Theoretical Aspects Concerning the Testing of the Significance of the Regression Model

Prof. Constantin ANGHELACHE, PhD.
Academy of Economic Studies, Bucharest
„Artifex" University of Bucharest
Alexandru URSACHE, PhD. Student
Academy of Economic Studies, Bucharest
Bogdan DRAGOMIR, PhD. Student
Academy of Economic Studies, Bucharest
Georgeta BARDAŞU (LIXANDRU), PhD. Student
Academy of Economic Studies, Bucharest
Marius POPOVICI, PhD. Student
Academy of Economic Studies, Bucharest

Abstract

Before being utilized, any model based on the regression function must be analysed and accommodated to the actual conditions implied by the performed analysis. The veracity of the outcomes resulting out of the utilization of the regression model implies that the significance of the model being taken into consideration is tested. Testing the significance of the regression linear model can be accomplished by applying statistical procedures, out of which we shall consider only the Student test and the variation analysis.

Key words: *significance, regression, test, parameter, model*

In submitting the two procedures applied to testing the hypotheses formulated on the parameters of the regression model, the following emphasizes are to be considered:

- the estimators of the parameters of the regression linear model are of minimum dispersion in the class of non-removed estimators;
- if the parameters of the model are estimated by means of the least squares method, then the dispersion of the residual is estimated through the relation:

$$\hat{\sigma}_{\varepsilon}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \qquad (1)$$

The estimator of the variance of the residual variable is a non-removed estimator:

$E\left(\hat{\sigma}_{\varepsilon}^2\right) = \sigma_{\varepsilon}^2$

- the residual variable is following up a normal repartition $N(0, \hat{\sigma}_{\varepsilon}^2)$.

Starting from the properties of the estimators of the regression linear model, the estimators $\hat{a}$ and $\hat{b}$ are linear combinations of randomly variable normally distributed[1].

In order to define the statistics, the follow two situations are to be considered:

- **the dispersion of the residual variable is known.**

Considering the expressions of the two estimators, it is resulting that these ones are meeting the following two properties:

$$\hat{a} \to N\left(a, \frac{\sigma_{\varepsilon}^2}{\sum_i (x_i - \bar{x})^2}\right), \qquad (2)$$
$$\hat{b} \to N\left(b, \frac{\sigma_{\varepsilon}^2}{n}\left(1 + \frac{\bar{x}^2}{\sigma_x^2}\right)\right).$$

Considering the properties of the normal distribution as well as the above outcomes, the following results are obtained:

$$\frac{\hat{a}-a}{\sigma_{\hat{a}}} = \frac{(\hat{a}-a)\sqrt{\sum_i (x_i - \bar{x})^2}}{\sigma_{\varepsilon}} \to N(0,1), \qquad (3)$$
$$\frac{\hat{b}-b}{\sigma_{\hat{b}}} = \frac{\left(\hat{b}-b\right)}{\sigma_{\varepsilon}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{n\sigma_x^2}}} \to N(0,1).$$

These outcomes are useful for testing some hypotheses formulated in respect of the parameters, as well as for defining the corresponding intervals of confidence.

- **the dispersion of the residual variable is unknown.**

[1] Anghel, M.G. (2014) – "*Econometric Model Applied in the Analysis of the Correlation between Some of the Macroeconomic Variables*", Revista Română de Statistică – Supliment/Nr. 1

In order to define the statistics used for testing the significance of the parameters of the regression linear models we have to keep in mind that:

- if $x_i \to N(0,1), i = 1,...,n$, then

$$z = \sum_{i=1}^{n} x_i^2 \to \lambda_n^2;$$

- if $x_i \to N(0,\sigma^2), i = 1,...,n$, then

$$z = \sum_{i=1}^{n} \left(\frac{x_i}{\sigma} \right)^2 \to \lambda_n^2;$$

- if $x_i \to N(0,1)$ and $z \to \chi_k^2$, then

$$\frac{x}{\sqrt{z/k}} \to t_k .$$

In terms of practical analysis, the dispersion of the residual variable is not known, this one being estimated through the relation (1). Taking into consideration the calculation relationship of the Student statistics and applying the three properties, the following results are obtained:

- *for the coefficient of the slope of the regression line:*

In order to test H0: $a = \hat{a}$, with the alternative $H_0: a \neq \hat{a}$, we have to keep in mind the fact that:

$$\frac{\hat{a}-a}{\sigma_{\hat{a}}} = \frac{(\hat{a}-a)\sqrt{\sum_i (x_i - \bar{x})^2}}{\sigma_\varepsilon} = \frac{(\hat{a}-a)\sigma_x}{\sigma_e}\sqrt{n-2} \to t_{n-2} \qquad (4)$$

- *for the free term*

In order to test the null hypothesis $H_0: b = \hat{b}$, with the alternative : $H_1: b \neq \hat{b}$, we have to keep in mind the fact that:

$$\frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}} = \frac{\left(\hat{b}-b\right)}{\hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}}} \to t_{n-2}. \qquad (5)$$

These two outcomes are useful for testing the significance and defining the intervals of confidence for the two parameters of the regression line[2].

[2] Manole, A. et. al. (2013) – "*Conditional Probability and Econometric Models*", Romanian Statistical Review Supplement., Issue 1/2013

- **Testing the null hypothesis** when there is an established significance threshold, if $\left|\frac{\hat{a}-a}{\hat{\sigma}_{\hat{a}}}\right| > t_{\alpha/2;n-2}$, then the null hypothesis is rejected. This test is used in order to set up whether the linear dependence between the two characteristics is a significant one. In this case the testing goes for H_0: a=0, with the alternative H_1: $a \neq 0$. The null hypothesis is rejected if $\left|\frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}\right| > t_{\alpha/2;n-2}$.

- **Defining the interval of confidence:** For a threshold of significance α established out of the Student repartition table the value $t_{\alpha/2;n-2}$ is set up for *n-2* degrees of liberty.

- A specific interval of confidence is defined for each parameter.

- For the parameter *a*, the interval of confidence is:

$$\hat{a} - t_{\alpha/2;n-2}\hat{\sigma}_{\hat{a}} \leq a \leq \hat{a} + t_{\alpha/2;n-2}\hat{\sigma}_{\hat{a}} \quad (6)$$

- For the free term the interval of confidence is defined as:

$$\hat{b} - t_{\alpha/2;n-2}\hat{\sigma}_{\hat{b}} \leq b \leq \hat{b} + t_{\alpha/2;n-2}\hat{\sigma}_{b} \quad (7)$$

In order to test if the linear dependence between the two variables is significant, namely if the value of the coefficient of the slope differs from zero, the dispersion analysis is applied as well.

Each parameter of the regression model would be separately tested or a procedure of a simultaneous testing could be applied. As the two estimators, $\hat{a}$ and $\hat{b}$, are not independent alleatory variables, it is considered that the successive testing of the two parameters is not exactly correct. Therefore, the simultaneous testing of the two parameters is recommended. The test hypothesis would be defined as follows:

$H_0 : a = a_0, b = b_0$

$H_1 : a \neq a_0, b \neq b_0$

If noting with $\hat{\Omega}_{(\hat{a},\hat{b})}$ the estimator of the covariance matrix of the parameters of the regression linear model then we define:

$$F_{a,b} = \frac{1}{2}\begin{pmatrix}\hat{a}-a\\ \hat{b}-b\end{pmatrix}^{-1}\left(\Omega^{(\hat{a},\hat{b})}\right)^{-1}\begin{pmatrix}\hat{a}-a\\ \hat{b}-b\end{pmatrix}$$
$$= \frac{1}{2\hat{\sigma}_{\varepsilon}^2}\left[n(\hat{a}-a)^2 + 2n\bar{x}(\hat{a}-a)(\hat{b}-b) + (\hat{b}-b)^2\sum_i x_i^2\right] \to F_{2;n-2} \quad (8)$$

For a simultaneous testing of the two parameters, we shall replace, within the expression of $F_{a,b,}$ the a,b by a_0, b_0. For a given threshold of

significance α, the value $F_{\alpha;2;n-2}$ is read from the Fisher – Snedecor table of distribution.

If the inequality $F_{calculated} > F_{tableted}$ is is fulfilled, then the null hypothesis is rejected, accepting that at least one parameter differs significantly from the specified value.

The dispersion analysis is a statistical procedure for testing the quality of the model, which has as a starting point the decomposition of the total variance of the dispersion due to the regression factor and the dispersion due to the action of the non-recorded factors[3].

We define the notions :

- $SPT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ representing the sum of the squares of the terms of the endogenous variable;

- $SPE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ quantifying the sum of the squares of the estimated terms deviations;

- $SPR = \sum_{i=1}^{n} (e_i)^2$ representing the sum of the squares of the estimating errors.

Between the three terms the equality

SPT = SPE + SPR (9)

is verified.

For each term out of the last equality the number of liberty degrees has to be set up. Thus, for the three terms, these are equal to *n-1*, *n-2*, *2-1*.

In order to define the test statistics, the property of the variables χ^2 has to be considered, namely:

If *x* and *z* are two alleatory independent variables with distributions χ^2 and k_2 liberty degrees, then:

$$F = \frac{x / k_1}{z / k_2} \rightarrow F_{k_1;k_2}$$

Out of the property of the estimator $\hat{a}$, it is resulting that:

$$\left(\frac{\hat{a} - a}{\sigma_{\hat{a}}^2} \right)^2 = \frac{(\hat{a} - a)}{\sigma_{\varepsilon}^2 / \sum (x_i - \bar{x})^2} \rightarrow \chi_1^2 \quad (10)$$

Out of the property of the residual variable, we get:

[3] Anghel, M.G. (2008) – *"Utilizarea modelului de regresie în analiza situaţiei pieţei de capital"*, Revista Română de Statistică – Supliment „România în procesul integrării europene", nr. 12/2008

$$\frac{\sum_i e_i^2}{\sigma_\varepsilon^2} \to \chi^2_{n-2} \tag{11}$$

In order to test the null hypothesis $H_0 : a = \hat{a}$ we define:

$$F = \frac{(\hat{a}-a)^2 \sum_i (x_i - \bar{x})^2}{\sum_i e_i^2 /(n-2)} \to F(1, n-2) \tag{12}$$

The null hypothesis *a=0* is tested, according to which the exogenous variable does not influence to a significant extent the values of the endogenous characteristic. The test relation is the following:

$$F = \frac{\hat{a}^2 \sum_i (x_i - \bar{x})^2}{\sum_i e_i^2 /(n-2)} \to F(1, n-2) \tag{13}$$

In order to set up an equivalent form for the last statistics, the fact that, under the null hypothesis conditions as to the independence of the two characteristics, the terms of the equality (9) are expressed: $SPE = \sum (\hat{y}_i - \bar{y})^2 = \hat{a}^2 \sum (xi - \bar{x})$, while $SPR = \sum (y_i - \hat{y}_i)^2$ has to be considered.

The *F* test is written as an equivalent form:

$$F = \frac{SPE/1}{SPR/(n-2)} \tag{14}$$

Out of the last relation the expression of the statistics *F* is deducted depending on the value of the determination ratio R^2:

$$F = \frac{R^2}{1-R^2}(n-2). \tag{15}$$

In order to set up whether the linear dependence between the two variables is a significant one, the estimated value *F* for the data series established for the two characteristics is compared with the tableted value of this statistics. If the inequality: $F > F_{1-\alpha;(1;n-2)}$ is observed, then the null hypothesis *H_0: a=0* is rejected.

For the significance threshold α it is ascertained that there is no significant linear dependence between the two variables[4].

[4] Anghelache, C., Anghel, M.G., Manole, A., Dincă (Nicola), Z. (2014) – *„The Regression Model used to Analyze the Correlation between Production and Labor*", Revista Română de Statistică - Supliment nr. 1/2014

If there is a significant linear dependence between the two variables, it has been demonstrated that $R^2 = r^2$. Under these circumstances, the relation (15) turns to:

$$F = \frac{r^2}{1-r^2}(n-2) = t^2_{n-2}.$$

There is a new statistics arising for testing the linear dependence between the two variables:

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-2} \to t_{n-2}.$$

In practice, there is an issue occurring, namely to set out whether the various regression linear models, which parameters have been estimated for the data recorded at the level of various populations, are significantly differing[5].

Let's consider the data series $(x_i, y_i)_{\overline{i=1,n}}$ and $(x'_i, y'_i)_{\overline{i=1,n'}}$ for the two statistical characteristics, in the case of two populations taken into account.

Based on the first series of values, the parameters of the regression linear model as well as the dispersion of the slope coefficient have been estimated:

- the regression linear model: $\hat{y}_i = \hat{b} + \hat{a}x_i$;
- the dispersion of the slope coefficient: $\hat{\sigma}^2_{\hat{a}}$.

For the second population, a similar procedure leads to the following outcomes:

- the estimated regression line is $\hat{y}'_i = \hat{b}' + \hat{a}' x'_i$
- the dispersion of the slope coefficient is $\hat{\sigma}^2_{\hat{a}'}$

The issue to consider is to set out whether the two regression models have different characteristics as against the coefficient of the regression line slope.

In order to test if the two coefficients of regression differ significantly we apply the relation:

$$t = \frac{\hat{a} - \hat{a}'}{\sqrt{\sigma^2_{\hat{a}} + \sigma^2_{\hat{a}'}}} \qquad (16)$$

[5] Anghel M.G. et al. (2014) – *Using the regression model for the portfolios analysis and management*, Theoretical and Applied Economics, Volume XXI, No.4

The alleatory variable $d = a - a'$ is defined in order to measure the difference between the slopes of the two regression lines. In order to establish if the two lines have the same value of the regression slopes, we define:

- the null hypothesis of the test H_0: $d = 0$, with the alternative H_1: $d \neq 0$;
- the test statistics: $t = \frac{\hat{d}}{\sigma_{\hat{d}}}$.

If considering that the two estimators are independent, then the test statistics is:

$$t = \hat{d} / \sqrt{\sigma_{\hat{a}}^2 + \sigma_{\hat{a}'}^2} \qquad (17)$$

In order to test the null hypothesis, a significance threshold α must be set out. Out of the Student table of distribution , the tableted value is set out as $t_{\alpha/2}$. If the value calculated through (17) is higher that $t_{\alpha/2}$ the null hypothesis is rejected. It is accepted that the two coefficients are significantly different.

Conclusion

The utilisation of the regression model is giving very good results for the economic analyses. In practice, there is an issue to be considered, namely, in case there are various regression models which statistical significance has been checked up, which one should we apply to? We are interested to get close positions for the parameters estimated for the recorded data In case they are significantly daggering, we use the test „t" given by the relation:

$$t = \frac{\hat{a} - \hat{a}'}{\sqrt{\sigma_{\hat{a}}^2 + \sigma_{\hat{a}'}^2}}$$

Then, we use the null hypothesis and finally we analyse the inequality:

F>T1-∞; (1; n-2)

References

Anghel, M.G. (2014) – "*Econometric Model Applied in the Analysis of the Correlation between Some of the Macroeconomic Variables*", Revista Română de Statistică – Supliment/Nr. 1

Anghel M.G. et al. (2014) – *Using the regression model for the portfolios analysis and management*, Theoretical and Applied Economics, Volume XXI, No.4

Anghel, M.G. (2013) – „*Modele de gestiune şi analiză a portofoliilor*", Editura Economică, Bucureşti

Anghel, M.G. (2008) – "*Utilizarea modelului de regresie în analiza situaţiei pieţei de capital"*, Revista Română de Statistică – Supliment „România în procesul integrării europene", nr. 12/2008

Anghelache, C., Anghel, M.G., Manole, A., Dincă (Nicola), Z. (2014) – „*The Regression Model used to Analyze the Correlation between Production and Labor*", Revista Română de Statistică - Supliment nr. 1/2014

Anghelache, C. (2013) – "*Elemente de econometrie teoretică*", Editura Artifex, Bucureşti

Anghelache, C., Lilea, F.P.C. (2012) – "*Econometrie*", Editura ARTIFEX, Bucureşti

Anghelache C., Isaic-Maniu AL., Mitruţ C., Voineagu V. (2011) - „*Sistemul conturilor naţionale: sinteze şi studii de caz*", Editura Economică, Bucureşti

Bardsen, G. et. al. (2005) – "*The Econometrics of Macroeconomic Modelling*", Oxford University Press

Benjamin, C. et.al. (2010) – „*Forecasting with an Econometric Model*", Springer

Dougherty, C. (2008) – "*Introduction to econometrics. Fourth edition*", Oxford University Press

Manole, A. et. al. (2013) – "*Conditional Probability and Econometric Models*", Romanian Statistical Review Supplement., Issue 1/2013

Mitruţ, C. (2008) – „*Basic econometrics for business administration"*, Editura ASE, Bucureşti