Все, что познается, имеет число,
ибо невозможно ни понять ничего,
ни познать без него.
Пифагор

Математичні методи, моделі та інформаційні технології в економіці

УДК 330.43(075.8) *JEL Classification: C52*

A NEW STABLE SOLUTION TO THE LINEAR REGRESSION PROBLEM UNDER MULTICOLLINEARITY

A. Tyzhnenko

Tyzhnenko A. A New Stable Solution to the Linear Regression Problem under Multicollinearity / A. Tyzhnenko // Економіка розвитку. – 2018. – № 2 (86). – С. 89–99.

The main shortcomings of the OLS (Ordinary Least Squares) solution to the multiple linear regression problem under multicollinearity which prevent from obtaining an adequate solution to the economic problem of evaluation of each regressor's contribution to the regressand have been considered.

The main causes of the OLS incorrectness of the economic problem solution have been revealed, these causes being related to a great variability in the OLS solution under considerable data multicollinearity.

The research has also shown that mathematically correct standard OLS solutions can become economically incorrect with data collinearity increasing which leads to diminishing of the OLS matrix codomain of physical correctness.

The current methods for overcoming the OLS solutions' great variability have been considered in both the economic and mathematical aspects. The current methods have been proved inefficient in overcoming multicollinearity by either mathematical or economic methods such as choosing the best regressions, lasso, etc.

The analysis has brought to a conclusion that the only way out is to create a new method of solving the OLS equation which would give a stable solution with small variability, as for example in the ridge method, but with a small bias. Precisely such method is the Modified OLS (MOLS) proposed in the paper.

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

The MOLS method has proved to give a stable and practically unbiassed 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 optimization of the regularization constant.

The proposed MOLS method has been verified 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 has been checked for biassedness and stability for both small and large samples.

Keywords: *multicollinearity, stable solution, negligeable biassedness, mathematical correctness, physical correctness, ridge regression.*

НОВИЙ МЕТОД СТАБІЛЬНОГО РОЗВ'ЯЗАННЯ ЗАДАЧІ ЛІНІЙНОЇ РЕГРЕСІЇ В УМОВАХ МУЛЬТИКОЛІНЕАРНОСТІ

Тижненко О. Г.

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

Виявлено основні причини появи некоректних розв'язків економічної задачі регресії математичним методом найменших квадратів, які пов'язані з великою варіабельністю МНК-рішення за значної колінеарності даних.

Показано, що некоректні з точки зору економіки математичні розв'язки стандартного МНК виникають у разі збільшення рівня колінеарності даних за рахунок зменшення області фізичної коректності МНК-матриці.

Розглянуто існуючі на сьогодні методи подолання великої варіабельності МНК-рішень як з економічної, так і з математичної точки зору. Отримано переконливі докази неефективності цих методів подолання проблеми мультиколінеарності як з боку математики, так і з боку економічного розгляду спрощення самої економічної проблеми: вибір найкращих регресій, lasso та ін.

Проведений аналіз дозволив зробити висновок, що єдиний вихід із існуючої ситуації – створити нові методи розв'язання МНК-рівняння, які б давали рішення з малою варіабельністю, як в рідж-методі, наприклад, але з малим зміщенням. Саме таким методом є новий модифікований метод найменших квадратів (ММНК), який подано в роботі.

ММНК є наближеним методом, в якому використано метод регуляризації Тіхонова і новий метод розв'язання регуляризованого МНК-рівняння, заснований на модифікованому методі Крамера, що запропоновано в статті.

Показано, що ММНК дає стійке та практично незміщене розв'язання задачі лінійної регресії за будь-якого рівня колінеарності даних. На відміну від методу рідж-регресії, ММНК показує дуже мале зміщення і не потребує оптимізації константи регуляризації.

Запропонований у роботі ММНК перевірено на адекватність за допомогою штучної генеральної сукупності (ШГС), яка створена за допомогою методу Монте-Карло. Завдяки використанню цієї ШГС показано як практичну незміщеність ММНК, так і високу стабільність розв'язання задачі регресії і для великих, і для малих вибірок.

Ключові слова: *мультиколінеарність, стабільний розв'язок, дуже мале зміщення, математична коректність, фізична коректність, рідж-регресія.*

НОВЫЙ МЕТОД СТАБИЛЬНОГО РЕШЕНИЯ ЗАДАЧИ ЛИНЕЙНОЙ РЕГРЕССИИ В УСЛОВИЯХ МУЛЬТИКОЛЛИНЕАРНОСТИ

Тыжненко А. Г.

Рассмотрены основные недостатки решения многофакторной задачи линейной регрессии в условиях мультиколлинеарности методом наименьших квадратов (МНК), которые не позволяют получить адекватное решение экономической проблемы оценки влияния каждого отдельного регрессора на отклик.

Выявлены основные причины некорректного решения экономической задачи регрессии математическим методом наименьших квадратов, которые 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. Получены убедительные доказательства

неэффективности существующих 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 единственный выход из существующей ситуации – создание новых методов решения МНК-уравнения, которые давали бы решения с малой вариабельностью, как в ридж-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 известный принцип 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 ридж-регрессии, ММНК показывает ничтожно малое смещение и не требует оптимизации константы регуляризации.

Предложенный в работе ММНК проверен на адекватность с помощью искусственной генеральной совокупности (ИГС), созданной с помощью метода Монте-Карло. Благодаря использованию этой ИГС показана как практическая несмещенность ММНК, так и высокая стабильность решений задачи линейной регрессии и для больших, и для малых выборок.

Ключевые слова: *мультиколлинеарность, стабильное решение, ничтожно малая смещенность, математическая корректность, физическая корректность, ридж-регрессия.*

An economic insight into the multiple linear regression solution can be provided as the obtaining of significant estimates of regression coefficients that represent the mean change in the response variable for one unit of change in the predictor variable with other predictors in the model being constant.

It is clear from the economic 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.

A new method for finding a solution to the linear regression problem has been proposed in which a common great instability of the OLS has been overcome.

This problem has been considered 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 to the 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 to an economic problem from being adequate 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 fact 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.

A very important source of errors in regression solutions is wrong model specification [9; 12], but this problem is connected with economic considerations and has not been 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 [4; 9 – 17; 19 – 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 a 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 in any OLS solution is a continuous process which depends on many parameters. This issue is also confirmed by the following further considerations of OLS solution properties.

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 the author calls 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 it 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 to the matrix equation $Ax = b$ is physically correct if $b \in D^c$; and it is incorrect otherwise.

It has been shown that this effect holds for any matrix equation with a square non-singular matrix. However, with the matrix conditional number increasing, the codomain of physical correctness, D^c, is becoming narrower. This issue may lead to the right-hand side (RHS) of the matrix equation being outside of D^c due to random errors in the matrix elements. If it is the case, there will be a change in signs of the solution components. This 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 the presence of near-collinearity and 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 out the appropriate estimates of regression coefficients in the population.

The advantage of the OLS is the unbiassedness 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 of 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 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 for solving the linear regression problem, which would give a small bias and acceptable solution variability for not very large samples.

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

The unbiassedness 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 out of a population converges in probability to the population solution (regression coefficients) with M increasing. This one is also used in the paper for testing the new methods for solving the linear regression problem with the aid of the artificial population (ADP) worked out for this purpose.

In this paper, a new method (MOLS) is proposed which produces stable solutions with 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 $(X'X + \alpha I)b = X'Y$ for $\alpha = 0.001$. This equation is solved further with the aid of a 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.

The only disadvantage of the MOLS is large computer loading that prevents from applying this method for a large number of regressors (more than 200 – 300).

The adequacy of the new method (MOLS) has been verified in this research 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 [24] in that it does not use a priori giving regression coefficients in a population. Thus, the ADP method simulates a population with unknown regression coefficients with values that can be precisely estimated by the OLS solution for a very large sample size, using its consistency.

The essence of the ADP method consists in 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 the 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 population should be equal. If two regressor vectors have angles with the regressand vector which differ by $\pi/2$, their regression coefficients should have opposite signs and be equal in the 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 new methods for solving the linear regression problem. For each modelled situation with population regression coefficients, one has an opportunity for estimation of the population regression coefficients with the aid of the asymptotic OLS solution. This one makes it possible to estimate both the biassedness and variability of any new linear regression solution 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 common simulation method [24] for multiple regression in which for a priori given regressors, $\{X_j\}_m$, and population regression coefficients, $\{b_j\}_m$, one set many times (M) a random residual error, $\{e\}_n$ for calculating M regressand realizations, $\{Y\}_n$.

The matter is that the given regressors, $\{X_j\}_m$, define exactly the OLS-matrix $X'X$ of the matrix equation $X'X\ b = X'Y$, which has a definite codomain of physical 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 $X'Y$. Such calculated RHS may not belong to this D^c. This one will lead to physically incorrect solution to the linear regression problem. The situation may change only if we take those a priori regression coefficients 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 biassedness of the new method for finding a solution to the linear regression problem but also to estimate the variability of sample regression coefficients. For this purpose, we can draw a large series of replicas from the ADP 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, the question of what real problem is to be solved is not raised. However it does not necesseraly result in an error. For instance, this is the case with the basis changing problem in the 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.

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 the codomain of physically correct solution of the matrix equation $Ax = b$ $(b \in D^c)$ and $\overline{D}^c$ is the codomain of physically incorrect solution of the matrix equation $Ax = b$ $(b \in \overline{D}^c)$.

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

Otherwise, if $b \in \overline{D}^c$, the signs of the exact solution 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 the equation $Ax = b$ changes the signs of some solution components when the 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 the equation $Ax = b$ is unstable in the ill-conditioning case, $cond(A) \gg 1$. Geometrically this one follows from projection properties. If the RHS is outside of D^c, projections of the RHS on the basis vectors increase with the RHS moving 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 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 the observed data.

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

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

$$Ax = b \Leftrightarrow a_1 x_1 + a_2 x_2 = b, \qquad (1)$$

where $a_1 = \begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix}$, $a_2 = \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}$, $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$.

From the 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.).

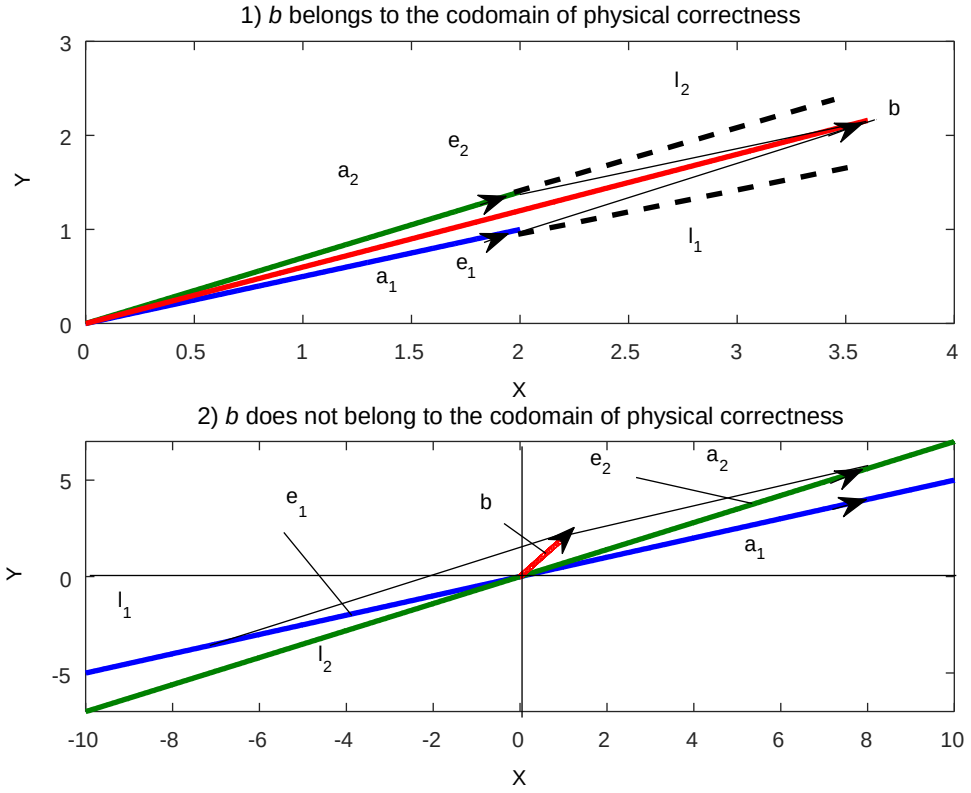

Fig. **The geometric interpretation of the appearance of physically incorrect solutions**

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

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., part 1). The vectors e_1 and e_2 in Fig., part 1 are the projections of vector b on the basis vectors a_1 and a_2. In this case, both equation (1) solution components are positive ($x_1 = |e_1|$, $x_2 = |e_2|$). Similarly, if for the same basis vectors the RHS has the inverse direction $(-b)$ and is located between $-a_1$ and $-a_2$, then both solution components are negative. In both of these cases the solutions have the same signs.

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

This simple example demonstrates the fundamental property of non-singular matrices: the 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. It is worth noting, that this property is not connected with the conditioning of the matrix equation.

Summarizing, 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.

A 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. The common sale is \$78.5. The common number of units sold is 87. How many hot dogs (x_1) and sodas (x_2) were sold separately? The system $(Ax = b)$:

$$\begin{cases} 1.5x_1 + 0.5x_2 = 78.5 \\ x_1 + x_2 = 87 \end{cases} \quad (2)$$

With det(A) = 1; cond(A) = 4.27. Clearly, it is a well-conditioned system. 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 its 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 is:

$$b_+ = (78.5; 157),\ b_- = (78.5; 52.\overline{3}).$$

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 that $b \in D^c$.

If $b = (78.5; 53)$, the solution is $x = (52; 1)$. This means that 52 hot dogs and one soda were sold. If $b = (78.5; 157)$, the solution is $x = (0; 157)$. This means that no hot dogs but 157 sodas were 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 the 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 $b(2) < 52.\overline{3}$ or $b(2) > 157$. For example, let the common number of units sold be $b(2) = 51$. 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 the solution is also incorrect. This means that 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 as regards the adequacy of the solution to the economic problem which is investigated with the aid of this linear system. If the RHS of a system does not belong to the system matrix codomain of physical correctness, a mathematically correct solution to this system will be not correct for the practical problem under investigation.

Because 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 are correct. This is the case with 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.

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 the 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."

The situation with multicollinearity, as described by [12], has not changed to date [4 – 6; 19; 22], as far as the author knows.

As can be seen from the above description, the OLS solution to the problems under multicollinearity is connected with the ill-conditioning of the OLS matrix equation $Ax = b$, which, as any non-singular matrix, has its codomains of physical correctness (D^c) and incorrectness $(\overline{D}^c)$. The increase in near-collinearity level in data tends to increase the condition number of the OLS-matrix and then leads to a contraction of the physical correctness (D^c) codomain. Besides that, the increasing of ill-conditioning level tends 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 the 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 so large a sample size that for any sample, the RHS of the OLS matrix equation will belong to the physical correctness (D^c) codomain and the OLS solution will be consistent and economically correct. For small and not very large samples this is not the case, as a rule.

It is also worth noting that a physically correct and consistent 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 this kind is proposed in this paper.

Summarizing, the only fruitful strategy for overcoming the near-collinearity in the linear regression problem is the construction of new methods for finding a solution to this problem which would provide consistent physically

correct solutions with standard deviations much less than the absolute values of the solution components (regression coefficients). Besides, such methods would be negligibly biassed.

For obtaining a stable solution to the linear regression problem in the standardized form, the author has proposed a modified OLS (MOLS) solution using the modified Cramer's rule, 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 $\alpha = 0.001$. This equation is solved further with the aid of the modified Cramer's rule.

The modified Cramer's rule is intended to solve definite ill-conditioned systems $X'Xb = X'Y$ that arise in the standardized linear regression problems (all variables are standardized). For brevity, let us write 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. Let us designate the new matrix by H:

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

where E is the identity matrix and $0 < \alpha \ll 1$ (an optimal value that gives a minimal residual sum of squares (RSS) is $\alpha = 0.001$). Let us replace 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

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

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

$\Delta = \sum_{k=1}^{n}(-1)^{j+k}H(k,j)\det(H(t_k, t_j))$, (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 the matrix H. That is, formulas (8 – 10) are figured 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 a matrix of this kind, 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 have been crossed out. For each j we multiply the numerator and denominator in (8) by a different determinant $\det(H_j^{-1})$. Using the determinant property:

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

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

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

$$\widetilde{\Delta} = \sum_{k=1}^{n}(-1)^{j+k}H_1(k,j)\det(H_j^{-1}H(t_k, t_j)), \tag{12}$$

with no changes in solution (8) to equation (7). Then, the approximate solution to equation (3) can be written as follows:

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

As the research has shown, such a simple transformation leads to substantial stabilization of the solution to the linear regression problem under near-collinearity if one chooses the regularization constant $\alpha = 0.001$ (the MOLS method). Regardless of the collinearity level, equation (13) gives a practically unbiassed 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 $\lambda = 0.5$.

A disadvantages of the MOLS is high complexity of calculations which prevents from using this method for big data analysis with the number of regressors larger than 200 – 300.

It should be noted that the outlined modified Cramer's rule 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).

It is worth noting, that both the MOLS and ridge regression are approximate methods for 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 $(A + \alpha I)$ [10; 25; 26]. 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 for finding a solution to the linear regression problem, we have to estimate both the biassedness of a solution and its variance. Let us consider these issues with the aid of the Monte Carlo simulation method.

To test any new method for solving a linear regression problem for adequacy, the author used the artificial data population (ADP) reconstruction. Such a population has some 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 unbiassedness and consistency of OLS solutions. In turn, the knowing of the population regression coefficients allows us to estimate the biassedness of the new method (using very large samples) and variances of its sample regression coefficients (using multiple drawing of samples of the given size).

Consider the creation of the mentioned artificial population. First, let us set a priori any regressand Y and an 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 is 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 the correspondent regression coefficient b_j in the modelled population increases also. For modelling 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 for solving a 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 is the constant that sets the near-collinearity level. This notation has been complemented with angles $\{\alpha_j\}$ which set the regression coefficient values in the artificial population $DSm(n, \alpha)$.

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 biassedness 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.

With the sample size (n) increasing, the MOLS solution should approach in probability the OLS mean solution if it is almost unbiassed. Exactly this issue is shown in the Table for $n = 10\,000$ for a single sample solution. We also see that in population $b_1 = b_2$ and other populations, the coefficients are decreasing in values. The mean OLS solution can be taken as the estimates of the population coefficients for $= 100\,000$: $b_1 = b_2 = 2.2860$; $b_3 = 0.2386$; $b_4 = 0.1154$; $b_5 = 0.0352$. If we look at the single MOLS solution for $n = 10$ in the Table, we can see practically the same values for regression coefficients as that for the OLS with $n = 100\,000$. The same thing can be seen for the ridge method $(\lambda = 0.5)$, except for being a bias.

Table

The OLS, MOLS and ridge solution means and their 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
1	2	3	4	5	6	7	8
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	10 000	-0.0004	2.2853	2.2859	0.2428	0.1238	0.0346
MOLS	10 000	0.0002	2.2862	2.2859	0.2384	0.1155	0.0353
Ridge	10 000	1.3671	2.0785	2.0785	0.2167	0.1050	0.0321

Table (the end)

1	2	3	4	5	6	7	8
Mean regression coefficients (10^4 sample replicas)							
OLS MOLS Ridge	10 10 10	0.0002 0.0008 1.3642	2.3190 2.2858 2.0781	2.3217 2.2859 2.0782	0.2346 0.2383 0.2167	0.1129 0.1155 0.1050	0.0356 0.0353 0.0321
OLS MOLS Ridge	10 000 10 000 10 000	0.0001 0.0007 1.3637	2.2860 2.2859 2.0782	2.2864 2.2859 2.0782	0.2386 0.2383 0.2167	0.1154 0.1155 0.1050	0.0352 0.0353 0.0321
Standard deviations of regression coefficients (10^4 sample replicas)							
OLS MOLS Ridge	10 10 10	0.0202 0.0129 0.0059	2.7236 0.0043 0.0039	2.7046 0.0044 0.0040	0.2757 0.0005 0.0004	0.1384 0.0002 0.0002	0.0411 0.0001 0.0001
OLS MOLS Ridge	10 000 10 000 10 000	0.0011 0.0059 0.0059	0.1455 0.0003 0.0003	0.1465 0.0003 0.0003	0.0152 0.0000 0.0000	0.0073 0.0000 0.0000	0.0022 0.0000 0.0000

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, the author uses only one value of the regularization constant, $\lambda = 0.5$, for all calculations in this research, that gives the most stable solution with a rather small bias as one can see in the Table. More than that, as a rule of thumb, one can obtain a practically unbiassed ridge solution with this λ, if they multiply all single sample ridge-solution components (except for b_0) by the number 1.1. This can be seen in the Table if we multiply the ridge solution by the value 1.1 for both $n = 10$ and $n = 10\,000$. This rule has produced an excellent result in all the investigations but it requires confirmation on a larger database.

It is worth noting that the considered simulation procedure also demonstrates the confidence intervals diminishing for the MOLS method compared to the OLS one. The significance of the sample regression coefficients can be also verified by this simulation method with the aid of z-test considering that the correspondent variances of the sample regression coefficients are precisely known.

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 for solving the linear regression problem for some observed sample, it provides an opportunity to verify any method for solving the linear regression problem under multicollinearity.

With this simulation method in hand the author has demonstrated almost complete unbiassedness of the MOLS and its very small variability in solving linear regression problems under near-collinearity of any level.

The mentioned simulation method (ADP) allows demonstrating a high proximity of the MOLS solutions to the solution in the population, which the author has estimated as the OLS solution for a very large sample. As one can see in the Table, a 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 with the aid of this notion the appearance of economically incorrect OLS-solutions in the presence of near-collinearity has been explained. It has been 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 diminishes the probability of the exit of the RHS from this domain. 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 unbiassed and does not need optimization of the regularization constant. These two advantages are decisive for practical applications.

The shortcomings of the MOLS are 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 regression method, in rare situations of a very

large partial regression residual error of some regressors with the regressand which is observed, for instance, in the presence of non-linear regressors.

References: 1. Seber G. A. F. Linear Regression Analysis / G. A. F. Seber. – New York : Wiley-Blackwell, 1977. – 456 p. 2. Seber G. A. F. Linear Regression Analysis / G. A. F. Seber, A. J. Lee. – 2nd ed. – New York : 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. – New York : McGraw-Hill, 2002. – 526 p. 5. Wooldridge J. M. Introductory Econometrics: Modern Approach / J. M. Wooldridge. – 5th ed. – Ohio : South-Western, 2009. – 633 p. 6. Badi Baltagi. Econometrics / Badi Baltagi. – New York : Springer, 2011. – 812 p. 7. Greene W. H. Econometric Analysis / W. H. Greene. – 7th ed. – New York : Pearson, 2012. – 1211 p. 8. Draper N. R. 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. – No. 49. – P. 92–107. 10. Hoerl A. E. Ridge regression: Biased estimation for nonorthogonal problems / A. E. Hoerl, R. W. Kennard // Technometrics. – 1970. – No. 12 (1). – P. 55–67. 11. Marquardt D. V. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation / D. V. Marquardt // Technometrics. – 1970. – No. 12. – P. 591–612. 12. Blanchard O. J. Comment / O. J. Blanchard // Journal of Business and Economic Statistics.– 1987. – No. 5. – 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. – P. 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. – No. 38 (2) . – P. 73–77. 16. Rao C. R. Linear Models: Least Squares and Alternatives / C. R. Rao, H. Toutenberg. – 2nd ed. – New York : Springer, 1999. – 301 p. 17. Spanos A. The Problem of Near-Multicollinearity Revisited: Erratic vs. Systematic Volatility / A. Spanos, A. McGuirk // Journal of Econometrics. – 2002. – No. 108. – P. 365–393. 18. Kabanichin S. I. Definitions and Examples of Inverse and Ill-posed Problems / S. I. Kabanichin // Journal of Inverse and Ill-Posed Problems. – 2008. – No. 16. – P. 317–357. 19. Adkins L. C. Collinearity Diagnostics in gretl / L. C. Adkins, M. S. Waters, R. C. Hill // Economics Working Paper Series 1506. – 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. – No. 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. – No. 26 (2). – P. 188–191. 24. Dougherty C. Introduction to Econometrics / C. Dougherty. – New York : Oxford University Press, 1992. – 402 p. 25. Tikhonov A. N. On the stability of inverse problems / A. N. Tikhonov // Doklady Acad. Sci. USSR. – 1943. – No. 39. – P.176–179. 26. Tikhonov A. N. Solutions of Ill-Posed Problems / A. N. Tikhonov, V. Y. Arsenin. – New York : Winston & Sons, 1977. – 287 p.

Information about the author

A. Tyzhnenko – PhD in Physical and Mathematical Sciences, Associate Professor of the Department of Higher Mathematics and Economic and Mathematical Methods of Simon Kuznets Kharkiv National University of Economics (9-A Nauky Ave., Kharkiv, Ukraine, 61166, e-mail: oleksandr.tyzhnenko@m.hneu.edu.ua).

Інформація про автора

Тижненко Олександр Григорович – канд. фіз.-мат. наук, доцент кафедри вищої математики та економіко-математичних методів Харківського національного економічного університету імені Семена Кузнеця (просп. Науки, 9-А, м. Харків, Україна, 61166, e-mail: oleksandr.tyzhnenko@m.hneu.edu.ua).

Информация об авторе

Тыжненко Александр Григорьевич – канд. физ.-мат. наук, доцент кафедры высшей математики и экономико-математических методов Харьковского национального экономического университета имени Семена Кузнеца (просп. Науки, 9-А, г. Харьков, Украина, 61166, e-mail: oleksandr.tyzhnenko@m.hneu.edu.ua).

Стаття надійшла до ред. 17.05.2018 р.