Scientific Research Publishing

Open Journal of Statistics, 2016, 6, 842-861
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

New Facts in Regression Estimation under Conditions of Multicollinearity

Anatoly Gordinsky

Berman Engineers LTD, Modiin, Israel
Email: ntlgrdnsk1@gmail.com

How to cite this paper: Gordinsky, A. (2016) New Facts in Regression Estimation under Conditions of Multicollinearity. *Open Journal of Statistics*, **6**, 842-861.
http://dx.doi.org/10.4236/ojs.2016.65070

Received: August 10, 2016
Accepted: October 18, 2016
Published: October 21, 2016

Abstract

This paper considers the approaches and methods for reducing the influence of multicollinearity. Great attention is paid to the question of using shrinkage estimators for this purpose. Two classes of regression models are investigated, the first of which corresponds to systems with a negative feedback, while the second class presents systems without the feedback. In the first case the use of shrinkage estimators, especially the Principal Component estimator, is inappropriate but is possible in the second case with the right choice of the regularization parameter or of the number of principal components included in the regression model. This fact is substantiated by the study of the distribution of the random variable $b'b-\beta'\beta$, where b is the LS estimate and β is the true coefficient, since the form of this distribution is the basic characteristic of the specified classes. For this study, a regression approximation of the distribution of the event $b'b-\beta'\beta<0$ based on the Edgeworth series was developed. Also, alternative approaches are examined to resolve the multicollinearity issue, including an application of the known Inequality Constrained Least Squares method and the Dual estimator method proposed by the author. It is shown that with a priori information the Euclidean distance between the estimates and the true coefficients can be significantly reduced.

Keywords

Linear Regression, Multicollinearity, Two Classes of Regression Models, Shrinkage Estimators, Inequality Constrained Least Squres Estimator, Dual Estimator

1. Introduction

In the statistical literature, the term "multicollinearity" is almost as popular as the term "regression". And this is natural since the regression analysis is one of the powerful tools to reveal dependencies which are hidden in the empirical data while multicolli-

nearity is one of the main pitfalls of this approach. Indeed, a high correlation between two or more of the explanatory variables (predictors) sharply increases the variance of estimators, which adversely affects the study of a degree and a direction of the predictor action on the response variable. Multicollinearity also impairs the predictive opportunities of the regression equation when the correlation in the new data significantly differs from the one in the training set. Numerous recent publications of researchers in various fields, such as medicine, ecology, economics, engineering, and others, devoted to the problem of multicollinearity, indicate that there are serious difficulties in this area until now. That is why the estimation of regression parameters under multicollinearity still remains one of the priorities of an applied and theoretical statistics.

The nature, methods of measurement, interpretation of the results, and methods of decreasing the effect of multicollinearity have been reviewed in numerous monographs and articles, of which we refer to the works have become classics [1]-[6] as well as the relatively recent articles [7]-[13].

In this paper, we focus on the parameters estimation methods in the multiple linear regression which corresponds to the following equation and assumptions

$$Y = X\beta + \epsilon, \tag{1}$$

where $Y, \epsilon \in \mathbb{R}^n, \epsilon \sim N\left(0, \sigma^2\right)$, σ^2 is the variance of ϵ, $\operatorname{cov}(\epsilon) = \sigma^2 I_n$, known $X \in \mathbb{R}^{n\times k}$ of rank k, and $\beta \in \mathbb{R}^k$ is unknown.

Our goal is to consider the empirical and mathematical methods of improving the statistical characteristics of estimators and, consequently, facilitate their interpretation in conditions of multicollinearity. We also will briefly touch upon the questions of a prediction by regression equations.

Recommendations present in the literature to decrease the influence of multicollinearity, in particular in sources mentioned above, reduce to the following.

1) Eliminate the one of the predictors that is most strongly correlated with the others.
2) Ignore the multicollinearity issue if the regression model is designed for prediction.
3) Standardize the data.
4) Increase the data sample size.
5) Use the shrinkage, and therefore biased, estimators.

The approach formulated in the first item requires a careful analysis of the nature of the correlation between the predictors. In some cases, the deletion of some predictor can distort the essence of the regression model. For example, in studying a quadratic trend the linear term of an equation may be strongly correlated with the quadratic term, but the exclusion of one of them, of course, is unacceptable. Next will be considered a class of regression models for which such an exclusion is always undesirable.

The second approach is only acceptable when the correlation matrix of the new dataset on which the prediction is being performed differs little from the same matrix of the training dataset. If this condition is not fulfilled, the variance of the prediction can increase many times. Unfortunately, experience in solving the real problems shows that the necessary proximity of the two indicated correlation matrices is rarely present.

Data standardization undoubtedly improves the conditionality of computational algorithms for regression, which is essential with a high degree of multicollinearity. This effect is particularly useful in the case of polynomial regression models, and models containing certain other functions. According to the literature, the issue of the effect of standardization on the interpretability of the estimation results is still controversial and is not considered in this article.

Increasing the dataset size always improves the quality of the estimation. Unfortunately, in this case, the indicator of the severity of multicollinearity, called the variance inflation factor (VIF) [5], reduces proportionately to $\sqrt{n}$, where n is the dataset size. This means, for example, that for an initial value of VIF = 50, one must increase n by a factor of 100 to reach the acceptable value VIF = 5. This is always expensive and is not always possible. Thus, given the above, it is advisable to study the last item of the recommendations to reduce the impact of multicollinearity, namely, the use of shrinkage estimators.

We shall discuss further some features of the use of penalized estimators, which include the ridge [14] and the Lasso [15] estimator, as well as the Principal Component estimator [16], and the James-Stein estimator [17]. The study of these estimators has been one of the key directions in statistics in recent decades. We show that there exists a class of regression models for which the application of the James-Stein estimator is useless, and employment of other mentioned estimators is dangerous. Moreover, this danger can be catastrophic for the PCR estimator. This statement is illustrated by the counterexample discussed in the second chapter. In the same chapter, on the basis of Monte Carlo trials, it is suggested that this situation is explained by the high probability of an event $b'b - \beta'\beta < 0$, where b is the LS estimator and β is the vector of the true regression coefficients.

This probability is studied in the third chapter. We obtained an analytical expression of the probability density function (pdf) for the simple regression, as well as analytical expressions for the four central moments for the multiple regression. Based on the last results, we derived an approximation of this pdf for multivariate regression, and its properties were examined.

The fourth chapter discusses the effect of the probability of the event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$ on the efficiency of the shrinkage estimators. This study is based on numerical modeling, the possibility of which is provided by the aforementioned approximation. The chapter introduces yet another class of regression models which is favorable for the application of shrinkage estimators. At the same time, the issue of setting the regularization parameter, which depends on the unknown coefficients β, remains open. In this case, the James-Stein estimator, which does not require a parameter of regularization, may provide only a meager improvement in efficiency.

Finally, the fifth chapter discusses alternative approaches. These approaches consist in an application of the known Inequality Constrained Least Squares method and the Dual estimator method proposed by the author in [18]. We show that in the presence of a priori information these approaches have significant advantages.

2. Methods

2.1. Counter Example for Shrinkage Estimators: A Degree of a Simultaneous Influence of Glucose and Insulin on the Blood Glucose Level

In this chapter, the author shows using counterexample that there are regression models with a high degree of multicollinearity for which the use of shrinkage estimators are useless or nearly useless, in the best case, and in the worst case the first three of the above methods give the mean squares error (MSE) exceeding the quadratic risk L^2 of the least square (LS) estimator. As far as the author knows, this fact was not considered in the literature, although it may be very important for researchers.

Let us now suppose that a researcher has experimental data corresponding to the assumptions of the normal multivariate linear regression (1), and has found the Least Square estimate

$$b = \left(X'X\right)^{-1} X'Y. \tag{2}$$

Let us also agree that the model is adequate, which means that F-criterion exceeds a critical value. This requirement is well founded as the estimate of coefficients does not make sense in the absence of adequacy. Moreover, the value of the F-test should exceed a critical value by several times [3], if one aims to ensure the acceptable quality of prediction with the regression model. This requirement for the value of the F-criterion would be more rigid, if one wishes to reach the sufficiently significance level of the estimates of the regression coefficients. The further necessary condition is that the structure of the regression model is found, and it does not change afterwards. Let there be multicollinearity existing in the experimental data. The researcher can apply four of the aforementioned shrinkage estimators to estimate the regression coefficients. The first three of them dominate only under certain conditions, whereas the fourth always dominates, the LS estimator. The researcher, however, is not interested in domination itself. Instead, he is interested in the quantitative characteristic of dominance, in particular, the ratio of the mean squares error (MSE) to the quadratic risk L^2. As is known [6]

$$MSE = E\left(b_{shr} - \beta\right)'\left(b_{shr} - \beta\right), \tag{3}$$

where b_{shr} is shrinkage estimator, and

$$L^2 = \sigma^2 Tr\left(\left(X'X\right)^{-1}\right) = \sigma^2 \sum_{i=1}^{k} \lambda_i, \tag{4}$$

where Tr is the trace, λ_i are the eigenvalues of the inverse matrix.

The shrinkage estimator will be almost useless, if the ratio MSE/L^2 is insignificantly less than 1.

Let us illustrate this assertion by the following example. Consider the problem of computing of the maximum blood glucose in patients with the diabetes mellitus under the simultaneous action of the glucose given peroral, and the insulin given subcutaneously. The physicians recommend [19] evaluating the influence of the insulin in order

to calculate its dose. Our approach provides a more realistic understanding of the interaction of glucose with insulin and, that is very important, the exclusion too large or too small the blood glucose levels. The regression model has the form:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{5}$$

where Y is maximum of the blood glucose, X_1 is quantity of the glucose in grams, X_2 is quantity of the insulin units, β_1, β_2 are the corresponding coefficients, β_0 is intercept, and ϵ is the random error. Let $\epsilon \sim N\left(0, \sigma^2 I_n\right)$, $\sigma^2 = 36$, $\beta_0 = 150$, $\beta_1 = 4$, $\beta_2 = -37$. These settings correspond to average values of patients with the diabetes mellitus of the first type. Data without a column containing only units are presented in Table 1. Let us emphasize that this table can serve as a prototype for examination of patients. A small number of experiments are important because every test is an injection.

Let us standardize these data, applying an approach somewhat different from that accepted in the literature, for example in [3] and numerous other sources. The fact is that, according to them, in order to obtain the standardized response after centering Y, providing Y_c, one must divide the Y_c by the random variable $\sqrt{Y_c'Y_c}$. But this procedure changes the distribution of the response, and we deviate from the scheme (1). Therefore hereinafter we will standardize according to [3] only the matrix X, obtain X_{st}, and use Y_c. In this case, the model coefficients are scaled in the obvious way:

$$\beta_{sti} = \sqrt{X_{ci}'X_{ci}}\,\beta_i, \tag{6}$$

where index i is a number of the coefficient and the column of the centered matrix X_c. Using this approach, we fully maintain the benefits of standardization, particularly, a convenience of a comparison of the scaled coefficients as well as the favorable eigenvectors of the matrix $X_{st}'X_{st}$. The σ^2, the response variable, and their properties do not change. After processing, our model gives the following results. The LS estimates of the standardized regression coefficients equal $b_{st1} = 93.020$, $b_{st2} = -116.922$ with exact values $\beta_{st1} = 167.8857$ and $\beta_{st2} = -199.5943$, and standard deviation $s = 6.5186$. The coefficient of determination $R^2 = 0.7200$, which is a rather modest value, and the

Table 1. Data for counterexample for shrinkage estimators.

Test number	Glucose	Insulin	Blood glucose
1	20	2	161
2	12	1	160
3	28	3.5	142
4	12	1	156
5	40	5	127
6	16	2	152
7	44	5	147
8	20	2	160
9	48	5.5	140
10	12	1	148

F-test at the significance level of 0.05 is equal to 8.99, which exceeds the critical value by a factor of 2.42. Finally, the t-tests respectively equal 1.898 and −2.386, while the critical value is 2.365 for the same significance level of 0.05. That is, we can see that the derived estimates of coefficients are almost not significant, despite the fact that the overall model is quite significant. The correlation matrix of the model is equal to $R = \begin{bmatrix} 1.0000 & 0.9911 \\ 0.9911 & 1.0000 \end{bmatrix}$, which indicates a high positive correlation of variables, variance inflation factor (VIF) is 56 for both variables, so that the presence of the considerable multicollinearity is not in doubt. Now we have complete data which are necessary for further studies.

Let us evaluate the potential capabilities of the shrinkage estimators for our task. Consider first the ridge estimator [14] the original form of which for our technique of standardization is written as

$$b_r = \left(R + rI_k\right)^{-1} X'_{st} Y_c, \tag{7}$$

where b_r is the ridge estimator, $r > 0$ is the regularization parameter, and we obtain the minimum MSE and the corresponding regularization parameter r_{opt} for (7).

For this we use approach [14] and derive the MSE expression in the matrix form:

$$MSE_{st} = \beta'_{st} P \left(\Lambda_n^{-1}\Lambda - I_k\right)^2 P' \beta_{st} + \sigma^2 Tr\left(\Lambda_n^{-2}\Lambda\right), \tag{8}$$

where Λ is the diagonal matrix of the eigenvalues of the correlation matrix sorted in descending order, P is the matrix of the corresponding eigenvectors, $\Lambda_n = \Lambda + rI_k$, and I_k is the identity matrix of order k. For our data, by minimizing MSE_{st} and using (4) we find $r_{opt} = 8.85 \times 10^{-5}$, and $MSE_{st} / L_{st}^2 \geq 0.9901$.

Thus in this task the ridge estimator is potentially useless.

If we were to apply under the same conditions the technique of the automatic selection of r [3], wherein $r = k\sigma^2 / b'_{st} b_{st}$, we would derive by the Monte Carlo simulation in MatLab that MSE_{st} / L_{st}^2 equals to 1.385.

As for the Lasso method, it is not possible to find the explicit expression for the MSE_{st}, and we again use the Monte Carlo approach in Matlab implementing the algorithm

$$\beta_l = \arg\min_\beta \left\{ \left\| Y - X\beta \right\|^2 + \gamma \left\| \beta \right\|_{L1} \right\}, \tag{9}$$

where β_l is the Lasso estimate, and $\gamma > 0$ is constant of regularization, and computing MSE_{st}, L_{st}^2 according to (3, 4) for big sets of ϵ and γ. The result is that $MSE_{st} / L_{st}^2 > 0.99$, *i.e.* this estimator is also potentially useless. It is also clear, if we estimate, using cross-validation as is customary [15], we find that the MSE_{st} exceeds the L_{st}^2. Here also note, that the exception of one of the predictors, which the Lasso may accomplish, in our case will distort the substance of the model.

Now we use the Monte Carlo method in an analogous manner to consider the possibilities of Principal Component Regression (PCR) [16]. For this case we derive $MSE_{st} / L_{st}^2 \geq 16.4$. This result should not be considered as too surprising, keeping in

mind the precautions in the literature (see e.g. [6]) that excluding components corresponding to small eigenvalues may sharply reduce the quality of the estimation and prediction. In particular, in the brief review of opportunities of the shrinkage estimators made in [18] this fact is confirmed by an analysis of the simulation outcomes in [20].

Finally, as to the James-Stein estimator, in our problem, we cannot apply this method in a canonical form [17], as the number of predictor variables is less than three. But nothing prevents us from deriving the MSE_{st} for the general kind of the shrinkage estimator that is employed in the James-Stein approach, exactly:

$$b_{shr} = \psi b, \tag{10}$$

where $\psi > 0$ is some constant.

Under known ψ and β_{st} after obvious calculations we obtain

$$MSE_{st} \geq \psi_{opt}^2 L_{st}^2 + \left(1-\psi_{opt}\right)^2 \beta_{st}'\beta_{st}, \tag{11}$$

where

$$\psi_{opt} = \beta_{st}'\beta_{st} \Big/ \left(\beta_{st}'\beta_{st} + \sigma^2 trace\left(\left(X_{st}'X_{st}\right)^{-1}\right)\right). \tag{12}$$

For our data we have $MSE_{st} \big/ L_{st}^2 \geq 0.944$.

Thus there are models with a high degree of multicollinearity for which the application of the shrinkage estimators is at least virtually useless. What is the reason for this phenomenon?

To answer this question, let us simulate the density of the random variable $b_{st}'b_{st} - \beta_{st}'\beta_{st}$ by the Monte Carlo method. The probability density function based on implementations of 10^6 tests is shown in **Figure 1**.

For the approximation presented, we obtained: $\Pr\left(b_{st}'b_{st} - \beta_{st}'\beta_{st} < 0\right) = 0.4996$, where $\Pr$ is probability. It was namely this fact which provides the answer to our question. It turns out that, contrary to established beliefs, significant multicollinearity does not always lead to an increase in the average norm of LS-estimates regarding the norm of the true coefficients. There are situations possible in which this norm is almost equal to the norm of the true coefficients. This results in an increase of the component MSE caused by the bias and a reduction of effectiveness of shrinkage estimators, which is precisely what we have observed.

It is evident that in order to ascertain the reasons for such a situation, it is necessary to explore the properties of the distribution of the random variable $b_{st}'b_{st} - \beta_{st}'\beta_{st}$, which is the purpose of the next chapter.

2.2. Properties of the Distribution of the Random Variable $b_{st}'b_{st} - \beta_{st}'\beta_{st}$

In this chapter, it will be shown that the random variable under study can be represented as a sum of the independent weighted central chi-square and normal variables. The author could not find the explorations of the distribution for this case, despite a vast literature on an approximation of the distribution of a sum of weighted chi-square variables (e.g., [21]-[23]). Let us define

$$\varphi = b'_{st} b_{st} - \beta'_{st} \beta_{st} \tag{13}$$

First we will get the exact expression of the probability density function in the one-dimensional case, if β_{st} and σ are known. Such a situation arises after a standardization of the simple regression.

Proposition 1. For the scheme (1) in the one-dimensional case and known β_{st}, σ the probability density function $g(\varphi)$ equals to

$$g(\varphi) = \begin{cases} \left(1/2\sigma\sqrt{2\pi\left(\beta_{st}^2 + \varphi\right)}\right)\left(\mathrm{e}^{-\theta_1^2(\varphi)/2\sigma^2} + \mathrm{e}^{-\theta_2^2(\varphi)/2\sigma^2}\right), & \text{if } \beta_{st}^2 + \varphi \geq 0, \\ 0, & \text{if } \beta_{st}^2 + \varphi < 0, \end{cases} \tag{14}$$

where $\theta_1(\varphi) = -\beta_{st} + \sqrt{\beta_{st}^2 + \varphi}, \theta_2(\varphi) = -\beta_{st} - \sqrt{\beta_{st}^2 + \varphi}.$

Proof.

Define

$$\delta_{st} = b_{st} - \beta_{st}. \tag{15}$$

One can easily obtain for the one-dimensional case that $\varphi = \delta_{st}^2 + 2\beta_{st}\delta_{st}$, where $\delta_{st} \sim N(0, \sigma)$. Consider φ as a nonmonotonic function of the δ_{st}. Then, we can use the well-known method given, for instance, in [24] to obtain (14). □

The result obtained shows that for dimensionality of two or more, the convolution of such a distribution would be extremely cumbersome and will require numerical solutions. For this reason, let us derive the equations for the first four central moments of the random variable φ for known β_{st} and σ.

Proposition 2. For the scheme (1) for $k \geq 2$ and known β_{st} and σ, the central moments of φ are equal to

$$\mu_1 = \sum_{i=1}^{k} c_i, \tag{16}$$

$$\mu_2 = \sum_{i=1}^{k} 2c_i^2 + p_i^2, \tag{17}$$

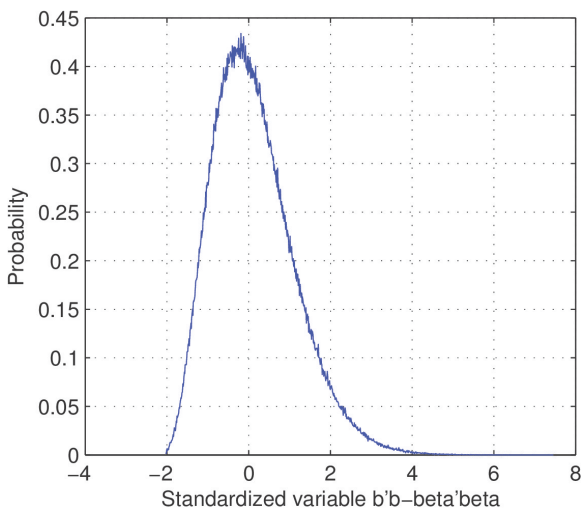

Figure 1. Simulated probability density function of the random variable $b'b - \beta'\beta$.

$$\mu_3 = \sum_{i=1}^{k} 8c_i^3 + 6c_i p_i^2, \tag{18}$$

$$\mu_4 = \sum_{i=1}^{k} 60c_i^4 + 60c_i^2 p_i^2 + 3p_i^4 + 6\sum_{i<j} \mu_{2i}\mu_{2j}, \tag{19}$$

$$c_i = \sigma^2 \lambda_i, \tag{20}$$

$$p_i = 2\beta'_{st} z_i \sigma \sqrt{\lambda_i}, \tag{21}$$

where λ_i is the eigenvalue of inverse of the correlation matrix R^{-1}, and z_i is the corresponding eigenvector of the same matrix.

Proof.

In the multidimensional case using (13), (15) we derive

$$\varphi = \delta'_{st}\delta_{st} - 2\beta'_{st}\delta_{st}. \tag{22}$$

Taking into account the orthogonality of the eigenvectors we represent φ as follows: $\varphi = \delta'_{st} zz' \delta_{st} - 2\beta'_{st} zz' \delta_{st}$. Examining the vector $z'\delta_{st}$, we derive the result that its components have normal distribution with zero expectation and variance of $\sigma^2\lambda_i$. Furthermore, these components are noncorrelated due to orthogonality of the eigenvectors and hence are independent together with functions of them. Therefore, using the notation (20), (21) we find:

$$\varphi = \sum_{i=1}^{k} c_i \alpha_i^2 + p_i \alpha_i, \tag{23}$$

where $\alpha_i \sim N(0,1)$ and $\operatorname{cov}(\alpha) = \sigma^2 I_k$. We find the moments about the origin of one of the independent variables $c_i\alpha_i^2 + p_i\alpha_i$ taking for k = 1 - 4 the definite integrals $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left(c_i x^2 + p_i x\right)^k \mathrm{e}^{-x^2/2}\mathrm{d}x$. Then we find the central moments, using the known relation for their calculations $\mu_n = \sum_{j=0}^{n}\binom{n}{j}(-1)^{n-j} M_j M_1^{n-j}$, where μ and M are the central and the initial moments. It remains to find the central moments for the sums of the independent random variables, which is performed according to the well-known rules. □

Once these moments are found, one can, if necessary, approximate the distribution by well-known methods using Pearson's or Johnson's families of curves [25] or by expansion into a certain series [26]. However, our ultimate goal is to estimate the probability of fulfillment of the inequality $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$, since it is specifically this which has a decisive influence on the expedience of the application of shrinkage estimators. Therefore let us simplify the task, and we will approximate this probability $\Pr\left(b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0\right)$. For the approximation we use variables that are included in the Edgeworth's series [25] [27] and we construct the regression on these variables. This approach is used because direct application of the specified series for our distribution does not provide an acceptable accuracy.

Let us designate

$$x = \mu_1 / \sqrt{\mu_2}, \tag{24}$$

$$q = \left(1/\sqrt{2\pi}\right)\mathrm{e}^{-x^2/2}, \tag{25}$$

$$P = 0.5\left(1 - erf\left(x/\sqrt{2}\right)\right), \tag{26}$$

$$P_3 = \left(x^2 - 1\right)q, \tag{27}$$

$$P_6 = -\left(x^5 - 10x^3 + 15x\right)q, \tag{28}$$

$$P_9 = \left(x^8 - 28x^6 + 210x^4 - 420x^2 + 105\right)q, \tag{29}$$

as well as the skewness, the kurtosis, and the relative fifth central moment

$$S_k = \mu_3/\mu_2^{1.5}, \tag{30}$$

$$Ku = \mu_4/\mu_2^2, \tag{31}$$

$$Ku_5 = \mu_5/\mu_2^{2.5}. \tag{32}$$

Note that the indexes in (27-29) are equal to the indexes of the corresponding Hermite polynomials. The vector-row U of the first seven variables included in the Edgeworth series appears as follows

$$U = \left[P \quad -\frac{S_k P_3}{6} \quad \frac{(Ku-3)P_4}{24} \quad \frac{S_k^2 P_6}{72} \quad \frac{-(Ku_5 - 10S_k)P_5}{120} \quad \frac{-S_k(Ku-3)P_7}{144} \quad -\frac{S_k^3 P_9}{1296}\right].$$

By using the probabilistic genetic algorithm, a model with interaction term is obtaining including the next four variables: intercept, $U(1), U(2), U(4)U(7)$. After transformation, model for the estimator $\hat{P}_r$ of the probability of the event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$ has the form:

$$\hat{P}_r = a(1) + a(2)P + a(3)S_k\left(P_3 + a(4)S_k^4 P_6 P_9\right), \tag{33}$$

where $a(i)$ is the corresponding element of the vector-row of coefficients a:

$$a = \left[-0.0498380531724143 \;\; 1.11882644043104 \;\; -0.17343711278363 \;\; 0.167210843039776\times10^{-3}\right].$$

In addition, if $\hat{P}_r < 0$, we take $\hat{P}_r = 0$, if $\hat{P}_r > 0.5$, we take $\hat{P}_r = 0.5$.

The above coefficients were established as follows. 15 tasks were generated and taken from literature, while for each of which 100 variants of the coefficients were used. That is, the volume of the training set was 1500. In the process we kept in mind the above mentioned restrictions to the coefficient of determination R^2, and the value of σ was selected accordingly. The true probability of fulfilment of the above mentioned inequality was calculated using a Monte Carlo method with a number of tests equals 2×10^6. The number of the predictors ranged from two to ten.

The maximum number of predictors taken requires some explanation. The literature data and the author's experience shows that the number of predictors permitting one to reach values of the coefficient of determination R^2 of 0.8 and above, with rare exceptions does not exceed ten, if the researcher properly use the methodology of the model selection as well as the nonlinear predictors (in particular, the multivariate polynomials). It seems reasonable to argue that the steadily functioning technical, biological or

other systems cannot be stable under a large number of degrees of freedom. As examples for such arguments, let us refer to the following works [28]-[33].

In all these works, despite the heterogeneity of the modeling objects, the number of variables in the model does not exceed nine, with variations of R^2 within the range 0.74 - 0.99. A rather characteristic situation has arisen during the research considered in [29]. The discussion in the journal Chemical Fibres in the former USSR, with the involvement of the leading theorists and practitioners, has identified 150 factors influencing the structure of the viscose fibres. At the same time the parametrically linear regression model of the indicator of the structure uniformity contained eight polynomial variables and had $R^2 = 0.85$. The last value proved to be the maximum achievable for the characteristics of the measurement errors in the experiment. A good confirmation of the efficiency of this model was the fact that a change of the parameters of the fabrication process carried out on the basis of the model equation provided very high increase of the highest-quality product. Note also, when constructing the model of the probability presented above (33), there was a set of 35 possible multivariate polynomials of the second order. However, only three variables entered the model, providing $R^2 = 0.985$.

Let us consider the validation of the received model (33). Similar to the formation of the training set, the testing set of the same size was created. For this set, the following characteristics of the errors in the calculation of the probability are obtained: mean value −0.00054 against 0.00022 in the training set, the standard deviation 0.01289 against 0.01598, skewness 0.64 and kurtosis 9.17. 95 percent of the errors lie in the interval [−0.024583 0.028161]. We add that Equation (33) has also been continuously tested in the subsequent investigations and corresponded to the above results. Thus one can consider that the model (33) provides acceptable accuracy in determining the probability of the event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$. Note, incidentally, that the error in the determination of the probability in the problem discussed in the introduction, is 0.00468.

Let us give some properties of the probability under study.

1) As can be seen from (16-21) the central moments are even functions of $\beta' z$. Thus the probability is also an even function. In other words, a simultaneous change of all signs of the coefficients does not change the probability.

2) For fixed values of the coefficient of determination R^2 the product $\beta' z$ has the decisive influence on the probability. We illustrate this assertion by an example on the data [5] (table D11, Example of multicollinearity). The average VIF for them equals 459, so there is a very high degree of multicollinearity. Let us set the values of the coefficients as follows: $\beta = \begin{bmatrix} 3.0 \\ -1.7 \\ -1.2 \end{bmatrix}$ (then standardized coefficients are equal to $\beta_{st} = \begin{bmatrix} 65.69 \\ -38.79 \\ -19.08 \end{bmatrix}$). Next, let us add to the response the normal noise providing $R^2 = 0.85$. The probability of the event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$ calculated from (33) equals 0.462. Now

we change the sign of the first coefficient to its opposite and add the normal noise with the different variance, so that R^2 remains the same. The Equation (33) gives a probability of 0.09.

Thus, under equal values of R^2 and $\beta'_{st}\beta_{st}$ but different signs of the coefficients and therefore the other value of the product $\beta' z$, the probability can be drastically different.

3) For any sample available to the researcher, even in the presence of a high degree of multicollinearity, the probability of the event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$ can range from 0 to 0.5. This probability depends on the coefficients β, their signs, and the properties of the matrix *X*, specifically its eigenvalues and eigenvectors. The value of the σ, from which the distribution under study depends, can be easily calculated, if X, Y, and β are known. However, since in real life the coefficients β and their signs are unknown, it is not possible to evaluate the probability without a priori information. The attempt to use the LS estimates derived from a sample can be misleading because the latter can sharply differ from the true coefficients by the magnitudes and signs.

4) The probability increases with the increase in the coefficient of determination R^2, if the magnitudes and signs of the true coefficients are unchanged.

Having derived the above results, in the next section the author will try to assess how the magnitude of the investigated probability affects the efficiency of the shrinkage estimators.

2.3. The Influence of the Probability Magnitude of the Event $b'_{st}b_{st} - \beta'_{st}\beta_{st} < 0$ on the Efficiency of the Shrinkage Estimators

In this chapter, the following shrinkage estimators will be considered: ridge estimator, the principal component estimator, and the James-Stein estimator. The Lasso method will not be considered separately, because this approach yields result that is close to the result of the ridge estimator, if the structure of the model is unchanged, as it was adopted in this paper. To calculate the above estimators we use the algorithms presented in [34]. Modeling of a number of tasks under the above stipulated conditions leads to the following conclusions.

1) The case of high probability of the considered event (0.4 and more).

First, note that in this case about half of the shrinkage point estimates will be worse than the LS point wise estimates. Further, the high probability unconditionally negatively affects the MSE of the shrinkage estimators. In this case, for the James-Stein estimator the ratio MSE_{st}/L_{st}^2 is very close to 1. In addition, this estimator cannot correct the false signs of the coefficients obtained by the LS method. The risk of the ridge regression is slightly less than the risk of the LS estimator or exceeds it, if we choose the regularization parameter in (7) as $r = ks^2/b'b$ [14]. In the latter case, the ratio MSE_{st}/L_{st}^2 can reach a value of 1.65. The Principal Component estimator may respond to a high probability dramatically. Recall, in the example discussed in the introduction $MSE_{st}/L_{st}^2 = 16.4$, while in the example of the second paragraph of the previous chapter this ratio is 5.5. In the generated example with 10 variables, sample size

70, and average VIF equal to 123, the ratio MSE_{st}/L_{st}^2 fluctuates in a small range around the value of 119 when we change the number of components from 1 to 9. It should be added that in the present case, the PCR estimator often gives such small absolute values of the coefficients that researcher may conclude that the predictors included in the model, do not in any way affect the response. But this may absolutely not correspond to reality. In addition, greater differences were observed in the value of MSE_{st}/L_{st}^2 when preserving a different number of components.

2) The case of low probability of the considered event (0.3 and less).

This case is more favorable for the use of shrinkage estimators. So, for the James-Stein estimator we successfully reached $MSE_{st}/L_{st}^2 = 0.87$, for the F criterion and R^2 in the range given above. The ridge estimator provides a significant reduction in the ratio MSE_{st}/L_{st}^2 which can reach a value of 0.4 or even less. The Principal Component estimator gives the similar results. In both cases however, despite the decrease in this ratio the risk may be unsatisfactory for practical application for a high degree of multicollinearity. As a result one should increase the R^2 and thus increase the probability, which can reach high values. It is important to emphasize the following. Paradoxically as it may seem, the risk of the LS estimator and simultaneously the risk of the shrinkage estimators for small probabilities may be considerably larger than under large probabilities even though the R^2 and the sum of squares of the true coefficients β are the same.

So, we found that the probability of the event $b_{st}'b_{st} - \beta_{st}'\beta_{st} < 0$ significantly affects the efficiency of the shrinkage estimators, especially the PCR estimator. But the researcher cannot assess this probability, since it depends on the unknown parameters β. Therefore, a reasonable way is to use a priori information. Apparently, the first thing to take into account is the following.

The analysis shows that a high probability of the event $b_{st}'b_{st} - \beta_{st}'\beta_{st} < 0$ occurs when the regression model describes a system in which there is a compensation of the influence of some factor by other factor. In the language of control theory this is called negative feedback. The example discussed in the introduction illustrates such a situation: the insulin compensates the glucose effect. If the researcher knows a priori that such compensation exists, then using shrinkage estimators is inexpedient.

Low probability of the event $b_{st}'b_{st} - \beta_{st}'\beta_{st} < 0$ corresponds to a different class of regression models in which there is multicollinearity but there is no compensation. An example of the such opposite situation is when the regression model establishes a relationship between the temperature at some point of the steam turbine path and two other temperatures ahead of this place [28]. Generally, last two temperatures significantly correlated, that is, multicollinearity is there exists, but compensation in the above sense is absent. In this case, there is reason to assume that the shrinkage estimators is appropriate. However, the difficulties of choosing the regularization parameter or amount of stored components remain. As for the James-Stein estimator, the benefit is small as shown above. Furthermore, as shown in [18], efficiency of this estimator decreases with increasing R^2.

2.4. The Inequality Constrained Least Squares and the Dual Estimator Methods in Conditions of Multicollinearity

In the previous chapter, we established that there are two classes of regression models. For the first class, the impact of some predictors on the response is compensated by other predictors. In the second class, there is no such compensation. In the language of specialists in control, this corresponds to models which describe systems with negative feedback and without. Let us emphasize, first, that strong negative feedback always results in multicollinearity, and second, that in this case the elimination of one of the compensated predictors is unacceptable. Statistically, the aforementioned models are fundamentally different. In the first case, the probability of the event $b_{st}'b_{st} - \beta_{st}'\beta_{st} < 0$ can be high, and hence utilizing shrinkage estimators is inadvisable. In the second case, we have the opposite situation. Belonging to a particular class depends on the values as well as the signs of the true coefficients. But since both are unknown, it is necessary to use a priori information. But in similar circumstances, we must add to the above four types of shrinkage estimators two additional approaches which may be beneficial. We are talking about the Inequality Constrained Least Squares method (ICLS), and the Dual estimator (DE) proposed by the author of this paper in [18].

The first method has a fairly long history [35] and is presented in mathematical packages, particularly in Matlab package as a lsqlin function. This function performs a constrained optimization of the following form

$$\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 \quad \text{such that} \begin{cases} A\beta \le b_L \\ A_{eq}\beta = beq \\ lb \le \beta \le ub \end{cases}, \tag{34}$$

where $A, b_L, A_{eq}, beq, lb, ub$ are a priori known matrices and vectors of appropriate sizes.

The essence of the second method is as follows. For the regression model (1), one can use the estimator

$$b_d = b - sign\left(z_1'\delta\right)\sqrt{\frac{\lambda_1}{\pi}}\frac{\Gamma\left(\left(n-k+1\right)/2\right)}{\Gamma\left(\left(n-k+2\right)/2\right)}\sqrt{e'e}z_1, \tag{35}$$

where δ defined by the (15) for standardized and nonstandardized values, $\Gamma(x)$ is the gamma-function, and λ_1, z_1 are the maximal eigenvalue of the matrix $\left(X'X\right)^{-1}$, and the normalized eigenvector corresponding to it, respectively, e is the regression residual vector. Taking into account the zero probability of the event $\delta = 0$, from (35) one has two solutions, corresponding to the signs plus and minus.

One of these estimators is [18] unbiased, consistent, and its quadratic risk L_{dual}^2 is

$$L_{dual}^2 = \left(1 - \frac{n-k}{\pi}\frac{\Gamma^2\left(\left(n-k+1\right)/2\right)}{\Gamma^2\left(\left(n-k+2\right)/2\right)}\frac{\lambda_1}{\sum_{i=1}^{k}\lambda_i}\right)L^2. \tag{36}$$

From (36) it is clear that the latter can be significantly lower than L^2 and with the

growth of a degree of multicollinearity the effectiveness of the Dual estimator increases with the ratio $\lambda_1 \Big/ \sum_{i=1}^{k} \lambda_i$ other factors being equal. Choosing the right solutions of the two alternatives (35) is carried out through the use of a priori information. It is also important that the method allows estimating of only part of the coefficients naturally for predictors that highly correlated with the others.

The value of any method of estimation based on the use of a priori information largely defined by its universality regarding possible forms of this information.

The shrinkage estimators allow only one kind of a priori information, which confirms the validity of the inequality $b'b > \beta'\beta$ and do not admit the nonstrict inequalities.

The Inequality Constrained Least Squares method has a high degree of flexibility, as is evident from (34) allowing to take into account constraints on the unknown regression coefficients in the form of systems of inequalities and equalities simultaneously. However, unlike the shrinkage estimators, this method permits only nonstrict inequalities, for example $lb \le \beta \le ub$. Therefore, inequalities of the type $\beta_i > \beta_j$ cannot be utilized. In addition, if the some element of the vector lb or ub equal to zero, due to the nature of the optimization algorithm one can obtain the estimate $\widehat{\beta}_i = 0$, which is often meaningless. A significant drawback is the fact that, if a priori information is related to the response, the researcher need to solve the inverse problem in order to obtain the inequalities for the regression coefficients. And it can be very difficult.

The possibility of using a priori information for the Dual estimator is entirely universal. The advantages of this method are most obvious in the two following situations. The first of these, which is considered in the previous chapter as an example of the regression model for temperatures, is characterized by the presence multicollinearity and an absence of mutual compensation of the influence of the predictors. In this case, Inequality Constrained Least Squares method is not applicable, the shrinkage estimators are biased and require the additional parameters, whereas the Dual estimator method gives the unbiased and consistent solution in the explicit form

$$b_d = b - sign\left(z_1'b\right)\sqrt{\frac{\lambda_1}{\pi}}\frac{\Gamma\left(\left(n-k+1\right)/2\right)}{\Gamma\left(\left(n-k+2\right)/2\right)}\sqrt{e'e}z_1. \tag{37}$$

As indicated, the example of the first situation given in the previous section is the dependence of some temperature on two others. If we use the method under study and (4), (33), (36), we obtain the probability $P_r = 0.266$ and $L^2_{dual\_st} \Big/ L^2_{st} = 0.376$, which is completely satisfactory. We point out that for this estimator there remains in force the above precaution with regard to the increase of the risk under the low probability.

Note also, that a priori information considered is qualitative, and thus in many cases it is the least burdensome and, hence, convenient for the researcher.

The second situation consists in confirmation of one of two competing theories with the help of an experiment. Application of the Dual estimator in this case is illustrated in details by a verification of the General Theory of Relativity based upon astronomical data [18].

Of interest are the possibilities of the methods considered for solving the problem

presented in the first chapter as a counterexample. Suppose the researcher have a priori information regarding the coefficient at a dose of insulin, which consists in the fact that $-45 \le \beta_2 \le -25$. For the standardized coefficient, one can obtain $-242.7 \le \beta_{2st} \le -134.8$. Let us use this information and find the ICLS estimate. In our case $lb = \begin{bmatrix} -\infty \\ -242.7 \end{bmatrix}$, $ub = \begin{bmatrix} \infty \\ -134.8 \end{bmatrix}$. The lsqlin function gives the solution $b_{icls} = \begin{bmatrix} 110.74 \\ -134.8 \end{bmatrix}$. Obviously, this solution is closer to the true values $\beta_{st} = \begin{bmatrix} 167.89 \\ -199.59 \end{bmatrix}$ than the LS estimate $b_{st} = \begin{bmatrix} 93.02 \\ -116.92 \end{bmatrix}$. Let us now find the Dual estimate. Calculating the standardized values b_{1st}, b_{2st} we derive using (35) $b_{1st} = \begin{bmatrix} 128.4 \\ -152.28 \end{bmatrix}$, $b_{2st} = \begin{bmatrix} 57.64 \\ -81.57 \end{bmatrix}$. It is evident that b_{2st} does not correspond to a priori information, and our solution will be b_{1st}. This solution is better than the previous one provided by lsqlin. We add that the length of individual confidence intervals, which is calculated by the expression (43) presented below, equals 0.636 times the same intervals for the LS estimate.

Let us now compare the capabilities of the two methods under study on a set of trials of the first, difficult for estimation, class of regression models with a constraint in the form of nonstrict inequality. For the Dual estimator method, we shall apply the following algorithm to account for a priori information.

Let the limits for some true regression coefficient $a_1 \le \beta_i \le a_2$ be known to the researcher. At this point, the reader may wonder why we are considering restrictions only for one coefficient. Our answer is that this is the most available constraint in many applications, since if we have a system of the inequalities for several coefficients then an intractable question of its consistency arises. In addition, under conditions of multicollinearity, when one coefficient is known the others are well established using Equation (41), presented below.

Let us find the estimate b_{di} for the coefficient β_i by the following rule

$$b_{di} = \begin{cases} b_i, & \text{if } a_1 \le b_i \le a_2 \\ b_i + p, & \text{if } a_1 - p < b_i < a_1 \\ b_i - p, & \text{if } a_2 < b_i < a_2 + p, \\ a_1, & \text{if } b_i < a_1 - p \\ a_2, & \text{if } b_i > a_2 + p \end{cases} \tag{38}$$

where

$$p = \sqrt{\frac{G_{i,i}}{\pi}} \frac{\Gamma\left((n-k+1)/2\right)}{\Gamma\left((n-k+2)/2\right)} \sqrt{e'e}, \tag{39}$$

and $G_{i,i}$ is a diagonal element with the number i of the inverse matrix G:

$$G = \left(X'X\right)^{-1}. \tag{40}$$

Having found b_{di} we can obtain the remaining coefficients using the known expres-

sion for the LS estimator with restrictions in the form of equation [3]:

$$b_d = b + \left(X'X\right)^{-1} H' \left(H\left(X'X\right)^{-1} H'\right)^{-1} \left(b_{di} - Hb\right), \tag{41}$$

in which H is a row vector of size k that contains 1 in the *ith* position and zeroes in all the remaining positions.

Let us illustrate this method on the counterexample from the second Chapter. Using Monte Carlo method let us compute Euclidean distances between the estimates and the true values of the coefficients for the two methods (ICLS and DE) and then we find the ratio of these distances to the same distance, given by the LS method. We obtain the ratio of 0.67 for the rule (38)-(41) and 0.78 for the Inequality Constrained Least Squares method. While extremely unsymmetrical a priori interval will give these ratios 0.70 and 0.71 respectively.

These results are consistent with the detailed research in [18]. For the first method, a priori information which is symmetric with respect to the parameter β_i gives on 15 - 20 percent less distance. For extremely asymmetric information distances are almost equal. *i.e.* an average the first method has an advantage.

Finally, consider the question relative to the individual confidence intervals in the Dual estimator. Let b_{d1} be the estimator which corresponds to the plus sign in (35) and b_{d2} be estimator which corresponds to the minus sign. As values a_1 and a_2 are known let us find b_{di} from (38) and the closest to this estimate *ith* element of the three possible vectors b_i, b_{d1i}, b_{d2i}.

The initial individual confidence intervals for these three values are [3] [18]:

$$b_i \pm t\left(n-k, 1-\alpha/2\right)\sqrt{\left(X'X\right)^{-1}_{i,i}}\, s, \tag{42}$$

$$b_{d1i} \pm t\left(n-k, 1-\alpha/2\right)\sqrt{Q_{i,i}}\, s, \tag{43}$$

$$b_{d2i} \pm t\left(n-k, 1-\alpha/2\right)\sqrt{Q_{i,i}}\, s, \tag{44}$$

$$Q = \left[\left(X'X\right)^{-1} - \left(n-k\right)\frac{\lambda_1}{\pi}\frac{\Gamma^2\left(\left(n-k+1\right)/2\right)}{\Gamma^2\left(\left(n-k+2\right)/2\right)} z_1 z_1'\right], \tag{45}$$

where $t\left(n-k, 1-\alpha/2\right)$ is the $1-\alpha/2$ point of the Student distribution with $n-k$ degrees of freedom, $Q_{i,i}$ is the corresponding diagonal element of the matrix Q from (45).

It remains to find the intersection of the confidence interval specified for the nearest element and a priori interval. This intersection will be the new confidence interval. The confidence intervals for the remaining coefficients are easily found with the aid of equation (41).

It is clear that one should compare the width of the new confidence intervals with the width of a priori interval. It is likewise evident that new confidence intervals cannot be wider than a priori intervals. Their ratio depends on the data sample properties, the value of a priori interval, and its location relative to the sought parameter. Tests on a large number of independent tasks have shown that a new confidence interval may be less by a factor of 1.3 than the a priori interval.

3. Results and Discussion

The article has analyzed the empirical approaches and the statistical methods which facilitate reducing the influence of multicollinearity on the estimation of coefficients in linear regression.

Cautions were expressed against the use without proper analysis of some unfortunately generally-accepted recommendations, such as excluding one of the correlated predictors and ignoring multicollinearity in the process of forecasting using the regression equation.

The concepts of the doubtful value of regression models with a large number of the predictors were presented.

The question of utilizing of shrinkage estimators for the purpose of reducing the influence of multicollinearity was considered in detail. It was shown that there are two classes of regression models, the first of which corresponds to systems with negative feedback, and the second of which corresponds to systems without this feedback. The use of shrinkage estimators is inappropriate in the first case. Particularly poor results may be obtained in using the Principal Component estimator. In the second case, the shrinkage estimators may be useful with the right choice of the regularization parameter or in the number of principal components included in the regression model, although this, generally speaking, is problematic. These facts were established by the study of the distribution of the random variable $b'b-\beta'\beta$, where b is the Least Squares estimate and β is the vector of true coefficients, since the form of this distribution is the basic characteristic of the specified classes.

For the purposes of this study, a regression approximation of the distribution of the event $b'b-\beta'\beta<0$ based on the Edgeworth series was developed.

The essential result is the investigation of alternative approaches to address the problem of multicollinearity. These approaches consist in application of the known Inequality Constrained Least Squares method and the Dual estimator method proposed by the author. It has been shown that for the models of both classes, with the presence of external information these methods can significantly reduce the Euclidean distance between vectors of estimates and true coefficients as well as the confidence intervals of the estimates. For the second class of models, the Dual estimator method gives unbiased and consistent solution in explicit form, and thus has no competitors. This method is also very effective in the problem of a confirmation of one of two competing theories with the help of an experiment.

Acknowledgements

The author is grateful to the anonymous referees and the editors for an excellent, constructive, and extremely helpful review of the paper.

References

[1] Sen, A.K. and Srivastava, M.S. (1990) Regression Analysis: Theory, Methods, and Applications. Springer-Verlag, New York, 347.

[2] Belsley, D.A., Kuh, T.D. and Welsch, R.E. (1980) Regression Diagnostics. John Wiley & Sons Inc., Hoboken, 291. http://dx.doi.org/10.1002/0471725153

[3] Draper, H.R. and Smith, H. (1998) Applied Regression Analysis. 3rd Edition, John Wiley & Sons Inc., New York, 713. http://dx.doi.org/10.1002/9781118625590

[4] Gruber, M.H.J. (1998) Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators. Marcel Dekker Inc., New York.

[5] Hocking, R.R. (2003) Methods and Applications of Linear Models. John Wiley and Sons Inc., Hoboken. http://dx.doi.org/10.1002/0471434159

[6] Rao, C.R. and Toutenburg, H. (1995) Linear Models: Least Squares and Alternatives. Springer, Berlin. http://dx.doi.org/10.1007/978-1-4899-0024-1

[7] Duzan, H. and Shariff, N.S.B.M. (2015) Ridge Regression for Solving the Multicollinearity Problem: Review of Methods and Models. *Journal of Applied Sciences*, **15**, 392-404. http://dx.doi.org/10.3923/jas.2015.392.404

[8] El-Dereny, M. and Rashwan, N.I. (2011) Solving Multicollinearity Problem Using Ridge Regression Models. *International Journal of Contemporary Mathematical Sciences*, **6**, 585-600.

[9] Graham, M.H. (2003) Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, **84**, 2809-2815. http://dx.doi.org/10.1890/02-3114

[10] Kraha1, A., Turner, H., Nimon, K., *et al.* (2012) Tools to Support Interpreting Multiple Regression in the Face of Multicollinearity. *Frontiers in Psychology*, **3**, 44.

[11] Vatcheva, K.P., Lee, M., McCormick, J.B. and Rahbar, M.H. (2016) Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology*, **6**, 227.

[12] Yoo, W., Mayberry, R., Bae, S., *et al.* (2014) A Study of Effects of Multicollinearity in the Multivariable Analysis. *International Journal of Applied Science and Technology*, **4**, 9-19.

[13] Bersten, A.D. (1998) Measurement of Overinflation by Multiple Linear Regression Analysis in Patients with Acute Lung Injury. *European Respiratory Journal*, **12**, 526-532. http://dx.doi.org/10.1183/09031936.98.12030526

[14] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression. Biased Estimation for Nonorthogonal Problems. *Technometrics*, **42**, 55-67. http://dx.doi.org/10.1080/00401706.1970.10488634

[15] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267-288.

[16] Jolliffe, I.T. (2002) Principal Component Analysis. Springer, Berlin, 405.

[17] James, W. and Stein, C. (1961) Estimation with Quadratic Loss. *Proceedings of the* 4*th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 361-379.

[18] Gordinsky, A. (2013) A Dual Estimator as a Tool for Solving Regression Problems. *Electronic Journal of Statistics*, **7**, 2372-2394. http://dx.doi.org/10.1214/13-EJS848

[19] Neithercott, T. (2010) A User's Guide to Insulin. Diabetes Forecast, 4. www.diabetesforecast.org

[20] Yoshioka, S. (1986) Multicollinearity and Avoidance in Regression Analysis. *Behaviormetrika*, **13**, 103-120. http://dx.doi.org/10.2333/bhmk.13.19_103

[21] Castano-Martineza, A. and Lopez-Blazquezb, F. (2006) Distribution of a Sum of Weighted Central Chi-Square Variables. *Communications in Statistics- Theory and Methods*, **34**, 515-524. http://dx.doi.org/10.1081/STA-200052148

[22] Withers, C.S. and Nadarajah, S. (2013) Expressions for the Distribution and Percentiles of the Sums and Products of Chi-Squares. *Statistics*, **47**, 1343-1362.

http://dx.doi.org/10.1080/02331888.2012.658399

[23] Wood, A.T.A. (1989) An F Approximation to the Distribution of a Linear Combination of Chi-Squared Variables. *Communication in Statistics Simulation and Computation*, **18**, 1439-1456. http://dx.doi.org/10.1080/03610918908812833

[24] Gnedenko, B. (1962) The Theory of Probability. Translated from the Russian, Chelsea, New York, 472. http://dx.doi.org/10.1063/1.3057804

[25] Cramer, H. (1946) Mathematical Methods of Statistics. Princeton Mathematical Series 9, Princeton University Press, Princeton, 575.

[26] Mnatsakanov, R.M. and Hakobyan, B.S. (2009) Recovery of Distributions via Moments. IMS Lecture Notes Monograph Series, *Optimality*: *The* 3*rd Erich L. Lehmann Symposium*, **57**, 252-265. http://dx.doi.org/10.1214/09-lnms5715

[27] Kendall, M.G. and Stuart, A. (1962) The Advanced Theory of Statistics, Vol. 1, Distribution Theory. 2th Edition, Griffin, London, 573.

[28] Gordinsky, A., Plotkin, E., Benenson, E. and Leizerovich, A. (2000) A New Approach to Statistic Processing of Steam Parameter Measurements in the Steam Turbine Path to Diagnose Its Condition. *Proceeding of the International Joint Power Generation Conference*, Miami Beach, 23-26 July 2000, 1-5.

[29] Gordinsky, A. (1996) Viscose Film and Textile Fibres Quality Investigation and Control in Industry. *The* 11*th International Conference of the Israel Society for Quality*, Jerusalem, 19-21 November 1996, 185-190.

[30] Kessler, V., Guttmann, J. and Newth, C.J.L. (2001) Dynamic Respiratory System Mechanics in Infants during Pressure and Volume Controlled Ventilation. *European Respiratory Journal*, **17**, 115-121. http://dx.doi.org/10.1183/09031936.01.17101150

[31] Leiphart, D.J. and Hart, B.S. (2001) Comparison of Linear Regression and a Probabilistic Neural Network to Predict Porosity from 3-D Seismic Attributes in Lower Brushy Canyon Channeled Sandstones, Southeast New Mexico. *Geophysics*, **66**, 1349-1358. http://dx.doi.org/10.1190/1.1487080

[32] Muramatsu, K., Yukitake, K., Nakamura, M., Matsumoto, I. and Motohiro, Y. (2001) Monitoring of Nonlinear Respiratory Elastance Using a Multiple Linear Regression Analysis. *European Respiratory Journal*, **17**, 1158-1166. http://dx.doi.org/10.1183/09031936.01.00017801

[33] Plotts, T. (2011) A Multiple Regression Analysis of Factors Concerning Superintendent Longevity and Continuity Relative to Student Achievement. Seton Hall University Dissertations and Theses (ETDs) Paper 484.

[34] Pantula, J.F. (1987) Optimal Prediction in Linear Regression Analysis. A Dissertation, the University of North Carolina, Chapel Hill, 194.

[35] Knopov, P.S. and Korkhin, A.S. (2012) Regression Analysis under a Priori Parameter Restrictions. Springer, Berlin. http://dx.doi.org/10.1007/978-1-4614-0574-0