Analysis of the multicollinear econometric model parameters with a rank deficient observation matrix

Viktor Kutovyi[1], Olga Katunina[2], Oleg Shutovskyi[3]

[1]Kyiv National Economic University named after Vadym Hetman
Peremohy avenue 54/1, Kyiv, Ukraine, 03057
skutova@ukr.net, orcid.org/0000-0001-7584-0037
[2]prommet@ukr.net, orcid.org/0000-0001-7584-0037
[3]Kiev National University of Civil Engineering and Architecture
Povitroflotsky avenue 31, Kyiv, Ukraine, 03037
shutovsk@ukr.net, orcid.org/0000-0003-2709-2059

Abstract. The topic of determining informative predictors, forming rational exogenous variables, substantiating the dimension and structure of predictor spaces is considered. The purpose of design and selection of characteristics is to prevent the effect of retraining, reduce the dimension in studying the processes apart from a master, build classifiers, reflect the process of dividing data into classes and determine the boundaries of solutions in limited space, as well as reasonable interpretation, provide in-depth understanding of the model and data for studying, visualization in spaces, the dimension of which is perceived by the researcher. The design predictor spaces and develop effective procedures problems for estimating the parameters of econometric models with multicollinear variables are developed. The study was made under alternative approaches to form the interdependencies models features.

A mathematical toolkit is proposed for calculating the parameters of a linear econometric model in case of rank deficient observation matrix, based on the study of singular expansions.

Using a singular toolkit for decomposing and analyzing the data matrix makes it possible to increase the operational efficiency and predictive quality of the procedures for estimating econometric models parameters. The mathematical approach to the construction of models of the interdependence of factors is intended to select characteristics and construct predictor spaces in the study of systems with multicollinear variables and rank deficient observation matrix.

Victor Kutovyi
Department of Economics and Mathematical Modeling
PhD, Ass. Prof.

Olga Katunina
Department of Economics and Mathematical Modeling
PhD, Ass. Prof.

Oleg Shutovskiy
Department of Information Technologies for Design and Applied Mathematics
PhD, Ass. Prof.

INTRODUCTION

The multicollinearity problem is relevant both for generalizing the factors interdependence of the complex systems simulation theoretical principles, and in terms of the operational aspect of the research. Applied modeling and mathematical description of cause and effect relationships as a form of regression dependencies is a leading, effective and practically demanded direction in a wide range of applications for solving various behavior analyses of technical, socio-economic, medical and biological systems problems. A model tool for identifying and formalizing interdependencies is used to solve the management problems of such systems and their components, in particular, for functioning regularities determining, decision-making support on determining the optimal parameters and operating modes, for their development trends forecast, as well as for justifying the strategies and scenarios of their behavior in the condition of uncertainty, indeterminacy, conflict, and, as a consequence, significant risk of external environment.

In mathematical model sciences dealing with the interdependence of factors, for example, engineering, study of operations, econometrics, biometrics etc., classical are considered the works by D. Farrar, R. Glauber, J. Johnston, M. Kendall, L. Klein, M. Bartlett, H. Tail, J. Meier, G. Orkat. The scientific works of these authors include fundamental results on the problems of estimating the parameters of regression models, in which, in particular, methodological approaches and tools for taking into account the phenomenon of multicollinearity in the input data are developed.

Thus, the work of D. Farrar and R. Glauber [17] on multicollinearity detection, diagnosis, evaluation, consideration and the subsequent correction of the design procedures for prediction space in order to eliminate its negative consequences is well known in the wide range of engineering, natural systems researchers, ecologists and econometricians. This work was completed in 1965 in cooperation with the Institute for Naval Research, which developed theoretical approaches and mathematical models to analyze the impact of the combination of factors on the maintenance condition and the cost characteristics functioning systems formation of the Navy. To the set of exogenous factors the researchers included, in particular, the age of the vessel and its dimensions, the intensity of use in marine logistics and defense projects, fuel consumption, the time between the scheduled maintenance and their cost.

Also the set of predictors includes the discrete characteristics of the system, such as the type of power plant (steam, diesel, nuclear fuel), engine operation, equipment operation and maintenance complexity, as well as a number of functional and security subsystems qualitative parameters for radar patrol, guided missiles etc. The spatial samples for the data studied were various naval purpose objects, in particular, the destroyers of the Atlantic Fleet and others.

Currently, the space of tools and technologies choice for estimating the models parameters of interdependence of factors is expanding intensively. For example, the researchers of the applied natural science systems use mostly OLS (Ordinary Least Squares), implemented in most traditional statistical packages, econometricians mainly focus on the maximum likelihood approaches, and the latest intellectual data analysis and machine study professionals use the wide range of regression models regularization.

It is notable that all these approaches are procedurally oriented and are chosen by the researchers according to prior arguments. At the same time the decision to choose the best method for analyzing the interdependence of the factors is practically not related to the specifics of the initial data.

Today the vast majority of regression tools application areas recognize the almost universal presence of multicollinearity, which is established as an indispensable characteristic of the data base. Modeling technologies for parameter estimation offer an advanced tool for diagnosing, taking into account and eliminating the negative effects of this phenomenon, in particular, lasso methods, combing regression, dimension diminishing, selection of optimal subset of variables etc.

It is important that in case of the implementation of the Big Data concept and digital economy technologies in reaching the broadest scope of human activity and the involvement of vast amounts of predominantly unstructured information, the relevance of the problem of multicollinearity will increase. Thus, the development of mathematical procedures for in-depth study of the mechanisms of interdependencies within the predictor spaces has not only an applied demand for forecasting the development of complex systems but also a significant operational relevance in terms of ensuring the adequacy of the constructed models and observing the target properties of the parameters estimates.

Here are the results of generalized approaches to the estimation of econometric models parameters with the multicollinearity and the rank deficient observation matrix, which were previously studied by the authors in a number of scientific works [5 − 7].

1. Let there be a linear relationship between the variable Y and m explaining the variables $X_1, X_2, ..., X_m$ and disturbance ε, ε − the random variable, we emphasize that we need only the existence of finite moments of the second order.

METHODOLOGY

If we have a sample of n observations on the modifications of Y i $X_j, j = 1, 2, ...m,$ then we can write

$$y_i = \sum_{j=1}^{m} x_{ij}\beta_j + \varepsilon_i, i = \overline{(1, n)}. \quad (1)$$

Equations (1) can be written in the matrix form

$$Y = X\beta + \varepsilon, \quad (2)$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Through X' and ε' we denote the matrices transposed to X and ε, respectively.

Let the conditions be:

1. $M\varepsilon = 0;$ (3)
2. $M(\varepsilon\varepsilon') = \sigma^2 E, -$ unit matrix;   (4)
3. $X -$ matrix whose elements are deterministic numbers   (5)
4. X range = m (matrix $- X$ full rank)   (6)

The task $(1 − 5)$ was considered by many authors [1, 2]. First of all, we give a very understandable version in the book of Wetherburg [13] and Davies [11]. Especially useful are the books of Plackett [12, 14] and Voyevodin [3].

The case when the condition (3) is not fulfilled was considered in [9] with $M\varepsilon \neq 0$. There is some compromise between the bias value $\hat{\beta}$ and value $D(\hat{\beta})$. Case when [2] is not fulfilled was considered by Aitken [10], who proposed a generalized least squares method on the condition that matrix X is of the full rank. In this study Aitken's method extends to task (1, 2), provided that (6) is not fulfilled, but takes place

$$rang X = t < m \quad (6')$$

at the same time takes place

$$M(\varepsilon\varepsilon') = D = \sigma^2 W, \quad (4')$$

where σ^2 is an unknown parameter, and D, w − are known symmetrically positive defined matrices of order $n \times n$. Then D admits the representation $D = PP'$, where the non-developed matrix P is positively determined. So $D = PP'$ that $P^{-1}DP^{-1} = E$ and $P^{-1}P^{-1} = D^{-1}$. Let's denote $P^{-1} = B$. D commutes with B.

From $(4)'$ follows

$$W^{-1} = D^{-1}\sigma^2, \ W^{-1} = \overline{B}\overline{B} = \sigma^2 BB.$$

So the matrix of covariance

$$M(\varepsilon\varepsilon') = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\sigma_{12} & \cdots & \sigma_1\sigma_n\sigma_{1n} \\ \sigma_1\sigma_2\sigma_{12} & \sigma_2^2 & \cdots & \sigma_2\sigma_n\sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1\sigma_n\sigma_{1n} & \sigma_2\sigma_n\sigma_{2n} & \cdots & \sigma_n^2 \end{bmatrix} =$$

$$= \sigma^2 W,$$

where $\sigma_i^2 = M\varepsilon_i^2 = Dy_i -$ dispersion y_i, $\sigma_i\sigma_j\rho_{ij} = M\varepsilon_i\varepsilon_j = \operatorname{cov} y_i y_j -$ covariance y_i and y_j. $W -$ the known weight matrix.

Let the vectors of the basis $e_1, e_2, ..., e_m \in H_1 -$ the linear space H_1 set in certain matching vectors $f_1, f_2, ..., f_m$ of linear space H_2. Then there exists a linear operator X operating with H_1 in H_2, which translates each vector e_k into a corresponding one vector $Xe_k = \sum_{i=1}^{n} f_i\alpha_{ki}$. We can determine the operator X by equality

$$X\beta = X\sum_k \beta_k e_k = \sum_k \beta_k Xe_k = \\ \sum_k \beta_k \sum_i \alpha_{ki} f_i = \sum_i f_i \sum_k \alpha_{ki}\beta_k. \tag{7}$$

Between the linear operators X and the matrices $\|\alpha_{ki}\|$ there is a one-to-one correspondence. Since the symbols and properties of operations over matrices and operators coincide, then any transformation of operator equality leads to the same matrix equality. Therefore, from a formal point of view, we do not care whether we are dealing with matrix or operator relations.

Let's consider rectangular matrices of $n \times m$ dimensions, the rank of which coincides with the minimum from numbers m, n. Such matrices are called full rank matrices. Their characteristic feature is that they do not change their rank under any small disturbances. Systems of linear algebraic equations with full

rank matrices have much in common with the non-degenerate matrices systems.

Let the matrix X of the system $(1 - 6)$ be a full rank matrix. In this case the system may be incompatible. Otherwise, it's always compatible, but it has not the only solution.

We will look for a normal pseudo-connection of the system $(1 - 6)$, that is such vector $\hat{\beta}$, which among all the vectors minimizing the functional of the discrepancy $F_0(\beta) = \|X\beta - Y\|^2$ has the smallest Euclidean norm.

The invariance of the Euclidean norm to unitary transformations allows us to reduce the problem of finding normal pseudo-connection of the general type system to a simpler task.

It is easy to verify that the task of determining normal system pseudo-connection (1) is equivalent to the solution of the same problem for another system. But transformations can always ensure that the matrix is sufficiently simple, for example, triangular, normalized, trapezoidal etc. On the construct rather effective numerical methods.

If any of the methods does not provide the necessary accuracy of the solution of linear algebraic equations system, then there is no reason to hope that another method will provide the same system with better results. Probably such a system can be regarded as unstable. It is known $[1 - 4]$ that the pseudo solution of a redefined system (2) with a full rank matrix is a common solution of the system

$$X'X\hat{\beta} = X'y,$$

with a square non degenerate matrix $X'X$ of order $m \times m$.

Normal solution of the system

$$\hat{\beta} = (X'X)^{-1}X'y.$$

It is necessary to demand

$$\sum_{i=1}^{m} \beta_i^2 \to \min. \tag{8}$$

Definition 1. The matrix $X^+(m \times n)$ is called the Moore-Penrose pseudo inverse for matrix X if it satisfies the following four con-

ditions:

1 $X^+ X X^+ = X^+$
2 $X X^+ X = X$
3 $X X^+$ − symmetric,
4 $X^+ X$ − symmetric.   (9)

From (9) condition follows

$$X^+ X X^+ = X^+ \Rightarrow X^+ X X^+ X = X^+ X,$$

if $X^+ X = P_1$, then $P_1^2 = P_1$.

In addition, from condition 4 P_1 is symmetric. So P_1 − orthoprojector. Similarly $X X^+ = P_2$ − orthoprojector.

It can be proved that such a matrix X^+ always exists and is unique [2]. If X is a nondegenerate square matrix, then it obviously satisfies the conditions (1 − 4), if X is rectangular and has a full rank, then $X^+ = (X'X)^{-1} X'$. It can be verified that the pseudo inverse to diagonal $n \times m$ matrix

$$\sum = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \cdots & \cdots & \sigma_m \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

is a diagonal $m \times n$ matrix.

$$\sum{}^+ = \begin{bmatrix} \sigma_1' & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2' & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_m' & 0 & \cdots & 0 \end{bmatrix}, \quad (10)$$

where $\sigma_1' = \begin{cases} \dfrac{1}{\sigma_1}, & \sigma_1 \neq 0 \\ \\ 0, & \sigma_1 = 0 \end{cases}$.

Further we use [3] a singular matrix X [3]

$$X = U \sum V', \qquad (11)$$

where U is the orthogonal $n \times n$ matrix, V' is the orthogonal $m \times m$ matrix, and $\sum$ − the diagonal matrix, at which $\sigma_{ij} = 0$ for, $i \neq j$, $\sigma_{ij} = \sigma_i \geq 0$. Columns of the matrix U are the intrinsic vectors of the matrix XX', and the columns of the matrix V are the vectors of the matrix XX'. Using (9), [2, 3] we obtain

$$X^+ = V \sum{}^+ U'.$$

RESULTS

The estimates of the least squares of the parameter β in (1) are defined as values $\beta_1, \beta_2, \ldots, \beta_m$, minimizing

$$L = \sum_{i=1}^n \sum_{k=1}^n (y_i - \sum_j x_{ij} \beta_j)(y_k - \sum_j x_{kj} \beta_j) \alpha_{ik} \tag{12}$$
$$\to \min_{(\beta)},$$

where the matrix $A = \| \alpha_{ik} \|$ is a symmetric positive definite matrix.

Solution (12) $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_m$ we will call task pseudo-solution (1 − 2). The solution will be linear towards y. In addition, under the condition (3), $P_1 \widehat{\beta}$ will be unbiased estimate $P_1 \beta$ in (1). That is

$$M(P_1 \widehat{\beta}) = P_1 \beta. \tag{13}$$

The solution in general will not be the only one. We will require that the minimum was the amount (8). Then the solution (12) is unambiguous.

2. Let be given two linear real Euclidean spaces H_1, H_2 and a linear operator X operating with H_1 in H_2. Operator X^*, acting with H_2 in H_1, conjugates to X. Then for arbitrary vectors $\beta \in H_1$ and $y \in H_2$ equality is per-

formed

$$(X, \beta, y) = (\beta, X^* y). \qquad (14)$$

For any arbitrary operator X there is a conjugate operator X^*, and thus it is only one. We choose in H_1 an orthonormal basis $e_1, e_2, \ldots, e_m$. Then for every one $\beta \in H_1$ there is a schedule

$$\beta = \sum_{k=1}^{m} (\beta e_k) e_k. \qquad (15)$$

Applying (15) to $X^* y$ we get

$$X^* y = \sum_{k=1}^{m} (X^* y, e_k) e_k.$$

Then, taking into account (14)

$$X^* y = \sum_{k=1}^{m} (y, X e_k) e_k.$$

Definition 1. The linear operator B acting in H_2 is called self-directed (symmetric), if for any $y_1, y_2 \in H_2 \Rightarrow (By_1, y_2) = (y_1, By_2)$

Definition 2. If, $(By, y) > 0$ for arbitrary $y \in H_2$, then the operator is positively definite. When $(By, y) \geq 0$, then operator B is called positive.

Self-directed operator B corresponds to a symmetric matrix $B' = B$.

Let's define the operators $X^* BBX$, $BXX^* B$ that act accordingly in H_1, H_2.

Operators $X^* BBX, BXX^* B$ are self-directed according to H_1, H_2. It is $X^* BBX$ positive in H_1, and the operator $BXX^* B$ is positive in H_2.

In fact, for any $\beta \in H_1$, $y \in H_2$ we have

$$(X^* BBX \beta, \beta) = (BX\beta, BX\beta) \geq 0$$
$$(BXX^* By, y) = (X^* By, X^* By) \geq 0.$$

Operators $X^* BBX$, $BXX^* B$ are self-directed, therefore they have a system of orthonormal vectors $e_1, e_2, \ldots, e_m$, which are their own vectors for $X^* BBX$ and $f_1, f_2, \ldots, f_n$. They are the vectors for $BXX^* B$.

The operator BX translates the system $e_1, e_2, \ldots, e_m$ into some orthogonal system $BXe_1, BXe_2, \ldots, BXe_m$.

Indeed,

$$(BXe_i, BXe_j) = (X^* BBXe_i, e_j) =$$
$$(\rho_i^2 e_i, e_j) = \rho_i^2 (e_i, e_j) = 0 \quad \text{at } i \neq j$$
$$(BXe_i, BXe_i) = (X^* BBXe_i, e_i) = \rho_i^2 (e_i, e_i) = \rho_i^2.$$

Consequently $|BXe_k| = \rho_k$, where ρ_k are the singular numbers.

Therefore, the vector BXe_k is non-zero, then and only if the proper value ρ_k^2 of the operator is not zero.

The vector $X^* Bf_k$ is the operator's vector $X^* BBX$. Indeed

$$(X^* BBX) X^* Bf_k = X^* B(BXX^* B)f_k = X^* B\rho_k^2 f_k$$
$$(X^* Bf_k X^* Bf_k) = (BXX^* Bf_k f_k) = \rho_k^2 (f_k f_k) = \rho_k^2.$$

Thus non-zero values of operators $X^* BBX$ and $BXX^* B$ always coincide. Denote our own values through $\rho_1^2, \rho_2^2, \ldots, \rho_t^2$. However, without limiting generality $\rho_1^2 \geq \rho_2^2 \geq, \ldots, \rho_t^2 > 0$, and the remaining eigenvalues ρ_k^2 are equal to zero.

Obviously, the eigenvaluesv of the operators $X^* BBX$, $BXX^* B$ differ only by the multiplicity of the null eigenvalue. Operator $X^* BBX$ has levels $m - t$, operator $BXX^* B$ has level $n - t$, which are called singular numbers of the operator X.

We take as a basis in H_1 the orthonormal system $e_1, e_2, \ldots, e_t$ of eigenvectors of the operator $X^* BBX$. Then the vectors $e_1, e_2, \ldots, e_t$ form a basis in the domain of values $X^* BBX$, and $e_{t+1}, e_{t+2}, \ldots, e_m$ the basis in the core of the

operator X^*BBX.

$f_1, f_2, \ldots, f_n$ − orthonormal basis in H_1. As $f_1, f_2, \ldots, f_t$ we take vectors obtained after normalization $BXe_1, BXe_2, \ldots, BXe_t$. That is

$$f_i = \frac{BXe_i}{\|BXe_i\|}, \ i = 1, 2, \ldots, t.$$ Let's take any basis

in the kernel BXX^*B. It is clear that $f_{t+1}, f_{t+2}, \ldots, f_n$ they are their own for the operator BXX^*B. Then we have

$$BXe_k = \begin{cases} \rho_k f_k, & k \le t \\ \\ 0, & k > t \end{cases} \quad (16)$$

$$X^*Bf_k = \begin{cases} \rho_k e_k, & k \le t \\ \\ 0, & k > t \end{cases}, \quad (17)$$

$e_1, e_2, \ldots, e_m$, $f_1, f_2, \ldots, f_n$ − singular bases.

Let's have $y = X\hat{\beta} + e \Rightarrow By = BX\beta + Be$.

We find a vector $\hat{\beta}$ for which

$$(B(y - X\hat{\beta}))'B(y - X\hat{\beta})) = e'BBe \to \min \quad (18)$$

Self-regulated, positively defined operator in H_2, where $B = p^{-1}, D = p^2, BB = D^{-1}$.

The pseudo-solution of the problem (18) is any vector $\beta \in H_1$ for which the function of the discrepancy (18) reaches its lowest value.

A normal task pseudo-solution is called a pseudo-solution, for which $\sum \beta_i^2 \to \min$.

We will prove that a normal pseudo-solution (18) exists as unique.

We will fix in spaces H_1, H_2 singular bases $e_1, e_2, \ldots, e_m$ and $f_1, f_2, \ldots, f_m$.

Let

$$\beta = \sum_{k=1}^{m} \beta_k e_k \in H_1 \quad (19)$$

$$By = \sum_{i=1}^{n} \gamma_i f_i \in H_2. \quad (20)$$

Then, taking into account (19), (20)

$$BX\beta - By = BX \sum_{k=1}^{m} \beta_k e_k - \sum_{i=1}^{n} \gamma_i f_i =$$

$$\sum_{k=1}^{m} \beta_k B(X e_k) - \sum_{i=1}^{m} \gamma_i f_i - \sum_{i=m+1}^{n} \gamma_i f_i =$$

$$\sum_{k=1}^{m} (\beta_k \rho_k f_k - \gamma_k f_k) - \sum_{k=m+1}^{n} \gamma_k f_k =$$

$$\sum_{k=1}^{t} (\beta_k \rho_k - \gamma_k) f_k + \sum_{k=t+1}^{m} (\beta_k \rho_k - \gamma_k) f_k - \sum_{k=m+1}^{n} \gamma_k f_k,$$

where f_k − orthonormal, so

$$F_0(\beta) = (e'BBe) = (y - X\beta)'BB(y - X\beta) =$$

$$= \sum_{k=1}^{t} (\beta_k \rho_k - \gamma_k)^2 + \sum_{k=t+1}^{m} (\beta_k \rho_k - \gamma_k)^2 + \quad (21)$$

$$+ \sum_{k=m+1}^{n} (\gamma_k)^2$$

Obviously, the smallest value of a non-compliance functional achieves under these values $\hat{\beta}$ in which the last m-t coefficients β_k are arbitrary, and the first t terms in (21) are equal to 0.

If $\sum \beta_i^2 \to \min$,

then $\beta_{t+1} = \beta_{t+1} = \ldots = \beta_m = 0$ and $\beta_k = \frac{\gamma_k}{\rho_k}$.

$(k = 1, 2, \ldots, t)$.

Normal problem solving (1)

$$\beta_0 = \sum_{k=1}^{t} \frac{\gamma_k}{\rho_k} e_k \quad (22)$$

Using (21) (22) we obtain that pseudo-solutions and only they satisfy

$$X^*BBX\hat{\beta} = X^*BBy. \quad (23)$$

Indeed, using (21)

$$X^*BBX\beta = X^*BBX\sum_{k=1}^{m}\beta_k e_k =$$

$$= \sum_{k=1}^{t}\rho_k^2\beta_k e_k + \sum_{k=t+1}^{m}\rho_k^2\beta_k e_k X^*BBy =$$

$$= X^*B\sum_{i=1}^{n}\gamma_i f_i = \sum_{i=1}^{n}\gamma_i X^*Bf_i = \sum_{i=1}^{t}\gamma_i\rho_i e_i$$

we obtain that $\widehat{\beta}_k = \dfrac{\gamma_k}{\rho_k}$ for $k = \overline{(1,t)}$ and exactly 0 for the other $m-t$ coordinates.

We have a pseudo-connection

$$\widehat{\beta} = (X^*BBX)^*X^*BBy \qquad (24)$$

the same

$$\widehat{\beta} = \sum_{k=1}^{t}\left\{\dfrac{\gamma_k}{\rho_k}\right\}e_k$$

Let operator X operate from space H_1 in H_2. Then we assign to each vector $y \in H_2$ a uniform vector β_0 − normal pseudo-solution of the equation (18), which means the solution $BY = BX\beta + B\varepsilon$. This correspondence defines some operator X_B^+ that acts from H_2 in H_1 and is called a normal pseudo-inverse to X .

Therefore, by definition $\beta_0 = \overline{X}_B^+Y$ for any $y \in H_2$ that is

$$\overline{X}_B^+ f_k = \begin{cases} \rho_k^{-1} e_k, & k \leq t \\ 0, & k > t. \end{cases}$$

Again, consider the equation (1, 2) under the condition $(6')$, $(4')$ and let

$$M\varepsilon\varepsilon' = D = B^{-1}B^{-1}, \text{ then } BY = BX\beta + B\varepsilon.$$
$$\text{var}(B\varepsilon) = \text{var}(\overline{\varepsilon}\cdot\overline{\varepsilon}'). \ M[B\varepsilon\varepsilon'B] =$$
$$BDB = E, \ \partial e \ \overline{\varepsilon} = B\varepsilon ,$$

where

$$\widehat{\beta} = (X^*BBX)^+X^*BBy = (X^*D^{-1}X)^+X^*D^{-1}Y = (X'W^{-1}X)^+X'W^{-1}Y,$$

$$\widehat{\beta} = (X^*BBX)^+X^*BBX\beta + (X^*BBX)^+X^*BB\varepsilon ,$$

$$\widehat{\beta} = P_1\beta + (X^*BBX)^+X^*BB\varepsilon , \qquad (25)$$

$$M\widehat{\beta} = P_1\beta .$$

Where $P_1 = (X^*BBX)^+X^*BBX$ and

$$P_1 e_i = \begin{cases} e_i, & i \leq t \\ 0, & i > t, \end{cases}$$

P_1 − orthoprojector on $L(e_1, e_2, \ldots, e_t)$.

Likewise $P_2 = (BXX^+B)^+(BXX^+B)$

$$P_2 f_i = \begin{cases} f_i, & i \leq t \\ 0, & i > t. \end{cases}$$

We have $P_1^2 = P_1, P_2^2 = P_2$, P_1, P_2 - orthorecomplexes, otherwise orthoprojectors

$$\widehat{\beta} - M\widehat{\beta} = (X^*BBX)X^*BB\varepsilon$$
We use (25) $\overline{B}\overline{B} = \sigma^2 BB$, $BB\varepsilon\varepsilon' = E$,then
$$\text{var}\,\widehat{\beta} = M((\widehat{\beta} - M\widehat{\beta})(\widehat{\beta} - M\widehat{\beta})) =$$
$$M(X\overline{B}\overline{B}X)^+X^*\overline{B}\overline{B}\sigma\varepsilon\varepsilon'\overline{B}\overline{B}X(X^*\overline{B}\overline{B}X)^+ =$$
$$\sigma^2(X^*\overline{B}\overline{B}X)^+X^*\overline{B}\overline{B}X(X^*\overline{B}\overline{B}X)^+ = \sigma^2(X^*\overline{B}\overline{B}X)^+$$

Let's calculate σ^2. Let

$$\begin{cases} Y = X\beta + \varepsilon \\ y = X\widehat{\beta} + e \end{cases} \Rightarrow \begin{cases} \overline{B}Y = \overline{B}X\beta + \overline{B}\varepsilon \\ \overline{B}y = \overline{B}X\widehat{\beta} + \overline{B}e \end{cases} \Rightarrow$$
$$\begin{cases} \overline{B}Y = \overline{B}X\beta + \overline{B}\varepsilon \\ \overline{B}e = \overline{B}y - \overline{B}X\widehat{\beta} ; \end{cases}$$
$$\overline{B}e = \overline{B}(X\beta + \varepsilon) - \overline{B}XX^+Y =$$
$$\overline{B}X\beta + \overline{B}\varepsilon - \overline{B}(XX^+X\beta) - \overline{B}XX^+\varepsilon =$$
$$\overline{B}(E - XX^+)\varepsilon = \overline{B}(E - P_2)\varepsilon ,$$

where $P_2 = XX^+$ - ortoprojector; P_2 − idempotent, symmetric matrix.

That's why $E - P_2$ is also an orthoprojector.

Let $P = E - P_2$, then $\bar{B}e = \bar{B}P\varepsilon$.

So $P = \|P_{ij}\|$ a matrix of size $n \times n$.

Denote $spP = \sum_{i=1}^{n} p_{ii}$.

Calculate $(\bar{B}e)'\bar{B}e$. We have $\bar{B}e = \bar{B}P\varepsilon$, so

$$B P\varepsilon'\varepsilon PB = PB\varepsilon'\varepsilon BP =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}((\bar{\varepsilon}_i\bar{\varepsilon}_j - M\bar{\varepsilon}_i M\bar{\varepsilon}_j) + \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij} M\bar{\varepsilon}_i M\bar{\varepsilon}_j .$$

Where $\bar{\varepsilon}_i = B\varepsilon_i, i = (1,\ldots,n)$, so

$$(\bar{B}e)'\bar{B}e = e'\bar{B}\bar{B}e = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}(\bar{\varepsilon}_i\bar{\varepsilon}_j - M\bar{\varepsilon}_i M\bar{\varepsilon}_j) =$$

$$= \sum_{i=j} P_{ii}(\bar{\varepsilon}_i^2 - (M\bar{\varepsilon}_i)^2) + \sum_{i \neq j} P_{ij}(\bar{\varepsilon}_i\bar{\varepsilon}_j - M\bar{\varepsilon}_i M\bar{\varepsilon}_j).$$

So

$$M(e'\bar{B}\bar{B}e) = \sum_{i=1}^{n} P_{ii}\sigma^2 + \sum_{i \neq j} P_{ij}\operatorname{cov}(\bar{\varepsilon}_i\bar{\varepsilon}_j) =$$

$$= \sigma^2 (spP).$$

Because $\operatorname{cov}(\bar{\varepsilon}_i\bar{\varepsilon}_j) = 0$, so $M(\varepsilon\varepsilon') = \sigma^2 E$

$$spP = sp(E - X(X'\bar{B}\bar{B}X)X'\bar{B}\bar{B} =$$

$$= spE_n - spX(X'\bar{B}\bar{B}X)^+ X'\bar{B}\bar{B} =$$

$$= spE_n - spX(X'\bar{B}\bar{B}X)^+ X'\bar{B}\bar{B} =$$

$$= n - sp(X'\bar{B}\bar{B}X)^+ X'\bar{B}\bar{B}X = n - spP_2 = n - r.$$

$Me'\bar{B}\bar{B}e = \sigma^2(n - r)$ and it means that it is

$$\sigma^2 = \frac{Me'\bar{B}\bar{B}e}{(n-r)} \text{ also } \frac{e'\bar{B}\bar{B}e}{(n-r)}$$

an unassigned estimate σ^2.

So
$X_B^+ = (X'BBX)^+ X'B$, then $\bar{X}_B^+ = X_B^+ B$.

We prove that X_B^+ satisfies the Moore-Penrose conditions (9), which are necessary

and sufficient for the matrix X_B^+ to be pseudo-inverse to BX [3].

Proving

1) $X_B^+ BXX_B^+ = (X'BBX)^+ X'BBX(X'BBX)^+ X'B =$

$$= (X'BBX)^+ X'B = X_B^+, \text{ and } X_B^+ BXX_B^+ = P_1 X_B^+ .$$

Where P_1 is the orthoprojector on the linear shell $L(e_1, e_2, \ldots, e_t)$

2) $\begin{aligned} BXX_B^+ BX &= BX(X'BBX)^+ X'BBX = \\ &= BX(X'BBX)^+ X'BBX = BXP_1 = BX \end{aligned}$,

3) $X_B^+ BX = (X'BBX)(X'BBX) = P_1$

$-$ orthoprojector on the linear shell $L(e_1, e_2, \ldots, e_t)$ self-preservation.

4) $BXX_B^+ = BX(X'BBX)X'B$ $-$ symmetric.

Thus X_B^+, it satisfies the Moore-Penrose conditions, and hence $X_B^+ = (BX)^+$, on the other hand, $(BX)^+ = X^+ B^{-1}$ [2] under the following condition: the set of values of B is an invariant subspace for $X'X$, and the set of values X' is an invariant subspace for BB'.

So $X_B^+ = X^+ B^{-1}$.

Definition 3. An estimation $\hat{\beta}$ of the parameter β is called X - unshielded, if $MX\hat{\beta} = X\beta$, i.e if $X\hat{\beta}$ - unmatched assessment $X\beta$.

Lemma 1. Evaluation $\hat{\beta} = X^+ BY$ is an X - unplanned estimate β.

Proving. We use (2). Then,

$$\hat{\beta} = X_B^+ BY = X_B^+ (BX\beta + B\varepsilon) =$$

$$= X^+ B^{-1} BX\beta + X^+ B^{-1} B\varepsilon =$$

$$= X^+ X\beta + X^+ \varepsilon$$

therefore, from (3) and (9)

$$M(X\widehat{\beta}) = M(XX^+X\beta + XX^+\varepsilon) =$$
$$XX^+X\beta = X\beta.$$

Lemma 2. The covariance matrix of the $D(X\widehat{\beta})$ parameter $- X\widehat{\beta}$ estimate in the model (2) is equal to $D(X\widehat{\beta}) = \sigma^2 X'(X'\overline{B}\overline{B}X)^+X$, where $(X'\overline{B}\overline{B}X)^+ -$ pseudo-return to $(X'\overline{B}\overline{B}X)$

Proving. We use [2]

$$X_B^+ = (X'BBX)^+X'B, X_B^+By =$$
$$(X'BBX)^+X'BBX\beta + X^+\varepsilon.$$

Then

$$X\widehat{\beta} = XX^+Y = XX^+X\beta + XX^+\varepsilon == X\beta + XX^+\varepsilon.$$

Then using (3), (4), (9) we obtain

$$D(X\widehat{\beta}) = M(X\widehat{\beta} - X\beta)(X\widehat{\beta} - X\beta)' =$$
$$M(XX^+\varepsilon, \varepsilon'(XX^+)') =$$
$$MX(X'\overline{B}\overline{B}X)^+X'\overline{B}\overline{B}\varepsilon\varepsilon'\overline{B}\overline{B}X(X'\overline{B}\overline{B}X)^+X =$$
$$\sigma^2 X'(X'\overline{B}\overline{B}X)^+X.$$

Let it $C = (C_1, C_2, ..., C_m)$.

Let $L(X_1, X_2, ..., X_n)$ the linear shell of the strings of matrix X.

Theorem 1.

$$C \in L(X_1, X_2, ..., X_n) \Rightarrow M(c\widehat{\beta}) = c\beta$$

Proving

$$C \in L(X_1, X_2, ..., X_n) \Rightarrow C = \gamma X,$$

where $\gamma = \{\gamma_1, \gamma_2, ..., \gamma_n\}$, $M(c\widehat{\beta}) = M(\gamma X\widehat{\beta})$

According to Lemma 2.

$$MX\widehat{\beta} = X\beta \text{ this } M(c\widehat{\beta}) = \gamma X\beta = c\beta.$$

That is $c\widehat{\beta}$ - a linear unmatched estimate $c\beta$. Calculate

$$D(c\widehat{\beta}) = M(c\widehat{\beta} - c\beta)(c\widehat{\beta} - c\beta)' =$$
$$M(\gamma X\widehat{\beta} - \gamma X\beta)(\gamma X\widehat{\beta} - \gamma X\beta)' =$$
$$M(\gamma XX^+\varepsilon\varepsilon'X^{+'}X'\gamma') =$$
$$M\gamma X(X'BBX)^+X'BB\varepsilon\varepsilon'BBX(X'BBX)^+X'\gamma' =$$
$$= \gamma X(X'B^2X)^+X'BBX(X'BBX)^+X\gamma =$$
$$= \gamma X'(X'BBX)^+X\gamma' = cX'BBXc'.$$

3. By assumptions in the linear model (1 − 5), rang $X < m$. The matrix $X'BBX$ has the order $m \times m$ and it is symmetric and inextricably defined. Therefore, it has m intrinsic eigenvalues $\rho_1^2, \rho_2^2, \rho_3^2, ..., \rho_m^2$, such as

$$\rho_1^2 \geq \rho_2^2 \geq ... \rho_t^2 \geq \rho_{t+1}^2 = \cdots = \rho_m^2 = 0.$$

Then takes place (11), where U is an orthogonal $n \times n$ matrix, and V is an orthogonal $m \times m$ matrix.

The columns of the matrix U are the intrinsic vectors of the matrix, and the columns of the matrix V are the own vectors of the matrix $X'BBX$. That is, the matrix $V = (v_1, v_2, ..., v_m)$, formed by its own vectors, v_i denotes a vector column corresponding to its own value ρ_i^2.

We have $(X'BBX)V_i = \rho_i^2 v_i$ in addition to $V'V = E_m$.

Let $L(V_1, V_2, V_3, ..., V_t)$ the linear sheath of vectors $V_1, V_2, V_3, ..., V_t$ and let

$$C' = (C_1, C_2, ..., C_m)' \in L(V_1, V_2, ..., V_t).$$

Then $C_i' = \sum_{i=1}^{t} v_i \alpha_i \Leftrightarrow C \Leftrightarrow \alpha V'$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_t)$.

Theorem 2.

$$D(\widehat{\beta}) = \sigma^2 (X\overline{B}\overline{B}X)^+ = \sigma^2 (XW^{-1}X)^+$$
$$= (X'BBX)^+ = (X'D^{-1}X)^+$$

Proving. Let $\widehat{\beta} = X^+BY$ a normal pseudo-solution. Also takes place (2), from where $\widehat{\beta} = X^+(X\beta+\varepsilon) = X^+X\beta + X^+\varepsilon$. From (3), in general, mathematical expectation $M\widehat{\beta} = X^+X\beta = \rho_1\beta$.

Scattering $\bar{D}(\widehat{\beta})$ is relatively $M(\widehat{\beta})$ equal

$$\bar{D}(\beta) = M(\widehat{\beta} - X^+X\beta)(\widehat{\beta} - X^+X\beta)' =$$
$$M(X^+\varepsilon,\varepsilon'X^{+'}) =$$
$$(X'BBX)^+X'(BB\varepsilon\varepsilon'BBX(X'BBX)^+ =$$
$$(X'BBX)^+(X'BBX)(X'BBX)^+ = (X'BBX)^+.$$

The theorem is proved.

We have $c = \alpha V'$. Let's calculate $D(c\widehat{\beta})$, where $c\widehat{\beta}$ − unmatched estimate $c\beta$.

$$M(c\widehat{\beta}) = M\alpha V'\widehat{\beta} = \alpha V'X^+X\beta \text{ and}$$
$$c\widehat{\beta} - Mc\widehat{\beta} = \alpha V'X^+\varepsilon$$
$$D(c\widehat{\beta}) = M\alpha V'X^+\varepsilon\varepsilon'X^+V\alpha' =$$
$$M\alpha V'(X'\bar{B}\bar{B}X)^+X'\bar{B}\bar{B}\varepsilon\varepsilon'\bar{B}\bar{B}X(X'\bar{B}\bar{B}X)^+V\alpha' =$$
$$\sigma^2\alpha V'(X'\bar{B}\bar{B}X)^+V\alpha' = D(c\widehat{\beta}) = \sigma^2\alpha\sum{}^+\alpha',$$

where

$$\sum{}^+ =$$
$$= \begin{bmatrix} (p_1^2)^+ & (0) & \cdots & (0) & (0) & \ldots & (0) \\ (0) & (p_2^2)^+ & \cdots & (0) & (0) & \ldots & (0) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ (0) & (0) & \cdots & (p_m^2)^+ & (0) & \ldots & (0) \end{bmatrix}$$

and where $(p_i^2)^+ = \begin{cases} \dfrac{1}{p_i^2}, & p_i^2 \neq 0 \\ 0, & p_i^2 = 0 \end{cases}$,

so that

$$D(c\widehat{\beta}) = \sigma^2(\frac{\alpha_1^2}{\rho_1^2} + \frac{\alpha_2^2}{\rho_2^2} + L + \frac{\alpha_t^2}{\rho_t^2}). \quad (26)$$

We see that the variance of a linear combination $c\widehat{\beta}$ depends on inverse eigenvalues.

And due to the smaller absolute value, the characteristic root has the greatest influence on this dispersion.

We are interested in a separate coefficient β_i', then let the vector $c = \alpha V' = (0,0,\ldots,1,\ldots,0)$ have a unit on the i position, and the remaining coordinate is equal to zero.

We require: $c' \in L(V_1,V_2,V_3,\ldots,V_t)$. We will

$$C = \alpha V' \text{ get with it } \alpha = e_i V = \begin{bmatrix} v_{i1} \\ v_{i2} \\ \vdots \\ v_{it} \end{bmatrix}',$$

where $v_{i1}, v_{i1}, \ldots, v_{i1}$ the elements of the line are matrix *V*.

Consequence 1.

$$D(\widehat{\beta}_i) = \sigma^2(\frac{v_{i1}^2}{\rho_1^2} + \frac{v_{i2}^2}{\rho_2^2} + \cdots + \frac{v_{it}^2}{\rho_t^2}) \ i = \overline{(1,m)}$$

With $e_i \in L(V_1,V_2,V_3,\ldots,V_t)$. $\quad (27)$

4. Let it

$C = \{C_1,C_2,\ldots,C_m\} \ \rho_1^2 \geq \rho_2^2 \geq \ldots\rho_t^2 > 0$ - own values $(X'BBX)$.

$L(V_1,V_2,\ldots,V_t)$ − a linear shell of vectors $V_1,V_2,V_3,\ldots,V_t$ for which - respectively $\rho_1^2 \geq \rho_2^2 \geq \ldots\rho_t^2 > 0$.

Let $C' = \{C_1,C_2,\ldots,C_m\}' \in$
$\in L(V_1,V_2,\ldots,V_t) \Rightarrow c = \alpha V', \alpha = (\alpha_1,\alpha_2,\ldots,\alpha_t)$

and is a solution of the system

$$\alpha_1(V_1) + \alpha_1(V_2) + \ldots\alpha_1(V_t) = c', \quad (28)$$

or in a matrix record $c = \alpha V'$.

Let $\bar{X}_B^+Y$ the normal pseudo-solution of the problem (2 − 5, 6). As shown above, takes place (28). We put in (28)

$$\alpha_{k+1}^2 + \alpha_{k+2}^2 + \ldots + \alpha_t^2 = 0 \qquad (29)$$

then $c_k = \sum_{i=1}^{k} \alpha_i V_i$ − shifted estimate c,

$\alpha_{k+1}(V_{k+1}) + \alpha_{k+2}(V_{k+2}) + \ldots \alpha_t(V_t)$ − bias, and $N_k = \alpha_{k+1}^2 + \alpha_{k+2}^2 + \ldots + \alpha_t^2$ − shift value $\hat{C}_k$.

Then from $(26) - (29)$

$$D(\hat{C}_k \hat{\beta}) = \sigma^2 \left(\frac{\hat{\alpha}_1^2}{\rho_1^2} + \frac{\alpha_1^2}{\rho_2^2} + \ldots + \frac{\alpha_k^2}{\rho_k^2} \right).$$

Let $M_k = N_k + D(\hat{C}_k \hat{\beta})$

The

$$M_k = \begin{cases} \sum_{i=1}^{t} \alpha_i^2, k = 0 \\ \sigma^2 \sum_{j=1}^{k} \frac{\alpha_j^2}{\rho_j^2} + \sum_{j=k+1}^{t} \alpha_j^2, 1 \le k \le t-1 \\ \sigma^2 \sum_{j=1}^{t} \frac{\alpha_j^2}{\rho_j^2}, k = t \end{cases}.$$

Looking for $\min M_k$ $(0 \le k \le t)$ so

$$M_{i-1} = \sigma^2 \sum_{j}^{i-1} \frac{\alpha_j^2}{\rho_j^2} + \sum_{j=1}^{t} \alpha_j^2$$

$$M_i = \sigma^2 \sum_{j}^{t} \frac{\alpha_j^2}{\rho_j^2} + \sum_{j=i+1}^{t} \alpha_j^2$$

$$M_{i+1} = \sigma^2 \sum_{j}^{i+1} \frac{\alpha_j^2}{\rho_j^2} + \sum_{j=2}^{t} \alpha_j^2$$

$$M_i - M_{i-1} = \frac{\alpha_i^2 \sigma^2}{\rho_1^2} - \alpha_i^2 = \alpha_i^2 \left[\frac{\sigma^2 - \rho_i^2}{\rho_i^2} \right]$$

$$M_{i+1} - M_i = \frac{\alpha_{i+1}^2 \sigma^2}{\rho_{i+1}^2} - \alpha_{i+1}^2 = \alpha_{i+1}^2 \left[\frac{\sigma^2 - \rho_i^2}{\rho_{i+1}^2} \right].$$

Minimum M_i when $M_i - M_{i-1}$ changing the minus sign to plus

$$\begin{cases} \sigma^2 - \rho_i^2 \le 0 \\ \sigma^2 - \rho_{i+1}^2 \ge 0 \end{cases} \Rightarrow \begin{cases} \rho_i^2 \ge \sigma^2 \\ \rho_{i+1}^2 \le \sigma^2 \end{cases} \Rightarrow \rho_{i+1}^2 \le \sigma^2 \le \rho_i^2.$$

The first case $\sigma^2 \ge \rho_1^2 \Rightarrow M_i - M_{i-1} > 0$.

Hence M_i grows. The minimum M_i for $i = 0$ and is equal, to

$$M_0 = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_t^2 \to \min$$

The second case

$$\rho_{i+1}^2 \le \sigma^2 < \rho_i^2.$$

The minimum $N_k + D(\hat{C}_k)$ under

$1 \le k \le t-1$ is equal to

$$N_i + D(\hat{C}_k \hat{\beta}) =$$

$$= \sigma^2 \left(\frac{\alpha_1^2}{p_1^2} + \frac{\alpha_1^2}{p_2^2} + \cdots + \frac{\alpha_i^2}{p_i^2} \right) + \sum_{j=i+1}^{t} \alpha_j^2 \to$$

$$\to \min \text{ при } 0 \le k \le t.$$

The third case $\sigma^2 < \rho_i^2$, M_i comes down. The minimum for $I = t$ is equal to

$$M_t = \sigma^2 \left(\frac{\alpha_1^2}{p_1^2} + \frac{\alpha_1^2}{p_2^2} + \cdots + \frac{\alpha_t^2}{p_t^2} \right) \to \min$$

Consequence 2.
Under conditions 1) $\sigma^2 < \rho_i^2$;

2) $e_i' = (0, \ldots, 1, 0, \ldots)' \in L(V_1, V_2, \ldots, V_t) \Rightarrow$

$$D(\hat{\beta}_i) = \sigma^2 \left(\frac{v_{i1}^2}{\rho_1^2} + \frac{v_{i2}^2}{\rho_2^2} + L + \frac{v_{it}^2}{\rho_t^2} \right) \to \min.$$

Consequence 2 is not valid for all $\hat{\beta}_i$, because all e_i cannot belong to $L(V_1, V_2, \ldots, V_t)$, since they are e_i-linearly independent (1, 2, ..., m), but $\text{rang} X < m$.

CONCLUSIONS

In econometric modeling, the design of the structure of the predictor space by the researcher is a tool for obtaining a prognostically effective specification of exogenous variables. According to the authors, the process of selecting signs and constructing prediktornym spac-

es is appropriate to apply iterative. An approach to the selection of explanatory variables based on the study of singular schedules is effective for constructing econometric models with multi-collinear variables and the matrix of incomplete observations. Despite the moderate increase in time for making decisions on the structure of the econometric model, the "personalization" of constructing patterns of interdependence of factors will contribute to the growth of their adequacy.

Currently, the toolkit of research analytics is intensively expanding with the latest procedures that provide high predictive efficiency. In econometric modeling, the problem of determining informative predictors, the formation of a rational set of exogenous variables, the justification of the dimension and structure of predictor spaces is relevant. The basic approaches to choosing the optimal set of features are, in particular, the overview of all combinations, direct selection and reverse exclusion, the weighting of signs in linear and logistic regression algorithms, the importance of features in decision trees and the variants of ensembles, for example, "random forest", etc.

Modern Data Technologies, Data Mining, Machine Learning (ML) provide a wide array of feature design techniques. The purpose of designing and selecting features is to prevent the effect of re-training, to achieve greater compactness of the model by eliminating excess regressors, reducing the dimensionality of the learning processes without a teacher, constructing classifiers, mapping the process of partitioning data into classes, and determining the boundary of solutions in the reduced space, as well as substantiated interpretation, providing in-depth understanding of the model and learning data, visualization in spaces, the dimension of which will be perceived by the researcher.

Determining a subset of signs is an important component of machine learning and greatly affects the accuracy of ML models. The concept of machine learning maximizes the ability to define templates in the data achieved, in particular, by aggregating a set of attributes. The informatively weak sign can significantly increase its own prognostic utility

and become strong in the presence of another effective set of features.

Reconstruction and selection of features contributes to increasing the automation of the learning process. When studying ML models in large numbers, there is a danger of retraining, but clear selection rules can reduce their number. Consequently, the purposeful variation in the number of signs should be used by researchers to calibrate and study the model, since it enables us to justify the choice of a rational set of independent variables that determine the structures in the data and subsequently successfully predict trends in the behavior of economic systems.

Note that even taking into account the representative set of algorithms for designing the features, implemented on the basis of platforms ML type R, Python, the researcher often has difficulties in designing sign spaces, and the choice of the conceptual approach and tools is ambiguous. When necessary to take into account the specifics of origin and data formats in Big Data technologies, along with their further unification and ensemblevization, there is a need for a researcher-driven "intervention" in a fast ensemble of large data analysis technology.

REFERENCES

1. **Johnston J., 1971.** Econometric Metods. MeGraw-Hill, 437.
2. **Lawson C.L., Hanson R.J., 1974.** Solving Least Squares Problems. Prentice – Hall, Inc., Englewood Cliffs N.J., 340.
3. **Voevodin V.V., 1977.** Vychislitel`nye osnovy lineinoi algebry. [numerical foundations of linear algebra]. Nauka, Moscow, 303(in Russian).
4. **Kutovyi V.O., 2001.** Pro teoremu Haussa-Markova u vypadku vyrodzhenoi matrytsi sposterezhen. Dopov. Dokl. Akad. Nauk Ukraine, No.5, 19-22 (in Ukrainian).
5. **Kutovyi V.O., 2000.** Pro zastosuvania instrumentalnyh zminnyh dlia vyznachenia parametriv zagalnoi liniynoi modeli Modeliuvayia ta informaciyni systey v economici.Kyiv.KNEU, No.64 (in Ukrainian) 168-173.
6. **Kutovyi V.O., Roskach O.S., 1997.** Matematyko-statystychne uzagalnenia pokrokovyh metodiv pobudovy predyktornyh

prostoriv. Mashynna obrobka informacii, No.59, 140-149 (in Ukrainian).

7. **Kutovyi V.O., Roskach O.S.**, **1997.** Pro zastosyvania na EOM algorytmu Farrara-Glaubera.Mashyna obrobka informacii. Kyiv, KNEU, No.61, 142-149 (in Ukrainian).

8. **Kutovyi V.O.**, **1999.** Pro umovy zastosuvania teoremy Gaussa-Markova. Vcheni zapysky Kyiv, KNEU, No.2C, 206-208 (in Ukrainian).

9. **Kutovyi V.O.**, **2001.** Pro efektyvnist zmishenyh ocinok parametriv economichnyh modelei. Kyiv, KNEU, No.3, 324-326 (in Ukrainian).

10. **Aitken A.C., 1993.** One Least-squares and Linear Combinetion of Observations. Proc., RoyalSoc., Edinburgh, No.55, 42-46.

11. **Pavies O., 1993.** Statistical momentpods in research and production, New York, 1957.

12. **Plackett R.**, **1960.** Principles of regression analysis. Oxfopd.

13. **Weatherburn C.E.**, **1961.** A first course in mathematical statistics. University Press, Cambridge, brosch, 18 s, 6 d, 278.

14. **Hamilton W.**, **1964.** Statistucs in physical science. New York, 1964.

15. **Jürgen Grob.**, **2004**. The general Gauss-Markov model with hjssibl singular dispersion matrix. Statistical Paper, No.45, 311-336.

16. **Farrar D.E., Glauber, R.R., 1967.** Multicollinearity in Regression Analysis: The Problem Revisited. Review of Economics and Statistics, 49(1), 92-107.

17. **Yangge Fian, Beisiegel M., Dagenais E., Haines C.**, **2008.** On the natural restrietions in the singular Grauss-Markov model. Statistical Papers, Vol.49, 553-564.

18. **Silvey S.D., 1969**. Multicallinearity and Imprecise Estimation. Jornal of the Real Statical Society, Series B, No.31, 539-552.

19. **Makhinko A., Makhinko N., 2017.** Osoblyvosti imovirnisnoho rozrakhunku vysotnykh sporud pry vrakhuvanni vypadkovosti obokh skladovykh vitrovoho vplyvu Probabilistic design of high-rise buildings with two stochastic components of wind velocity. Underwater Technologies, Iss.6, 16-27 (in Ukrainian).

20. **Bogdanov V.**, **2017.** Impact of a hard cylinder with flat surface on the elastic layer. Underwater Technologies, Iss.5, 8-15 (in Ukrainian).

21. **Skochko V.**, **2017.** Shaping of the frameworks of technical forms, which defined by implicit functions on a plane. Underwater Technologies, Iss.7, 3-17 (in Ukrainian),

Анализ параметров мультиколлинеарных эконометрических моделей с матрицей наблюдений неполного ранга

Виктор Кутовой, Ольга Катунина, Олег Шутовский

Аннотация. Рассмотрена проблема определения информативных предикторов, формирования рациональной совокупности экзогенных переменных, обоснования размерности и структуры предикторных пространств. Целью проектирования и отбора признаков является предупреждение эффекта переобучения, снижение размерности в процессах обучения без учителя, построение классификаторов, отражение процесса разбиения данных на классы и определения границ решений в редуцированном пространстве, а также обоснованная интерпретация, обеспечение углубленного понимания модели и данных для обучения, визуализация в пространствах, размерность которых воспринимается исследователем. Рассмотрены вопросы проектирования предикторных пространств и разработки эффективных процедур оценивания параметров эконометрических моделей с мультиколлинеарными переменными. Проведено исследование альтернативных подходов к формированию совокупности признаков в моделях взаимозависимостей.

Предложен математический инструментарий для вычисления параметров линейной эконометрической модели в случае матрицы наблюдений неполного ранга, базирующийся на исследовании сингулярных разложений.

Использование сингулярного инструментария для декомпозиции и анализа матрицы данных позволяет повысить операционную эффективность и прогностическое качество процедур оценивания параметров эконометрических моделей. Математический подход к построению моделей взаимообусловленности факторов предназначен для выбора признаков и конструирования предикторных пространств при исследовании систем с мультиколлинеаными переменными и матрицей наблюдений неполного ранга.

Ключевые слова: проектиррование признаков, эконометрическая модель, мультиколлинеарность, матрица наблюдений неполного ранга, сингулярное разложение, собственные числа.