UDC 330.43 (075.8)
JEL Classification: C52

ORDINARY LEAST SQUARES: THE ADEQUACY OF LINEAR REGRESSION SOLUTIONS UNDER MULTICOLLINEARITY AND WITHOUT IT

© 2019 TYZHNENKO A. G., RYEZNIK Y. V.

UDC 330.43 (075.8)
JEL Classification: C52

Tyzhnenko A. G., Ryeznik Y. V.

Ordinary Least Squares: the Adequacy of Linear Regression Solutions under Multicollinearity and without it

The article deals with the problem of economic adequacy of solving a linear regression problem by the OLS method. The study uses the following definition of adequacy: a linear regression solution is considered adequate if it not only has correct signs but also correctly reflects the relationship between coefficients of regression in the population. If in this case the coefficient of determination is greater than 0.8, the solution is considered economically adequate. As an indicator of adequacy of a linear regression problem solution it is proposed to use a 10 % level of the coefficient of variability (CV) of the regression coefficients. It is shown that OLS solutions may be not adequate to the solution in the population, although they may be physically correct (with correct signs) and statistically significant. The mentioned result is obtained by using the artificial data population (ADP) algorithm. The ADP allows generating data of any size with known regression coefficients in the whole population, which can be calculated with the aid of the OLS solution for a very large sample. The ADP algorithm makes it possible to change the regular component of the influence of the regressors on the response. Besides, the random changes of regressors in the ADP are divided into two parts. The first part is coherent to the response changes, but the second part is completely random (incoherent). This one allows changing the near-collinearity level of the data by changing the variance of the incoherent noise in regressors. Studies using ADP have shown that with a high probability the OLS solutions are physically incorrect if the sample sizes (n) are less than 23; physically correct but not adequate for 23 < n < 400; adequate for n > 400. Furthermore, it is noted that if the elimination of strongly correlated regressors is not economically justified but is rather a measure of lowering the value of the VIF-factor, the results may be far from the reality. In this regard, it is stated that the use of the MOLS eliminates the need to exclude strongly correlated regressors at all, since the accuracy of the MOLS solution increases with an increase in the VIF.

Keywords: *multicollinearity, OLS, data simulation, artificial population, physical correctness, adequacy.*

DOI: *https://doi.org/10.32983/2222-0712-2019-1-217-227*

Tbl.: *3.* ***Formulae:*** *5.* ***Bibl.:*** *29.*

Tyzhnenko Alexander G. *– Ph.D. (Physics and Mathematcs), Associate Professor, Department of Mathematics and Mathematcal Methods in Economics, Simon Kuznets Kharkiv Natonal University of Economics (9a Nauky Ave., Kharkiv, 61166, Ukraine)*

E-mail: *olersandr.tyzhnenko@m.hneu.edu.ua*

ORCID: *http://orcid.org/0000-0001-8508-7341*

Ryeznik Yevgen V. *– Ph.D. (Applied Mathematics and Statistics) , Department of Mathematcs, Uppsala University (Lagerhyddsvagen 1, Uppsala, 75106, Sweden)*

E-mail: *yevgen.ryeznik@math.uu.se*

ORCID: *https://orcid.org/0000-0003-2997-8566*

УДК 330.43 (075.8)
JEL Classification: C52

Тижненко О. Г., Резнік Є. В. Метод найменших квадратів: адекватність рішень задачі лінійної регресії за наявності мультиколінеарності і без неї

Статтю присвячено проблемі економічної адекватності рішення задачі лінійної регресії методом найменших квадратів (МНК). Використано таке означення адекватності: рішення задачі регресії вважається адекватним, якщо воно не тільки має коректні знаки, а й вірно відображає взаємовідношення між коефіцієнтами регресії в генеральній сукупності (ГС). Якщо при цьому коефіцієнт детермінації більший за 0.8, рішення вважається економічно адекватним. Як показник адекватності рішення задачі регресії запропоновано використати 10 %-ний рівень коефіцієнта варіабельності коефіцієнтів регресії. Показано, що МНК рішення можуть бути не адекватними рішенню в ГС, хоча бути фізично коректними (з вірними знаками) і статистично значущими. Зазначений результат був отриманим за допомогою алгоритму штучної генеральної сукупності (artificial data population – ADP). ADP дозволяє генерувати вибірки будь-якого розміру з відомими коефіцієнтами регресії в ГС, які можуть бути розрахованими за допомогою МНК рішення для дуже великої вибірки. Алгоритм ADP дозволяє змінятии регулярну компоненту впливу регресора на відгук. Крім цього, випадкова складова регресорів в ADP розділена на дві частини. Перша частина когерентна змінам відгуку, а друга є повністю випадковою (некогерентною).

UDC 330.43 (075.8)
JEL Classification: C52

Тыжненко А. Г., Резник Е. В. Метод наименьших квадратов: адекватность решений задачи линейной регрессии при мультиколлинеарности и без нее

Статья посвящена проблеме экономической адекватности решения задачи линейной регрессии методом наименьших квадратов (МНК). В статье использовано следующее определение адекватности: решение задачи регрессии считается адекватной, если оно не только имеет корректные знаки, но и верно отражает взаимоотношения между коэффициентами регрессии в генеральной совокупности (ГС). Если при этом коэффициент детерминации больше 0.8, решение считается экономически адекватным. В качестве показателя адекватности решения задачи регрессии предложено использовать 10 % уровень коэффициента вариабельности коэффициентов регрессии. Показано, что МНК решения могут быть неадекватными решению в ГС, хотя быть физически корректными (с верными знаками) и статистически значимыми. Указанный результат был получен с помощью алгоритма искусственной генеральной совокупности (artificial data population – ADP). ADP позволяет генерировать выборки любого размера с известными коэффициентами регрессии в ГС, которые могут быть рассчитаны с помощью МНК решения для очень большой выборки. Алгоритм ADP позволяет менять регулярную компоненту воздействия регрессоров на отзыв. Кроме этого, случайная составляющая регрессоров в ADP разделена на две части.

Саме це дозволяє змінювати рівень майже-колінеарності за допомогою зміни дисперсії некогерентного шуму в регресорах. Дослідження за допомогою ADP показали, що з високою ймовірністю МНК рішення можуть бути фізично некоректними при розмірі вибірки (n) менших, ніж 23; фізично коректними, але не адекватними при 23 < n < 400; адекватними при n > 400. Зазначено, що виключення сильно корелюючих регресорів, якщо це невиправдано з економічної точки зору, а диктується тільки необхідністю зменшити VIF-фактор, може привести до результатів, далеких від реальності. У зв'язку з цим зазначено, що використання модифікованого МНК (ММНК) взагалі звільняє від необхідності виключення сильно корелюючих регресорів, оскільки точність ММНК тільки зростає зі зростанням VIF-фактора.

Ключові слова: *мультиколінеарність, МНК, моделювання даних, штучна генеральна сукупність, фізична коректність, адекватність.*

Табл.: *3.* ***Формул:*** *5.* ***Бібл.:*** *29.*

Тижненко Олександр Григорович *– кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики та економіко-математичних методів, Харківський національний економічний університет ім. С. Кузнеця (пр. Науки, 9а, Харків, 61166, Україна)*

E-mail: *olersandr.tyzhnenko@m.hneu.edu.ua*

ORCID: *http://orcid.org/0000-0001-8508-7341*

Рєзнік Євген Володимирович *– кандидат фізико-математичних наук, кафедра математики, Уппсальський університет (Легерхюддсвеген, 1, Уппсала, 75106, Швеція)*

E-mail: *yevgen.ryeznik@math.uu.se*

ORCID: *https://orcid.org/0000-0003-2997-8566*

Первая часть когерентная изменениям отклика, а вторая является полностью случайной (некогерентной). Именно это позволяет изменять уровень почти-коллинеарности, с помощью изменения дисперсии некогерентного шума в регрессоров. Исследования с помощью ADP показали, что с высокой вероятностью МНК решения могут быть физически некорректными при размере выборки (n) меньших, чем 23; физически корректными, но не адекватными при 23 < n < 400; адекватными при n > 400 Отмечено, что исключение сильно коррелирующих регрессоров, если это неоправданно с экономической точки зрения, а диктуется только необходимостью уменьшить VIF-фактор, может привести к результатам, далеким от реальности. В связи с этим отмечено, что использование модифицированного МНК (ММНК) вообще освобождает от необходимости исключения сильно коррелирующих регрессоров, поскольку точность ММНК только растет с ростом VIF-фактора.

Ключевые слова: *мультиколлинеарность, МНК, моделирование данных, искусственная генеральная совокупность, физическая корректность, адекватность.*

Табл.: *3.* ***Формул:*** *5.* ***Библ.:*** *29.*

Тыжненко Александр Григорьевич *– кандидат физико-математических наук, доцент, доцент кафедры высшей математики и экономико-математических методов, Харьковский национальный экономический университет им. С. Кузнеца (пр. Науки, 9а, Харьков, 61166, Украина)*

E-mail: *olersandr.tyzhnenko@m.hneu.edu.ua*

ORCID: *http://orcid.org/0000-0001-8508-7341*

Резник Евгений Владимирович *– кандидат физико-математических наук, кафедра математики, Уппсальский университет (Легерхюддсвеген, 1, Уппсала, 75106, Швеция)*

E-mail: *yevgen.ryeznik@math.uu.se*

ORCID: *https://orcid.org/0000-0003-2997-8566*

In general, solving the linear regression problem by the OLS method is clearly divided into two parts [1]. 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 [9-19]. As shown in [1], this issue is related to an attempt of finding exact solutions to problems by means of the OLS.

Therefore, it is clear that a method of solving the economic problem of linear regression (part 2) must be approximate but rather accurate. It is precisely such method, the MOLS, is proposed in [1]. It is shown that the MOLS method gives a 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 optimization of the regularization constant.

The MOLS permits to obtain a stable and adequate solution to the linear regression problem without extracting from the model strongly correlated regressors, which have to remain in the model, since they may have different economic content.

In principle, the economic indices can be strictly proportional or even equal, and this should not prevent solving the economic task of determining the degree of influence of regressors on the response.

Therefore, formally, the method of solving a regression problem should allow finding solutions even in the case when two (or more) regressors are simply equal. In this case, the validity of the method can be verified by the equality of the regression coefficients for the same regressors.

Although it is clear that mathematical methods by themselves cannot give recipes for an adequate compilation of a regression model, at the same time it is necessary to create a mathematical program that could find an adequate solution to the economic problem of linear regression under conditions of the near-collinearity of data. Such a method is presented in [1]. In this paper, we want to show that the widely used OLS method is not always suitable for these purposes.

For this purpose, we consider in more details solving the economic linear regression problem by the OLS method.

All problems that are associated with solving the economic problem of linear regression in the presence of near-collinearity arise when solving the matrix OLS-equation:

$$X^{'}Xb = X^{'}Y \Leftrightarrow Ab = B. \tag{1}$$

Recall that this does not mean the goodness of fit problem in the linear regression, but the problem of an adequate estimation of the regression coefficients b_j used in the economics to quantify the impact of regressors X_j on the Y (response).

For any degree of the data near-collinearity, the OLS solution to equation (1) is mathematically correct, as shown by the high accuracy of the approximation (goodness of fit) problem solving [2-8]. At the same time, the finding of adequate estimation of the regression coefficients by the OLS method in the presence of near-collinearity encounters serious difficulties.

Firstly, the solutions do not always have proper signs, i.e., they can be physically incorrect (the 1st problem).

Secondly, physically correct solutions may not correspond to economic suppositions about the power of the influence of the respective regressors on the response in the population, i.e., it may be inadequate (the 2nd problem).

General problems of solving linear equations systems of any level of ill-conditionality are considered in [1], where it is shown that the codomain of any non-singular square matrix A consists of two parts: the codomain of physical correctness, D^c in which all signs of the solution to the equation $Ax = B$ correspond to the content of the problem being investigated, and the codomain of physical incorrectness $\bar{D}^c$, in which not all the solution signs have a sense.

If the right-hand side B of the matrix equation (1) belongs to D^c, all the solution' signs correspond to the content of the problem being investigated, i.e., the solution is physically correct. If the right-hand side of the matrix equation $B \in \bar{D}^c$, then some signs of the solution are necessarily incorrect and all the solution is physically incorrect [1].

In [1], it is also shown that this property of non-singular matrix equations is observed at any level of the *matrix A* conditioning. It is also shown that with the increasing of the ill-conditioning level of the *matrix A*, the area of physical correctness D^c is narrowed.

Using the example of a simple economic problem, in [1], it is shown that the well-conditioned matrix equation of the second order, which arises in the problem of the sale of two products, for some values of the parameters of the problem has an economically incorrect solution. This happens if the right-hand side B of the matrix equation belongs to the codomain of physical incorrectness ($\bar{D}^c$) of the matrix.

On the basis of the studies carried out in [1], one can state that the problems of the OLS method in the presence of near-collinearity are as follows.

As the ill-conditioning level of the matrix $X'X$ grows with the data near-collinearity level, the physical correctness codomain D^c narrows.

Under the influence of random errors in the regressors and response, both D^c and the right-hand side B are changed. This leads to changing in the position of the vector B inside D^c, which makes it possible an exit of the vector B from D^c. It results in changing some of the signs of the solution, and the components of the solution increase by the module, the higher the level of ill-conditioning the more the increase is.

High instability of the OLS method in the presence of near-collinear data is primarily due to a poor conditioning of the OLS matrix equation. In this case, the codomain of physical correctness (D^c) of the OLS matrix is very narrow and varies significantly with random data changes, enforcing the right side of the OLS equation to exit out of the codomain of physical correctness of the OLS matrix.

However, as shown in [1], the most important aspect of the instability of the OLS method is the method of solving the matrix equation itself. The OLS uses the exact method of solving the linear systems (Gauss' or Cramer'), which is very unstable under the data near-collinearity.

Numerous studies of solving poorly conditioned equations [20–29] have shown that exact methods cannot give a solution with acceptable variability, although approximate methods that would give a solution with acceptable variability and accuracy, too, do not exist to date.

At present, the best method for finding an approximate solution to the linear regression problem under the ill-conditioning remains the ridge-regression method [11; 12].

This method gives a stable solution for a not very small value of the ridge parameter, but its accuracy and stability depend on this parameter value itself. This one leads to the ridge parameter optimization problem, which is also has been solved with an accuracy which cannot always be correctly estimated theoretically. In this regard, in practice, the OLS is used, as before, in the case when the solution has the correct signs [19].

In this work, first and foremost, it has been shown what kind of problems can arise when using the OLS in cases where the OLS-solution has the correct signs. We believe that such an investigation will allow researchers to decide for themselves whether they are satisfied with the accuracy given by the OLS, or there is a need to apply a more precise method, namely, the MOLS [1].

In the literature, the appearance of unreasonably large OLS solutions is associated with the poor conditioning of the OLS matrix in the presence of near-collinearity. As for the reason for the appearance of incorrect signs of the OLS solution, there is no well-founded opinion on this matter in the literature [13; 23; 28; 29]. In article [29], the author gives many reasons for the appearance of incorrect signs in the OLS solutions, mainly of an economic nature, and provides recommendations for their elimination. The same concerns works [13; 23; 28], which consider similar reasons for the possible incorrectness of OLS solutions but without mentioning the reasons for the appearance of incorrect signs.

In these works, the methods for eliminating the incorrectness of the OLS solution for all the mentioned authors are practically identical and have an economic direction. However, it should be taken into account that, as clearly stated in [13], measures based on economic theory do not always eliminate the appearance of incorrect OLS solutions, including incorrect signs. But it remains unclarified what causes exactly lead to the wrong signs of the OLS solutions.

This question is considered in [1], where it is shown that both the appearance of very large values in the OLS solution and the appearance of incorrect signs of the solution are associated only with the extremely large variability of the method of solving matrix equation (1).

Namely, due to the high variability of the OLS method, small changes in the data for not very large sample sizes lead to significant changes in the solution to the matrix equation (1).

Due to the impact of unrecorded factors, which increases the incoherent noise in the regressors, vector B in (1) may come out from the codomain of matrix A, namely, D^c. Therefore, both the incorrect signs in the solution to the economic regression problem and the increase of the amplitude of solutions in the presence of near-collinearity are appearing (the 1st problem).

If, despite the influence of random factors, the vector B remains in D^c, the OLS solution will have true signs and meaningful absolute values. In this case, the solution is perceived by the researcher as adequate, although it may be not very close to the solution in the population due to the instability of the mathematical method itself (the 2nd problem).

With the advent of the first problem, it is clear to everyone that the solution is not appropriate, and that it is necessary to take some measures. Basically, these measures include the removal of some regressors according to a certain principle, see, e.g., [13; 23; 28]. After this, it is believed that everything is in order and, after checking for the significance by Student's *t*-test, the solution is used for further research.

Herewith, one does not take into account the fact that the OLS solution can be unstable even when $B \in D^c$, and the physically correct solution obtained can incorrectly reflect the relationship between the regression coefficients in the population. In this case, the OLS solution may have the correct signs and be significant by Student's *t*-test but be inadequate to the solution in the population. In addition, for another sample from the same population, some regression coefficients may become insignificant or have incorrect signs.

In this work it is shown that both problems (physical incorrectness and inadequacy of solutions) have the same source – the high instability of solutions of the matrix equation (1).

Moreover, if the problem of physical incorrectness (the 1st problem) is easily diagnosed, then the problem of the inadequacy of physically correct solutions (the 2nd problem) does not manifest itself at all in individual solutions and, as far as the authors know, has not even been discussed in the literature.

The existence of the problem of the adequacy of physically correct solutions is shown in this work with the help of artificial data population (ADP) algorithm proposed in [1] for simple and multiple linear regression models. The ADP algorithm, which is used in the work for simulating data, allows to generate linear regressors for a given a priori response.

Recall the basic principles of the ADP. At first, we set a priory any response Y: $Y = a + s * randn$. Here, α and s are the arbitrary numbers, $randn(n,1)$ – the n-size pseudo-random vector. Based on this response, we set m regressors X_j, $j = 1 : m$ as follows: $X_j = = k_j * (Y + d_j * s * \alpha * randn(n,1))$, where, $k_j = \tan(\beta_j * pi / 180)$ is the slope (angular coefficient) of a trend of the simple regression of X_j on Y. Accordingly, β_j is the angle between this trend and the *OY* axis. The angular coefficient k_j determines the linear influence of the regressor X_j on the response Y. The coefficient α allows to change the level of incoherent noise in all regressors simultaneously. The coefficients d_j make it possible to change the level of incoherent noise in the individual regressors X_j. The coefficient s allows to get rid of the dependence of the parameters of the artificial population on the choice of the variance of the response (s), which is important for comparing the simulation results with actual data.

The value $k_j Y$, changes coherently with the response and depends on the economic law of the influence of the regressor on the response (k_j). The second term in regressors is the incoherent noise. With the diminishing of the parameter β_j, the regular influence of the regressor X_j on the response Y increases. Because of that, the correspondent regression coefficient b_j in the modeled population increases as well. It should also be noted that while modeling the stochastic regressors, the pseudo-random function $randn(n,1)$ in (3) restarts for each replica.

The regressors are consisting of a coherent part that is generated by regular random changes under the influence of economic laws that are the same for all objects, the incoherent part, i.e., random noise, which is a consequence of the influence on the regressors of unaccounted factors and the regular part, which accounts a linear impact of the regressor on the regressand.

The ADP allows to generate data from a limited population (a population, which consists of all samples of a given size), in which it is possible to regulate the values of regression coefficients, by changing the law of the influence of factors on the response (the regular part of the regressor that contains also the coherent changes), and the value of the variance of random noise (the incoherent part of the regressor).

This allows us to investigate various methods of solving the linear regression problem for variability, depending on the size of the sample and the level of the regressors' near-collinearity. The increase or decrease of the level of near-collinearity is carried out by increasing or decreasing of the incoherent noise variance.

With the aid of the ADP, an artificial population of limited size (a limited population) is simulated. With this ADP in hand, we investigate in the paper the most common method of solving the linear regression problems, namely, the OLS.

It has been studied the variability of the OLS solution, depending on the size of the sample and the level of regressors' near-collinearity. For this, the level of regressors' near-collinearity is estimated by using the VIF-factor .

The presence of an artificial population (ADP), from which it is possible to take any number of samples of a certain size, helps us clarify a lot of details related to solving the linear regression problem.

First of all, this concerns the question of the diagnosis of multicollinearity [13-19], which is still discussed in the hope of finding the value of the VIF-factor , which delimits data on multicollinear and non-multicollinear, although some researchers believe that multicollinearity is a continuous process for which the criterial number does not exist [12; 23].

In the paper, data simulations with the help of the ADP show a high variability of the VIF-factor itself for samples of any size and for any level of data near-collinearity, which, in principle, does not imply the existence of a certain criterion that would distinguish between multicollinear and non-multicollinear data.

This means that we not always can find out in advance whether there is or not the multicollinearity and whether to take any measures to reduce it or not. This is especially important in cases where the VIF-factor is not very large.

For large VIFs it becomes clear, as will be seen further, that the near-collinearity should occur despite the large variability of the VIF itself.

There is, however, an opportunity as if to bypass the issue of big VIF if, after solving the regression problem, all signs of the regression coefficients are correct and all coefficients are significant by Student's *t*-test.

In this case, the researcher is compelled to consider the obtained estimates of regression coefficients to be adequate to their values in the population. However, it is necessary to take into account the opportunity that all the correct signs of the regression coefficients could turn out randomly, and for another sample, the signs may be incorrect. Therefore, if the critical value of the VIF-factor were known, the researcher could make a more informed decision about the possible adequacy or inadequacy of the estimates obtained.

Since the critical value of the VIF-factor does not exist, according to our investigation, the researcher is forced to make a decision only on the basis of Student's criterion in the case of correct signs for all regression coefficients. In this regard, the question arises of the stability of Student's criterion itself from sample to sample. Because of this, a study of the variability of the *t*-statistic for regression coefficients obtained by the OLS is conducted in the paper. For this purpose, the variances of the regression coefficients are calculated by data modeling with the ADP and by the theoretical formulas obtained for the normal distribution of the residual error with practically the same result.

Therefore, the usual linear regression problem has been solved many times by the OLS method for samples drawn from the artificial data population (ADP), and each time the *t*-statistics were calculated for all regression coefficients using the known standard deviations of the regression coefficients. After that, the obtained values of *t*-statistics were averaged and their coefficients of variation were determined.

The ADP data modeling also has allowed us to reveal a high variability of the values of *t*-statistic both for large and non-very large samples and to study the variability of the values of *t*-statistic depending on the sample size and the level of near-collinearity and to find the areas of parameters where the use of the OLS is unacceptable despite the OLS-solutions may be physically correct, i.e., have all correct signs.

The study of the variability of the *t*-statistics calls into question the existing methodology for making decisions about the adequacy of OLS solutions due to the significance of the regression coefficients and their correct signs received by using the only one not very large sample.

Further application of the ADP simulation made it possible to investigate the variability of the regression coefficients themselves and to find out those ranges of the sample size and the level of near-collinearity in which the OLS gives an adequate solution. The paper shows that the level of adequacy depends on the degree of regression coefficients variability, which should be set by the researcher a priori as a certain value of their coefficients of variation. For example, the acceptable value, to our mind, of the critical coefficient of variation may be: $CV = 10$ %.

Thus, in the present work, we will consider the solution to the linear regression problem to be adequate to the solution in the population if the coefficient of variation of the solution does not exceed 10 %.

The article also shows that the regression data simulation using the ADP allows estimating the sample size, starting from which the OLS provides an acceptable accuracy of the regression coefficients at a given level of random noise in the regressors, which determines the near-collinearity level of the regressors.

Note that special attention in these studies is paid to the study of the variability of the solution to the linear regression problem with a high level of random noise in the regressors, i.e., with a low level of near-collinearity up to the practically non-correlated regressors.

It turned out that weakly correlating regressors with a high level of random noise in the regressors have an increased variance due to the random errors, which reduce the regression coefficient and increase its variability. This one casts doubt on the expediency of excluding strongly correlating regressors from the model with the only goal to reduce the level of the VIF-factor .

The authors agree with [23] in the point that the level of variability of OLS solutions is associated with the level of regressors' near-collinearity, which is determined, as a rule, by the VIF-factor [8]. In other words, the idea expressed in work [23] is as follows: the variability of the least squares solution in a continuous manner depends on the level of near-collinearity and there is no any critical value that would separate the data into "multicollinear" and "non-multicollinear".

This assumption contradicts many theoretical works in which the authors try to find the critical value of the VIF-factor, e.g., 10 in [24] and 5 in [25], or the condition number of the OLS matrix, 20 as the critical one [8, p. 130].

Whether or not there is a critical value of a factor that divides data into multicollinear and non-multicollinear ones can be checked directly from the data generated by the ADP. As an indicator, we take the VIF-factor.

To do this, we use $M = 10^4$ samples from the population $DS5(n, \alpha)$, that consists of 5 regressors with size $n = 10$ (small), $n = 40$ (medium), $n = 100$ (fairly large) and so on, with different values of the *alpha*-parameters: 3; 1; 0.9, 0.8; 0.7; 0.6; 0.5; 0.4; 0.3; 0.2; 0.1, 0.01. In this $DS5(n, \alpha)$ we take the following β_j : {1, 1, 5, 5, 5} in degrees and d_j: {1, 1, 1, 1, 1}. In this case, in the population, the values of the first two regression coefficients should be the same and large, the other three should also be the same but smaller. The variance of the incoherent noise of each regressor is given by the vector *d*. In this paper, it is taken the same for all regressors. The variances of the first two regressors should be the same, the last three should also be the same but have a smaller value.

As an example of data simulations, Table 1 shows the 95 % confidence intervals of the VIF-factor, the average values of the VIF-factor and the coefficient of variation of the VIF-factor for each α-value and for $n = 10$.

Table 1

Sample size, n = 10; 5 regressors

α	CI_{VIF}	mean *VIF*	CI_{VIF}(%)
3	(1.4; 11.1)	3.6	90.0
1	(1.8; 22.3)	6.7	133.1
0.9	(1.9; 27.3)	7.8	131.6
0.8	(2.1; 31.8)	8.8	122.0
0.7	(2.4; 39.3)	10.8	109.2
0.6	(2.8; 49.7)	13.6	125.9
0.5	(3.5; 69.6)	18.7	133.0
0.4	(4.8; 104.1)	27.7	152.0
0.3	(7.5; 177.0)	46.6	229.2
0.2	(16.2; 405.4)	101.0	130.0
0.1	(54.0; 1512.0)	407.0	184.0
0.01	($0.6\ 10^4$; $14.5\ 10^4$)	$3.9 \cdot 10^4$	119.0

When the α-parameter decreases, a near-collinearity arises due to the first two and the last three regressors, in which the angular coefficients are the same. As the α-parameter increases, the near-collinearity level decreases due to a growth in the incoherent components of regressors.

When the α-parameter is equal to 3, the regressors practically do not correlate with each other and their VIF-factor is close to 1.

In this case, the regressors behave like the orthogonal ones. On the other hand, when the alpha parameter value is equal to 0.01, the regressors become near-collinear. In this case, the VIF-factor is about 10^4.

What should be noted first of all is that for a sample of any size the values of the VIF-factor vary considerably from sample to sample.

For small samples (n = 10, Table 1), the case of α = 3 really corresponds (as we will see later) to the absence of near-collinearity, according to the estimate in [24] and our investigations of mutual correlations in artificial data (ADP). However, in this case, the VIF-factor can vary within fairly wide limits from 1.4 to 11.1.

On the other hand, the case of VIF = 10 from this interval, for instance, can be also realized even at much smaller alphas up to α = 0.3, when, as we will see later, the near-collinearity can no longer be considered unimportant.

Thus, for α = 0.3, the 95 % confidence interval is (7.5; 177.0), wherefrom we can see that the probability of finding the VIF = 10 in the interval (7.5; 10) is significantly less than the probability of finding the VIF in the interval (10; 177.0). This probability is small (~ 0.015), but it is not equal to zero.

This means, from the diagnostic point of view, that the VIF-value obtained in an experiment, e.g., VIF = 10, can correspond both to the case of the absence of near-collinearity, and the case of its presence. Clearly, this applies not only to the value VIF = 10. If we obtain in an experiment, e.g., VIF = 5, then it could happen with α from 3 to 0.4. For these α, the variability of the VIF-factor is of about (90-150) %.

Thus, a very large variability of the VIF-factor of small samples does not allow us to speak about the existence of some specific critical value, which determines the existence or absence the near-collinearity of regressors in the limited population.

Note that the MOLS solutions coincide with good accuracy with the OLS solutions for large values of the α-parameter, i.e., for small VIFs (~ 1) for samples of any size.

Summing up the above considerations, we can state that using the ADP and the tables of correspondence between the VIF and α-parameter, similar to Table 1, one can completely determine the range of applicability and, most importantly, the inapplicability of the OLS, which is determined by its possible inadequacy.

Investigation of the OLS solutions for adequacy. Before checking the solutions to the linear regression problem for adequacy, we will discuss the criterion of adequacy of a solution. As mentioned above, adequacy of a solution, in our opinion, can be determined by the smallness of the variability level of the regression coefficients.

Then, the adequacy of the solution to the regression problem can be determined by setting the level of the coefficient of variation (CV) of the regression coefficients. In the present work, for this purpose, the 10 % level of the CV is used, although it is clear that this level may vary depending on the practical problem being solved.

It should also be added that for the fruitful application of the results of regression analysis in the economy, besides the adequacy of solving the regression problem, a sufficiently high value of the coefficient of determination is also required [26].

The commonly used condition is $R^2 \geq 0.8$. However, it should be noted that R^2 also changes from sample to sample and, therefore, it is necessary to estimate the coefficient of variation of R^2 .

It is clear that we cannot calculate the coefficient of variation of R^2 for only one sample. But this one can be done approximately by finding out, using tables similar to Table 1, what value of the ADP parameter α corresponds to the observed sample VIF-factor for the given n.

This study also has shown that it is necessary to make a decision on the significance of regression coefficients, using the observed value of t-statistic, with caution, since its coefficient of variation may be unacceptably large.

The 1st and 2nd problems of the OLS. Since the OLS is essentially the main method for solving the linear regression problems in practice, consider in detail two aspects of its solutions.

First, (the 1st problem), we consider at which sample sizes the OLS solution gives, with a given probability, physically correct solutions, i.e., solutions with correct signs. Assuming the law of distribution of the regression coefficients is normal and using a 95 % confidence interval for a regression coefficient, it is easy to obtain that the condition of the positivity of the regression coefficients is satisfied, with a probability of 95 %, if CV < 50% for each regression coefficient.

Really, using a 95 % confidence interval for a regression coefficient

$$P(m_b - 2s_b < b < m_b + 2s_b) = 0.95, \tag{2}$$

we can write down (2) via the coefficient of variation $CV = s_b / m_b$:

$$P(1 - 2CV < b / m_b < 1 + 2CV) = 0.95. \tag{3}$$

We can see from (3) that a regression coefficient will be positive with a probability of 0.95 % if CV = 0.5. For this, the error in estimating the regression coefficient is 100 %:

$$b = m_b \pm m_b. \tag{4}$$

Here, m_b and s_b are estimates of expectation and standard deviation in the limited population (all samples of size n).

If we want to estimate the regression coefficients more accurately, e.g., with an accuracy of up to 20 %, then it is necessary that the CV does not exceed 10 %:

$$CV = 0.1 \Rightarrow b = m_b \pm 0.2 m_b. \tag{5}$$

Second, (the 2nd problem), we consider at which sample sizes the OLS solution gives, with a given probability, adequate solutions, i.e., solutions, which correctly reflects the relationships between the coefficients of regression in the population. In this study, we believe this is the case if $CV \leq 10$ %. Although, depending on the economic problem being solved, a 20 % error in estimating the regression coefficients may be too large.

These two problems to research, we consider the OLS solutions for different sample sizes with different levels of regressors' near-collinearity, namely, according to above considerations: α = 3 (no collinearity, VIF ~ 1), α = 0.5 (weak collinearity, VIF ~ 10), α= 0.1 (medium collinearity, VIF ~ 100), α = 0.01 (strong collinearity, VIF ~ 10^4).

Tables 2-3 show the solutions of the 5-factor linear regression model obtained by the OLS for data taken from the *DS*5(*n*, α) population for different values of *n* and α. With such values of the *DS*5(*n*, α) parameters, the first two and last three regression coefficients have to be equal in the *n*-size limited population; their variances and coefficients of variation have to be the same inside each group but different between groups; the regression coefficients in the first group are more in absolute value.

Absence of collinearity (VIF~1). Let us now consider the results of solving a linear regression problem using the OLS method given in Table 2 for the case of absence of the near-collinearity between the regressors (VIF~1, α = 3) for different values of the sample size (*n*). Samples was drawn from *DS*5(*n*, α) with parameters β_j = {1, 1, 5, 5, 5} and d_j = {1, 1, 1, 1, 1}.

Table 2

OLS solutions under no collinearity (α = 3, VIF~1)

One solution	b_0	b_1	b_2	b_3	b_4	b_5
n =10^5, VIF=1.03	9.6377	4.1122	4.0367	0.8202	0.8142	0.8240
Theoretical *t*	412.74	88.26	86.44	87.87	87.34	88.30
n =10^6, VIF=1.03	9.6472	4.0767	4.1042	0.8177	0.8181	0.8116
Theoretical *t*	1307.79	275.86	278.08	277.44	277.63	275.58
Simulation, $n=10$; $CI_{VIF.0,95}=(1.41;11.45)$; $t_c=2.77$						
$mean\, b_j$	9.5944	4.1536	4.0491	0.8332	0.8193	0.8251
CV_{b_j} ,%	40.49	189.51	194.01	188.62	186.71	186.50
$\overline{R}^2=0.69,\ CV_{R^2}=26.8$ %; $\overline{F}=3.99,\ CV_F=199$ %; $F_{0.05}=6.26.$ $\overline{t}\cong 0.5;\ CV_t\cong 210\%$						
Simulation, $n=10$; $CI_{VIF.0,95}=(1.07;1.63)$; $t_c=2.03$						
$mean\, b_j$	9.6319	4.0994	4.0414	0.8288	0.8278	0.8079
CV_{b_j} ,%	13.13	62.05	64.14	61.65	61.67	62.68
$\overline{R}^2=0.43,\ CV_{R^2}=26.8$ %; $\overline{F}=5.65,\ CV_F=49\%$; $F_{0.05}=2.49,\ \overline{t}\cong 1.6;\ CV_t\cong 62\%$						
Simulation, $n=60$; $CI_{VIF.0,95}=(1.06;1.41)$; $t_c=2.01$						
$mean\, b_j$	9.6406	4.1335	4.1030	0.8141	0.8132	0.8143
CV_{b_j} ,%	10.5236	49.4922	50.0838	49.6141	49.3576	49.9902
$\overline{R}^2=0.40,\ CV_{R^2}=24.0$ %; $\overline{F}=7.84,\ CV_F=42$ %; $F_{0.05}=2.39,\ \overline{t}\cong 2.0;\ CV_t\cong 50$ %						
Simulation, $n=100$; $CI_{VIF.0,95}=(1.0417;1.2630)$; $t_c=1.99$						
$mean\, b_j$	9.6483	4.0762	4.0855	0.8194	0.8171	0.8132
CV_{b_j} ,%	7.84	37.35	37.48	37.29	37.91	37.41
$\overline{R}^2=0.38,\ CV_{R^2}=19.8$ %; $\overline{F}=12.25,\ CV_F=32$ %; $F_{0.05}=2.49,\ \overline{t}\cong 2.7;\ CV_t\cong 37$ %						
Simulation, $n=1000$; $CI_{VIF.0,95}=(1.03;1.07)$; $t_c=1.96$						
$mean\, b_j$	9.6441	4.0952	4.0848	0.8166	0.8160	0.8168
CV_{b_j} ,%	2.43	11.37	11.50	11.44	11.51	11.41
$\overline{R}^2=0.38,\ CV_{R^2}=19.8$ %; $\overline{F}=112.17,\ CV_F=11\%$; $F_{0.05}=2.22\ \overline{t}\cong 8.7;\ CV_t\cong 11\%$						

Due to the consistency property of the OLS solutions, the coefficients of the regression of a sample tend in probability to regression coefficients in the population if $n \to \infty$. Similarly, for cross-sectional data drawn from an n-size limited population, the average from sample to sample value of the regression coefficients also tends in probability to the regression coefficients in the whole population with the number of repetitions $M \to \infty$.

This property is shown in Table 2, the first two lines of which represent usual OLS solutions for large samples, namely $n = 10^5$ and $n = 10^6$. We see that the solutions for $n = 10^5$ and $n = 10^6$ are statistically identical and any of them can be taken as a solution in the population. On the other hand, the averaging of the multiple repeated OLS solutions for a sample size of $n = 10$ (the number of repetitions is $M = 10^4$) also leads to a statistically close result with the first two rows.

Thus, we can get an estimate of the solution (regression coefficients) in the population $DS5(n, \alpha)$ for given n and α using the OLS solution either with samples of large size or by repeating samples of the same size (n)many times.

To get, however, the dispersion of the regression coefficients (b_j) and calculate their coefficients of variation (CV_j), as well as the average values and coefficients of variation of the regression parameters (R^2 – coefficient of determination, F – Fisher's statistic, t – Student's statistic and others, if necessary, including the VIF) for samples of a given size (n), it is necessary to use many times ($M = 10^4$ in our investigation) data generation with the aid of the algorithm $DS5(n, \alpha)$. The results of such calculations for the case of the absence of near-collinearity ($\alpha = 3$, VIF ~ 1) for different sample sizes are shown in Table 2.

Analysis of the results shown in Table 2 can be summarized as follows:

A. The mean values of the solutions (regression coefficients) of the regression problem with M-times resampling from $DS5(n, \alpha)$ for a given value of parameters (n, α) coincide in probability with the solution for one very large sample drawn from $DS5(n, \alpha)$ for $n = 10^6$. This means that we can determine with any accuracy the regression coefficients in a limited population with any parameters with the aid of the same algorithm $DS5(n, \alpha)$.

B. In the absence of a near-collinearity (VIF~1), the OLS solution gives physically correct solutions ($CV < 50$ %) for samples only larger than 60. For these cases, with a probability of 95 %, all regression coefficients will be positive and can be Student's significant but maybe not adequate. As can be seen from the calculations in Table 2, the OLS solutions become adequate (and statistically significant) starting with sample size more than 1000. It should be added that with small samples ($n < 60$), with a probability of 95 %, the researcher will not receive even just a physically correct solution, i.e., a solution with correct signs.

However, if we look at the values of the coefficient of determination (R^2), we will see that its value increases with sample size decreasing and becomes quite acceptable for $n \le 10$. If, at the same time, in the experiment, randomly, all solutions of the OLS will have the correct signs and will be significant, then the researcher may mistakenly consider the solution to the regression problem to be economically correct.

C. With a sample size increase, the coefficient of determination decreases in average. As can be seen from Table 2, for $n = 10$ the coefficient of determination is of a moderate effect size [26], ($\bar{R}^2 = 0.69$, $CV_{R^2} = 26.8$ %), i.e., more or less acceptable. Already for n = 40, it is of a low effect size [26] ($\bar{R}^2 = 0.43$, $CV_{R^2} = 26.8$ %), i.e., unacceptable from an economic point of view. The same issue holds for larger samples, which means that in the absence of near-collinearity, for whatever size of the sample, the solution to the linear regression problem cannot be useful in economic analysis (a very small coefficient of determination indicates a small regular effect of regressors on the response).

It should be noted that a decrease in the coefficient of determination with an increase in the sample size is not related to the solution method but is only due to the presence of large non-coherent noise in the regressors.

Summarizing the above results, we can draw the following inference: solving the linear regression problem with non-correlating stochastic regressors does not have an economic sense, no matter what method we use.

This allows to make another conclusion: a real linear regression problem under near-collinearity should not be reduced to no-correlated regressors by discarding a part of strongly correlated regressors without an economic necessity.

Medium collinearity (VIF~100, $\alpha = 0.1$). Let us further consider what happens to the OLS solution with an increase in the level of near-collinearity. In Table 3, we consider the properties of the least squares solution with a medium level of near-collinearity ($\alpha = 0.1$, VIF~100) using the same parameters of $DS5(n, \alpha)$.

From the calculations given in Table 3 we can draw the following conclusions:

A. With a decrease in incoherent noise in the regressors (with increasing the VIF-factor) under the same economic laws (the same β_j angular coefficients), the influence of the regressors on the response increases (the regression coefficients b_j increase in the population).

B. With an increase in the collinearity level, the possibility of obtaining an adequate solution to the linear regression problem by the OLS method opens up. We see from Table 3 that with a VIF ~ 100 the OLS solutions are adequate starting with the sample size of 400.

C. The coefficient of determination (R^2) remains high (~ 0.95) for all sample sizes. Such a situation with the coefficient of determination opens up the possibility of obtaining an economically adequate solution to the regression problem by the OLS method using a sample size larger than 400. In this case, the value of t-statistics and its variability for all regression coefficients is also quite acceptable.

Considering the above results, we can draw the following inference: a solution to the linear regression problem for weak-correlating stochastic regressors can be adequate and have an economic sense when using samples larger than ~ 400.

In the range of sample sizes from ~ 23 to ~ 400, an OLS solution may have correct signs and be statistically significant but inadequate. Such a solution may not correspond to the relationship between regression coefficients in the population and some of the solution's components may be insignificant.

For samples smaller than 23, the OLS method is likely to give a physically incorrect solution, i.e., a solution with wrong signs.

Table 3

OLS solutions under medium collinearity ($\alpha = 0.1$, VIF~100)

One solution	b_0	b_1	b_2	b_3	b_4	b_5
$n = 10^6$, VIF=1.03	0.0246	11.3827	11.4061	2.2942	2.2615	2.3089
Theoretical *t*	11.46	156.32	157.28	157.87	155.57	159.15
Simulation, $n=10; \quad CI_{VIF.0,95}=(73;1730); \quad t_c=2.77$						
$mean\, b_j$	0.0291	11.3771	11.6084	2.2865	2.2456	2.2901
CV_{b_j} ,% $\cdot 10^{-3}$	1.3334	0.1175	0.1137	0.1166	0.1156	0.1162
$\overline{R}^2=0.9989,\ CV_{R^2}=0.12$ %; $\overline{F}=1832,\ CV_F=223$ %; $F_{0.05}=6.26,\ \overline{t}\cong 0.9;\ CV_t\cong 115$ %						
Simulation, $n=23; \quad CI_{VIF.0,95}=(64;327); \quad t_c=2.11$						
$mean\, b_j$	0.0298	11.4726	11.2944	2.2905	2.2782	2.2960
CV_{b_j} ,%	565.49	50.36	51.15	50.19	49.50	50.61
$\overline{R}^2=0.9983,\ CV_{R^2}=0.08$ %; $\overline{F}=2497,\ CV_F=53$ %; $F_{0.05}=2.81,\ \overline{t}\cong 2,0;\ CV_t\cong 50$ %						
Simulation, $n=400; \quad CI_{VIF.0,95}=(75;106); \quad t_c=1.97$						
$mean\, b_j$	0.0299	11.4132	11.4358	2.2801	2.2870	2.2815
CV_{b_j} ,%	114.16	10.19	10.03	10.02	10.10	10.03
$\overline{R}^2=0.9980,\ CV_{R^2}=0.02$ %; $\overline{F}=4\cdot 10^4,\ CV_F=10$ %; $F_{0.05}=2.24,\ \overline{t}\cong 10;\ CV_t\cong 10$ %						

In addition, it is necessary to take into account the fact that there already exists a method for solving a linear regression problem adequately under any degree of near-collinearity of the regressors (see [1]).

Conclusions. Summing up the study of the applicability of the OLS in economic research, we can note the following.

A. In the paper, a new algorithm for modeling data (ADP), which constitute a population of limited size with an adjustable level of near-collinearity of the regressors and their influence on the response, is used. This algorithm does not use predefined regression coefficients in the population. This one makes it possible to correctly simulate a multiple regression of any dimension and near-collinearity level. This issue fundamentally distinguishes the ADP from the standard method [27], which, as shown in the article, cannot be applied for simulating multiple regression problems at all.

B. The mathematical and economic correctness of the data modeling algorithm (ADP) has been justified. This modeling takes into account a regular influence on the response of the regressors and not a regular but coherent influence, which is a consequence of economic laws, as well as a random (incoherent) noise in regressors, which is a consequence of the influence on the regressors of random factors.

C. With the help of the ADP, the variability of OLS solutions (CV_{b_j}) is investigated depending on the sample size and the level of near-collinearity of the data (VIF), as well as the variability of the VIF itself and the most important characteristics of the regression problem: the coefficient of determination (R^2), *t*- and *F*-statistic. High variability of these parameters, especially the VIF, has been found.

D. Due to the high variability of the VIF, it is concluded that there is no critical value for this parameter, which divides the data into multicollinear and non-multicollinear ones.

E. Due to the fact that the VIF value found from the results of observations can vary greatly from sample to sample, a qualitative scale of the level of collinearity of data is proposed, namely: "no collinearity", VIF ~ 1; "weak collinearity", VIF ~ 10; "medium collinearity", VIF ~ 100 and "strong collinearity", VIF ~ 10^4. These values of the VIF-factor correspond approximately to the following values of the α-parameter in the ADP algorithm: 3; 0.5; 0.1 and 0.01.

The tables like Tables 1-3 for the given sample size, allows determining to which of these four cases the observed data are relating and to approximately estimate, using the ADP with corresponding α-parameter, the statistical characteristics of the population, from which, presumably, data were extracted.

F. A *qualitative scale* of the level of conformity of a mathematical solution to a linear regression problem to its economic meaning is proposed: a solution is physically incorrect (not all signs of the solution are correct); a solution is physically correct but not adequate; a solution is adequate; a solution is economically adequate.

G. A *quantitative scale* of the level of conformity of a mathematical solution to a linear regression problem to its economic meaning is proposed: a solution is physically incorrect with a probability of 0.95 if the coefficient of variation of the solution is more than 50 %; a solution is physically correct but not adequate (with the same probability) if the coefficient of variation of the solution is less than 50 % but greater than 10 %; a solution is adequate (with the same probability) if the coefficient of variation of the solution is less than 10 % (solution error is less than 20%); a solution is economically adequate if it is adequate and $R^2 \geq 0.8$.

H. The variability of the OLS solution to the 5-factors regression problem in the absence of data collinearity (VIF ~ 1, α = 3) is investigated. It is shown that in this case, solutions to the regression problem with any sample size cannot be used in economic studies either due to a large CV of the solution (for small samples) or due to a small R^2 (for large samples). Thus, in some cases, the OLS solution can be physically correct and even adequate but have a small R^2, i.e., to be economically inadequate.

I. It is noted that with an increase in the near-collinearity level it becomes possible to correctly use the OLS solution in the economy. The solutions become economically adequate, starting with the sample size of ~ 400. With sample sizes from ~23 to ~ 400, an OLS-solution may be physically correct and significant but not adequate, which means that the solution may be far from the solution in the population.

J. In the case of a "strong" near-collinearity (VIF ~ 10^4, α = 0.01), an OLS solution and its properties practically do not differ from the case of the "medium" near-collinearity (VIF ~ 100).

Summing up the results of the study of the 5-factor regression model *DS*5(*n*, α), it can be stated that the OLS is likely to give an inadequate solution for sample sizes smaller than ~ 400.

Physically correct OLS-solutions for samples ranging in size from ~23 to ~ 400 create the illusion of economically correct solutions, but, in fact, the solutions obtained may be far from the solution in the population. For samples smaller than ~ 23, the OLS with a high probability gives a physically incorrect solution (with incorrect signs).

Note that the properties of the OLS solution do not change significantly depending on the number of regressors: the qualitative picture remains the same.

In connection with the foregoing, the authors believe that the conducted research is sufficient to show the necessity of using the MOLS [1] instead of the common OLS, especially because the MOLS only improves its accuracy with the growth of the regressors near-collinearity level and eliminates the need for removing strongly correlated regressors at all.

LITERATURE

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

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

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

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

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

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

7. Baltagi B. Econometrics. NY : Springer, 2011. 812 p.

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

9. Draper, N., Smith H. Applied Regression Analysis. NY : Wiley, 1966. 445 p.

10. Farrar D., Glauber R. Multicollinearity in Regression Analysis: The problem revisited. *Review of Economics and Statistics*. 1967. Vol. 49. P. 92–107.

11. Hoerl, A. E., Kennard R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics.* 1970. No. 12 (1). P. 55–67.

12. Marquardt D. V. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. *Technometrics.* 1970. No. 12. P. 591–612.

13. Blanchard O. J. Comment. *Journal of Business and Economic Statistics.* 1987. No. *5*. P. 449–51.

14. Adkins L. C., Hill R. C. Collinearity. Companion in Theoretical Econometrics, edited by Badi Baltagi. Oxford : Blackwell Publishers, Ltd., 2001. P. 256–278.

15. Belsley D. A., Kun E., Welsh R. T. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. NY : Wiley, 2004. P. 651.

16. Belsley D. A. Demeaning conditioning diagnostics through centering. *The American Statistician.* 1984. Vol. 38 (2). P. 73–77.

17. Rao C., Toutenberg H. Linear Models: Least Squares and Alternatives, 2nd ed. NY : Springer, 1999. P. 301.

18. Spanos A., McGuirk A. The Problem of Near-Multicollinearity Revisited: erratic vs. systematic volatility. *Journal of Econometrics.* 2002. Vol. 108. P. 365–393.

19. Adkins L. C., Waters M. S., Hill R. C. Collinearity Diagnostics in gretl // Economics Working Paper Series 1506. Oklahoma : Oklahoma State University, Department of Economics and Legal Studies in Business, 2015. 452 p.

20. Tikhonov A. N. On the stability of inverse problems // Doklady Acad. Sci. USSR. 1943. Vol. 39. P. 176–179.

21. Tikhonov A. N., Arsenin, V. Y. Solutions of Ill-Posed Problems. NY : Winston & Sons, 1977. 287 p.

22. Kabanichin S. I. Definitions and Examples of Inverse and Ill-posed Problems. *J. Inv. Ill-Posed Problems.* 2008. Vol. 16. P. 317–357.

23. Harvey A. C. Some Comments on Multicollinearity in Regression. *Applied Statistics*. 1977. Vol. 26 (2). P. 188–191.

24. Kutner, M. H., Nachtsheim C. J., Neter J. Applied Linear Regression Models. NY : McGraw-Hill / Irwin, 2004. P. 701.

25. Sheather S. J. A modern approach to regression with R. NY : Springer, 2009. P. 393.

26. Moore D. S., Notz W. I., Flinger M. A. The basic practice of statistics. NY : W. H. Freeman and Company, 2013. P. 138.

27. Dougherty C. Introduction to Econometrics. NY : Oxford University Press, 1992. 402 p.

28. Maddalla G. S. Introduction to Economics. NY : Macmillan, 1992. 396 p.

29. Kennedy P. E. Oh no! I got the wrong sign! What should I do? *The Journal of Economic Education.* 2005. Vol. 36. No. 1. P. 77–92.

REFERENCES

Adkins, L. C., and Hill, R. C. "Collinearity". In *Companion in Theoretical Econometrics*, 256-278. Oxford: Blackwell Publishers, Ltd., 2001.

Adkins, L. C., Waters, M. S., and Hill, R. C. "Collinearity Diagnostics in gretl". In *Economics Working Paper Series 1506*. Oklahoma: Oklahoma State University, Department of Economics and Legal Studies in Business, 2015.

Baltagi, B. *Econometrics*. New York: Springer, 2011.

Belsley, D. A. "Demeaning conditioning diagnostics through centering". *The American Statistician*, vol. 38 (2) (1984): 73-77.

Belsley, D. A., Kun, E., and Welsh, R. T. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 2004.

Blanchard, O. J. "Comment". *Journal of Business and Economic Statistics*, no. 5 (1987): 449-451.

Dougherty, C. *Introduction to Econometrics*. New York: Oxford University Press, 1992.

Draper, N., and Smith, H. *Applied Regression Analysis*. New York: Wiley, 1966.

Farrar, D., and Glauber, R. "Multicollinearity in Regression Analysis: The problem revisited". *Review of Economics and Statistics*, vol. 49 (1967): 92-107.

Greene, W. H. *Econometric Analysis*. New York: Pearson, 2012.

Gujarati, D. N. *Basic econometrics*. New York: McGraw-Hill, 2002.

Harvey, A. C. "Some Comments on Multicollinearity in Regression". *Applied Statistics*, vol. 26 (2) (1977): 188-191.

Hoerl, A. E., and Kennard, R. W. "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics*, no. 12 (1) (1970): 55-67.

Kabanichin, S. I. "Definitions and Examples of Inverse and Ill-posed Problems". *J. Inv. Ill-Posed Problems*, vol. 16 (2008): 317-357.

Kennedy, P. E. "Oh no! I got the wrong sign! What should I do?" *The Journal of Economic Education*, vol. 36, no. 1 (2005): 77-92.

Kutner, M. H., Nachtsheim, C. J., and Neter, J. *Applied Linear Regression Models*. New York: McGraw-Hill / Irwin, 2004.

Maddalla, G. S. *Introduction to Economics*. New York: Macmillan, 1992.

Marquardt, D. V. "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation". *Technometrics*, no. 12 (1970): 591-612.

Moore, D. S., Notz, W. I., and Flinger, M. A. *The basic practice of statistics*. New York: W. H. Freeman and Company, 2013.

Rao, C., and Toutenberg, H. *Linear Models: Least Squares and Alternatives*. New York: Springer, 1999.

Seber, G. A. F. *Linear Regression Analysis*. New York: Wiley-Blackwell, 1977.

Seber, G. A. F., and Lee, A. J. *Linear Regression Analysis*. New York: Wiley, 2003.

Sheather, S. J. *A modern approach to regression with R*. New York: Springer, 2009.

Spanos, A. *Probability Theory and Statistical Inference: econometric modeling with observational data*. Cambridge: Cambridge University Press, 1999.

Spanos, A., and McGuirk, A. "The Problem of Near-Multicollinearity Revisited: erratic vs. systematic volatility". *Journal of Econometrics*, vol. 108 (2002): 365-393.

Tikhonov, A. N. "On the stability of inverse problems". In *Doklady Acad. Sci. USSR*, vol. 39 (1943): 176-179.

Tikhonov, A. N., and Arsenin, V. Y. *Solutions of Ill-Posed Problems*. New York: Winston & Sons, 1977.

Tyzhnenko, A. G. "A new stable solution to the linear regression problem under multicollinearity". Economics of Development. 2018. http://www.ed.ksue.edu.ua/ER/knt/ee182_86/e182tyz.pdf

Wooldridge, J. M. *Introductory Econometrics: Modern Approach*. Ohio: South-Western, 2009.