JEL Classification: C52 УДК 330.43 (075.8)

A NEW STABLE SOLUTION TO THE LINEAR REGRESSION PROBLEM UNDER MULTICOLLINEARITY

Tyzhnenko A. G.

Abstract. The main shortcomings of the OLS solution to the linear regression problem under multicollinearity, which prevent from obtaining an adequate contribution of each regressor to the regressand, have been considered.

It is shown that the main cause of a common incorrectness regarding the economic aspect of OLS solutions is their great variability in the presence of data multicollinearity.

It is also shown that mathematically correct OLS-solutions can become economically incorrect with data collinearity increasing which leads to a diminishing of the physically correct codomain of the OLS-matrix.

Existing in the literature methods to overcome the OLS-solutions great variability are considered both from economical and mathematical aspects. The considering provided shows with confidence the impossibility of existing methods to overcome the data multicollinearity problems both from mathematical and economic considerations such as choosing the best regressions, "lasso" and so on.

The detailed analysis of the situation with multicollinearity provided in the paper allows concluding that the only way out of this situation is to create a new method solution of the OLS-equation which should give a stable solution with small variability, as in the ridge-method, and small bias. Precisely such method is the Modified OLS (MOLS) which is proposed in the paper.

The MOLS is an approximate method which uses the known Tikhonov's regularization principle and a new method solution to the regularized OLS-equation, which is based on the modified Cramer's rule, which is proposed in the paper.

It is 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 method gives a negligible bias and does not require the optimization of the regularization constant.

The proposed MOLS method is verified in the paper for adequacy with the aid of the artificial data population (ADP), which is based on the Monte Carlo simulation method. Using the ADP, the new MOLS method is checked for the biasedness and stability for both small and large samples.

Key words: multicollinearity, stable solution, almost unbiasedness, mathematical correctness, physical correctness, ridge-regression

НОВИЙ МЕТОД СТАБІЛЬНОГО РІШЕННЯ ЗАДАЧІ ЛІНІЙНОЇ РЕГРЕСІЇ В УМОВАХ МУЛЬТІКОЛІНЕАРНОСТІ

Тижненко О. Г.

Анотація. Розглянуто існуючі проблеми рішення багатофакторної задачі лінійної регресії за наявністю мультиколінеарності методом найменших квадратів (МНК), які не дозволяють отримати адекватне рішення економічної проблеми оцінки впливу кожного окремого регресора на відгук.

Виявлені причини появи некоректних рішень економічної задачі регресії математичним методом найменших квадратів, які пов’язані з великою варіабельністю МНК-рішення при значної колінеарності даних.

Показано, що некоректні з точки зору економіки математичні рішення стандартного МНК виникають при збільшенні рівня колінеарності даних за рахунок зменшення області фізичної коректності МНК-матриці.

Розглянуто існуючі на сьогодення методи подолання великої варіабельності МНК-рішень як з економічної, так і з математичної точки зору. Проведений розгляд з очевидністю показує неспроможність існуючих на сьогодні методів подолання проблеми мультиколінеарності як з боку

математиці, так і з боку економічного розгляду спрощення самої економічної проблеми: вибір найкращих регресій, lasso і т. і.

Проведений аналіз дозволив зробити висновок, що єдиним виходом з існуючій ситуації є створення нових методів розв'язку МНК-рівняння які б давали рішення з малою варіабельністю, як в ridge-методі, наприклад, але з малим зміщенням. Саме таким методом є новий модифікований метод найменших квадратів (ММНК), який є представленим в роботі.

ММНК є наближеним методом, в якому є використаним метод регуляризації Тіхонова і новий метод рішення регуляризованого МНК-рівняння, заснований на модифікованому методі Крамера, який запропонован в статті.

Показано, що ММНК дає стійке та практично незміщене рішення задачі лінійної регресії при будь-якому рівні колінеарності даних. На відміну від методу ridge-регресії, ММНК не потребує оптимізації константи регуляризації.

Запропонований в роботі ММНК перевіряється на адекватність за допомогою штучної генеральної сукупності, яка створена за допомогою методу Монте-Карло. З використанням цієї генеральної сукупності в роботі показана як практична незміщеність ММНК, так і висока стабільність рішень задачі регресії як для великих, так і для малих вибірок.

Ключові слова: мультиколінеарність, стабільне рішення, майже незміщеність, математична коректність, фізична коректність, рідж-регресія

НОВИЙ МЕТОД СТАБИЛЬНОГО РЕШЕНИЯ ЗАДАЧИ ЛИНЕЙНОЙ РЕГРЕССИИ В УСЛОВИЯХ МУЛЬТИКОЛЛИНЕАРНОСТИ

Тыжненко А. Г.

Аннотация. Рассмотрены существующие проблемы решения многофакторной задачи линейной регрессии в условиях мультиколлинеарности методом наименьших квадратов (МНК), которые не

позволяеют получить адекватное решение экономической проблемы оценки влияния каждого отдельного регрессора на отклик.

Выявлены причины некорректного решения экономической задачи регрессии математическим методом наименьших квадратов. Эти причины связаны с большой вариабельностью МНК-решений при значительной коллинеарности данных. Показано, что некорректные, с точки зрения экономики, математические решения стандартного МНК возникают при увеличении уровня коллинеарности данных за счет уменьшения области физической корректности МНК-матрицы.

Рассмотрены существующие на сегодняшний день методы борьбы с большой вариабельностью МНК-решений как с экономической, так и с математической точек зрения. Проведенное рассмотрение убедительно показывает недееспособность существующих на сегодняшний день методов преодоления мультиколлинеарности как со стороны математики, так и со стороны экономического рассмотрения способов упрощения самой экономической проблемы: выбор наилучших регрессий, lasso, и т.д.

Проведенный анализ позволяет сделать вывод о том, что единственным выходом из существующей ситуации есть создание новых методов решения МНК-уравнения, которые давали бы решения с малой вариабельностью, как в ridge-методе, например, но с малым смещением. Именно таким методом является новый модифицированный метод наименьших квадратов (ММНК), который представлен в работе.

ММНК является приближенным методом, в котором использован метод регуляризации Тихонова и новый метод решения регуляризованого МНК-уравнения, основанный на модифицированном методе Крамера, который предложен в статье.

Показано, что ММНК дает устойчивое и практически несмещенное решение задачи линейной регрессии при любом уровне коллинеарности данных. В отличие от метода ridge-регрессии, ММНК не требует оптимизации константы регуляризации.

Предложенный в работе ММНК, проверяется на адекватность с помощью искусственной генеральной совокупности, созданной с помощью метода Монте-Карло. С использованием этой генеральной совокупности, в работе показана как практическая несмещенность ММНК, так и высокая стабильность решений задачи линейной регрессии, как для больших, так и для малых выборок.

Ключевые слова: мультиколлинеарность, стабильное решение, практическая несмещенность, математическая корректность, физическая корректность, ридж-регрессия

Preamble

An economic insight into the multiple linear regression solutions can be figured out as the obtaining of significant estimates of regression coefficients that represent the mean change in the response variable for the unit changing in the predictor variable while holding other predictors in the model constant.

It is clear from the economical point of view that the mathematical solution to the linear regression problem must be stable and the regression coefficients obtained must have the same signs that the partial regression coefficients between the regressand and regressors have. It is known that this is frequently not the case if the regression problem is solved with the aid of the common OLS method. In this paper a new method solution to the linear regression problem is proposed in which a common great instability of the OLS is overcome.

This problem is considered in the paper under the following assumptions: the residual error is normal, $\varepsilon \sim \mathbb{N}(0, \sigma\mathbf{I})$; the relationships between variables are linear in the population; all assumptions of the Gauss-Markov theorem are fulfilled; non-stochastic regressors are considered.

From a mathematical point of view, the linear regression problem is formulated as the curve fitting problem [1-7]. The OLS method of solving the linear regression problem can give an adequate solution of economic regression problem if the regressors are near-orthogonal. Unfortunately, this is not the case in practice.

The main drawback that prevents the OLS solution from being adequate to an economic problem is the near-collinearity of regressors [8-19].

In terms of just a curve fitting problem, the OLS always gives mathematically correct result regardless of the regressors' collinearity level (the VIF-factor, for example). However, with the VIF-factor increasing, the variability of the OLS-solution drastically increases for not very large samples. This issue prevents from getting an adequate economic solution to the regression problem in practice.

The data near-collinearity is not the single source of the regression solution errors. Another source of errors is the non-linearity of the population that is investigated. This problem concerns the regression model inadequacy and may, in principle, be eliminated with the aid of appropriate data transformations.

Very important source of errors in regression solutions is the wrong model specification [9, 12], but this problem is connected with economic considerations and is not considered in this paper.

Different remedies have been proposed to dealing with ill-conditioning and near-collinearity including regularization and ridge regression, omitting variables, grouping variables in blocks, collecting additional data and so on; among others see [4, 9 - 21].

However, these remedies may be time-consuming, costly, impossible to achieve or controversial [22]. Also, the diagnostic tools that signal the presence of near-collinearity are crucial. More than that, the author agrees with [23] that any signal of multicollinearity does not exist at all because “multicollinearity is a matter of degree rather than one of kind.”

Despite the theoretical warnings about the inadmissibility of using OLS in the presence of near-collinearity of any level, this technique is still in use in practice, in economic and other studies, with attempts to reduce somehow the level of collinearity. Many years of efforts did not yield any results in the search for a critical level of near-collinearity. It seems that A. C. Harvey in [23] was right that there is no such a critical level at all and the influence of near-collinearity at any OLS solution is a continuous process which depends on many parameters. This issue is also confirmed by the following further considerations of OLS-solutions properties.

In reality, with a sample size decreasing, the presence of the data near-collinearity usually leads to an unacceptable increase in the OLS estimates dispersion which makes the OLS solution inadequate in terms of economic content. For instance, signs of the OLS estimates may be incompatible with their economic meanings.

Such a behavior of OLS-solutions immediately follows from the Cramer's rule and the determinant decomposition by matrix eigenvalues. The presence of small eigenvalues in the numerator and the denominator of the Cramer's formula can lead to significant changes in the solution due to random changes in the data observed and then in eigenvalues.

As for the appearance of incorrect signs in the OLS solutions, that is when solutions have no economic (in general, physical) sense, this phenomenon, as shown in the paper, is connected with the fundamental property of nonsingular square matrices.

It has been revealed that any non-singular matrix operator has a codomain that consists of two parts, which we called codomains of physical correctness (D^c) and incorrectness ($\overline{D}^c$) of the corresponding matrix equation solution.

That is, any matrix equation *Ax* = *b* always has a mathematically correct solution but such a solution may be either physically correct or physically incorrect. A solution is physically correct if it has an economic (a physical) meaning. The solution of the same equation is physically incorrect if its solution has no physical meaning. In the latter case, the solution necessarily changes the signs of some solution components. This issue depends on the RHS (b) only. That is, a solution of the matrix equation *Ax* = *b* is physically correct if $b \in D^c$ and is incorrect otherwise.

It has been shown in this paper that this effect holds for any matrix equation with square non-singular matrix. However, with the matrix conditional number increasing, the codomain of physical correctness, D^c, is becoming more narrow. This issue may lead to the RHS of matrix equation being outside of D^c due to random errors in matrix elements. If it is the case, there will be a change in signs of the solution components. This issue is often observed in OLS solutions to the regression problem due to their large variability.

Thus, the main drawback of the OLS method is a narrow codomain of physical correctness in presence of near-collinearity and the high variability of a solution in the case when an observed sample size is not very large. Both these effects promote the exit of the RHS from the D^c under the influence of random errors. This one does not allow finding the suitable estimates of regression coefficients in the population.

The advantage of the OLS is the unbiasedness and consistency of a solution and its variance, i.e. a reduction of the sample regression coefficients variance with a sample size increasing and the approach of the mean value of the OLS solutions to the regression coefficients in the population, which makes it possible, in general, to estimate, with the aid of data modeling, the adequacy of the regression problem solution when this problem is solved by any other method.

Thus, due to the properties of OLS-solutions which are proved theoretically, one can test, in principle, the regression problem solution results obtained by other methods for which the closeness of the estimated coefficients to the population ones cannot be proved theoretically.

As to the influence of near-collinearity on the variability of the OLS solutions, the prior investigations unambiguously show the need to create new methods solution the linear regression problem, which would give a small bias and acceptable solution variability for not very large samples.

Any new method solution to the linear regression problem should give the regression coefficients which have to approach in probability the coefficients of the OLS solution when sample size increases unlimitedly. This issue is a consequence of OLS solutions unbiasedness and consistency, which allows us to test a new method with the aid of data simulation.

The unbiasedness and consistency of OLS solutions is also manifested in the fact that the mean of many times ($M \gg 1$) repeated OLS-solutions for samples of limited size (n) drawn from a population converges in probability to the population solution (regression coefficients) with M increasing. This one is also used in the paper for testing new methods solution to the linear regression problem with the aid of the artificial population (ADP) worked out in the paper for this purpose.

In this paper, a new method (MOLS) is proposed which produces stable solutions with a negligibly small bias to the linear regression problem under near-collinearity of any level for samples of any size.

The MOLS is based on the OLS for standardized variables with some modifications. The OLS matrix equation $X'Xb = X'Y$ is replaced by the regularized equation $(X'X + \alpha I)b = X'Y$ with very small regularization constant $\alpha = 0.001$. This equation is solved further with the aid of modified Cramer's rule which is suggested in the paper.

Unlike the ridge regression, the modified OLS (MOLS) method gives practically zero bias and does not require the regularization constant (α) adjustment for any collinearity level.

To disadvantages of the MOLS can be attributed a large computer loading that prevents applying this method for a large number of regressors (more than 200-300).

The new method' (MOLS) adequacy has been verified with the aid of a special Artificial Data Population (ADP) developed in the paper. The linear regression problem modeling with such a population differs from the standard one [26] in that it does not use a priori giving regression coefficients in a population. Instead of this, the ADP method simulates a population with unknown regression coefficients which values can be precisely estimated by the OLS solution for a very large sample size, using its consistency property.

The essence of the ADP method checking for adequacy consists of a priori giving a regressand vector *Y* and setting the regressor vectors $\{X_j\}$ geometrically with given angles to the regressand vector. With the angle between regressor and regressand vectors diminishing, the absolute value of the corresponding regressor' coefficient should increase. If two regressors, for instance, have the same angle to the regressand, the corresponding regression coefficients in the population should be equal. If two regressor vectors have angles with regressand vector which differ by $\pi/2$, their regression coefficients should have opposite signs and be equal in absolute value.

Then, with such an artificial population in hand, one can construct various situations for the population regression coefficients which allow estimating the adequacy of a new method solution to the linear regression problem. For each modeled situation with population regression coefficients, one has an opportunity for estimating the population regression coefficients with the aid of the asymptotic OLS solution. This one allows estimating both the biasedness and variability of any new linear regression solution method for any sample size.

Creating a new method data simulation (ADP) for testing the linear regression problem solutions is connected with the incorrectness of the conventional simulation method [26] for multiple regression in which for a priori given regressors, $\{X_j\}_m$, and population regression coefficients, $\{b_j\}_m$, one sets many times (M) a random residual error, $\{e\}_n$ for calculating M regressand realizations, $\{Y\}_n$.

The matter is that the given regressors, $\{X_j\}_m$, defines exactly the OLS-matrix $X'X$ of the matrix equation $X'X\, b = X'Y$, which has a definite codomain of physically correctness, D^c. In order to make this matrix equation have a correct solution, the RHS $X'Y$ should belong to this D^c. If we set the population regression coefficients arbitrarily, we make the regressand Y arbitrary as well and so the RHS $X'Y$. A such calculated RHS may not belong to this D^c. This one will lead to a physically incorrect solution to the linear regression problem. The situation may change only if we take as a priori regression coefficients those ones which are close to the population coefficients. However, it is not probable to guess randomly the regression coefficients which are close to the true ones.

It is worth noting again that the developed ADP permits us not only to test the biasedness of a new method solution to the linear regression problem but also to estimate the variability of sample regression coefficients. For this one, we can draw from the ADP a large series of replicas with a given size and calculate the standard deviation of regression coefficients obtained by the OLS and the new method.

From a mathematical point of view, the solution of the equation $Ax = b$ is a vector x that gives a zero discrepancy: $\| Ax - b \| = 0$. Herewith, not raised the question of what a real problem is solved. It is revealed in this paper that such a situation is not always admissible. This is the case, for instance, in the basis changing problem in a vector space.

In most real problems connected with the equation $Ax = b$ one has to consider the context of the problem. It has been revealed that a common condition of zero discrepancy does not indicate the correctness of a real problem solution.

Codomain of physical correctness

Definition: any *n*-dimensional nonsingular square matrix A over the reals has the R^n as its codomain which consists of two parts, D^c and $\overline{D}^c$ ($D^c \cup \overline{D}^c = R^n$), where D^c is a codomain of physically correct solution of matrix equation $Ax = b$ ($b \in D^c$) and $\overline{D}^c$ is a codomain of physically incorrect solution of matrix equation $Ax = b$ ($b \in \overline{D}^c$).

If $b \in D^c$, the signs of exact solution components of the equation $Ax = b$ are consistent with the signs needed for the real problem which is investigated and the solution is stable for any condition number of matrix A if the random errors of matrix elements do not remove the RHS vector b from D^c.

Otherwise, if $b \in \overline{D}^c$, the signs of exact solution components of the equation $Ax = b$ are not consistent with the signs needed for the real problem that is investigated with this equation.

In general, the exact solution of equation $Ax = b$ changes the signs of some solution components when vector b passes from D^c to $\overline{D}^c$. The absolute values of solution components depend on the orientation of the RHS vector and increase with the matrix A condition number increasing if $b \in \overline{D}^c$.

Both these issues are inappropriate for a real problem that is investigated. More than that, if $b \in \overline{D}^c$, the exact solution of equation $Ax = b$ is unstable in the ill-conditioning case, $cond(A) \gg 1$. Geometrically this one follows from projection properties. If a RHS is outside of D^c, projections of the RHS on the bases vectors increase with the RHS moves away from D^c. The more is the condition number of the matrix, the narrower is D^c. This one enlarges the projection values and, consequently, the absolute values of solution components.

Because of the linear regression problem OLS-solutions are based on the matrix equation solution, in the case of near-collinearity, such solutions may reveal both the appearance of unexpected signs of the regression coefficients and their abnormal absolute values.

The first problem is connected with the exit of the OLS-equation RHS from the codomain of physical correctness (D^c) due to random errors in observed data.

The second problem relates geometrically to narrow D^c and mathematically to a small determinant of the OLS-matrix in presence of random errors in matrix elements.

Geometrical considerations to system solution in 2D

Consider the first problem of unexpected signs of the regression coefficients using the example of two-dimensional full rank matrix equation:

$$Ax = b \Leftrightarrow a_1x_1 + a_2x_2 = b, \tag{1}$$

where

$$a_1 = \binom{a_{11}}{a_{21}}, a_2 = \binom{a_{12}}{a_{22}}, b = \binom{b_1}{b_2}.$$

From geometrical point of view, equation (1) is the coordinate-wise representation of the vector b with respect to the basis $\{a_1, a_2\}$.

Suppose the angle between the basis vectors a_1 and a_2 is significantly smaller than 90°; vector b is located between them; all vectors belong to the 1st quadrant (Fig. 1(1), Appendix A).

Let a small 2D area between vectors (a_1, a_2) and ($-a_1$, $-a_2$) be the matrix A codomain of physical correctness, D^c (Fig. 1(1)). The vectors e_1 and e_2 in Fig. 1(1) are the projects of vector b on basis vectors a_1 and a_2. From this drawing we can see that in this case both equation (1) solution components are positive ($x_1 = |e_1|$, $x_2 = |e_2|$). Seemingly, if for the same basis vectors the RHS has the inverse direction $(-b)$ and is located between $-a_1$ and $-a_2$ vectors, both solution components are negative. In both this cases the solutions components have the same signs.

Another situation is shown in Fig. 1(2), where the RHS vector b is located between vectors $-a_1$ and a_2 (in the wide 2D area that we denoted as $\overline{D}^c$). In this case the solution components have different signs, ($a_1 = -|e_1|, a_2 = |e_2|$). Analogous situation will be if the only basis vector a_2 changes the direction.

This simple example demonstrates the fundamental property of non-singular matrices: a matrix equation $Ax = b$ has fundamentally different solutions for $b \in D^c$ and $b \in \overline{D}^c$. If $\in \overline{D}^c$, the solution of this equation is mathematically correct, but has no physical meaning. Which part of the whole codomain (R^n) is D^c should be determined from economic considerations. Worth noting, that this property is not connected with the conditioning of matrix equation.

Summarizing, we can state that any determined matrix equation $Ax = b$ has both physically correct and physically incorrect solutions depending on the RHS. In both cases solutions are mathematically correct.

Let us demonstrate the existence of the fundamental property of non-singular matrices using the example of a simple economic problem.

2D selling problem

Suppose you are selling hot dogs and sodas. Each hot dog costs \$1.50 and each soda costs \$0.50. At the end of the session you made a total of \$78.50. You sold a total of 87 hot dogs and sodas combined. You must report the number of hot dogs sold and the number of sodas sold. How many hot dogs and sodas were sold separately? Shortly: one hot dog costs \$1.5; one soda costs \$0.5. A common sale \$78.5. A common number of units sold is 87. How many hot dogs (x_1) and sodas (x_2) ware sold separately? System ($Ax = b$):

$$\begin{cases} 1.5x_1 + 0.5x_2 = 78.5 \\ x_1 + x_2 = 87 \end{cases} . \tag{2}$$

Here, det(A)=1; cond(A)=4.27. The system (2) is well-conditioned. The Gauss' solution in the Matlab is: $x = A\backslash b = (35; 52)$. The basis vectors have the following coordinates: $a_1 = (1.5;1), a_2 = (0.5;1)$. Let us write down the RHS as follows: $b = 87(0.9023; 1)$. Then, it is clear that the RHS vector lies between a_1 and a_2 and belongs to the codomain of physical correctness, D^c, since there projections on a_1 and a_2 are positive as it should be from economic considerations.

Let us consider further the system (2) solution behavior with the RHS vector *b* changing if the common sale is fixed, $b(1) =$ \$78.5.

The marginal values of the RHS vector *b* inside the D^c can be determined from the parallel conditions: $b||a_1$ and $b||a_2$. That are: $b_+ = (78.5; 157), b_- = (78.5; 52.\bar{3})$. According to the economic meaning of the problem, we take $b_-(2) = 53$. So, the common number of units sold can vary within [53; 157] in order to $b \in D^c$.

If $b = (78.5; 53)$, the solution is $x = (52; 1)$. This means that 52 hot-dogs and one soda was sold. If $b = (78.5; 157)$, the solution is $x = (0; 157)$. This means that no hot-dogs but 157 sodas ware sold. It is clear, that inside D^c there are also other RHSs that give the whole solutions. For example, for $b = (78.5; 109)$ we have the solution: $x = (24; 109)$.

In any other practical situation, one can also sell a part of the unit and then the solutions do not have to be whole numbers. In general case, we can find a solution of such an equation in real numbers.

Suppose the right-hand side vector *b* does not belong to the D^c. It is the case if $b(2) < 52.\bar{3}$ or $b(2) > 157$. For example, let the common number of units sold is $b(2) = 51$. Then the solution is $x = (53; -2)$. This solution is incorrect relative to the investigated problem. Suppose now that $b(2) = 159$. Then, $x = (-1;\ 160)$ and is also incorrect. This means we cannot set arbitrary the RHS of equation (1) if we investigate any practical problem. If we do that we can obtain a solution with wrong signs despite the fact that the system is well conditioned.

This example demonstrates an important property of a linear system solution regarding of its adequacy to the economic problem which is investigated with the aid of such system. If the RHS of a system does not belong to this system' matrix codomain of physical correctness, a mathematically correct solution to this system will be not correct for the practical problem under investigation.

Because of a linear system solution necessarily changes the signs of some solution components when the RHS passes from one codomain to another one, the codomain of physical correctness can be easily determined if a practitioner knows exactly what signs of solution components are correct. This is the case for the regression problem, for example, in which one knows that the regression coefficients must have the same signs as the correspondent partial regression coefficients for the regressand and regressors.

The problem of physical correctness in linear regression

There is no consensus in the economic literature on the discussion of the OLS-solution properties. As an aside, the matrix equation $Ax = b$ with a non-singular square matrix A should always have a unique solution for any $b \in R^n$ from mathematical point of view. As another aside, it is clear that for ill-conditioned matrices (A) the matrix equation ($Ax = b$) solution may be not always true from the point of view of the applied problem that is investigated.

This situation was characterized by [12] as follows: "Multicollinearity is God's will, not a problem with OLS or statistical technique in general." "Only use of more economic theory in the form of additional restrictions may help alleviate the multicollinearity problem." "One should not, however, expect miracles; multicollinearity is likely to prevent the data to speak loudly on some issues, even when all of the resources of economic theory have been exhausted."

Denoted by [12] such a situation with multicollinearity has not changed to date [4, 5, 6, 19, 22], as far as the authors know.

As we can see from the above, the OLS solution problems under multicollinearity are connected with the ill-conditioning of OLS matrix equation $Ax = b$, which matrix, as any non-singular matrix, has its codomains of physical correctness (D^c) and incorrectness ($\overline{D}^c$). The increasing of near-collinearity level in data leads to increasing the condition number of the OLS-matrix and then to a contraction of the physical correctness (D^c) codomain. Besides that, the increasing of ill-conditioning level leads to increase variability of OLS-solutions due to decreasing of the OLS-matrix minimal singular number. Both these issues can drastically spoil the OLS solution to the linear regression problem: the regression coefficients may be too large in values and become of wrong signs if the errors in data remove the RHS of matrix equation from the physical correctness (D^c) codomain, which has become too narrow.

In general, the instability of the OLS solutions to the linear regression problem depends on two parameters only: the VIF-factor and the sample size. With the VIF-factor increasing, the volatility of the OLS solutions increases. With the sample size increasing, the volatility of the OLS solutions decreases. For any value of the VIF-factor one can find such a large sample size that for any sample, the RHS of the OLS matrix equation will belong to the physical correctness codomain (D^c) and the OLS solution will be stable and economically correct. For small and not very large samples this is not the case, as a rule.

It is also worth to note that a physically correct and stable OLS solution may have yet sufficiently large standard deviation, because it is desirable to estimate the standard deviation of the obtained solution by the Monte Carlo simulation. The method of such a kind is proposed in this paper.

Summarizing, we can state that the only fruitful strategy of struggle with the near-collinearity in the linear regression problem is the construction of new methods solution to this problem which would provide stable physically correct solutions with standard deviations much less than the absolute values of solution components (regression coefficients). Besides that, such methods would be negligibly biased. Such a method is proposed in the paper.

Modified OLS method

For obtaining a stable solution to the linear regression problem in the standardized form, we propose here the modified OLS (MOLS) that used the modified Cramer’s rule, which has been invented in the paper, instead of the Gauss’ method solution to algebraic systems.

The MOLS is based on the OLS for standardized variables with some modification: the OLS matrix equation

$$X'Xb = X'Y \tag{3}$$

is multiplied by $(X'X)'$ and replaced by the regularized equation

$$((X'X)'(X'X) + \alpha I)b = (X'X)'X'Y \tag{4}$$

with small α $(\alpha = 0.001)$. This equation is solved further with the aid of the modified Cramer's rule.

The modified Cramer's rule is intended to solve the definite ill-conditioned systems $X'Xb = X'Y$ that arise in the standardized linear regression problems (all variables are standardized). For brevity, let us wright down this system as usual:

$$Ax = b, \tag{5}$$

with $A = X'X$, $x = b$, $b = X'Y$. Multiply further (5) by A':

$$A'Ax = A'b.$$

Let us denote further: $A'A = H_1, A'b = b_1$ and solve the equation

$$H_1 x = b_1. \tag{6}$$

Taking into account a possible ill-conditioning of the matrix H_1, let us reduce the conditioning level by adding a regularizer to H_1. A new matrix let us designate by H:

$$H = H_1 + \alpha E,$$

where E is the identity matrix and $0 < \alpha \ll 1$ (optimal value that gives a minimal RSS is $\alpha = 0.001$). Let us replace the equation (6) by a regularized equation (basic and single approximation):

$$Hx = b_1. \tag{7}$$

According to the Cramer's rule, the solution of this equation can be written as

$$\tilde{x}_j = \frac{\Delta_j}{\Delta}\,, \tag{8}$$

where

$$\Delta_j = \sum_{k=1}^{n} (-1)^{j+k} B_1(k) \det\left(H\left(t_k, t_j\right)\right)\,, \tag{9}$$

$$\Delta = \sum_{k=1}^{n} (-1)^{j+k} H(k, j) \det(H(t_k, t_j))\,, \tag{10}$$

and $t_k = 1, 2, \dots, k-1, k+1, \dots, n$. Here, $H(t_k, t_j)$ is the matrix H, from which the k-th row and j-th column are crossed out, $H(k, j)$ is the (*k*, *j*) element of matrix H. That is, the formulas (8-10) figure out as the common Cramer's rule, in which the Laplace' formula is used.

In (8) we always can multiply the numerator and denominator by any determinant of some nonsingular matrix. As such a matrix we take H_j^{-1} – the inverse of the matrix H_j, where H_j is the matrix H, from which the *j*-th row and *j*-th column are crossed out. For each j we multiply the numerator and denominator in (8) by different determinant $\det\left(H_j^{-1}\right)$. Using the determinant property:

$$\det(AB) = \det(A)\det(B),$$

we can write down the determinants (9, 10) as follows:

$$\tilde{\Delta}_j = \sum_{k=1}^{n} (-1)^{j+k} B_1(k) \det\left(H_j^{-1} H\left(t_k, t_j\right)\right) \tag{11}$$

$$\tilde{\Delta} = \sum_{k=1}^{n} (-1)^{j+k} H(k,j) \det(H_j^{-1} H(t_k, t_j)) \ , \tag{12}$$

without changing in solution (8) to the equation (7). Then, the approximate solution to the equation (3) we can record as follows:

$$\tilde{x}_j = \tilde{\Delta}_j / \tilde{\Delta} \ . \tag{13}$$

As our investigations have shown, such a simple transformation leads to substantial stabilization of solution to the linear regression problem under near-collinearity, if we choose the regularization constant as $\alpha = 0.001$ (MOLS method). Regardless of the collinearity level, the solution (13) gives practically unbiased solution using the MOLS in the linear regression problem and practically the same small variance of the regression coefficients as in the correspondent ridge regression solution with regularization constant $\lambda = 0.5$.

To the disadvantages of the MOLS one can attribute high complexity of calculations. This one prevents from using this method for big data analysis with number of regressors larger than 200-300.

It should be noted that the modified Cramer's rule outlined above should not be used for solving common ill-conditioned linear systems of large size. For common systems this method is subject to accumulation of computational errors while computing the determinants.

In standardized regression problems the computational errors are mutually neutralized due to data centering. This issue makes it possible to solve the linear systems with sufficient accuracy up to the order of 200-300 (regressors).

Worth noting, that both the MOLS and ridge-regression are approximate methods of stabilization of the OLS under near-collinearity. Both methods use the same idea of regularization of the ill-conditioned OLS matrix equation with the aid of replacement of the original matrix (A) by the one that is close to it ($A + \alpha I$) [24, 25] and [10]. However, the difference between the MOLS and

Ridge methods is that in the MOLS a neutralization of ill-conditioning is used with the aid of multiplication of the matrix $H(t_k, t_j)$ by the inverse matrix H_j^{-1} in (11, 12). Such neutralization drastically improves the situation with ill-conditioning for a very small regularization constant ($\alpha = 0.001$) and reduces significantly the dependence of the solution on this constant. The latter one allows us not to search for the optimal value of this constant. In all cases one can use the only constant value of $\alpha = 0.001$. That one allows obtaining in all cases practically the same RSS for the MOLS as for the OLS.

The disadvantage of the MOLS, as well as of the ridge-regression, is the lack of the ability to estimate theoretically the variance of the regression coefficients obtained with the aid of the observed sample.

Then, for any new method solution to the linear regression problem, we have to estimate both the biasedness of a solution and its variance. Let us consider these issues with the aid of the Monte Carlo simulation method.

Artificial data population

To test any new method solution to the linear regression problem for adequacy, we use in this paper the artificial data population (ADP) reconstruction. Such a population has given parameters of contained variables but a priori unknown regression coefficients. All variables of this population have linear relationships between themselves and are normal. The last condition is optional.

From such a population one can draw samples of any size. Very large samples from this population allow us to estimate the regression coefficients in the population with a priori given accuracy due to the unbiasedness and consistency of OLS solutions. In turn, the knowing of the population regression coefficients allow us to estimate the biasness of a new method solution (using very large samples) and variances of its sample regression coefficients (using multiple drawing of samples of the given size).

Consider the creation of an artificial population mentioned above. At first, let us set a priory any regressand Y and auxiliary vector $T = Y + \alpha\, randn$, where $randn \sim N(0,1)$ is a standard normal vector of size Y taken from the MATLAB pseudo-random generator and α – a constant which a priori sets the near-collinearity level. Regressors $\{X_j\}$ are constructed with the aid of the auxiliary vector T: $X_j = k_j T$, where $k_j = \tan(\alpha_j)$ and α_j is the angle between Y and X_j vectors. With diminishing of α_j, the mean influence of X_j on Y increases as the projection of a unit increment along the trend and that, the correspondent regression coefficient b_j in the modeled population increases also. For modeling stochastic regressors, the pseudo-random function $randn$ restarts for each replica.

This method makes it possible to create a population, in which all regression coefficients are the same, for instance, or these ones are decreasing (increasing) in a given manner, or these are having the given signs. These issues allow one to test a new method solution to the linear regression problem for adequacy.

The data simulated with this method have been denoted as $DSm(n,\alpha)$, where m is the number of regressors, n is the sample size and α sets the near-collinearity level. To this notation, we should add a set of angles $\{\alpha_j\}$ which sets the regression coefficient values in the artificial population $DSm(n,\alpha)$.

The modified OLS method testing

Let us consider the artificial population $DS5(n, 0.01)$ with the set of angles $\{\alpha_j\} = \{5, 5{,}40{,}60{,}80\}$ for demonstrating the stability and small biasedness of the modified OLS (MOLS) method. With this set of angles $\{\alpha_j\}$, the first two regression coefficients in the population should be equal and much more in value than the others. Other coefficients are descending in magnitude. All coefficients have to be positive.

Table 1. OLS, MOLS and Ridge solution means and theirs standard deviations via sample size n under severe near-collinearity ($\alpha = 0.01, VIF = 57107$).

Method	n	b_0	b_1	b_2	b_3	b_4	b_5
Single sample solutions							
OLS	10	-0.0099	5.9015	6.4643	-0.4212	-0.0539	0.0644
MOLS	10	-0.0124	2.2868	2.2910	0.2350	0.1153	0.0353
Ridge	10	1.3452	2.0815	2.0791	0.2168	0.1050	0.0320

OLS	10000	-0.0004	2.2853	2.2859	0.2428	0.1238	0.0346
MOLS	10000	0.0002	2.2862	2.2859	0.2384	0.1155	0.0353
Ridge	10000	1.3671	2.0785	2.0785	0.2167	0.1050	0.0321
Mean regression coefficients (10^4 sample replics)							
OLS	10	0.0002	2.3190	2.3217	0.2346	0.1129	0.0356
MOLS	10	0.0008	2.2858	2.2859	0.2383	0.1155	0.0353
Ridge	10	1.3642	2.0781	2.0782	0.2167	0.1050	0.0321
OLS	10000	0.0001	2.2860	2.2864	0.2386	0.1154	0.0352
MOLS	10000	0.0007	2.2859	2.2859	0.2383	0.1155	0.0353
Ridge	10000	1.3637	2.0782	2.0782	0.2167	0.1050	0.0321
Standard deviations of regression coefficients (10^4 sample replics)							
OLS	10	0.0202	2.7236	2.7046	0.2757	0.1384	0.0411
MOLS	10	0.0129	0.0043	0.0044	0.0005	0.0002	0.0001
Ridge	10	0.0059	0.0039	0.0040	0.0004	0.0002	0.0001
OLS	10000	0.0011	0.1455	0.1465	0.0152	0.0073	0.0022
MOLS	10000	0.0059	0.0003	0.0003	0.0000	0.0000	0.0000
Ridge	10000	0.0059	0.0003	0.0003	0.0000	0.0000	0.0000

With the sample size (n) increasing the MOLS solution should approach in probability to the OLS solution mean if it is almost unbiased. Exactly this issue one can see in the Table 1 for $n = 10000$ for a single sample solution. We also see that in population $b_1 = b_2$ and the other coefficients are decreasing in value. As the estimates of the population coefficients we can take the mean OLS solution for $n = 100000$: $b_1 = b_2 = 2.2860$; $b_3 = 0.2386$; $b_4 = 0.1154$; $b_5 = 0.0352$. If we look at the single MOLS-solution for $n = 10$ in the Table 1, we can see practically the same values for regression coefficients as that for the OLS with $n = 100000$. The same thing can be seen for the ridge-method ($\lambda = 0.5$), except for being a bias.

Solution' standard deviations for both MOLS and Ridge methods are equal and drastically smaller than those of the OLS method. So, this comparison confirms the stability and small bias of MOLS solutions, and demonstrates a possibility of finding a correct solution to the linear regression problem both for small and large samples.

As for the ridge-method, we use only one value of the regularization constant, $\lambda = 0.5$, for all calculations in this paper, that gives the most stable solution with not very large bias as we can see in Table 1. More than that, as a rule of thumb one can obtain practically unbiased ridge-solution with this λ, if she or he multiplies all single sample ridge-solution components (except for b_0) by the number 1.1. This one we can see in Table 1 if we multiply the ridge solution by the value of 1.1 for both $n = 10$ and $n = 10000$.This rule has produced an excellent result in all our investigations but it requires confirmation on a larger database.

It is worth noting that the considered simulation procedure allows also demonstrating the confidence intervals diminishing for the MOLS method compared to the OLS one. One can also verify the significance of sample regression coefficients by the proposed simulation method with the aid of the z-test considering that she or he knows precisely the correspondent variances of the sample regression coefficients.

In general, the developed simulation method allows comparing the common OLS with the new MOLS method for demonstrating the advantages of the latter. Although this simulation method is not intended to solve the linear regression problem for some observed sample, it provides an opportunity to verify any method solution to the linear regression problem under multicollinearity.

With this simulation method in hand, the author has demonstrated in the paper a practical unbiasedness of the MOLS and its very small variability in solving the linear regression problems under near-collinearity of any level.

The mentioned simulation method (ADP) allows demonstrating the high proximity of the MOLS solutions to the solution in the population, which we estimate as the OLS solution for a very large sample. As one can see from Table 1, an MOLS solution is very close to the population solution even for a very small sample size.

In general, the ADP has made it possible to affirm that the developed MOLS method gives an adequate solution to the linear regression problem under near-collinearity and multicollinearity.

Summarizing, the obtained results can be characterized as follows.

The notion of physically correct and physically incorrect codomain of any non-singular matrix has been introduced and explained with the aid of this notion the appearance of economically incorrect OLS-solutions in the presence of near-collinearity. It was clarified that the incorrectness of the OLS solutions is a consequence of the exit of the OLS matrix equation' RHS from the codomain of physical correctness due to random errors in the data and great variability of the OLS-solutions.

The new method presented in the paper, that is the MOLS, is based on the OLS matrix regularization, which enlarges the codomain of physical correctness and then diminished the probability of exit of the RHS from this codomain. More than that, the modified Cramer's formulas give a more stable solution than the Gauss' method. Both these factors lead to a more stable and economically adequate solution of the MOLS than the OLS. Relatively to the ridge-method, the MOLS is practically unbiased and does not need to optimize the regularization constant. These two advantages are decisive for practical applications.

To shortcomings of the MOLS, one can attribute the intensive computer loading of the algorithm and possible OLS-like behavior of the solution for the MOLS method, as well as for the ridge-method, in a rare situation of a very large partial regression' residual error of some regressors with the regressand. Such situation is observed, for instance, in the presence of non-linear regressors.

References: 1. Seber, G.A.F. Linear Regression Analysis / G.A.F. Seber. – NY: Wiley-Blackwell, 1977. – 456 p. 2. Seber, G.A.F. Linear Regression Analysis, 2nd edition / G.A.F. Seber, A.J. Lee. – NY: Wiley, 2003. – 341 p. 3. Spanos, A., Probability Theory and Statistical Inference: econometric modeling with observational data / A. Spanos. – Cambridge: Cambridge University Press, 1999. – 401 p. 4. Gujarati, D. N. Basic econometrics / D. N. Gujarati. – NY: McGraw-Hill, 2002. – 526 p. 5. Wooldridge, J.M. Introductory Econometrics: Modern Approach, 5th ed. / J.M. Wooldridge. – Ohio: South-Western, 2009. – 633 p. 6. Badi Baltagi. Econometrics / Badi Baltagi. – NY: Springer, 2011. – 812 p. 7. Greene, W.H. Econometric Analysis, 7th ed. / W.H. Greene,. – NY: Pearson. 2012. – 1211 p. 8. Draper, N. R. H. Applied Regression Analysis / N. R. Draper, H. Smith. – New York: Wiley. – 1966. – 445 p. 9. Farrar, D.

Multicollinearity in regression Analysis: The problem revisited / D. Farrar, R. R. Glauber // Review of Economics and Statistics. – 1967. – 49. – P. 92-107. 10. Hoerl, A. E. Ridge regression: Biased estimation for nonorthogonal problems / A. E. Hoerl, R. W. Kennard // Technometrics. – 1970. – 12(1) . – P. 55–67. 11. Marquardt D.V. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation / D.V. Marquardt // Technometrics. – 1970. – 12. – P. 591-612. 12. Blanchard O.J. Comment. / O.J. Blanchard // Journal of Business and Economic Statistics 5. – 1987. – P. 449–51. 13. Adkins L. C., Collinearity. Companion in Theoretical Econometrics, edited by Badi Baltagi / L. C. Adkins, R. C. Hill. – Oxford: Blackwell Publishers, Ltd. – 2001. – pp. 256-278. 14. Belsley D. A. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity / D. A. Belsley, E. Kun, R. T. Welsh. – New York: Wiley. – 2004. – P. 651. 15. Belsley, D.A. Demeaning conditioning diagnostics through centering / D.A. Belsley // The American Statistician. – 1984. –38(2) . – P. 73-77. 16. Rao, C. R. Linear Models: Least Squares and Alternatives, 2^{nd} ed. / C. R. Rao, , H. Toutenberg. – NY: Springer, 1999. – P. 301. 17. Spanos, A. The Problem of Near-Multicollinearity Revisited: erratic vs. systematic volatility / A. Spanos, A. McGuirk // Journal of Econometrics. – 2002. – 108. – P. 365-393. 18. Kabanichin S. I. Definitions and Examples of Inverse and Ill-posed Problems / S. I. Kabanichin // J. Inv. Ill-Posed Problems. – 2008. – 16. – P. 317 – 357. 19. Adkins L. C. Collinearity Diagnostics in gretl, Economics Working Paper Series 1506 / M. S. Waters, R. C. Hill. – Oklahoma: Oklahoma State University, Department of Economics and Legal Studies in Business. – 2015. – 452 p. 20. Fox, J. Applied regression analysis, linear models, and related methods / J. Fox. – Thousand Oaks, CA: Sage Publications. – 1997. – 742 p. 21. Cook, R.D. Comment on Demeaning Conditioning Diagnostics through Centering, by Belsley, D.A. / R.D. Cook // The American Statistician. – 1984. – 38. –P. 78-79. 22. Maddalla G.S. Introduction to Economics / G.S. Maddalla. – New York: Macmillan, 1992. – 396 p. 23. Harvey, A.C. Some Comments on Multicollinearity in Regression / A.C. Harvey // Applied Statistics. – 1977. – 26(2). – P. 188-191. 24. Tikhonov A. N. On the stability of inverse problems / A. N. Tikhonov // Doklady Acad. Sci. USSR. – 1943. – 39. – P.176–179. 25.

Tikhonov, A. N. Solutions of Ill-Posed Problems / Tikhonov A. N., Arsenin, V. Y. – New York: Winston & Sons, 1977. – 287 p. 26. Dougherty C. Introduction to Econometrics / C. Dougherty. – NY: Oxford University Press, 1992. – 402 p.

Appendix A

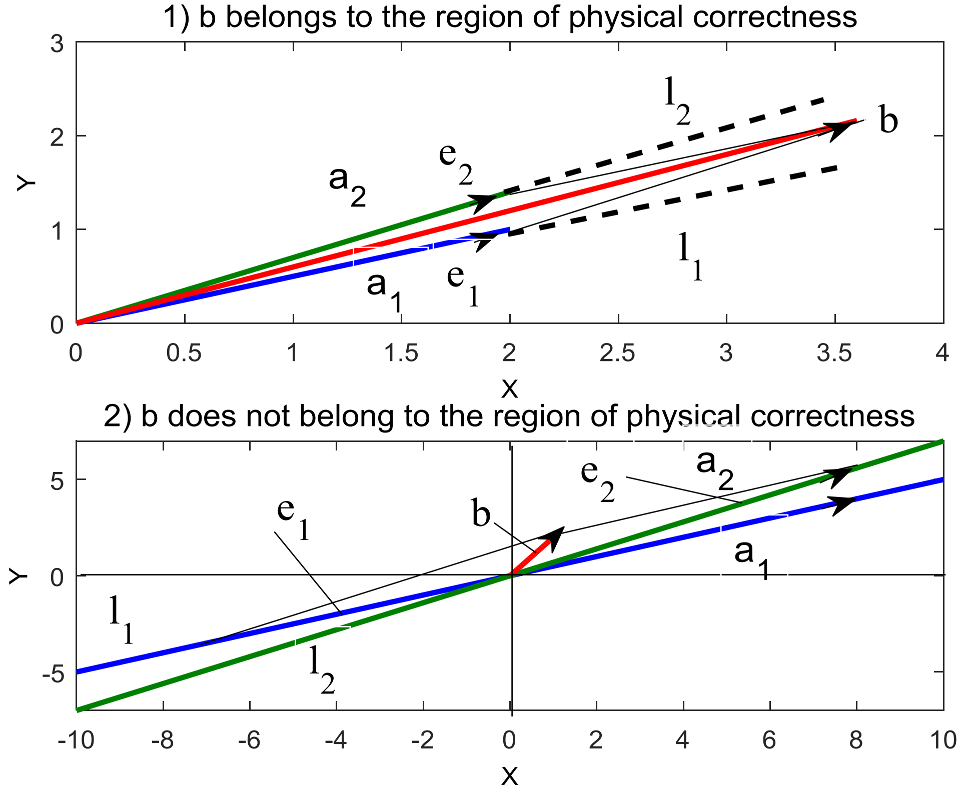

Figure 1. At both figures the right-hand side vector b is represented as the sum of two components, e_1 and e_2, in the basis (a_1, a_2) of matrix columns. The bold lines limit from the outside the 2D region D^s of physical correctness in both directions from the origin. 1) The right-hand side b belongs to the region of physical correctness. Both solutions have the same signs. 2) The right-hand side b does not belong to D^s. Both solutions, e_1 and e_2, become larger in value and have different signs.