THE DETERMINATION OF THE OPTIMAL STRUCTURE REGRESSION MODEL

Vezhichanina K.
ksuvezh@gmail.com

Introduction. It is known that the inclusion of an econometric model of additional factors (explanatory variables) always leads to an increasing the coefficient of determination R2 and to reducing the sum of squared residuals model RSS. Thus, it is an "improvement" of a model. However, the extra factors reduce the efficiency of the model parameter estimations, and lead to the expansion of the confidence intervals of these estimations.

Aim. The purpose of this paper is to analyze the ways to optimize the number of explanatory variables in the regression models.

Materials and methods. The advisability of adding of a new group of regressors is usually determined by one of embodiments of Wald's test. We compare long and short regression (long contains additional explanatory variables). There is a decrease of sum of squares of residues in the transition from short to long regression (from RSS_k to RSS_m). It means regression becomes more accurate. The essential question is whether it compensates the decline? The null hypothesis assumes the simultaneous vanishing of the coefficients of these variables. The corresponding F-statistics is as follows:

$$F = \frac{(RSS_k - RSS_m)/(m-k)}{RSS_m/(n-m)}$$

where k - the number of regressors in the short (restricted) regression, m - quantity of regressors in a long (unrestricted) regression, n - number of observations.

If the calculated value of F-statistics is larger than $F_{кр} = F_{1-\alpha}(m-k, n-m)$, where α - significance level, long regression is significantly better (null hypothesis is rejected).

The present approach is quite cumbersome in terms of its implementation. At the same time to solve the problem, you can use the so-called punitive criteria for determining the quality of the constructed regression models, which are easier to use. The most popular criteria are the Akaike information criterion AIC and Bayesian information criterion (Schwarz criterion) BIC.

For the linear, multiple regression model Akaike criterion value is calculated using the following formula:

$$AIC = n \ln(RSS/n) + 2p$$

where *n* - number of observations, *p* - the number of model parameters, *RSS* - the sum of squared residuals model obtained in the evaluation of the coefficients of the least squares' method.

Schwartz criterion is:

$$BIC = n \ln(RSS/n) + \ln(n)\, p$$

Results and discussion. By increasing the number of explanatory variables, the first term on the right side is reduced and the second is increased. Consequently, the criterion rewards approach for the quality of approximation and penalizes for excess of explanatory variables. Among several alternative models, one is preferred in which value of *AIC* and *BIC* are smaller.

The prediction model is the better, the lower the *AIC* and *BIC* are. Reduction of the residual dispersion positively influences on these criteria and the number of enabled parameters influences negatively. The main difference between them is the degree of stiffness, i.e. the value of the penalty for a large quantity of factors in the model.

BIC is more stringent criterion. As we can see from the formula above, its stiffness increases with increasing of n. Thus *AIC* is more focused on the accuracy of the forecast, and *BIC* - to minimize the number of covariates.

The analyzed approaches about solving of problem of obtaining the optimal structure of the regression models have their drawbacks. As a part of the algorithm associated with the consideration of statistical hypotheses, it is possible only pairwise comparison of models with different numbers of factors. Thus, analysis of ~ m^2 pairs of hypotheses is needed (*m* - the maximum quantity of factors). Approach, connected with using of punitive criteria, has no such a drawback. Each set of explanatory variables meets one criterion value. Therefore in the analysis we have to deal with a relatively small number (*m*) of values of used criterion. At the same time this approach is too simplistic and not sufficiently well-grounded.

Conclusions. It is presented that the penal criteria can be used in the initial phase of construction of a regression model by the primary determination of its structure.