JEL: C01; C13; C52

OLS: THE ADEQUACY OF LINEAR REGRESSION SOLUTIONS

Alexander G. Tyzhnenko
Simon Kuznets National University of Economics, Kharkiv, Ukraine

Abstract – The article is devoted to the problem of the economic adequacy of solving a linear regression problem (that is, finding the economic adequate regression coefficients) by the common OLS and the modified OLS (MOLS) [1] methods.

Key Terms – Adequacy, Economical correctness, Modified Ordinary Least Square, Ordinary Least Square, Regression modeling.

The article is devoted to the problem of the economic adequacy of solving a linear regression problem (that is, finding the economic adequate regression coefficients) by the common OLS and the modified OLS (MOLS) [1] methods.

For solving the problem of economic adequacy of the solution of the linear regression problem we use the following definition of the economic interpretation of the regressor coefficients in the linear model: a regression coefficient is the value of the contribution of a given regressor to the regressand when this regressor changes by a unit, with the rest of regressors unchanged.

Because of that, the main objective of the work is the study of the adequacy of a mathematical solution of the linear regression problem to the economic content of the influence of regressors on the response.

With this goal in hand, in our investigation was obtained the following result: the linear regression solution should not only have the correct signs but should also correctly reflect the relationship between the coefficients of regression in the population. Otherwise, the researcher may be incorrectly informed about which regressor contributes more to the response, and which one less.

The solution of the linear regression problem, which has the correct signs and correctly reflects the relationship between the regression coefficients in the population, is called in the paper as adequate.

Our investigation shows that the physical correctness of a solution of the regression problem by an OLS-like method (i.e. a solution with correct signs), which has been considered in the paper [1], is only a necessary condition for the adequacy of a solution based on the observed data. A sufficient condition for the adequacy of a solution is a correct reflection of the relationships between the regression coefficients in the population.

First of all, this problem of adequacy is investigated in our work on examples of solving the linear regression problem by the method of least squares (OLS), as the most applicable method while solving the regression problems by practitioners.

It has been shown that a requirement of the adequacy of an OLS-solution of the regression problem is more stringent than a requirement of the physical correctness in the sense of variability of a solution, that is, requires a less solution variability.

Study conducted has been shown that OLS solutions for not very large samples may be not adequate to the solution in the population, although they may be physically correct (with correct signs) and statistically significant.

The OLS solutions are compared with the MOLS solutions, obtained in [1]. It has been shown that MOLS solutions are essentially less variable than OLS solutions. In contrast to the OLS method, the variability of the MOLS solutions decreases with the growth of the VIF factor. This one allows obtaining an adequate solution in those cases where the OLS gives an inadequate solution. Thus, the application of MOLS significantly expands the possibility of finding economically adequate solutions of the linear regression problem at any level of near-collinearity.

In our investigation the variability of the VIF factor and the Student's statistic are also considered. It is shown that both indicators change very much both in small and large samples. Thus, it has been confirmed that there is no definite critical value of the VFF factor that separates multicollinear and non-multicollinear data. For the same reason, the results of t-tests for individual samples may be questionable.

All mentioned results are obtained by using the artificial population algorithm (APA) proposed in [1] and modified in further investigations. The APA allows you generating data of any size with known regression coefficients in the whole population.

Random changes of regressors in APA are divided into two parts. The first part is coherent to the response changes, but the second part is random (incoherent). This one allows us changing the near-collinearity level by changing the variance of the incoherent noise in regressors.

With the APA in hand we investigate both positive and negative sides of both methods, the OLS and MOLS. To the very important advantage of the OLS can be attributed its unbiasedness and consistency. Both of these properties allow us determining with the required accuracy the values of regression coefficients in a population generated by APA using a large sample ($\sim 10^5 - 10^6$). The knowledge of the regression problem solution in a population allows us estimating the biasedness of the solution and their variability by both the OLS and MOLS methods for samples of any size.

Using the APA, it has been investigated in our investigations the problem of finding the best regressions. We show that the elimination of strong collinear regressors leads to an increase in the incoherent noise in the regressors, which in turn leads to a decrease in the regression coefficients and their significance.

It has been also shown that the elimination of strongly correlated regressors is not economically justified, but is rather a measure of lowering the value of the VIF factor. Given that the MOLS only improves its accuracy with the growth of the regressors near-collinearity level, the new method of solving the regression problem proposed in [1], avoids the need for the elimination of strongly correlated regressors. Moreover, as shown in our investigations it should be eliminated, most likely, the weakly correlated regressors after the appropriate economic analysis.

In general, the solution of the linear regression problem by the OLS method is clearly divided into two parts. Part 1 is a purely mathematical problem of the approximation of a response (the goodness of fit problem in the linear regression) [2-8], which the OLS solves flawlessly. Part 2 is an economic (in the general sense, physical) task of evaluating the influence of regressors on the regressand. It is this task that the OLS solves unsatisfactorily. As will be seen from the following considerations, this issue is connected with the attempt of the OLS to solve the problem exactly.

Therefore, it is clear that the method of solving the economic problem of linear regression (part 2) must be approximate, but rather accurate. It is precisely such method, the MOLS, has been proposed in [1]. It has been shown that the MOLS method gives stable and practically unbiased solution to the linear regression problem regardless of the near-collinearity level of the data used. Unlike the ridge-method, the MOLS gives a negligible bias and does not require an optimization of the regularization constant.

In general, the economic data, which are of interest to research, are usually rather collinear. This is mainly due to the fact that the regressors are indicators (factors, indexes) of economic activity, and the very task of a linear regression is designed to identify the existence of economic laws, for which the factors selected as regressors (X_j), affect a certain economic factor, which is selected by the researcher as a response (Y).

If the economic laws for all economic objects under investigation are the same, then it is clear that the values of all selected regressors (economic factors) X_{ij} of the i-th object must be formed under the same economic laws.

According to the same economic laws, the factors of other objects ($k \neq i$) must also be formed. That is, all the values of the factors X_{ij}, depending on the index i, should vary to a certain extent coherently, if the indexes are chosen correctly and they obey the same economic laws. It is believed that this is exactly the case for statistical series (cross-sectional data), but this is even more obvious for time series, which deals with the economic factors of one object or a group of objects in time, when economic indices are formed under the influence of the same laws that vary in time.

If, when compiling a linear regression model, all factors that influence the response (Y) are taken into account, then all the values of the factors X_{ij}, depending on the index i, must change strictly coherently. It is clear that such a condition is never fulfilled because the factors that influence the response (Y) are actually much a lot and they cannot be fully taken into account.

Due to a very large number of unaccounted factors, it can be assumed that changes in the values of the factors X_{ij}, depending on index i, will contain, in addition to the coherent component, the random noise, which reduces the correlation between the regressors and the response, as well as the value of the regressors themselves.

It is important also to add that exactly the coherent component of the regressors determines the level of influence of the regressor on the response, which occurs due to regular relationships between economic factors.

Increasing the incoherent component increases the regressor dispersion and reduces the correlation coefficient between the regressor and the response, as well as between this regressor and other regressors. As a result, the regression coefficient of this regressor and the VIF-factor will decrease and, as a consequence, the regressor' influence on the response also will decrease according to the linear regression model.

Indeed, if some factor should have a significant influence on the response from theoretical considerations, but its effect is really small due to the high level of the incoherent component in this regressor, then this is most likely due to the factors not included in the model or due to the incorrectness of the theory in this case.

From these general considerations follows that exclusion from the model of strongly correlated regressors to achieve small values of the VIF factor may be economically unwarranted. This one also reduces the values of regression coefficients by module and their significance. On the other hand, exclusion from the model of weakly correlated regressors increases by module the regression coefficients of the other regressors and their significance. But whether to remove or not a weakly correlated regressor must decide the economist.

Note, that the procedure for exclusion from the model of strongly correlated regressors cannot be justified from an economic point of view. It is clear that all strongly correlated regressors must remain in the model, since they have different economic content.

On the other hand, the removal of weakly correlated regressors should also not be done without detailed economic consideration. It should be investigated whether this regressor can be influential on the regressand. If not, this regressor must be discarded. If so, it is necessary to review the experiment in search of unregistered regressors, the absence of which leads to increased noise in the weakly correlated regressor.

References

1. Tyzhnenko A. G. A new stable solution to the linear regression problem under multicollinearity / A. G. Tyzhnenko // Economics of Development. – 2018. – 2(86). – P. 89-99. Available at: http://www.ed.ksue.edu.ua/ER/knt/ee182_86/e182tyz.pdf.

2. Seber, G.A.F. Linear Regression Analysis / G.A.F. Seber. – NY: Wiley-Blackwell, 1977. – 456 p.

3. Seber, G.A.F. Linear Regression Analysis, 2nd edition / G.A.F. Seber, A.J. Lee. – NY: Wiley, 2003. – 341 p.

4. Spanos, A., Probability Theory and Statistical Inference: econometric modeling with observational data / A. Spanos. – Cambridge: Cambridge University Press, 1999. – 401 p.

5. Gujarati, D. N. Basic econometrics / D. N. Gujarati. – NY: McGraw-Hill, 2002. – 526 p.

6. Wooldridge, J.M. Introductory Econometrics: Modern Approach, 5th ed. / J.M. Wooldridge. – Ohio: South-Western, 2009. – 633 p.

7. Badi Baltagi. Econometrics / Badi Baltagi. – NY: Springer, 2011. – 812 p.

8. Greene, W.H. Econometric Analysis, 7th ed. / W.H. Greene,. – NY: Pearson. 2012. – 1211 p.

Author:

Alexander G. Tyzhnenko – PhD in Physical and Mathematical Sciences, Associate Professor of the Department of Higher Mathematics and Economic and Mathematical Methods of Simon Kuznets Kharkiv National University of Economics (oleksandr.tyzhnenko@m.hneu.edu.ua).

Manuscript received 20 January 2019.

Published as submitted by the author.