USE OF THE SIMPLE LINEAR REGRESSION MODEL IN MACRO-ECONOMICAL ANALYSES. - theoretical aspects

PhDProfessor Constantin ANGHELACHE
Academy of Economic Studies, Bucharest
"ARTIFEX" University, Bucharest

Abstract

The article presents the fundamental aspects of the linear regression, as a toolbox which can be used in macroeconomic analyses. The article describes the estimation of the parameters, the statistical tests used, the homoscesasticity and heteroskedasticity. The use of econometrics instrument in macroeconomics is an important factor that guarantees the quality of the models, analyses, results and possible interpretation that can be drawn at this level.

Key words: *linear regression, macroeconomics, EViews, finance, Keynes model, indicator*

A series of econometric models are used in the macro-economical analyses in the European Union states, and we can appreciate that these are already standard. We shall emphasize the theoretical analysis regarding the simple and multiple regression, and the linear correlation quotient.

The mathematical relationship of the consumption function used for the Keynes model is the following:

$K_t=\alpha+\beta \cdot Y_t$

where:

K_t = consumption for a period of time (one year, usually);

Y_t = income for the same period;

α,β = parameters of the regression model.

In designing a linear regression model for macroeconomic analyses, the following statements must be made:

a) Establishing the independent variables (Y = resultant variable, the data series is noted by $(y_i)_{i=1,n}$ and X = the explicative factorial, variable, define by the series $(x_i)_{i=1,n}$)

Subsequently, a determinist dependency between the two variables is:

Y= b + aX

The calculated estimators for the two parameters are $\hat{b}$ and $\hat{a}$, established in a stochastic manner.

b) The identification of the residual variable, noted by ε. The residual variable is normally distributed, with the average 0 and constant dispersion. The residual value is included in the model because in economy there is not always available a functional linear dependency between the two variables, but a probabilistic one, the data series

are not affected by measurement errors that influence the estimation of the two parameters, the data series are established through observation on some samples.

c) Establishing the usage conditions for the regression model. After the nature of the data series, there are two domains of use for the linear regression model: in the analysis of dependency between variables, if the data series are recorded at the levels of population statistical units for an interval or moment, by using the formula

$y_i = b + a \cdot x_i + \varepsilon_i$,

where:

y_i = the resultant (explicated) characteristic;

x_i = factorial (explicative) characteristic;

and to emphasize the dependence between the two variables in a certain time horizon, the time series are used.

a) In order to estimate the parameters and to use the regression, a series of hypotheses are applied to identify if the data series are affected or not by measurement errors, the residual variable, the dispersion etc.

- I_1: the data series are not affected by measurement errors;
- I_2: the residual variable has the value 0;
- I_3: the dispersion of the residual variable is non-variant in time, that is, it is homoscedastic;
- I_4: the residuals are not self-correlated;
- I_5: the factorial (explicative) variable is not correlated with the residual one;
- I_6: $\varepsilon_t \rightarrow N(0, \sigma_\varepsilon^2)$.

To test the hypotheses, statistical tests are used, respectively: if the linear dependency is found following transformations on the two variables, then the regression model is linear on its parameters.

Emphasis of these aspects is given by the dependency between the available income and the population consumption, which is a linear shape (the slope quotient is positive).

In macroeconomic analyses, uni-factorial non-linear models can be encountered, that can be linear through transformations applied to the variables of the regression model.

Such non-linear models transformed into linear models are: $y_i = a \cdot x_i^b$ is transformed into a linear model through the logarithm of the terms of the above relationship: $\log y_i = \log a + b \cdot \log x_i$. A linear model exists in report to the variables $\log y_i$ and $\log x_i$, *the exponential model or the log model,* defined by the relationship: $y_i = a * b^x$, that is aligned by logarithm, thus resulting the linear model: $\log y_i = \log a + x_i \cdot \log b$. The use of the model is recommended when the points $(x_i, \log y_i)_{i=1,n}$ are along a straight line.

There are also non-linear models that cannot be written as linear models by applying elementary transformations, and the estimation of parameters is made through other estimation techniques. In these cases, the estimation of parameters can be made by numerical methods.

To estimate a linear regression model, the data series for the two characteristics are used, represented by two vectors:

$$a = \begin{bmatrix} a_1 \\ a_2 \\ . \\ . \\ . \\ a_n \end{bmatrix}$$ for the factorial characteristic.

$$b = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix}$$ for the resultant characteristic

The linear regression model assumes the knowledge on the methods used for the estimation of the two parameters, test of the properties for the regression model estimators and of the elements implies the knowledge on the methods used for the estimation of the two parameters, the test of the properties of the regression model estimators, and of the elements of regression used in forecasts.

If we will consider the relation $y_i = ab^x_i$, it is observed that the estimated value of the resultant variable, of the estimators of the model parameters and their properties depend on the characteristics of the independent variable and the properties of the residual variable. We present the hypotheses related to the variables that define the model or that regard the residual variable.

- An initial hypothesis regards the fact that data series are not affected by recording errors, this hypothesis postulates the characteristics of series of values that are used for the estimation of parameters. The parameter estimation is made based on a set of values $(x_i, y_i)_{i=1,n}$, for the two variables. The function used to analyze the dependency between the two variables includes a large number of statistical observations, so the estimation of the parameters is based on the law of large numbers. The values for the two variables are not affected by significant values able to distort the accuracy of parameter estimation.

This hypothesis is important in establishing the properties of the linear regression model. The values of the factorial characteristic are non-stochastic if each value is related to a family of values of the resulting characteristic. For each value x_i of the factorial characteristic, an average is calculated per the family of the resultant characteristic and the series of values is determined.

For each fixed value of the factorial characteristic, the residual variable has a zero, respectively:

$$E\left[\varepsilon_i \middle| X = x_i\right] = 0 \text{ , for any } i$$

Based on this affirmation, it results that the other factors, non-recorded, with the exception of the factorial characteristic, does not have a significant influence on

the average of the resultant characteristic. If the hypothesis is satisfied by the linear regression model, we can write:

$$E\left[Y \mid X = x_i\right] = b + ax_i$$

• The second hypothesis is the homoscedasticity, which assumes the constant distribution of the residual.

The property emphasizes that the conditional distributions have the same dispersion, namely $\text{var}\left[\varepsilon_i \mid X = x_i\right] = \sigma_\varepsilon^2$,, constant for any *i*.

If the residual variables do not comply with that property, we consider that the regression model is heteroskedastic. In conclusion, the residual variables have different variances.

• The residuals are not correlated, meaning that between the residual factors the co-variance phenomenon is not present, respectively:

$\text{cov}(\varepsilon_i, \varepsilon_j) = 0$, for any $i \neq j$.

If the residual variable fulfills the "b" and "c" hypotheses, the following relationship results:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, i \neq j \\ \sigma_\varepsilon^2, i = j \end{cases}$$

A different situation is when the residual variable presents a order-one autocorrelation, that is:

$\varepsilon_t = \rho\varepsilon_{t-1} + u_t$.

where u_t is a white noise.

• The residual variable is not correlated to the independent variable, situation in which: $\text{cov}(X, \varepsilon_i) = 0$, for any *j*. This relation shows that an increase of the values of the factorial variable does not lead automatically to the increase of the values of the residual variable.

The distribution of the residual variables is made on a normal repartition, with zero average and dispersion σ_x^2, respectively $\varepsilon_i \in N(0, \sigma^2)$.

In compliance with those presented above, the linear regression model $y_i = b + a \cdot x_i + \varepsilon_i$, $i = 1,..,n$ can be defined under two equivalent forms, depending on the stated hypotheses, respectively:

• When the hypotheses are formulated regarding the residual variable:

$$\begin{cases} E(\varepsilon_i) = 0 \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, i \neq j \\ \sigma_\varepsilon^2, i = j \end{cases}; \\ \varepsilon_i \to N(0, \sigma_\varepsilon^2) \end{cases}$$

• If the hypotheses are defined for the resultant variable:

$$\begin{cases} E\left(y_i \middle| X = x_i\right) = b + a \cdot x_i \\ \operatorname{cov}(yi, yj) = \begin{cases} 0, i \neq j \\ \sigma_\varepsilon^2, i = j \end{cases} \\ y_i \to N\left(b + a \cdot x_i \cdot \sigma_\varepsilon^2\right) \end{cases}.$$

Regardless the form in that the linear dependence (y_i, x_i) $i = \overline{1,n}$, is defined inside the simple linear regression model, we estimate the values of the resultant variable as the equation $\hat{y}_i = \hat{b} + \hat{a}x_i$. The residuals are estimated by the relation $e_i = \hat{y}_i - (\hat{b} + \hat{a}x_i)$, satisfying the equality: $\sum_{i=1}^{n} e_i = 0$.

• The parameters of the linear regression model can be estimated by using either the least squares or the maximum likelihood method.

a) By using the least squares method, the values of the resultant characteristic are estimated with the relation:

$\hat{y}_i = \hat{b} + \hat{a}x_i$, (where $\hat{a}$ and $\hat{b}$ are the parameters estimators for the regression line)

The real values of the resultant characteristic are equal with the estimation achieved with the help of the regression model, corrected with the residual error $(y_i = \hat{y}_i + e_i)$

To estimate the parameters, we enforce the condition that the sum of the squares of differences between the real value and the value estimated by regression is minimal.

The optimum conditions of the function lead to the following equations:

$$\begin{cases} \dfrac{\partial\left(\hat{a}, \hat{b}\right)}{\partial\left(\hat{b}\right)} = -\sum_i 2\left(y\hat{i} - \hat{b} - \hat{a}x_i\right) = 0 \\ \dfrac{\partial\left(\hat{a}, \hat{b}\right)}{\partial\left(\hat{a}\right)} = -2\sum_i \left(y\hat{i} - \hat{b} - \hat{a}x_i\right) \cdot x_i = 0 \end{cases}$$

The equations that must be solved to obtain the values of the parameters are established by applying the moment's method.

The two equations are obtained:

- The first equation results from the condition $E(\varepsilon_i) = 0$, defining the equality:

$$\frac{1}{n}\sum_i e_i = 0 \text{ or } \sum_i e_i = 0;$$

- The second equation of the system is established by starting at the hypothesis of non-correlation of the series of values of the factorial variable with the series of the residual variable $(\operatorname{cov}(X, \varepsilon) = 0)$, having the equality:

$\frac{1}{n}\sum_i x_i e_i = 0$.

The two estimators are determined by using the linear equation system:

$$\begin{cases} n\hat{b} + \hat{a}\left(\sum_i x_i\right) = \sum_i y_i \\ \left(\sum_i x_i\right)\cdot\hat{b} + \hat{a}\left(\sum_i x_i\right) = \sum_i y_i \end{cases}.$$

The test for the solution of the system if it complies with the second order conditions is made by using the second order derivatives of the function:

$$\begin{pmatrix} \partial^2\varphi(\hat{a},\hat{b})/\partial\hat{a}^2 & \partial^2\varphi(\hat{a},\hat{b})/\partial\hat{a}\partial\hat{b} \\ \partial^2\varphi(\hat{a},\hat{b})/\partial\hat{a}\partial\hat{b}^2 & \partial^2\varphi(\hat{a},\hat{b})/\partial\hat{b}^2 \end{pmatrix} = \begin{pmatrix} 2\sum_i x_i^2 & 2\sum_i x_i \\ 2\sum_i x_i & 2n \end{pmatrix}$$

The matrix so defined has two properties:

- Is positively defined;
- The determinant of the matrix is positive:

$$\Delta = 4n\sum_i x_i^2 - 4\left(\sum_i xi\right)^2 = 4n\left[\sum_i (x_i - \bar{x})^2\right] > 0 .$$

The calculation relationships of the two estimators, $\hat{a}$ and $\hat{b}$, result from the solving of the linear equations system.

The computation of the regression line slope quotient is made with the equality:

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

The estimator of the regression line slope is a linear combination of the values for the resultant characteristic: $\hat{a} = \sum_{i=1}^{n} w_i y_i$.

The series $(w_i)_{i=1,n}$ fulfills three properties, respectively $\sum_{i=1}^{n} w_i = 0$, $\sum_{i=1}^{n} w_i^2 = 0$, $\sum_{i=1}^{n} w_i x_i = 1$.

Property 1: $\sum_{i=1}^{n} w_i = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\cdot\sum_{i=1}^{n}(x_i - \bar{x}) = 0$;

Property 2: $\sum_{i=1}^{n} w_i^2 = \frac{1}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} \cdot \sum_{i=1}^{n}(x_i-\bar{x})^2 = 0$;

Property 3: $\sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i (x_i - \bar{x}) = 1$.

The calculation relationship for the estimator of the free term of the regression line is determined by solving the equation system or by taking into account that the regression line passes through the center of the points cloud, respectively the equality $\hat{b} + \hat{a}\bar{x} = \bar{y}$.

So, for the *b* parameter, the estimation is achieved by using the relationship:

$\hat{b} = \bar{y} - \hat{a}\bar{x}$.

The use of the least squares method has some inconveniences, the most important are: it does not offer acceptable results if the formulated hypotheses are not satisfied; if $\hat{a}^n, \hat{b}^n$ are the estimators determined based on the series (x_i, y_i), $i = \overline{1,n}$ and $\hat{a}^{n+1}, \hat{b}^{n+1}$ are those evaluated for the series (x_i, y_i), $i = \overline{1,n+1}$, it results that between the two pair of estimators there is no simple recurrence relationship; if the data series presents major changes, the estimators are distorted.

a) The use of the least squares method took into consideration a series of hypotheses on the residual variable ε_i, that did not refer to the form of the random variable ε_i repartition. Surpassing this inconvenient is achieved by using the maximum likelihood method.

The residual value has the property given by the relation:

$$\varepsilon_i \in N(0, \sigma_\varepsilon) \Leftrightarrow f(\varepsilon_i) = \frac{1}{\sigma_\varepsilon \sqrt{2\pi}} e^{-\frac{\varepsilon_i^2}{2\sigma_\varepsilon^2}}$$

out of which results $yi \in N\left(\tilde{b}, \tilde{a}x_i, \tilde{\sigma}_\varepsilon\right)$.

For the case of the linear regression model, the likelihood function is given by the formula:

$$\ell\left(\tilde{a}, \tilde{b}, \tilde{\sigma}_\varepsilon^2\right) = \prod_{i=1}^{n} f(y_i / x_i).$$

The maximum conditions of the likelihood function can be synthesized in:

$$\frac{\partial l}{\partial a} = 0, \frac{\partial l}{\partial b} = 0, \frac{\partial l}{\partial \sigma_\varepsilon^2} = 0.$$

The form of the estimators is made by using the maximum condition for the logarithm of the likelihood function.

The relation below is applied:

$$L(a,b,\sigma_\varepsilon^2) = \ln \ell(a,b,\sigma_\varepsilon^2) = -\frac{n}{2}\left[\ln(2\pi) - \ln \sigma_\varepsilon^2\right] - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(y_i - b - ax_i)^2$$

By using the properties of the logarithm function, we determine:

$$\max_{\tilde{a},\tilde{b},\tilde{\sigma}_\varepsilon^2} \ell\left(\tilde{a},\tilde{b},\tilde{\sigma}_\varepsilon^2\right) \Leftrightarrow \max_{\tilde{a},\tilde{b},\tilde{\sigma}_\varepsilon^2} L\left(\tilde{a},\tilde{b},\tilde{\sigma}_\varepsilon^2\right)$$

The maximum likelihood method leads to the achievement of the same set of estimators for the model parameters as the least squares method. Also, the maximum likelihood method allows the direct achievement of the estimator of the residual variable dispersion.

The expression of this estimator results from the condition:

$$\frac{\partial \log \ell\left(\tilde{\sigma}_\varepsilon^2\right)}{\partial \tilde{\sigma}_\varepsilon^2} = 0$$

After calculating, it is achieved the equation for the determination of the relation of the limit of the residual variable variance estimator.

By taking into view the calculation formula for the adjustment errors, in the end, the dispersion of the residual variable is determined from the relation: $\tilde{\sigma} = \frac{1}{n}\sum_i e_i^2$.

• The regression line – basis of the regression function

Starting from the generalized simple linear regression function, ($y_i = a + bx_i$), the determination of the formula for the estimator do the regression line slope quotient is made on the relation: $a = \frac{\text{cov}(x,y)}{\text{var}(x)}$.

It is achieved the equality starting from the calculation relationship of the estimator, by dividing both the numerator and the denominator to the volume of the sample. From the previous relationship, it results that the estimator and the covariance calculated for the two variables have the same sign.

Between the parameters of the regression slope, the following relationship exists:

$$\frac{a}{a'} = \frac{\text{var}(y)}{\text{var}(x)}$$

The two lines, in the same plan, intersect each other in the gravity center of the point cloud, so the two lines pass through the point $G(\bar{x}, \bar{y})$.

The hypothesis can be demonstrated if we take into account the fact that for the two regression models the equalities $\sum_{i=1}^{n}(y_i - b - ax_i) = 0$ and $\sum_{i=1}^{n}(x_i - b' - a' y_i) = 0$ are valid.

By dividing to *n* (the number of terms for the two equalities), that passes through the $G(\bar{x}, \bar{y})$ point, we have a system of two equations.

The size of the angle formed by the two lines shows the intensity of the connection between the two variables. If the lines are identical in the case of the reciprocal connection between the two variables, it results that, as the size of the angle is smaller, the reciprocal linear link between the two characteristics is stronger.

Subsequently, we obtain the calculation formulas fot the free terms of the two lines if the two quotients of the regression slopes are known: $b = \bar{y} - a\bar{x}$ and $b' = \bar{x} - a'\bar{y}$.

From the equations of the two lines and from the relationships above we obtain the relationships for the two regression lines.

The calculation relationships are:

$$y_i = \bar{y} + \frac{\text{cov}(x, y)}{\text{var}(x)}(x_i - \bar{x}),$$

$$x_i = \bar{x} + \frac{\text{cov}(x, y)}{\text{var}(y)}(y_i - \bar{y}).$$

The algebraic value of the quotients of the slopes in the regression model and the reciprocal regression model are the same and express the sense of dependency between the two variables.

In report to the sign of the estimator for the *a* parameter we distinguish:

- If $\hat{a}$ >0, the dependency between the two variables is direct;
- If the estimation of *a* is zero, no linear dependence exists between the two variables;
- If the quotient of the regression slope is $\hat{a}$ <0, then a reversed linear dependency manifests between the two variables.

We will observe that the sign of the quotient of the slope for the regression line is identical to the sign of the variance calculated for the two variables.

The estimator of the quotient of the regression line slope determined through least squares method is a non-displaced estimator and with minimal dispersion.

If $E(\varepsilon_i) = 0$, for any *i*, the following equalities are obtained progressively:

$$E(\hat{a}) = E(a) + E(\sum_i w_i \varepsilon_i) = a + \sum_i w_i E(\varepsilon_i) = a$$

By taking into consideration the calculation relationship of the estimator, it is observed that is a linear combination of the series of values $y_1, y_2, \ldots, y_{1n}$. by using the three properties of the value series $(w_i)_{i=\overline{1,n}}$, it results the equivalent relation of the estimator, $E(\hat{a}) = E(a) + E(\sum_i w_i \varepsilon_i)$.

To outline the hypothesis that the estimator obtained through least squares is non-displaced, the average operator is applied to the terms of the equality.

For the $\hat{a}$ estimator, the following equalities are enforced:

$$E(\hat{a}) = a$$

$$\text{var}(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \ .$$

The second equality in the „**b"** property is realized by calculating the dispersion of the estimator. In this respect, the following relation is considered:

$$\text{var}(\hat{a}) = E(a - \hat{a})^2 = \text{var}(\sum_i w_i \varepsilon_i) = E((\sum_i w_i \varepsilon)^2 = E(\sum_i w_i^2 \varepsilon_i^2 + \sum_{\substack{i,j \\ i \neq j}} w_i w_j \varepsilon_i \varepsilon_j)$$

$$= \sum_i w_i^2 E(\varepsilon_i^2) + \sum_{\substack{i,j \\ i \neq j}} w_i w_j E(\varepsilon_i \varepsilon_j)$$

By using the hypothesis regarding the non-correlated residual variables, and the hypothesis of homoscedasticity of the residual variable, we obtain:

$$\text{var}(\hat{a}) = \sum_i w_i^2 \sigma_\varepsilon^2 = \sigma_\varepsilon^2 \sum_i w_i^2 = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})}$$

Out of which results that the dispersion of the estimator is smaller as the dispersion of the factorial characteristic is higher.

The demonstration of the Gauss-Markov theorem can be made by considering the $\hat{a}_*$ estimator as a linear combination of the series of values recorded for the resultant characteristic, thus resulting the equality: $\hat{a}_* = \sum a_i y_i$. We ascertain that the weights of the linear combination from the last relationship are identical to the ones of the $(w_i)_{i=\overline{1,n}}$ series, because $y_i = b + ax_i + \varepsilon_i$.

The second restriction of the $\hat{a}_*$ estimator refers the fact that is non-displaced, and from this two properties result, that are satisfied by the weight system $(a_i)_{i=\overline{1,n}}$: $\sum_i a_i = 0$, $\sum_i a_i x_i = 0$.

From the two equalities, results that the estimator is obtained by the relationship:

$$\hat{a}_* = a + \sum_i a_i \varepsilon_i \ .\ .$$

If we compare the dispersions of the two non-displaced estimators that are expressed as linear functions of the values of the resultant variable, results that between the series of weights of the two estimators, the relationships $a_i = w_i + d_i$ are verified for any i.

The third sum of the last relationship is null, by taking into account the properties of the system of weights of the first estimator and the restrictions imposed on the weights system for the second estimator.

- When the residual variable follows the normal repartition, the estimator follows a normal repartition too, with average a and standard deviance $\frac{\sigma_\varepsilon}{\sqrt{n}} \cdot \frac{1}{\sigma_x}$.

The notations signify: σ_x standard deviation of the factorial variable and σ_ε the standard deviation of the residual variable.

The best estimation of the regression line is obtained by reducing the standard deviation of the estimator for the regression slope.

The reduction of this measure is based on the possibility to write the indicator as:

$$\sigma_{\hat{a}} = \frac{\sigma_\varepsilon}{\sqrt{n}} \cdot \frac{1}{\sigma_x}.$$

The standard deviation is directly proportional with the dispersion of the y_1, y_2, …, y_{1n} observations around the regression line and reversely proportional to the number of observations and the dispersion of x_1, x_2, …, x_{1n} values.

The dispersion degree of the series of values for the exogenous characteristics is measured by the average standard deviation of the series and in this context, as the values of the factorial variable are more dispersed, the precision of the estimation is higher.

The estimator of the free term of the regression line achieved through the least squares method is a non-displaced estimator with minimum dispersion.

The following relations are defined:

$$E(\hat{b}) = b \text{ şi } \operatorname{var}(\hat{b}) = \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{\bar{x}^2}{\sigma_x^2}\right)$$

The point cloud determines the possibility to write the equalities:

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = \frac{1}{n}\sum_i y_i - \hat{a}\bar{x} = \frac{1}{n}\sum_i (b + ax_i + \varepsilon_i) - \hat{a}\bar{x} = b + (a - \hat{a})\bar{x} + \frac{1}{n}\sum_i \varepsilon_i$$

The demonstration of the properties of the free term estimator of the linear regression model is made by considering the properties of the $(C_i)_{i=\overline{1,n}}$ series of values, that is:

- $\sum_i C_i = x\sum_i w_i - 1 = -1$;
- $E(C_i) = -\frac{1}{n}$;
- $\sum_i C_i = \frac{\bar{x}^2 \cdot n}{\sigma_x^2} + \frac{1}{n}$;
- $\text{cov}(C_i, \varepsilon_i) = 0$.

The co-variance matrix of the estimators for the linear regression model, „$\hat{a}$" and „$\hat{b}$" is subsequently presented under the form:

$$\Omega_{(\hat{a},\hat{b})} = \begin{pmatrix} \text{var}(\hat{a}) & \text{cov}(\hat{a},\hat{b}) \\ \text{cov}(\hat{b},\hat{a}) & \text{var}(\hat{b}) \end{pmatrix} = \sigma_\varepsilon^2 \begin{pmatrix} \frac{1}{S_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \end{pmatrix}.$$

The co-variance matrix of the estimators is defined by the relations:

$$\text{cov}(\hat{a},\hat{a}) = \text{var}(\hat{a}), \text{cov}(\hat{b},\hat{b}) = \text{var}(\hat{b}), \text{cov}(\hat{a},\hat{b}) = \text{cov}(\hat{b},\hat{a})$$

The calculation model for the estimators covariance takes into account the hypotheses of the classical regression model, resulting:

$$\text{cov}(\hat{a},\hat{b}) = E(a-\hat{a})(b-\hat{b})^2 = E\left[\left(-\sum w_i\varepsilon_i\right)\left(\sum C_i\varepsilon_i\right)\right] = -\sigma_\varepsilon^2 \sum w_i C_i$$

$$= -\sigma_\varepsilon^2 \sum w_i \left(w_i\bar{x} - \frac{1}{n}\right) = -\sigma_\varepsilon^2 \bar{x}\sum_i w_i^2 + \frac{\sigma_\varepsilon^2}{n}\sum_i w_i = \frac{\sigma_\varepsilon^2 \bar{x}}{S_{xx}}$$

It can be demonstrated the fact that the „$\hat{a}$" estimator converges in probability to the „*a*" parameter. Similarly, the estimator of the free term of the classical regression model, , „$\hat{b}$", tends in probability towards „*b*".

The affirmations are evident if we take into view that:

$$\text{var}(\hat{a}) = \frac{\sigma_\varepsilon^2}{n\sigma_n^2} \xrightarrow[n\to\infty]{} 0$$

$$\text{var}(\hat{b}) = \frac{\sigma_\varepsilon^2}{n}\left(1 + \frac{\bar{x}^2}{\sigma_x^2}\right) \xrightarrow[n\to\infty]{} 0$$

The covariance of $\hat{a}$ and $\hat{y}$, for x_i fixed, is null:

$$\text{cov}(\hat{a},\bar{y}) = \text{cov}\left(\sum_i w_i y_i, \bar{y}\right) = \sum_i w_i \text{ cov}(y_i, \bar{y}).$$

But $\operatorname{cov}(y_i, \bar{y}) = \operatorname{cov}\left(y_i, \frac{1}{n}\sum_j y_j \right) = \sum_j \frac{1}{n}\operatorname{cov}(y_i, y_j) = \frac{\sigma_\varepsilon^2}{n}$, because y_i and y_j are independent variables, if i ≠ j. we shall obtain, by considering the properties of the series $(w_i)_{i=\overline{1,n}}$, the following equalities:

$$\operatorname{cov}(\hat{a}, \bar{y}) = \sum_i w_i \frac{\sigma_y^2}{n} = \frac{\sigma_y^2}{n}\sum_i w_i = 0$$

References

- Anghelache C., Mitruț. C, Bugudui, E., Deatcu, C. (2010) – *„Econometrics"*, Artifex Publishing House, Bucharest, ISBN 978-973-7631-76-3

- Anghelache C., Anghelache, G.V. (2010) – „*Equilibrium models for macroeconomic forecasts*", Romanian Statistical Review, Issue 9/2010, supplement

- Anghelache C., Dinu M., Barbu C.M. (2010) - „*Estimation of the regression function's transformation*", Metalurgia Internațional Vol. XV (2010)

- Buşe, L., Ganea, M., Cîrciumaru, D. (2010) – „*Using Linear Regression In The Analysis Of Financial-Economic Performances*", Annals of Computational Economics, Issue 38/2010

- Giles, E. (2010) – „*Bayesian Estimation of a Possibly Mis-Specified Linear Regression Model*", Department of Economics, University of Victoria in Econometrics Working Papers

- Schlag, K., Gossner, O. (2009) – „*Finite Sample Nonparametric Tests for Linear Regressions*", Department of Economics and Business, Universitat Pompeu Fabra, Economics Working Papers

- Søren Johansen & Morten Ørregaard Nielsen (2010) - "*Likelihood inference for a fractionally cointegrated vector autoregressive model*", CREATES Research Papers 2010-24, School of Economics and Management, University of Aarhus