Aspects Concerning the Verification of the Residual Normality and the Prediction of the Regression Model

Prof. Constantin ANGHELACHE, PhD.
Academy of Economic Studies, Bucharest
„Artifex" University of Bucharest
Prof. Radu Titus MARINESCU, PhD.
„Artifex" University of Bucharest
Assoc. prof. Alexandru MANOLE, PhD.
„Artifex" University of Bucharest
Ec. Emilia STANCIU

Abstract

The tests used for verifying certain hypotheses formulated on the parameters of the regression model as well as for defining the intervals of confidence for these ones and elaborating predictions have as starting point the verification of the hypothesis concerning the normal distribution of the residual.

Key words: *residual, prediction, trimming, coefficient, distribution*

The relations applied for testing the characteristics of the residual distribution are defined by taking into account the asymmetry and the trimming of the normal distribution.

For an alleatory variable with a normal distribution, the asymmetry coefficient is zero while the trimming one is three.

Let's consider the regression linear model $y_i = b + ax_i + \varepsilon_i, i = 1,...,n$, and the series of the estimated residual $(e_i)_{i=\overline{1,n}}$, cu $e_i = y_i - (\hat{b} + \hat{a}x_i)$. For the residual series there are two indicators to define, used by the descriptive statistics in order to analyze the asymmetry and the trimming of a distribution series:

- asymmetry coefficient

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

- trimming coefficient

$$\beta_2 = \frac{\mu_4}{\mu_2}$$

In order to define the statistical tests used for the verification of the residual distribution according to a normal distribution, the following property of the symmetry and trimming coefficients:

„Consider the alleatory variable $X \to N(m, \sigma_x^2)$. The asymmetry and trimming coefficients calculated for a series of data with *n* values, which is defined for this variable, are meeting the following properties:

$$\hat{\beta}_1^{1/2} \to N\left(0, \sqrt{\frac{6}{n}}\right),$$

$$\hat{\beta}_2 \to N\left(3, \sqrt{\frac{24}{n}}\right)$$

In order to verify the null hypothesis of the normal distribution of the residual $(e_i)_{\overline{i=1,n}}$ we have to resort to one of the tests:

- tests for verifying the asymmetry and trimming for the distribution of the estimated residual;
- tests for verifying the asymmetry and the test Jarque – Bera.

On the ground of the estimated series it is verified whether this distribution is normally divided. By using this series, the two coefficients are estimated, as $\hat{\beta}_1^{1/2}$, respectively $\hat{\beta}_2$.

Under the null hypothesis H_0: $\beta_1 = 0$, it is resulting:

$$S = \frac{\hat{\beta}_1^{1/2}}{\sqrt{\frac{6}{n}}} \to N(0,1).$$

Similarly, if defining the null hypothesis on the second coefficient, as H_0:β_2=3, then:

$$K = \frac{\hat{\beta}_2 - 3}{\sqrt{\frac{24}{n}}} \to N(0,1)$$

The null hypothesis according to which the residual is uniformly distributed is accepted provided the following inequities are simultaneously met:

$|S| < t_{1-\frac{\alpha}{2}}$ and $|K| < t_{1-\frac{\alpha}{2}}$,

where $t_{1-\frac{\alpha}{2}}$ is the value of the distribution quartile *N(0,1)*for the significance threshold α.

The test Jarque – Bera allows the simultaneous verification of the properties of asymmetry and trimming of the residuals series.

The test is defined as against the two coefficients $\hat{\beta}_1^{1/2}$ and $\hat{\beta}_2$, taking into consideration the distribution of their estimators, resulting:

$$J - B = \left(\frac{\hat{\beta}_1^{1/2}}{\sqrt{\frac{6}{n}}} \right)^2 + \left(\frac{\hat{\beta}_2 - 3}{\sqrt{\frac{24}{n}}} \right)^2 \to \chi_2^2 .$$

Or, as the equivalent form:

$$J - B = \frac{n}{6}\beta_1 + \frac{n}{24}\left(\hat{\beta}_2 - 3\right)^2$$

For significance threshold α the null hypothesis of the normal distribution of the residual variable is rejected if the inequality below is met:

$J - B > \chi^2_{2;1-\alpha}$.

Prediction through the regression model

On the basis of the data series $(x_i, y_i)_{i=\overline{1,n}}$ the parameters of the regression line have been estimated. Thus we get the series of the estimated values for the endogenous variable through the relation:

$\hat{y}_i = \hat{b} + \hat{a}x_i, i = \overline{1,n}$

Within the prediction process, using the regression linear model, there is the question mark on how to solve the following two issues:

- **accomplishing predictions either punctually or through intervals of confidence.** For accomplishing the first prediction the punctual method is applied while for the second situation, the prediction is made through an interval of confidence;
- **verifying the framing of certain points within the tendency postulated by a regression model.** If there are named values for

the two characteristics of the regression model, under the form of the pair (x_0,y_0), the issue to settle consists of setting out whether they line up with the trend defined by the regression model. We shall verify if he value of the endogenous characteristic is taking part with the interval of prediction being defined for a level of the exogenous characteristic and a threshold of significance.

We shall make punctual or through an interval of confidence predictions for a value of the endogenous characteristic y_0 or for its mean, $E(y_0)$. For each and every case there are various calculation formulas being established for the punctual prediction and the prediction through an interval of confidence.

For the regression linear model, the real value of the endogenous characteristic is specified through the relation:

$$y_0 = b + ax_0 + \varepsilon_0 \qquad (1)$$

where ε_0 is the accomplishment of a normal distribution of mean zero and dispersion equal to one.

The punctual value estimated through the regression linear model is defined by the relation:

$$y_0 = \hat{b} + \hat{a}x_0$$

As a rule, this value is utilized for defining an interval of confidence. In order to define the interval of confidence, in the conditions that a level of the significance threshold is specified, we must take into consideration the fact that, by utilizing the regression linear model for defining the punctual prediction, a prediction error is made, equalling to:

$$e_0 = y_0 - \hat{y}_0 = (b - \hat{b}) + (a - \hat{a})x_0 + \varepsilon_0 \qquad (2)$$

Considering the properties of the two estimators of the regression line estimators, we shall submit below the main properties of the prediction error.

The mean of the prediction error equals to zero. We define the equality:

$E(e_0)=0$

The above result is obvious if applying the mean operator to the terms of the equality (2), taking into account the properties of the two estimators and the hypothesis formulated on the residual variable.

The dispersion of the prediction error made in the case when the purpose is to make a prediction for the value of the endogenous characteristic y_0 is:

$$\operatorname{var}(e_0) = \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{\left(\bar{x} - x_0\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \right] \qquad (3)$$

In order to obtain the expression of the variance of the prediction error, the dispersion of the terms of the equality (2) is applied. The following outcomes are obtained:

$$\operatorname{var}(e_0) = E(e_0^2) = \operatorname{var}(\hat{b}) + x_0^2 \operatorname{var}(\hat{a}) + \operatorname{var}(e_0) + 2x0 \operatorname{cov}(\hat{a}, \hat{b})$$

$$= \sigma_\varepsilon^2 \left[\frac{1}{S_{xx}} + x_0^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) - 2x_0 \frac{\bar{x}}{S_{xx}} \right]$$

$$= \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{\left(\bar{x} - x_0\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \right]$$

For building up an interval of prediction for the value of the endogenous variable, in the conditions of a fixed level of the exogenous characteristic, the following two results are to be taken into consideration:

$$\frac{y_0 - \hat{y}_0}{\sigma} \to N(0,1)$$
$$\frac{y_0 - \hat{y}_0}{\hat{\sigma}_p} \to t_{n-2} \qquad (4)$$

We noted by $\hat{\sigma}_p$ the estimator of the average standard deviation of the prediction error made for the value y_0. This is calculated through the following relation:

$$\hat{\sigma} = \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}} \qquad (5)$$

If a certain significance threshold α is fixed, then we shall define the interval of prediction for y_0:

$$\hat{y} - t_{\alpha/2} \cdot \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}} < y_0 < \hat{y} + t_{\alpha/2} \cdot \sigma_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2}} \qquad (6)$$

For a stable significance threshold, the size of the prediction interval is function of the following measurements:

- the value of the exogenous for which the value of the endogenous characteristic is predicted. This factor is quantified through the term $\left(x_0 - \bar{x}\right)^2$;
- the number of terms of the series which have been used for estimating the parameters of the regression linear model. The prediction error is proportionally inverse to n;
- the quality of the regression model being quantified through the dispersion of the residual variable;
- the value of the significance threshold.

In the situation where a prediction on the average values $E(y_0)$ is made, under the conditions of an established values for the exogenous characteristic, the dispersion of the prediction error is:

$$\text{var}(e_0) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{\sum_i \left(x_i - \bar{x}\right)^2} \right] \qquad (7)$$

For proving the last relation it must be considered that the prediction error made in this case is:

$$e_0 = E(y_0) - \hat{y}_0$$
$$= (b - \hat{b}) + (a - \hat{a})x_0 + \varepsilon_0$$

By applying the mean operator to the terms of the equality above, we get the formula (7).

References

Andrei, T., Bourbonais, R. (2008) - „*Econometrie*", Editura Economică, Bucureşti

Andrei, T., Stancu, S., Iacob A.I., Tusa, E., - „*Introducere în econometrie utilizând Eviews*", Editura Economică, Bucureşti

Anghelache, C. şi alţii (2012) – „*Elemente de econometrie teoretică şi aplicată*", Editura Artifex, Bucureşti

Anghelache, C., Cruceru, D., Marinescu, R.T. (2011) - „*Modele econometrice utilizate în analiza performanţelor financiare*", Scientific Research Themes/Studies Communications at the National Seminary „Octav Onicescu", Romanian Statistical Review Trim. 2/2011, pp. 94-100

Dougherty, C. (2008) – "*Introduction to econometrics. Fourth edition*", Oxford University Press
Hendry, D.F. (2002) – „*Applied econometrics without sinning*", Journal of Economic Surveys, 16
Voineagu, V., Ţiţan, E. şi colectiv (2007) – "*Teorie şi practică econometrică*", Editura Meteor Press