Estimation of Parameters in Regression Analysis Based on *QR* Decomposition of Rectangular Matrices by Householder Reflections

Oleksandr Dorokhov*[1], Lyudmyla Malyarets[2], Kadri Ukrainski[1] and Dmytro Yevstrat[3]
[1]Chair of Public Economics and Policy, University of Tarty Narva mnt 18, 51009, Tarty, Estonia
[2]Department of Higher Mathematics and Economic and Mathematical Methods, Simon Kuznets Kharkiv National University of Economics Nauka Ave 9A, 61166, Kharkiv, Ukraine
[3]Department of Information Systems, Simon Kuznets Kharkiv National University of Economics Nauka Ave 9A, 61166, Kharkiv, Ukraine
E-mail: oleksandr.dorokhov@ut.ee, malyarets@ukr.net, kadri.ukrainski@ut.ee, dmitryyevstrat@gmail.com
*Corresponding author

Keywords: regression analysis, Householder reflection, *QR* decomposition

Received: February 04, 2022

An approach to eliminate multicollinearity problems in regression analysis using QR decomposition of rectangular matrices by Householder reflection has been proposed. The reliability of this computational procedure has been proved.

Povzetek: V regresijski analizi je narejena ocena parametrov s pomočjo dekompozicije QR.

1 Introduction

In substantiating the decision taken in the management of various socio-economic systems, it is important to choose an analytical tool that analyzes the state of systems and predicts their further development.

In the presence of various kinds of uncertainties and a large amount of various data, in the problems of economic modeling, two main approaches are most often used recently: fuzzy modeling [1-5] and various statistical methods [6-8].

Among the last, most often, multiple regression analysis is used as such a tool [9-13].

Despite its prevalence in economics and duration in applications, many problems still need to be solved, since the existing algorithms for multiple regression analysis are far from perfect [14-17].

To solve many problems of mathematical methods in economics, the so-called *QR* decomposition of rectangular matrices is useful [18-21].

An $n \times m$ matrix X is decomposed into two factors $X = QR$, where Q is an $n \times n$ orthogonal matrix and R is an upper triangular $n \times m$ matrix with zero entries below the main diagonal.

Recall that the determinant of the orthogonal matrix is equal to one $|Q| = 1$; its inverse matrix coincides with the transposed $Q^{-1} = Q'$ (i.e. $QQ' = Q'Q = I$), and orthogonal transformation of vectors a, b does not change their scalar products: $(Qa,Qb) = (a,b)$; $|Qa| = |a| = a$; $|Qb| = |b| = b$.

It can be argued that numerical algorithms with orthogonal transformations do not introduce additional errors into the solution of the problem.

2 *QR* decomposition to solve the problem of multicollinearity

Consider how *QR* decomposition helps to overcome the problem of multicollinearity in regression analysis.

Suppose we need to find the best estimates of the parameters b of a linear three factor regression model from 6 observations ($n = 6$; $m = 3+1 = 4$):

$$Y = Xb + E,$$

which in expanded form reduces to solving the overvalue value of a system of 6 linear equations with respect to 4 unknowns (b_0, b_1, b_2, b_3):

$$\begin{aligned}
y_1 &= b_0 + b_1 \cdot x_{11} + b_2 \cdot x_{21} + b_3 \cdot x_{31} + e_1\\
y_2 &= b_0 + b_1 \cdot x_{12} + b_2 \cdot x_{22} + b_3 \cdot x_{32} + e_2\\
y_3 &= b_0 + b_1 \cdot x_{13} + b_2 \cdot x_{23} + b_3 \cdot x_{33} + e_3\\
y_4 &= b_0 + b_1 \cdot x_{14} + b_2 \cdot x_{24} + b_3 \cdot x_{34} + e_4\\
y_5 &= b_0 + b_1 \cdot x_{15} + b_2 \cdot x_{25} + b_3 \cdot x_{35} + e_5\\
y_6 &= b_0 + b_1 \cdot x_{16} + b_2 \cdot x_{26} + b_3 \cdot x_{36} + e_6
\end{aligned}$$

Here E – vector of errors (residuals e_i), which can be found after determining the estimates of the model parameters b_0, b_1, b_2, b_3.

If the *QR* decomposition of the matrix $X = QR$ is known, then the above problem is solved as follows.

Should multiply the matrix equation $Y = QRb + E$ leftward to the orthogonal matrix Q' and we obtain an equivalent equation $Z = Rb + \Xi$, where marked $Z = Q'Y$, $\Xi = Q'E$.

In expanded form we have a system of linear equations with a triangular matrix R:

$$z_1 = b_0 \cdot r_{01} + b_1 \cdot r_{11} + b_2 \cdot r_{21} + b_3 \cdot r_{31} + \xi_1$$
$$z_2 = b_1 \cdot r_{12} + b_2 \cdot r_{22} + b_3 \cdot r_{32} + \xi_2$$
$$z_3 = b_2 \cdot r_{23} + b_3 \cdot r_{33} + \xi_3$$
$$z_4 = b_3 \cdot r_{34} + \xi_4$$
$$z_5 = \xi_5$$
$$z_6 = \xi_6$$

Due to the orthogonality of the matrix transformation, the following relation is preserved $\|\Xi\| = \|E\|$, that is, the sum of the squares of the converted errors $\sum \xi_i^2$ (vector norm Ξ) is always equal to the sum of the squares of the residuals of the original system of equations $\sum e_i^2$ (vector norm E).

Due to the successful determination of the model parameters b_i you can equate to zero the first few components of the vector Ξ and obtain the minimum value of the sum of the squares of the residuals

$$\sum e_i^2 \to min\,.$$

So it is possible to obtain estimates of the parameters of the model by the least squares method (but in a slightly different, non-standard computational way, without first drawing up a system of normal equations).

If the last diagonal element of the triangular matrix R is nonzero $r_{34} \neq 0$, then you can equate to zero the maximum number (m) of the first components ξ_i and find estimates of the model parameters from the triangular system of equations.

This also determines the (minimum) sum of the squares of the residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=m+1}^{n} \xi_i^2 = \sum_{i=m+1}^{n} z_i^2\,.$$

If the last diagonal element of the triangular matrix R is exactly equal to zero $r_{34} = 0$, then this means that the last variable x_4 is not independent, but it is a linear combination of other argument variables $x_4 = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3$.

In this case, the estimate $b_4 = 0$ should be equated to zero.

That is, it is not necessary to include in the model a combination of already taken into account variables, and estimates of other parameters of the model should be obtained from the conditions of zero of the smaller number of the first components $\xi_i = 0$.

If the last diagonal element of the triangular matrix R is not exactly zero, but close to it $|r_{34}| \approx 0$, this means that there is a multicollinear relationship between the source variables x_i.

In this case, it makes sense to equate the score $\boldsymbol{b_4} = 0$ to zero and remove the questionable term from the model (otherwise an unstable solution with large errors will be obtained).

So, if we have a QR decomposition of the matrix X, the problem of multi-linearity is solved quite simply.

At the same time, the strong side of the QR decomposition is that it allows you to calculate (find a numerical) solution to the least squares problem.

As is well known, the classical, ordinary least squares method gives us a closed solution in the form of normal equations. But this solution is not always suitable for practical, specific applications.

Therefore, if you need to find the actual numerical solution, the least squares method is not suitable, at the same time, the QR decomposition easily gives such numerical values.

It should be noted that today the most well-known and used methods for obtaining an orthogonal matrix Q are the Gram-Schmidt process, the Householder transformation, and the Givens rotation.

An orthogonal matrix Q (and a triangular matrix R) can be obtained by successive operations with matrices $H_k = I - 2\omega_k\omega_k'$, where I – single matrix, ω_k – normalized vector ($\omega_k'\omega_k$=1), in which the first $(k\text{-}1)$ components are zero.

It's not hard to see that $H_k'H_k = H_kH_k = I$, that is, the matrix H_k is orthogonal. This matrix is also called as the Householder reflection matrix.

Any vector can be represented as $a = a_1\,\omega + a_2\,\varpi$, where ω, ϖ – orthonormal vectors ($\omega'\varpi = 0$; $\omega'\omega = 1$; $\varpi'\varpi = 1$).

Householder reflection $Ha = -a_1\,\omega + a_2\,\varpi$ changes the sign of the first component to the opposite.

We have to take into account some important properties of the Householder matrix: it is Hermitian $H=H^*$ and unitary $HH^*=I$, and therefore it is an involution $H^2=I$. In this case, the transformation $H_u(x)$ displays (reflects) point x to point $x\text{-}2(u, x)u$.

The Householder matrix has one eigenvalue equal to (-1), which corresponds to the eigenvector u, while all other eigenvalues are equal to (+1).

In this case, the determinant of the Householder matrix is (-1), and the Householder transformation in the metric space preserves distances.

3 Converting a rectangular matrix to upper triangular

Consider the process of sequential transformation of a rectangular matrix $A = [a_1, a_2, \ldots, a_m]$ to the upper triangular shape (possibly with permutations of columns).

The first transformation with a matrix H_1 must transform the first vector a_1 (the first column of the matrix A) to vector $\pm a_1\,e_1$, where e_1 a coordinate vector in which only one first element is nonzero (this element is equal to 1, the other elements are zero); as a_1 is marked length of the vector a_1 (because the orthogonal transformation does not change the length of the vector).

Consider the approach of how to find a vector ω_1, which defines the whole matrix H_1.

To do this, consider that again (re-)reflection restores vector a_1:

$a_1 = H_1 (\pm a_1 e_1) = (I - 2\omega_1\omega_1')(\pm a_1 e_1) = \pm a_1 e_1 + 2a_1\omega_{11}\omega_1$,

hence the vector ω_1 proportional to the vector a_1, to the first component of which a value $\pm a_1$ is added; sign (+ or –) should be selected by the sign of the element a_{11}.

So we get a (still normalized) vector Ω_1 in the form $\Omega_1 = a_1 + a_1 \cdot \mathrm{sgn}(a_{11}) \cdot e_1$.

The square of the length of this vector is equal to $\Omega_1'\Omega_1 = 2a_1 (a_1 + |a_{11}|)$.

Thus the first Hausholder matrix is constructed $H_1 = I - \frac{\Omega_1\Omega_1'}{a_1(a_1+|a_{11}|)}$.

In the transformed matrix A in the first column will have only one non-zero element: $H_1 a_1 = a_1 - \Omega_1 = -a_1 \cdot \mathrm{sgn}(a_{11}) \cdot e_1$.

Other converted columns of matrix A (and the dependent variable column Y) have the form: $H_1 a_j = a_j - \lambda_j\Omega_1$, where λ_j – numerical coefficients $\lambda_j = \frac{(a_1,a_j)+a_1 \cdot sgn(a_{11}) \cdot a_{j1}}{a_1(a_1+|a_{11}|)}$

The first elements of the converted columns will no longer change in subsequent ones Householder reflections.

Without these first elements of the norms of all vectors (columns of the transformed matrix A) are reduced on the value a_{jk}^2 (by now $k = 1$).

Let find the column a_q with the highest residual norm (relative to the original norm):

$$\frac{a_q^2 - a_{qk}^2}{a_q^2} = \max\left(\frac{a_j^2 - a_{jk}^2}{a_j^2}\right).$$

If the largest relative residual norm is less than some limit value (for instance, 0,01), further transformations are canceled, the process ends prematurely due to the detection of multicollinear connections.

In the second stage we find the matrix H_2, which transform vector a_q (without first component) to vector $\pm(a_q)^* e_2$, where e_2 – the second coordinate vector; by $(a_q)^*$ marked shortened vector a_q length without its first component.

Thus, after the second transformation, the vector a_q will have two non-zero components.

Then again we find the residual norms and determine the next vector a_p, which (without the first two components) will be converted to $\pm(a_p)^* e_3$.

The process will end after m iterations, or early if the residual norms become less than accepted limit level.

Usually in regression analysis the first column of the matrix X (matrix A) there is a column of ones $x_0 \equiv 1$ (to take into account in the regression model the obligatory free term).

Therefore, at the first stage (of Householder reflections), it is the first The order of selection of the following columns for conversion is determined by the values of their relative residual norms.

4 An example of the implementation of the algorithm

The described above algorithm was implemented in an electronic spreadsheet *Excel* in the macros form on the *VBA* language (*Visual Basic for Application*).

Consider numerical example of the analysis of the regression dependence of the resultant feature on 5-factors, the corresponding initial data for this are presented in Table.1.

Table 1. Data for regression analysis

№	y	x_0	x_1	x_2	x_3	x_4	x_5
1	1075,3	1	32,06	17,9	12,08	35,61	8,33
2	1002,7	1	27,57	10,23	14,06	37,48	10,63
3	995,6	1	27,88	10,29	11,26	37,77	12,72
4	872,4	1	31,65	11,72	7,32	34,98	14,25
5	909,3	1	34,81	12,64	1,68	39,04	11,75
6	1009,9	1	29,47	10,87	1,31	46,14	12,15
7	919,7	1	34,42	12,77	1,28	41,04	10,48
8	876,2	1	32,76	12,26	1,06	42,53	11,32
9	908,6	1	31,24	11,65	4,49	41,27	11,28
10	935,8	1	30,4	11,33	6,88	40,07	11,26
11	949,9	1	29,96	11,18	8,84	39,48	10,5
12	927,4	1	30,49	11,41	7,73	39,55	10,78
13	1003,9	1	29,71	11,05	13,08	29,46	16,68
14	1017,6	1	29,02	10,79	14,34	29,06	16,77
15	997,6	1	29,55	10,99	11,75	30,07	17,63
16	958,2	1	30,79	11,44	7,94	31,29	18,54
17	907,5	1	32,55	12,08	1,45	33,03	20,87
18	928	1	33,27	12,35	1,41	32,32	20,63
19	930,1	1	35,34	13,42	0,76	32,24	18,23
20	892,6	1	33,71	12,79	0,78	33,58	19,13
21	917,7	1	32,3	12,03	4,26	32,68	18,7
22	947,2	1	31,32	11,64	6,99	31,65	18,38
23	959,6	1	30,97	11,55	8,94	31,24	17,3
24	943,4	1	31,52	11,7	7,63	31,64	17,5

By using Householder reflections matrix X was transformed into a triangular form (to matrix $R = Q'X$).

The process of transformation ended prematurely.

Variables x_1, x_4 was not connected in the model since their residual norms decreased to values of about 0.1% of the original norms as shown in Table 2 and Table 3.

Let move converted columns x_1 and x_4 (with small relative residual rates less than 0.01) to the left to the column Z.

As result, we can obtain a system of 4 equations with 3 columns of free terms (Z, x_1 and x_4) relative to 4 parameters of the considered model b_0, b_2, b_3, b_5 as shown in Table 4.

Table 2: Data after householder reflections

№	$z = Q'y$	x_0	x_1	x_2	x_3	x_4	x_5
1	-4651,2	-4,8	-153,6	-58,3	-32,1	-174,1	-72,6
2	-168,7		7,6	1,4	-22,2	8,3	3,2
3	-18,0		0,7	-1,4		-19,2	18,2
4	62,9		4,5	6,8		-6,2	
5	-3,9		2,0			-2,5	
6	135,5		-2,4			3,4	
7	5,7		1,6			-2,1	
8	-24,3		0,1			-0,2	
9	-12,4		-0,1			0,1	
10	-1,6		-0,0			-0,0	
11	-4,2		0,2			-0,4	
12	-20,3		0,3			-0,5	
13	21,9		0,5			-0,7	
14	28,9		0,3			-0,4	
15	30,8		-0,1			0,1	
16	20,0		-0,3			0,5	
17	21,8		-1,1			1,7	
18	37,0		-0,6			0,8	
19	20,4		1,0			-1,6	
20	-3,4		-0,3			0,3	
21	3,0		-0,2			0,4	
22	13,8		-0,2			0,3	
23	7,9		0,1			-0,2	
24	1,4		0,2			-0,3	

Table 3: Norms

Output	21691,3	24	23707,1	3461,7	1524,9	30845,7	5617,9
Residual	24827,4	0	16,928	0	0	31,7126	0
Relative	0,00114	0	0,00071	0	0	0,00102	0

Table 4: System of equations with a triangular matrix

y	x_1	x_4	x_0	x_2	x_3	x_5
-4651,2	-153,6	-174,1	-4,8	-58,3	-32,1	-72,6
-168,7	7,6	8,3		1,4	-22,2	3,2
-18,0	0,7	-19,2		-1,4		18,2
62,9	4,5	-6,2		6,8		

The solution to this system is given in Table 5. Explanatory variables are shown in the model x_2, x_3, x_5. Other variables (including y) are expressed through these explanatory variables.

Table 5: Inverse matrix and system solution

Model parameters			**Inverse matrix**			
y	x_1	x_4	x_0	x_2	x_3	x_5
790,78	24,03	67,07	-0,20	0,29	-0,86	-1,97
9,14	0,65	-0,90				0,14
8,14	-0,28	-0,60		-0,04	0,01	0,01
-0,25	0,09	-1,13			0,05	0,01

5 The results of calculations

As a result, the following regression models can be obtained (together with the coefficients of determination):

$y = 790{,}78 + 9{,}14 \cdot x_2 + 8{,}14 \cdot x_3 - 0{,}25 \cdot x_5$; $R^2 = 0{,}56$

$x_1 = 24{,}03 + 0{,}65 \cdot x_2 - 0{,}28 \cdot x_3 + 0{,}09 \cdot x_5$; $R^2 = 0{,}82$

$x_4 = 67{,}07 - 0{,}90 \cdot x_2 - 0{,}60 \cdot x_3 - 1{,}13 \cdot x_5$; $R^2 = 0{,}93$

Coefficients of determination R^2 can be calculated, for example, as follows:

$$R^2 = 1 - \frac{||\boldsymbol{e}||}{||\boldsymbol{y}|| - N \cdot (\bar{y})^2} = 1 - 0{,}4312 = 0{,}5688$$

So we have multicollinear relationships of variables x_1 and x_4 with explanatory variables x_2, x_3, x_5 and with multiple correlation coefficients $R_1 = \sqrt{0{,}8251} = 0{,}9084$ and $R_4 = \sqrt{0{,}9382} = 0{,}9686$, which exceeds the closeness of the relationship of the explanatory variables with the resultant feature $R_y = \sqrt{0{,}5686} = 0{,}7542$.

Note that the matrix of paired correlation coefficients r_{xy} does not show (not demonstrate) any effect of multicollinearity at once.

But it turns out that the determinant of this correlation matrix is almost zero $|r_{xy}| = 0{,}0000656$, that is, all explanatory variables are interconnected by a precise multicollinear relationship as shown in Table 6.

So, with the help of QR decomposition of X rectangular matrix it is possible (without first compiling a system of normal equations) to obtain least square methods' estimates of the parameters of the regression model together with all multicollinear connections.

Table 6: Correlation matrix

r_{xy}	x_1	x_2	x_3	x_4	X_5	y
x_1	1	0,5817	-0,7796	-0,0065	0,2102	**-0,5863**
x_2	0,5817	1	-0,2002	-0,0160	-0,1664	**0,1259**
x_3	-0,7796	-0,2002	1	-0,3694	-0,1748	**0,7031**
x_4	-0,0065	-0,0160	-0,3694	1	-0,7743	**-0,1736**
x_5	0,2102	-0,1664	-0,1748	-0,7743	1	**-0,1969**

6 Analysis of the stability of the obtained results

We turn to consider a very important problem of the stability of the results obtained from some particular observations, which have an excessive effect on the values of the model parameters. Using a matrix $\boldsymbol{Q}$ you can detect all such questionable observations, such as observation №1 in the above tables.

If you delete the observation №1, the values of the model parameters will change significantly:

With observation №1:

$y = 790{,}78422 + 9{,}14172 \cdot x_2 + 8{,}14859 \cdot x_3 - 0{,}25244 \cdot x_5$;

Without observation №1:

$y = 1300{,}5080 - 34{,}4492 \cdot x_2 + 1{,}75898 \cdot x_3 + 2{,}25143 \cdot x_5$.

This unpleasant effect is shown in the graphs of component effects in Figure 1 (with), Figure 2 (without).

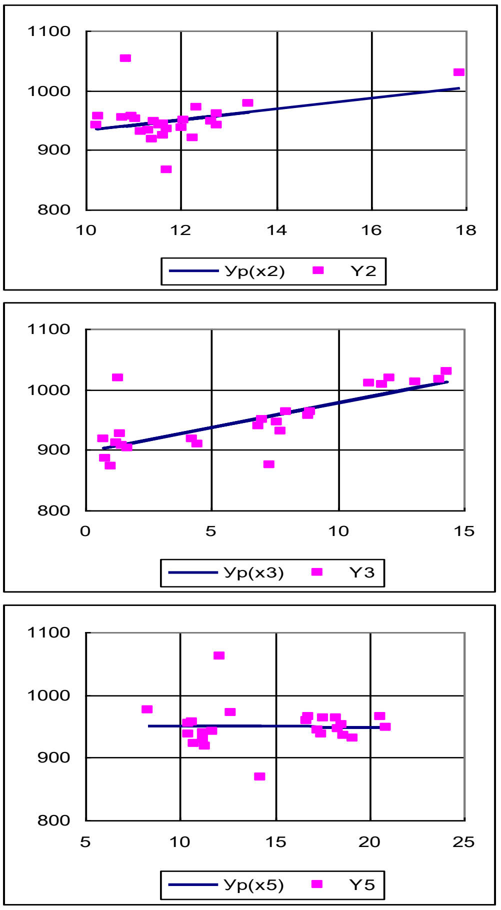
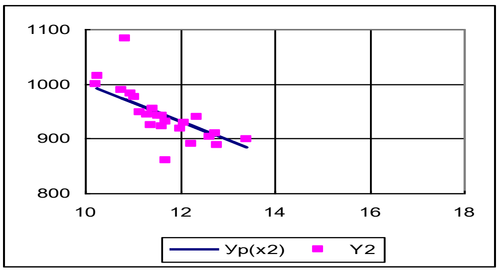


Figure 1: Along with observation № 1.

On the graphs empirical points are superimposed on theoretical regression lines (calculated values). It helps to find out which dependencies are significant and which are not. Indeed, it is always possible to choose such scales of coordinate axes that all theoretical lines will have the same slope in 45°.

The presence of empirical points (or 95% confidence bands) will not allow incorrect conclusions in this case.

However, the question arises, how to determine such empirical points so that only one variable varies on each graph, and the rest is fixed at the average levels?



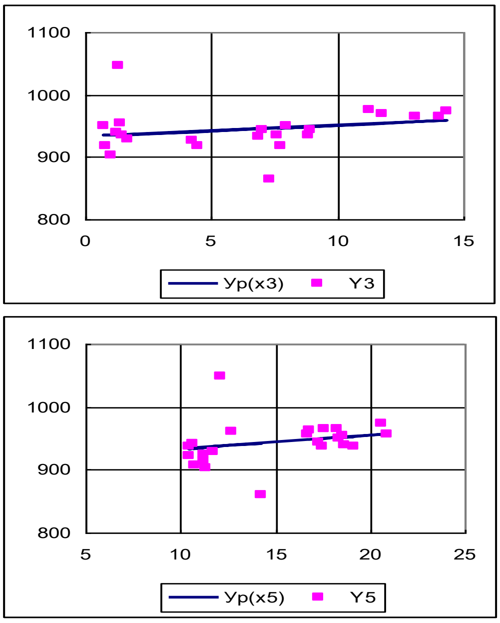


Figure 2: Without observation № 1.

It turns out that this is a problem whose solution is insufficiently covered in the literature. Therefore, we propose such a solution.

After determining the parameters of the model, you can find the deviation of each observation from the theoretical values (residuals): $e = y - y_p$.

Now, for each observation, you can write down the identity:

$$y_i = \bar{y} + b_2(x_{2i} - \bar{x}_2) + b_3(x_{3i} - \bar{x}_3) + b_5(x_{5i} - \bar{x}_5) + e_i$$

Members $b_k(x_k - \bar{x}_k)$ are called "component effects".

Sum $\bar{y}$ and the corresponding component effect is the equation of the theoretical dependence of the resultant feature y on one variable x_k at the average values of the remaining explanatory variables.

We will fix in identities all variables (except one) on average values and we will receive the corrected data as the sum of the total mean $\bar{y}$, the corresponding component effect $b_k(x_k - \bar{x}_k)$ and the residual:

$$Y_2 = Y_p(x_2) + e = \bar{y} + b_2(x_2 - \bar{x}_2) + e;$$
$$Y_3 = Y_p(x_3) + e = \bar{y} + b_3(x_3 - \bar{x}_3) + e;$$
$$Y_5 = Y_p(x_5) + e = \bar{y} + b_5(x_5 - \bar{x}_5) + e.$$

7 Conclusions and further research

Since the regression analysis does not end only with the assessment of the parameters of the model, it is necessary to identify all the most influential observations and assess their negative contribution.

Modern mathematical theory offers methods for identifying such components and assessing their acceptability in the data. These methods are based on the previous *QR* decomposition of the data matrix.

The extra-large amount of computing work is no longer an obstacle in the presence of modern computers.

Comparing the results of calculations of multifactor regression analysis by the method of *QR* decomposition of rectangular matrices by House-Holder mappings with the calculations performed in the *StatGraphics* package, we obtained a 100% match.

Checking the significance of the parameters of the equations by the *student's* criterion shows that not all of them are significant:

$$y = 899{,}649 + 7{,}59366 * x_3;$$
$$x_1 = 26{,}1328 + 0{,}60714 * x_2 - 0{,}305855 * x_3;$$
$$x_4 = 67{,}0703 - 0{,}906909 * x_2 - 0{,}600053 * x_3 - 1{,}13156 * x_5.$$

In the following, we can consider how the procedures for estimating the parameters of the equation can be performed in the *StatGraphics* package.

Summarizing all the above, as a conclusion, we can say that the use of *QR* decomposition matrices has significant advantages over the standard procedure of the least squares' method in the presence of multicollinearity of data and is a reliable computational procedure.

References

[1] Dorokhov, O., Dorokhova, L. (2011). Fuzzy model in fuzzy-tech environment for the evaluation of transportation's quality for cargo enterprises in Ukraine. *Transport and Telecommunication*. 12(1):25–33

[2] Dorokhova, L., Dorokhov, O. (2017). Computer fuzzy model regarding pharmacies integral perceptions by visitors. *Bulletin of the Transilvania University of Brasov, Series III: Mathematics, Informatics, Physics*. 10(2):155–170 https://doi.org/10.31926/but.mif.2019.12.61.2.23

[3] Omelchenko, O., Dorokhov, O., Kolodiziev, O., Dorokhova, L. (2018). Fuzzy modeling of the creditworthiness assessments of bank's potential borrowers in Ukraine. *Ikonomicheski Izsledvania*. 27(4):100–125

[4] Susanto, S., Utama, D. (2022). Fuzzy based decision support model for health insurance claim. *Informatica*. 46:119–130 https://doi.org/10.31449/inf.v46i7.4325

[5] Narang, M., Joshi, M., Pal, A. (2022). A hesitant fuzzy multiplicative base-criterion multi-criteria group decision making method. *Informatica*. 46:235–242 https://doi.org/10.31449/inf.v46i2.3452

[6] Malyarets, L., Kovaleva, K., Lebedeva, I., Misiura, I., Dorokhov, O. (2018). The Heteroskedasticity Test Implementation for Linear Regression Model Using MATLAB. *Informatica*. 42(4):545–553 https://doi.org/10.31449/inf.v42i4.1862

[7] Vigneau, E. (2021). Clustering of variables for enhanced interpretability of predictive models. *Informatica*. 45:507–516 https://doi.org/10.31449/inf.v45i4.3283

[8] Malyaretz, L., Dorokhov, O., Dorokhova, L. (2018). Method of constructing the fuzzy regression model of bank competitiveness. *Journal of Central Banking Theory and Practice*. 7(2):139–164 https://doi.org/10.2478/jcbtp-2018-0016

[9] Ishikawa, A., Fujimoto, S., Mizuno, T. (2021). Why does production function take the Cobb–Douglas form? *Evolutionary and Institutional Economics Review*. 18(1):79-102. https://doi.org/10.1007/s40844-020-00180-3

[10] Yang, Y., Schmidt, P. (2021). An econometric approach to the estimation of multi-level models. *Journal of Econometrics.* 220(2):532-543. https://doi.org/10.1016/j.jeconom.2020.04.012

[11] Liu, L., Fan, T. (2021). The Impact of Identity on Satisfaction toward Entrepreneurial Environment: A Case Study. *Discrete Dynamics in Nature and Society*. 2021:1-8. https://doi.org/10.1155/2021/7234635

[12] Borisov, I. (2021). Agricultural personnel and rural labour: How is their reproduction related? *Journal of New Economy*. 22(3):161-183. https://doi.org/10.29141/2658-5081-2021-22-3-9

[13] Varbanova, V., Beutels, P. (2020). Recent quantitative research on determinants of health in high income countries: A scoping review. *PLOS ONE*. 15(9):1-21. https://doi.org/10.1371/journal.pone.0239031

[14] Wang, W., Zhou, Y. (2021). Eigenvector-based sparse canonical correlation analysis: Fast computation for estimation of multiple canonical vectors. *Journal of Multivariate Analysis*. 185(C): 30-41. https://doi.org/10.1016/j.jmva.2021.104781

[15] Cheng, H. (2021). Importance sampling imputation algorithms in quantile regression with their application in CGSS data. *Mathematics and Computers in Simulation (MATCOM)*. 188(C):498-508. https://doi.org/10.1016/j.matcom.2021.04.014

[16] Li, T., Song, X., Zhang, Y., Zhu, H., Zhu, Z. (2021). Clusterwise functional linear regression models. *Computational Statistics & Data Analysis*. 158(C): 101-112. https://doi.org/10.1016/j.csda.2021.107192

[17] Jian, D., Yu, C., Aihua, G., Jingwei, C., Lili, L., Qinling, C., Shujun, L., Qifeng, X. (2020). Potential Trend for Online Shopping Data Based on the Linear Regression and Sentiment Analysis. *Mathematical Problems in Engineering*. 2020:1-11. https://doi.org/10.1155/2020/4591260

[18] Bagnato, L., Punzo, A. (2021). Unconstrained representation of orthogonal matrices with application to common principal components. *Computational Statistics*. 36(2):1177-1195. https://doi.org/10.1007/s00180-020-01041-8

[19] Wei, C., Bo, Z. (2021). Controllability of Flow-Conservation Transportation Networks with Fractional-Order Dynamics. *Complexity*. 2021:1-13. https://doi.org/10.1155/2021/8524984

[20] Wireko, F., Barnes, B., Sebil, C., Ackora-Prah, J. (2021). The Eigenspace Spectral Regularization Method for Solving Discrete Ill-Posed Systems. *Journal of Applied Mathematics*. 2021:1-11. https://doi.org/10.22541/au.161212952.20018992/v1

[21] Jiayan, W., Lanlan, G., Zongmin, L., Xueqin, W., Zhengqing, F. (2021). An Efficient Algorithm for Ill-Conditioned Separable Nonlinear Least Squares. *Journal of Mathematics.* 2021:1-8.