100	110	130	100	80	80	90	120	120	90

y
10
8
7
6
13
6
12
7
9
15

X'y
93
9260

(2*10)   (0*1)   (2*1)   Result 2

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 **(X'X)⁻¹**

X'X
10	1020
1020	106800

(X'X)⁻¹
3,8696	-0,0370
-0,0370	0,0004

Result 1 (2*2)   Result 3 (2*2)

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 $\hat{\boldsymbol{\beta}}$

(X'X)⁻¹
3,8696	-0,0370
-0,0370	0,0004

X'y
93
9260

$\hat{\beta}$
17,6522
-0,0819

Result 3 (2*2)   Result 2 (2*1)   Result 4 (2*1)

4) Calculation of $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ and σ^2

y'
10	8	7	6	13	6	12	7	9	15

y
10
8
7
6
13
6
12
7
9
15

y'y
953

(1*10)   (10*1)   (1*1)   Result 5

$\hat{\beta}$
17,6522	-0,0819

X'y
93
9260

$\hat{\beta}'X'y$
883

Result 4 (1*2)   Result 2 (2*1)   Result 6 (1*1)

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{y}'\mathbf{y} - \hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = R.5 - R.6 = 953 - 883 = 70$$

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-2} = \frac{70}{8} = 8.6993$$

5) Calculation of covariance matrix of $\hat{\boldsymbol{\beta}}$

$$\text{var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left[\mathbf{X}'\mathbf{X}\right]^{-1} = 8.6993\begin{pmatrix} 3.8696 & -0.0370 \\ -0.0370 & 0.0004 \end{pmatrix} = \begin{pmatrix} 33.6624 & -0.3215 \\ -0.3215 & 0.0032 \end{pmatrix}$$

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: $\alpha = 0$, $\beta = 1$, and $\gamma = 1$

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

$$\ln(y)=\beta_1+\beta_2\ln(x)+u \tag{1}$$
$$\ln(y)=\beta_1+\beta_2\ln(x)+\beta_3\ln(z)+u \tag{2}$$
$$\ln(y)=\beta_1+\beta_2\ln(z)+u \tag{3}$$
$$y=\beta_1+\beta_2 z+u \tag{4}$$

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

$$y=X\beta+u$$

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:

$$\mathbf{X'X} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{bmatrix} \qquad \mathbf{\hat{u}'\hat{u}} = 22$$

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 $\hat{\beta}$, 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:

$$\widehat{sleep} = 1440 - 1 \times \ total\_work - 1 \times leisure$$

$$R^2 = 1.000 \quad n = 1000$$

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:

$$\widehat{\ln(wage)} = 1.565 + 0.0730educ + 0.0177tenure + 0.0065age$$

$$R^2 = 0.337 \quad n = 800$$

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:

$$\widehat{price} = -2418 + 5827bedrooms + 19750bathrms + 5.411lotsize$$

$$R^2 = 0.486 \quad n = 546$$

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:

$$\widehat{\ln(salary)} = 4.641 + 0.0054roa + 0.2893\ln(sales) + 0.0000564profits + 0.0122tenure$$

$$R^2 = 0.232 \quad n = 447$$

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:

$$\widehat{rdintens} = -1.8168 + 0.1482\ln(sales) + 0.0110exponsal$$

$$R^2 = 0.048 \quad n = 1983$$

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, $\beta_{3,} \beta_4$ 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

$$conspc_t = \beta_1 + \beta_2 incpc_t + \beta_3 conspc_{t-1} + u_t$$

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

$$conspc^e = \beta_1 + \beta_2 incpc^e + \beta_3 conspc^e$$

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:

$$sales = \beta_1 + \beta_2 advert + \beta_3 incent + u \qquad (1)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(advert) + \beta_3 \ln(incent) + u \qquad (2)$$

$$\ln(sales) = \beta_1 + \beta_2 advert + \beta_3 incent + u \qquad (3)$$

$$sales = \beta_2 advert + \beta_3 incent + u \qquad (4)$$

$$\ln(sales) = \beta_1 + \beta_2 \ln(incent) + u \qquad (5)$$

$$sales = \beta_1 + \beta_2 incent + u \tag{6}$$

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.

       b1)   (1) and (6)
       b2)   (2) and (3)
       b3)   (1) and (4)
       b4)   (2), (3) and (5)
       b5)   (1), (4) and (6)
       b6)   (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 $\mathbf{y}$ (remember that $\hat{\boldsymbol{\beta}}$ is also a function of $\mathbf{y}$), given by

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X'X} \right]^{-1} \mathbf{X'} + \mathbf{A} \right] \mathbf{y} \tag{3-93}$$

where $\mathbf{A}$ is $k \times n$ arbitrary matrix, that is a function of $\mathbf{X}$ and/or other non-stochastic variables, but it is not a function of $\mathbf{y}$. For $\tilde{\boldsymbol{\beta}}$ to be unbiased, certain conditions must be accomplished.

Taking (3-52) into account, we have

$$\tilde{\boldsymbol{\beta}} = \left[\left[\mathbf{X'X} \right]^{-1} \mathbf{X'} + \mathbf{A} \right] \left[\mathbf{X\beta} + \mathbf{u} \right] = \boldsymbol{\beta} + \mathbf{AX\beta} + \left[\left[\mathbf{X'X} \right]^{-1} \mathbf{X'} + \mathbf{A} \right] \mathbf{u} \tag{3-94}$$

Taking expectations on both sides of (3-94), we have

$$E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{AX\beta} + \left[\left[\mathbf{X'X} \right]^{-1} \mathbf{X'} + \mathbf{A} \right] E(\mathbf{u}) = \boldsymbol{\beta} + \mathbf{AX\beta} \tag{3-95}$$

For $\tilde{\boldsymbol{\beta}}$ to be unbiased, that is to say, $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, the following must be accomplished:

$$\mathbf{AX} = \mathbf{I} \tag{3-96}$$

Consequently,

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left[\left[\mathbf{X'X} \right]^{-1} \mathbf{X'} + \mathbf{A} \right] \mathbf{u} \tag{3-97}$$

Taking into account assumptions 7 and 8, and (3-96), the $Var(\tilde{\boldsymbol{\beta}})$ is equal to

$$Var(\tilde{\beta}) = E((\tilde{\beta}-\beta)(\tilde{\beta}-\beta)') = E\left[\left[\left[X'X\right]^{-1}X' + A\right]uu'\left[X\left[X'X\right]^{-1} + A'\right]\right]$$
$$= E\left[\left[\left[X'X\right]^{-1}X'\right]uu'\left[X\left[X'X\right]^{-1}\right] + AA'\right] = \sigma^2\left[\left[X'X\right]^{-1} + AA'\right] \tag{3-98}$$

The difference between both variances is the following:

$$Var(\tilde{\beta}) - Var(\hat{\beta}) = \sigma^2\left[\left[X'X\right]^{-1} + AA' - \left[X'X\right]^{-1}\right] = \sigma^2 AA' \tag{3-99}$$

The product of a matrix by its transpose is always a semi-positive definite matrix. Therefore,

$$Var(\tilde{\beta}) - Var(\hat{\beta}) = \sigma^2 AA' \geq 0 \tag{3-100}$$

The difference between the variance of an estimator $\tilde{\beta}$ - arbitrary but linear and unbiased – and the variance of the estimator $\hat{\beta}$ is a semi positive definite matrix. Consequently, $\hat{\beta}$ is a <u>B</u>est <u>U</u>nbiased <u>L</u>inear <u>E</u>stimator; 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

$$\hat{u} = y - X\hat{\beta} = y - X\left[X'X\right]^{-1}X'y = \left[I - X\left[X'X\right]^{-1}X'\right]y = My \tag{3-101}$$

where **M** is an idempotent matrix.

Alternatively, the vector of residuals can be expressed as a function of the disturbance vector:

$$\hat{u} = \left[I - X\left[X'X\right]^{-1}X'\right]y = \left[I - X\left[X'X\right]^{-1}X'\right]\left[X\beta + u\right]$$
$$= X\beta - X\left[X'X\right]^{-1}X'X\beta + u - X\left[X'X\right]^{-1}X'\beta u$$
$$= X\beta - X\beta + \left[I - X\left[X'X\right]^{-1}X'\right]u = \left[I - X\left[X'X\right]^{-1}X'\right]u \tag{3-102}$$
$$= Mu$$

Taking into account (3-102), the sum of squared residuals (*SSR*) can be expressed in the following form:

$$\hat{u}'\hat{u} = u'M'Mu = u'Mu \tag{3-103}$$

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:

$$E[\hat{\mathbf{u}}'\hat{\mathbf{u}}] = E[\mathbf{u}'\mathbf{Mu}] = trE[\mathbf{u}'\mathbf{Mu}] = E[tr\mathbf{u}'\mathbf{Mu}]$$
$$= E[tr\mathbf{Muu}'] = tr\mathbf{M}E[\mathbf{uu}'] = tr\mathbf{M}\sigma^2\mathbf{I} \qquad (3\text{-}104)$$
$$= \sigma^2 tr\mathbf{M} = \sigma^2(n-k)$$

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:

$$tr\mathbf{M} = tr\left[\mathbf{I}_{n\times n} - \mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'\right] = tr\mathbf{I}_{n\times n} - tr\mathbf{X}[\mathbf{X}'\mathbf{X}]^{-1}\mathbf{X}'$$
$$= tr\mathbf{I}_{n\times n} - tr\mathbf{I}_{k\times k} = n - k$$

According to (3-104), it holds that

$$\sigma^2 = \frac{E[\hat{\mathbf{u}}'\hat{\mathbf{u}}]}{n-k} \qquad (3\text{-}105)$$

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 $\hat{\beta}_2$ 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 $\mathbf{Q}$ to the result:

$$\lim_{n\to\infty}\frac{1}{n}\mathbf{X}'\mathbf{X} = \mathbf{Q} \qquad (3\text{-}109)$$

If $\mathbf{X}$ 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

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}\left[\frac{1}{n}\mathbf{X'u}\right]$$

The last term of (3-108) can be written as

$$\frac{1}{n}\mathbf{X'u} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ x_{21} & x_{22} & \cdots & x_{2i} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{ji} & \cdots & x_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \cdots & x_{ki} & \cdots & x_{kn} \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix}$$

(3-110)

$$= \frac{1}{n}\begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_i & \cdots & \mathbf{x}_n \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i u_i = \overline{\mathbf{x}_i u_i}$$

where $\mathbf{x}_i$ is the column vector corresponding to the ith observation

Now, we are going to calculate the expectation and the variance (3-110),

$$E\left[\overline{\mathbf{x}_i u_i}\right] = \frac{1}{n}\sum_{i=1}^{n}E\left[\mathbf{x}_i u_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i E\left[u_i\right] = \frac{1}{n}\mathbf{X'}E\left[\mathbf{u}\right] = \mathbf{0} \tag{3-111}$$

$$var\left[\overline{\mathbf{x}_i u_i}\right] = E\left[\overline{\mathbf{x}_i u_i}(\overline{\mathbf{x}_i u_i})'\right] = \frac{1}{n}\mathbf{X'}E\left[\mathbf{uu'}\right]\mathbf{X} = \frac{1}{n}\frac{\sigma^2}{n}\frac{\mathbf{X'X}}{n} = \frac{\sigma^2}{n^2}\mathbf{Q} \tag{3-112}$$

since $E\left[\mathbf{uu'}\right] = \sigma^2\mathbf{I}$, according to assumptions 7 and 8.

Taking limits in (3-112), it then follows that

$$\lim_{n\to\infty} var\left[\overline{\mathbf{x}_i \mathbf{u}_i}\right] = \lim_{n\to\infty}\frac{\sigma^2}{n^2}\mathbf{Q} = 0(\mathbf{Q}) = \mathbf{0} \tag{3-113}$$

Since the expectation of $\overline{\mathbf{x}_i \mathbf{u}_i}$ is identically zero and its variance converges to zero, $\overline{\mathbf{x}_i \mathbf{u}_i}$ converges in mean square to zero. Convergence in mean square implies convergence in probability, and so $\text{plim}(\overline{\mathbf{x}_i \mathbf{u}_i})=0$. Therefore,

$$\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}(\mathbf{x}_i\mathbf{u}_i) = \boldsymbol{\beta} + \mathbf{Q}^{-1}\text{plim}\left[\frac{1}{n}\mathbf{X'u}\right] = \boldsymbol{\beta} + \mathbf{Q}^{-1}\times 0 = \boldsymbol{\beta} \tag{3-114}$$

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:

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \tag{3-115}$$

The expectation and variance of the distribution of $\mathbf{y}$ are given by

$$E(\mathbf{y}) = E[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}] = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{u}) = \mathbf{X}\boldsymbol{\beta} \tag{3-116}$$

$$\text{var}(\mathbf{y}) = E\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\right] = E\left[\mathbf{u}\mathbf{u}'\right] = \sigma^2 \mathbf{I} \tag{3-117}$$

Therefore,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \tag{3-118}$$

The probability density of $\mathbf{y}$ (or likelihood function), considering $\mathbf{X}$ and $\mathbf{y}$ fixed and $\boldsymbol{\beta}$ and σ^2 variable, will be in accordance with (3-118) equal to

$$L = f(y \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\left(1/2\sigma^2\right)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \tag{3-119}$$

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:

$$\frac{\delta \ln(L)}{\delta \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) \tag{3-121}$$

$$\frac{\delta \ln(L)}{\delta \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^4} \tag{3-122}$$

Equating (3-121) to zero, we see that the maximum likelihood estimator of $\boldsymbol{\beta}$, denoted by $\tilde{\boldsymbol{\beta}}$, satisfies that

$$\mathbf{X}'\mathbf{X}\tilde{\beta} = \mathbf{X}'\mathbf{y} \tag{3-123}$$

Because we assume that $\mathbf{X}'\mathbf{X}$ is invertible,

$$\tilde{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{X}\right]^{-1} = \mathbf{X}'\mathbf{y} \tag{3-124}$$

Consequently, the maximum likelihood estimator of $\boldsymbol{\beta}$, under the assumptions of the *CLM*, coincides with *OLS* estimator, that is to say,

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} \tag{3-125}$$

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 $\boldsymbol{\beta}$ 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

$$\lim_{n\to\infty}\frac{n-k}{n} = 1 \qquad (3\text{-}130)$$