SOME MEASURES OF AMOUNT OF INFORMATION FOR THE LINEAR REGRESSION MODEL

Mihoc Ion, Fătu Cristina-Ioana

Abstract.Econometric theory concerns the study and development of tools and methods for applied econometric applications. The ordinary least squares (OLS) estimator is the most basic estimation procedure in econometrics.

In this paper we will present the ordinary linear regression model and discussed how to estimate linear regression models by using the method of least squares. Also, we present some properties of the optimal estimators (OLS estimators) in the case of the linear regression model as well as some measures of the amount of information associated to these estimators.

Keywords and Phrases: Econometrics, regression model, regressand, regressors, linear model, ordinary least squares, OLS estimator, measure of the information, Fisher's information.

2000 Mathematics Subject Classifications: 62H10, 62H12, 62J05, 94A17.

1. The classical linear regression model

The most elementary type of regression model is **the multiple linear regression model**, which can be expressed by the following equation

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + u_t, t = \overline{1,n} \tag{1.1}$$

where subscript t is used to index the observation of a sample and n represents the **sample size**. The relation (1.1) links the observations on a dependent and the explanatory (independent) variables for each observation in terms of $(k+1)$ **unknown parameters**, $\beta_0, \beta_1, \beta_2, ..., \beta_k$, and an **unobserved error term**, u_t.In the context of econometrics, equation (1.1) is usually thought of as a model of economic behaviors. The variable Y_t typically represents the response of economic agents to a collection of "stimulus" variables x_t.

Remark 1.1. If we wish to make sense of the regression model (1.1), then, we must make some assumptions about the properties of the error term u_t. Precisely what those assumptions are will vary from case to case. In all cases, thought, it is assumed that u_t is a random variable.

Before to present the assumptions that comprise the classical linear model, we introduce the vector and matrix notation. Thus, the matrix form for the classical linear regression model (then when $\beta_0 = 0$) is represented by

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u}, \tag{1.2}$$

where

$$\mathbf{Y}=[Y_1\ Y_2\ ...\ Y_n]^{\intercal},\ dim\ \mathbf{Y} = n \times 1;\ \mathbf{Y}-\text{ the vector of observations on} \tag{1.3}$$

the explained variable;

$$\mathbf{X}=\begin{bmatrix} \mathbf{X}_1^{\intercal} \\ \mathbf{X}_2^{\intercal} \\ ... \\ \mathbf{X}_n^{\intercal} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1, & \mathbf{X}_2, & ..., & \mathbf{X}_k \end{bmatrix}, \tag{1.4}$$

$dim\ \mathbf{X} = n \times k;$ $\mathbf{X}-$ the matrix of observations on the explanatory variables which is assumed to be "fixed" or deterministic;

$$\mathbf{X}_i^{\intercal} = [x_{i1}, x_{i2}, ..., x_{ik}],\ \text{dim}\ \mathbf{X}_i^{\intercal} = 1 \times k,\ i = \overline{1,n}; \tag{1.5}$$

$$\mathbf{X}_j = [x_{1j}, x_{2j}, ..., x_{nj}]^{\intercal},\ \text{dim}\ \mathbf{X}_j = n \times 1,\ j = \overline{1,k}; \tag{1.6}$$

and the column vectors β and $\mathbf{u}$ have the forms

$$\beta = \begin{bmatrix} \beta_1 & \beta_2 & ... & \beta_k \end{bmatrix}^T, dim\ \ \beta = k \times 1; \beta-\text{the vector of parameters} \tag{1.7}$$

to be estimated;

$$\mathbf{u} = \begin{bmatrix} u_1 & u_2 & ... & u_n \end{bmatrix}^T = \mathbf{Y} - \mathbf{X}\beta, dim\ \ \mathbf{u} = n \times 1; \mathbf{u}-\text{ the vector of errors.} \tag{1.8}$$

When a regression model is written in the form (1.2), then the separate columns of the matrix $\mathbf{X}$ are called **regressors,** and the column vector $\mathbf{Y}$ is called the **regressand.** Also, we observe that the element $x_{tj} \in \mathbf{X}$ represents the t^{th} observation on the j^{th} regressor, $j = \overline{1,k}$, $t = \overline{1,n}$. The $(n \times 1) - vector$ $\mathbf{Y}$ and $(n \times k) - matrix$ $\mathbf{X}$ are sometimes called the **data vector** and the **data matrix.**

Remark 1.2. The classical linear regression model with **nonstochastic regressor variables** is characterized by the following **assumptions** (denoted by $\mathbf{A}_{1j}, j = \overline{1,7}$):

$$\begin{cases} \mathbf{A}_{11}) & E(u_i) = 0, i = \overline{1,n},\ that\ is,\ E(\mathbf{u}) = \mathbf{0}; \\ \mathbf{A}_{12}) & \begin{cases} Var(u_i) = E(u_i^2) = \sigma^2, i = \overline{1,n},\ that\ is, Var(\mathbf{u}) = E(\mathbf{u}\,\mathbf{u}^T) = \\ = \sigma^2 \mathbf{I}_n, where:\ \mathbf{I}_n\ is\ an\ \ n \times n\ identity\ matrix,\ Var(\mathbf{u})\ is \\ the\ variance - covariance\ matrix\ of\ \ disturbances\ (errors); \end{cases} \\ \mathbf{A}_{13}) & \begin{cases} Errors\ u_i, i = \overline{1,n} - are\ random\ variables\ independently \\ and\ identically\ distributed,\ that\ is,\ u_i \sim IID(0, \sigma^2),\ i = \overline{1,n}; \\ or\ \mathbf{u} \sim IID(0, \sigma^2 \mathbf{I}_n); \end{cases} \\ \mathbf{A}_{14}) & \begin{cases} The\ regressors\ \mathbf{X}_j\ and\ the\ errors\ u_i\ \ are\ independently, \\ \ that\ is,\ cov(\mathbf{X}_j, u_i) = 0, i - fixed,\ i = \overline{1,n};\ \forall j, j = \overline{1,k}; \end{cases} \\ \mathbf{A}_{15}) & \begin{array}{l} The\ \mathbf{X}\ matrix\ is\ nonsingular\ and\ its\ columns\ vector\ \mathbf{X}_j, \\ j = \overline{1,k}\ are\ linearly\ independent,\ that\ is, \\ \ rank\ (\mathbf{X}^T\mathbf{X}) = rank\ \mathbf{X} = k \Rightarrow (\mathbf{X}^T\mathbf{X})^{-1} - exists. \end{array} \\ \mathbf{A}_{16}) & \begin{array}{l} The\ errors\ are\ nonautocorrelated,\ that\ is,\ E(u_i u_s) = 0, \\ \ i, s = \overline{1,n};\ i \neq s, \end{array} \\ \mathbf{A}_{17}) & \begin{cases} \textbf{Obtionally}, we\ will\ sometimes\ assume\ that\ the\ errors \\ are\ normally,\ independently,\ and\ identically\ distributed, \\ that\ is,\ u_i \sim IIN(0, \sigma^2),\ i = \overline{1,n}\ \ or\ \mathbf{u} \sim IIN(0, \sigma^2 \mathbf{I}_n). \end{cases} \end{cases} \tag{1.9}$$

Remark 1.3. Using the assumption $\mathbf{A}_{11}$), it follows that

$$E(\mathbf{Y}) = \mathbf{X}\beta = \beta_1 \mathbf{X_1} + \beta_2 \mathbf{X_2} + ... + \beta_k \mathbf{X}_k \tag{1.10}$$

and the statistics

$$\widehat{\mathbf{Y}} = \widehat{E(\mathbf{Y})} = \mathbf{X}\widehat{\beta} = \widehat{\beta}_1 \mathbf{X_1} + \widehat{\beta}_2 \mathbf{X_2} + ... + \widehat{\beta}_k \mathbf{X}_k \tag{1.11}$$

represents an **absolute correct estimator** for the expected value $E(\mathbf{Y})$, where we denoted by

$$\widehat{\beta} = \left[\ \widehat{\beta}_1\ \widehat{\beta}_2\ ...\widehat{\beta}_k\ \right]^{\mathsf{T}},\ dim\ \widehat{\beta} = k \times 1 \tag{1.12}$$

the **best estimator** for the column vector of parameters β.

2. The method of least squares

The method of least squares is the standard technique for extracting an estimator of β from a **sample (data)** of n observations. According to the principle of **the least-squares method,** the problem to determine the best vector estimator $\widehat{\beta}$, for the unknown vector parameter β, is the same with the problem to **minimize the criterion function** $S(\beta) = S(\beta_1, ..., \beta_k)$, where

$$S(\beta) = \mathbf{SSR}(\beta) = \sum_{i=1}^{n}[Y_i - E(Y_i)]^2 = \sum_{i=1}^{n} u_i^2, \tag{2.1}$$

and

$$u_i^2 = [Y_i - E(Y_i)]^2 = [Y_i - (\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})]^2, \ i = \overline{1, n} \tag{2.2}$$

represents the squares deviations (errors).

The **ordinary least squares (**or **OLS) estimator,** for this **linear regression model**, is defined as the value that minimizes the sum of the squared errors, that is,

$$\widehat{\beta} = \arg\min_{\beta} S(\beta). \tag{2.2a}$$

Because the criterion function (2.1) can be rewritten as

$$\begin{aligned} S(\beta) &= (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \\ &= (\mathbf{Y}^T - (\mathbf{X}\beta)^T)(\mathbf{Y} - \mathbf{X}\beta) = \\ &= (\mathbf{Y}^T - \beta^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{X}\beta) = \\ &= \mathbf{Y}^T\mathbf{Y} - \beta^T\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta = \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{2}\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta = \\ &= \|\mathbf{Y} - \mathbf{X}\beta\|^2, \end{aligned}$$

then when we have in view the equality

$$\beta^T \mathbf{X}^T \mathbf{Y} = \mathbf{Y}^T \mathbf{X}\beta, \tag{2.3}$$

(that is, the scalar $\beta^T \mathbf{X}^T \mathbf{Y}$ equals its transpose $\mathbf{Y}^T \mathbf{X}\beta$), it follows that the **final form of the criterion function** will be

$$S(\beta) = \mathbf{Y}^T \mathbf{Y} - \mathbf{2}\beta^T \mathbf{X}^T \mathbf{Y} + \beta^T \mathbf{X}^T \mathbf{X}\beta = \tag{2.4}$$

$$= \|\mathbf{Y} - \mathbf{X}\beta\|^2, \tag{2.4a}$$

that is, **the OLS estimator** must to be chosen such that: it to minimize the **Euclidian distance** between $\mathbf{Y}$ and $\mathbf{X}\beta$.

Because, in this last relation, the term $\mathbf{Y}^T \mathbf{Y}$ does not depend on β, the first-order conditions, for minimization of the criterion function $S(\beta)$, can be write as

$$\begin{aligned}\frac{\partial S(\beta)}{\partial \beta} &= \frac{\partial}{\partial \beta}[\mathbf{Y}^T \mathbf{Y} - \mathbf{2}\mathbf{Y}^T \mathbf{X}\beta + \beta^T \mathbf{X}^T \mathbf{X}\beta] = \\ &= \frac{\partial}{\partial \beta}[\mathbf{Y}^T \mathbf{Y}] - \frac{\partial}{\partial \beta}[\mathbf{2}\mathbf{Y}^T \mathbf{X}\beta] + \frac{\partial}{\partial \beta}[\beta^T \mathbf{X}^T \mathbf{X}\beta] = \\ &= -\mathbf{2}\mathbf{Y}^T \mathbf{X} + \mathbf{2}\beta^T \mathbf{X}^T \mathbf{X} = \mathbf{0}\end{aligned} \tag{2.5}$$

or in the form

$$(\mathbf{X}^T \mathbf{X})\widehat{\beta} = \mathbf{X}^T \mathbf{Y} \tag{2.6}$$

which, evidently, represents just the **system of the normal equations of least squares.**

Then, because the coefficient matrix $\mathbf{X}^T \mathbf{X}$ is positive definite, (that is, $\mathbf{X}^T \mathbf{X}$ is nonsingular), from (2.6) we obtain the vector solution as

$$\widehat{\beta} = \left[\ \widehat{\beta}_1\ \widehat{\beta}_2\ ...\widehat{\beta}_k\ \right]^{\mathsf{T}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}. \tag{2.7}$$

or as

$$\widehat{\beta} = \frac{\mathbf{S_{XY}}}{\mathbf{S_{XX}}}, \tag{2.8}$$

where

$$\mathbf{S_{XX}} = \frac{1}{n}\mathbf{X}^T \mathbf{X} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{X}_i,\ \mathbf{S_{XY}} = \frac{1}{n}\mathbf{X}^T \mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i^T Y_i. \tag{2.8a}$$

Viewed as a function of the sample $(\mathbf{Y}, \mathbf{X})$, this **optimal solution** is sometimes called the **OLS estimator** and, for any given sample $(\mathbf{Y}, \mathbf{X})$, the value of this **OLS estimator** will be an **OLS estimate**.

In the next, to verify that $\widehat{\beta}$ corresponds to a minimum point for the function $S(\beta)$, we check the second-order sufficient condition

$$\frac{\partial^2 S(\beta)}{\partial \beta^2} = 2\mathbf{X}^T\mathbf{X}. \tag{2.9}$$

Because $rank(\mathbf{X}^T\mathbf{X}) = k$ (that is, the matrix $\mathbf{X}^T\mathbf{X}$ has full rank), it follows that this matrix is positive definite, so $\widehat{\beta}$ is in fact a minimizer and we have

$$S_{\min}(\beta) = S(\widehat{\beta}) = \sum_{i=1}^{n}[\underbrace{Y_i - (\widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + ... + \widehat{\beta}_k x_{ik})}_{=\widehat{u}_i}]^2 = \sum_{i=1}^{n} \widehat{u}_i^2 =$$
$$= (\mathbf{Y} - \mathbf{X}\widehat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \widehat{\mathbf{u}}^T\widehat{\mathbf{u}} = \parallel \widehat{\mathbf{u}}^2 \parallel, \tag{2.10}$$

where

$$\widehat{\mathbf{u}} = \mathbf{u}(\widehat{\beta}) = \mathbf{Y} - \mathbf{X}\widehat{\beta} \tag{2.11}$$

represents the **vector of residuals (**or the **vector of least-squares residuals) $\mathbf{u} = \mathbf{Y} - \mathbf{X}\beta$**, evaluated at $\widehat{\beta}$, and it is often denoted by $\widehat{\mathbf{u}}$.

The vector

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} \tag{2.12}$$

is referred to as the **vector of fitted values**.

Remark 2.1. Because the relation (2.6) can be written as

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \mathbf{0} \tag{2.13}$$

or in the form of the **scalar product** as

$$\mathbf{X}_j^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = (\mathbf{X}_j, \mathbf{Y} - \mathbf{X}\widehat{\beta}) = (\mathbf{X}_j, \widehat{\mathbf{u}}) = 0,\ i = \overline{1, k}, \tag{2.13a}$$

it follows that the vector $\widehat{\mathbf{u}} = \mathbf{Y} - \mathbf{X}\widehat{\beta}$ is **orthogonal** to all of the regressors $\mathbf{X}_j$ that represent the explanatory variables. For this reason, equations like (2.13) or (2.13a) are often referred to as **orthogonality conditions**.

3. Informational characterization of the simple linear regression model

3.1. The model and its OLS estimator

The most elementary type of regression model is the simple linear regression model, can be obtained from the model (1.1) if we consider $k = 1$, namely

$$Y_t = \beta_0 + \beta_1 x_t + u_t, t = \overline{1,n}, \tag{3.1.1}$$

which can easily be written in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{3.1.2}$$

where

$$\mathbf{Y} = \left[\begin{array}{c} Y_1 Y_2 ... Y_n \end{array}\right]^{\mathsf{T}}, dim\ \mathbf{Y} = n \times 1; \tag{3.1.3}$$

$$\mathbf{X} = [\mathbf{X}_0\ \mathbf{X}_1] = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{bmatrix}^{\mathsf{T}}, dim\ \mathbf{X} = n \times 2, \tag{3.1.3a}$$

$dim\ \mathbf{X}_j = n \times 1,\ j = \overline{0,1};$

$$\mathbf{u} = \left[\begin{array}{c} u_1\ u_2\ ...\ u_n \end{array}\right]^{\mathsf{T}}, dim\ \mathbf{u} = n \times 1; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, dim\ \beta = 2 \times 1. \tag{3.1.4}$$

The typical row of the equation (3.1.2) is

$$\mathbf{Y}_t = \mathbf{X}_t\beta + u_t = \sum_{i=0}^{1} \beta_i x_i + u_t,\ \mathbf{X}_t = [x_0\ x_t],\ t = \overline{1,n}, \tag{3.1.2a}$$

where we have used $\mathbf{X}_t$ to denote the t^{th} row of the matrix $\mathbf{X}$.

Since

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ ... & ... \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum\limits_{i=1}^{n} x_i \\ \sum\limits_{i=1}^{n} x_i & \sum\limits_{i=1}^{n} x_i^2 \end{bmatrix} \tag{3.1.5}$$

and

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{bmatrix}, \tag{3.1.5a}$$

the the **system of the normal equations** has the form

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i & = \sum_{i=1}^{n} Y_i, \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 & = \sum_{i=1}^{n} x_i Y_i \end{cases} \tag{3.1.6}$$

or the matrix form

$$(\mathbf{X}^T\mathbf{X})\widehat{\beta} = \mathbf{X}^T\mathbf{Y}, \tag{3.1.7}$$

where

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}. \tag{3.1.7a}$$

If the matrix $\mathbf{X}^T\mathbf{X})^{-1}$ exist (that is, if *rank* $\mathbf{X} = 2$), then the matrix solution has the form

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} \overline{Y} - \widehat{\beta}\overline{x} \\ \frac{\sum_{i=1}^{n} Y_i(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \end{bmatrix}, \tag{3.1.8}$$

and, from here, it follows

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x} \tag{3.1.9}$$

$$\widehat{\beta}_1 = \frac{S_{xY}}{S_{xx}}, \tag{3.1.10}$$

where

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2; \tag{3.1.10a}$$

$$S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}); \qquad (3.1.10b)$$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \qquad (3.1.10c)$$

3.2. The statistical properties of the OLS estimator

In the next we will recall and examine some statistical properties of ordinary least squares (OLS) estimator which depend upon the **assumptions** which were specified in the **Remark 1.3** (see [5]).

Theorem 3.2.1. The estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are linear functions of the dependent variables $Y_1, Y_2, ..., Y_n$, that is,

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i,\ c_i = \frac{(x_i - \overline{x})}{nS_{xx}},\ c_i - real\ constants,\ i = \overline{1,n} \qquad (3.2.1)$$

$$\widehat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i,\ d_i = \frac{1}{n} - \overline{x}\frac{(x_i - \overline{x})}{nS_{xx}}, d_i - real\ constants, i = \overline{1,n}. \qquad (3.2.2)$$

Theorem 3.2.2. *The estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are* ***unbiased estimators*** *for the unknown parameters β_0 and β_1, that is, we have*

$$\begin{cases} M(\widehat{\beta}_1) & = & \beta_1, \\ D^2(\widehat{\beta}_1) & = & \dfrac{\sigma^2}{nS_{xx}} \to 0, \quad n \to \infty \end{cases} \qquad (3.2.3)$$

and

$$\begin{cases} a) & E(\widehat{\beta}_0) & = & \beta_0, \\ b) & Var(\widehat{\beta}_0) & = & \frac{\sigma^2}{n}\left[1 + \frac{\overline{x}^2}{S_{xx}}\right] \to 0,\ dacă\ n \to \infty. \end{cases} \qquad (3.2.4)$$

More, when $n \to \infty$, the statistics $\widehat{\beta}_0$ and $\widehat{\beta}_1$ represent ***absolute correct estimators*** *for the unknown parameter β_0 and β_1.*

Theorem 3.2.3. *The estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are correlated (and therefore dependent) unless $\overline{x} = 0$, that is*

$$Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{nS_{xx}}, \ if \ \overline{x} \neq 0. \tag{3.2.5}$$

Theorem 3.2.4. *The* ***unbiased estimator*** *for the unknown parameter σ^2 is represented by the statistic*

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left[Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right]^2 = \tag{3.2.6}$$

$$= \frac{\| \widehat{\mathbf{u}} \|^2}{n-2}, \tag{3.2.6a}$$

where n is the number of the observations, $k = 2$ represents the number of the parameters estimated in the model and

$$\| \widehat{\mathbf{u}} \|^2 = \widehat{\mathbf{u}}^T \widehat{\mathbf{u}} = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^{\intercal}(\mathbf{Y} - \mathbf{X}\widehat{\beta}) = \sum_{i=1}^{n}\left[Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right]^2 = \sum_{i=1}^{n} \widehat{u}_i^2 \tag{3.2.6b}$$

represents the ***sum of squared residuals,*** *or* ***SSR.***

Theorem 3.2.5. *If the disturbances (errors) u_i, $i = \overline{1,n}$ are random variable independently, identically and distributed normally, then the maximum-likelihood estimators, denoted by β_0^* and β_1^*, for the unknown parameters β_0 and β_1, are the same as the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$.*

Remark 3.2.1. In the preceding expressions given the variances for the least squares estimators in terms of σ^2.Usually the value of σ^2 will be unknown, and we will need to make use of the sample observations to estimate σ^2.

Theorem 3.2.6. *In the hypothesis of the Theorem 3.2.5, then when the parameter σ^2 is* ***unknown****, the maximum-likelihood estimator (the ML estimator), denoted by $(\sigma^2)^*$, is different from the unbiased estimator for σ^2 and we have*

$$(\sigma^2)^* = \frac{\| \widehat{\mathbf{u}} \|^2}{n} \neq \frac{\| \widehat{\mathbf{u}} \|^2}{n-2} = \widehat{\sigma}^2. \tag{3.2.7}$$

3.3. Fisher's information measures

Let Z be a continuous random variable with the probability density function $f(z;\theta)$,where $\theta = (\theta_1, \theta_2, ..., \theta_k)$, $\theta \in \mathbf{D}_\theta$, $\mathbf{D}_\theta \subseteq \mathbb{R}^k, k \geq 1$, $\mathbf{D}_\theta -$ the parameter space (which, sometimes, is called the set of admissible values of θ). To each value of θ, $\theta \in \mathbf{D}_\theta$, we have one member $f(z;\theta)$ of the family which will be denoted by the symbol $\{f(z;\theta); \theta \in \mathbf{D}_\theta\}$.

In the next we wish to estimate a specified function of θ, $g(\theta)$, with the help of statistic

$$t = t(Z_1, Z_2, ..., Z_n), \tag{3.3.1}$$

where $S_n(Z) = (Z_1, Z_2, ..., Z_n)$ is a random sample of size n and $Z_1, Z_2, ..., Z_n$ are sample random variables statistically independent and identically distributed as the random variable Z, that is, we have

$$f(z;\theta) = f(z_i;\theta); i = \overline{1,n}, \theta \in \mathbf{D}_\theta. \tag{3.3.2}$$

Let

$$L_n(z_1, z_2, ..., z_n; \theta_1, \theta_2, ..., \theta_k) = L_n(\mathbf{z}; \theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(z_i; \theta_1, \theta_2, ..., \theta_k) \tag{3.3.3}$$

be the maximum-likelihood function of $\theta = (\theta_1, \theta_2, ..., \theta_k)$ of the random sample $S_n(Z) = (Z_1, Z_2, ..., Z_n)$where $\mathbf{z} =(z_1, z_2, ..., z_n)$.

Lemma 3.3.1. *If* $\mathbf{W}$ *is an* $k-$ *dimensional random vector as*

$$\mathbf{W} = (W_1, W_2, ..., W_k)^{\intercal} \tag{3.3.4}$$

and its components have the forms

$$W_j = \frac{\partial \ln L_n(\mathbf{z}; \theta_1, \theta_2, ..., \theta_k)}{\partial \theta_j}, j = \overline{1,k}, \tag{3.3.5}$$

then, we have

$$E(W_j) = E\left(\frac{\partial \ln L_n(\mathbf{z}; \theta_1, \theta_2, ..., \theta_k)}{\partial \theta_j}\right) = 0, j = \overline{1,k}. \tag{3.3.6}$$

Proof. Indeed, since the maximum-likelihood function $L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)$ is a probability density function then, if are satisfied some regularity conditions, the equality

$$\int\limits_{\mathbb{R}^n} L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)\mathbf{dz} = 1, \tag{3.3.7}$$

implies the following relations

$$\frac{\partial}{\partial\theta_j}\left[\int\limits_{\mathbb{R}^n} L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)\mathbf{dz}\right] = \int\limits_{\mathbb{R}^n}\frac{\partial}{\partial\theta_j}\left[L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)\right]\mathbf{dz} = \tag{3.3.7a}$$

$$= \int\limits_{\mathbb{R}^n}\frac{\partial}{\partial\theta_j}\left[\ln L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)\right] L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)\mathbf{dz} = \tag{3.3.7b}$$

$$= E\left\{\frac{\partial \ln L_n(\mathbf{z};\theta_1,\theta_2,...,\theta_k)}{\partial\theta_j}\right\} = E(W_j) = 0, j = \overline{1,k}, \tag{3.3.7c}$$

which represents just the equalities (3.3.6).

Lemma 3.3 2. *The maximum-likelihood function (3.3.3), when $k = 1$, has the following property*

$$E\left\{\left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2\right\} = -E\left\{\frac{\partial^2 \ln L_n(\mathbf{z};\theta)}{\partial\theta^2}\right\}. \tag{3.3.8}$$

Proof. First, using the Lemma 3.3.1, we obtain the relation

$$E\left[\frac{\partial}{\partial\theta}\left(\ln L_n(\mathbf{z};\theta)\right)\right] = \int_{\mathbf{R}^n}\frac{\partial}{\partial\theta}\left(\ln L_n(\mathbf{z};\theta)\right) L_n(\mathbf{z};\theta)\mathbf{dz} = 0 \tag{3.3.9}$$

which, making use of (3.3.9), implies the relations

$$\frac{\partial}{\partial\theta}\left[\int_{\mathbf{R}^n}\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}L_n(\mathbf{z};\theta)\mathbf{dz}\right] = \int_{\mathbf{R}^n}\frac{\partial}{\partial\theta}\left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}L_n(\mathbf{z};\theta)\right]\mathbf{dz} =$$

$$= \int_{\mathbf{R}^n} \frac{\partial^2 \ln L_n(\mathbf{z};\theta)}{\partial\theta^2} L_n(\mathbf{z};\theta)\mathbf{dz} + \int_{\mathbf{R}^n} \left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2 L_n(\mathbf{z};\theta)\mathbf{dz} =$$

$$= E\left[\frac{\partial^2 \ln L_n(\mathbf{z};\theta)}{\partial\theta^2}\right] + E\left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2 = 0 \qquad (3.3.10)$$

and, hence, it follows just the relation (3.3.8).

Using these two lemmas we can to present the next definitions.

Definition 3.3.1. *The quantity*

$$\mathbf{I}_1(\theta) = E\left\{\left[\frac{\partial \ln f(z;\theta)}{\partial\theta}\right]^2\right\} = \int_{-\infty}^{+\infty} \left[\frac{\partial \ln f(z;\theta)}{\partial\theta}\right]^2 f(z;\theta)dz = \qquad (3.3.11)$$

$$= -\int_{-\infty}^{+\infty} \left[\frac{\partial^2 \ln f(z;\theta)}{\partial\theta^2}\right] f(z;\theta)dz = -E\left\{\left[\frac{\partial^2 \ln f(z;\theta)}{\partial\theta^2}\right]\right\} \qquad (3.3.11a)$$

represents the ***Fisher information*** *which measures the information about the unknown parameter* θ *which is contained in an observation of the random variable* Z.

Definition 3.3.2. *The quantity* $\mathbf{I}_n(\theta)$, *defined by the relation*

$$\mathbf{I}_n(\theta) = E\left\{\left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2\right\} = \int_{\mathbb{R}^n} \left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2 L_n(\mathbf{z};\theta)d\mathbf{z} = \qquad (3.3.12)$$

$$= -\int_{\mathbb{R}^n} \left[\frac{\partial^2 \ln L_n(\mathbf{z};\theta)}{\partial\theta^2}\right] L_n(\mathbf{z};\theta)d\mathbf{z} = -E\left\{\left[\frac{\partial^2 \ln L_n(\mathbf{z};\theta)}{\partial\theta^2}\right]\right\}, \qquad (3.3.12a)$$

represents the Fisher information measure which measures the information about unknown parameter θ *contained in a random sample* $S_n(Z) = (Z_1, Z_2, ..., Z_n)$.

Remark 3.3.1. Because the sample random variables $Z_1, Z_2, ..., Z_n$ are independent and identically distributed, with the probability density function $f(z;\theta), \theta \in \mathbf{D}_\theta$, we get

$$\mathbf{I}_n(\theta) = \int_{\mathbb{R}^n} \left[\frac{\partial \ln L_n(\mathbf{z};\theta)}{\partial\theta}\right]^2 L_n(\mathbf{z};\theta)d\mathbf{z} = \int_{\mathbb{R}^n} \left[\frac{\frac{\partial L_n(\mathbf{z};\theta)}{\partial\theta}}{L_n(\mathbf{z};\theta)}\right]^2 L_n(\mathbf{z};\theta)\mathbf{dz} =$$

$$= n \int_{-\infty}^{+\infty} \left[\frac{\partial \ln f(z;\theta)}{\partial \theta} \right]^2 f(z;\theta) dz = -n \int_{-\infty}^{+\infty} \left[\frac{\partial^2 \ln f(z;\theta)}{\partial \theta^2} \right] f(z;\theta) dz \quad (3.3.13)$$

Definition 3.3.3. *The Fisher information matrix (or the information matrix), denoted by* $\mathbf{I}(\mathbf{Z};\theta)$*, when* $\theta = (\theta_1, \theta_2, ..., \theta_k)$*, has the form*

$$\mathbf{I}(\mathbf{Z};\theta) = \begin{bmatrix} I_n^{(1,1)}(\theta) & I_n^{(1,2)}(\theta) & ... & I_n^{(1,k)}(\theta) \\ I_n^{(2,1)}(\theta) & I_n^{(2,2)}(\theta) & ... & I_n^{(2,k)}(\theta) \\ ... & ... & ... & ... \\ I_n^{(k,1)}(\theta) & I_n^{(k,2)}(\theta) & ... & I_n^{(k,k)}(\theta) \end{bmatrix} \quad (3.3.14)$$

where the generic element of the information matrix, in the $ijth$ *position, is*

$$I_n^{(i,j)}(\theta) = E\left[\frac{\partial \, \ln L_n(\mathbf{z};\theta)}{\partial \theta_i} \times \frac{\partial \, \ln L_n(\mathbf{z};\theta)}{\partial \theta_j} \right] = \quad (3.3.15)$$

$$= -E\left[\frac{\partial^2 \, \ln L_n(\mathbf{z};\theta)}{\partial \theta_i \partial \theta_j} \right], i = \overline{1,k}; j = \overline{1,k}. \quad (3.3.15a)$$

Theorem 3.3.1. *The* ***informational matrix*** $\mathbf{I}(\mathbf{Z};\theta)$ *has the following property*

$$\mathbf{I}(\mathbf{X};\theta) = \mathbf{K}_{\mathbf{W}}(\theta), \quad (3.3.16)$$

where

$$\mathbf{K}_{\mathbf{W}}(\theta) = \begin{bmatrix} Cov(W_1, W_1) & Cov(W_1, W_2) & ... & Cov(W_1, W_k) \\ Cov(W_2, W_1) & Cov(W_2, W_2) & ... & Cov(W_2, W_k) \\ ... & ... & ... & ... \\ Cov(W_k, W_1) & Cov(W_k, W_2) & ... & Cov(W_k, W_k) \end{bmatrix}, \quad (3.3.17)$$

is the ***covariance matrix*** *associated with the real random vector* $\mathbf{W} = (W_1, W_2, ..., W_k)$ *which was defined above.*

Proof. To prove this theorem we will use the relations (3.3.6), the Definition 3.3.3, as well as, the relations (3.3.15). Thus, we obtain

$$I_n^{(i,j)}(\theta) = E\left[\frac{\partial \, \ln L_n(\mathbf{z};\theta)}{\partial \theta_i} \times \frac{\partial \, \ln L_n(\mathbf{z};\theta)}{\partial \theta_j} \right] =$$

$$= E\left\{\left[\frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_i} - E\left(\frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_i}\right)\right]\times\right.$$
$$\left.\times\left[\frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_j} - E\left(\frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_j}\right)\right]\right\} =$$

$$= Cov\left(\frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_i}, \frac{\partial\ \ln L_n(\mathbf{z};\theta)}{\partial\theta_j}\right) =$$
$$= Cov(W_i, W_j) = K_{ij},\ i,j = \overline{1,k}.$$

Therefore, $I_n^{(i,i)}(\theta)$, where

$$\begin{aligned} I_n^{(i,i)}(\theta) &= Cov(W_i, W_i) = E(W_i^2) = Var(W_i) = \\ &= -E\left(\frac{\partial^2 \ln L_n(z;\theta)}{\partial\theta_i^2}\right) = -\int_{\mathbb{R}^n}\left(\frac{\partial^2 \ln L_n(z;\theta)}{\partial\theta_i^2}\right) L_n(\mathbf{z};\theta)\mathbf{dz}, \end{aligned} \tag{3.3.18}$$

represents the Fisher information measure concerning the unknown parameter θ_i, $i = \overline{1,k}$, obtaining using the sample vector $S_n(Z) = (Z_1, Z_2, ..., Z_n)$.

Analogous, the quantities $I_n^{(i,j)}(\theta)$, where

$$\begin{aligned} I_n^{(i,j)}(\theta) &= Cov(W_i, W_j) = M\left[\frac{\partial lnL_n(\mathbf{z};\theta)}{\partial\theta_i}\times\frac{\partial lnL_n(\mathbf{z};\theta)}{\partial\theta_j}\right] = \\ &= -M\left[\frac{\partial^2 lnL_n(\mathbf{z};\theta)}{\partial\theta_i\partial\theta_j}\right] = -\int_{\mathbb{R}^n}\left(\frac{\partial^2 L_n(\mathbf{z};\theta)}{\partial\theta_i\partial\theta_j}\right) L_n(\mathbf{z};\theta)\mathbf{dz}, \end{aligned} \tag{3.3.19}$$

represents the Fisher information measure concerning the unknown vector parameter (θ_i, θ_j), $i,j = \overline{1,k}$, $i \neq j$, obtaining using the sample vector $S(Z) = (Z_1, Z_2, ..., Z_n)$.

3.4.Fisher's information measures for the simple linear regression model

In the next we will present an informational characterization for the simple linear regression model using the Fisher information measure.

First of all, we recall that all assumptions, denoted by $\mathbf{A}_{1j}, j = \overline{1,7}$, from the Remark 1.3, are satisfied and, more, the hypothesis of the Theorem 3.2.4 are satisfied. In these conditions the dependent variable Y, from the simple linear regression model

$$Y = \beta_0 + \beta_1 x + u, \tag{3.4.1}$$

is a random variable which follows a normal distribution with

$$E(Y) = m_Y = \beta_0 + \beta_1 x, Var(Y) = \sigma^2, \sigma^2 \in \mathbb{R}^+, \tag{3.4.1a}$$

where β_0 and β_1 are the components of the unknown parameter vector $\theta = \beta = (\beta_0,\beta_1)^\intercal$, β_0, $\beta_1 \in \mathbb{R}$.

The probability density function associated to Y is as

$$f(y; m_Y, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{[y-(\beta_0+\beta_1 x)]^2}{2\sigma^2}\right\}. \tag{3.4.2}$$

The likelihood function of $\beta_{0,}\beta_1$ (when σ^2 is **known**) of the random sample $S_n(Y) = (Y_1, Y_2, ..., Y_n)$ has the form

$$L_n(y_1, y_2, ..., y_n; \beta_0, \beta_1) = L_n(\mathbf{y}; \beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1) =$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1 x_i)]^2\right\} \tag{3.4.3}$$

because the sample random variables $Y_1, Y_2, ..., Y_n$ are normally, statistically independent and identically distributed.

Theorem 3.4.1. *The Fisher information matrix concerning the unknown vector parameter $\beta = (\beta_0, \beta_1)$, obtaining using the sample vector $S_n(\mathbf{Y}) = (Y_1, Y_2, ..., Y_n)$, will be as*

$$\mathbf{I}_n(\beta_0, \beta_1) = n.\mathbf{I}_1(\beta_0, \beta_1), \tag{3.4.4}$$

where

$$\mathbf{I}_1(\beta_0, \beta_1) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & S_{xx} + \overline{x}^2 \end{pmatrix} \tag{3.4.4a}$$

Proof. To proof this theorem we shall use the relation of the definition (3.3.14) for the information matrix which, in our case, has the following form

$$\mathbf{I}_n(\beta_0, \beta_1) = \begin{bmatrix} I_n^{(0,0)}(\beta_0, \beta_1) & I_n^{(0,1)}(\beta_0, \beta_1) \\ I_n^{(1,0)}(\beta_0, \beta_1) & I_n^{(1,1)}(\beta_0, \beta_1) \end{bmatrix}, \tag{3.4.5}$$

and its elements follows to be establish.

For this, we consider the log of the likelihood function associated with the likelihood function (3.4.3), namely

$$L* = L^*(\beta_0, \beta_1, \sigma^2) = \log_e L_n(\beta_0, \beta_1, \sigma^2) = \\ = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 \tag{3.4.6}$$

for which the second-order derivatives are as follows

$$\begin{cases} \dfrac{\partial^2\ L^*}{\partial\beta_1^2} & = & -\dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n} x_i^2 \\ \dfrac{\partial^2\ L^*}{\partial\beta_0^2} & = & -\dfrac{n}{\sigma^2} \\ \dfrac{\partial^2\ L^*}{\partial\beta_0\partial\beta_1} & = & -\dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n} x_i, \end{cases} . \tag{3.4.6a}$$

Now, using these second-order derivatives, as well as, the relations (3.3.18) and (3.3.19), we obtain just the elements of the information matrix (3.3.4a).

References

[1] Creel, M., Econometrics, Internet, 2005.

[2] Koch,K.R., Parameter estimation and hypothesis testing in linear models, Springer-Verlag Berlin Heidelberg,1999.

[4] Mihoc, I., Fătu, C.I., Calculul probabilităţilor şi statistică matematică, Casa de editură Transilvania Press, Cluj-Napoca, 2003.

[5] Mihoc, I., Fătu, C.I., Introducere în econometrie, Editura RISO-PRINT, 2007.

[7] Pollack, D.G.S., A Cours of Econometrics, Internet, 2005.

[9] Wilks, S.S., Mathematical statistics, John Wiley & Sons, Inc., New York, London, 1963.

Authors:

Mihoc Ion, Fătu Cristina-Ioana
Faculty of Economics at Christian University
"Dimitrie Cantemir", 3400 Cluj-Napoca
e-mail: *imihoc@cantemir.cluj.astral.ro*