Electronic Journal of Statistics
Vol. 7 (2013) 2372–2394
ISSN: 1935-7524
DOI: 10.1214/13-EJS848

A Dual estimator as a tool for solving regression problems

Anatoly Gordinsky

Berman Engineering Ltd, Israel
e-mail: agordinsky@gmail.com

Abstract: This paper discusses a parameter estimation method that employs an unusual estimator called the Dual estimator. For a linear regression model, we obtain two alternative estimators by subtracting or adding a certain vector to the vector of the Ordinary Least Squares Estimator (OLSE). One of them strictly dominates the latter. Moreover, under the normality assumption this estimator is unbiased, and consistent, and has significantly smaller variance than the OLSE. The use of a priori information is a universal way to choose a better alternative. An important property of the proposed method is the possibility of using the strict inequalities as a priori information. In particular, if the external information is that the L2-norm of the OLS estimate exceeds the same norm of a vector of true coefficients, one can choose a better alternative without additional parameters. If it is known that the parameter is restricted by a linear non-strict inequality, the method has a smaller Mean Squared Error than a Constrained Least Squares technique. Finally, a priori information on two possible parameter values can be successfully used for the experimental confirmation of one of two alternative theories, which is illustrated by a verification of the General Theory of Relativity based upon astronomical data.

AMS 2000 subject classifications: Primary 62G05; secondary 62J05.
Keywords and phrases: Linear regression, Dual estimator, strictly dominating, unbiasedness, consistency, advantage in variance, robustness, comparative analysis.

Received August 2013.

Contents

1. Introduction

Let us consider the problem of parameters estimation for the following linear regression model:

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

where $Y, \epsilon \in \mathbb{R}^n,\ E(\epsilon) = 0,\ \mathrm{cov}(\epsilon) = \sigma^2 I_n,\ \sigma^2$ is the variance of ϵ, known $X \in \mathbb{R}^{n\times k}$ of rank k, and $\beta \in \mathbb{R}^k$ is unknown. The above stated conditions correspond fully with the premises of the Gauss-Markov theorem [31]. According to the theorem, the Ordinary Least Squares Estimator (OLS estimator), defined as

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

is unbiased and has the smallest variance in the class of unbiased estimators that are linear relative to Y. Under conditions of normality the estimator properties are improved.

However, the fact that the estimator is optimal in one sense does not yet guarantee that it is good enough for practical purposes. Input data multicollinearity is one of the most important and well-known problems. Let us use quadratic risk as a criterion for quality of the regression parameters estimation

$$L^2 = E((b-\beta)'(b-\beta)). \tag{3}$$

From the known relation [32]

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

where Tr is the trace, and λ_i are the eigenvalues of the inverse matrix, it is easy to see that when input data has multicollinearity, the quadratic risk (4) can dramatically increase.

However, despite all the importance of decreasing the effect of multicollinearity, resolving this issue does not remove all the problems of regression parameters estimation. There are many problems with orthogonal or almost orthogonal columns of the design matrix X (i.e. without multicollinearity) for which, considering the available characteristics of the data, the least squares method cannot provide the estimation quality required for the task. These include one-dimensional problems of measurement processing, such as evaluation of fundamental constants in physics and astrophysics [35, 30], many problems in biology, medical and technical diagnostics, analytical chemistry, quality control, etc. In particular, orthogonal problems arising in experimental design fall into this category. It can be assumed that, as scientific and technical problems to be solved become more complex, the need for better quality regression estimates will still be preserved and will even increase. As an example of a strict requirement for estimation accuracy of regression model parameters, let us consider the development of a regression model for diagnostics of a powerful steam turbine condition. In [12] it was shown that for a reliable detection of a stage failure, the regression model must have a multiple correlation coefficient $R \geq 0.995$.

In the past decades, there have been hundreds of studies on improving the quality of regression estimation (1), and we shall refer only to selected ones.

Shrinkage and therefore biased estimators of different types were proposed and investigated: examples include the James-Stein estimator [19, 26, 3], Ridge Regression [17, 27], Principal Component Regression (PCR) [14, 3, 22], Least Absolute Shrinkage and Selection Operator (LASSO) [37, 15], and their variants. A Bayesian approach to regression estimation was developed and studied in [11, 4, 34], as well as the extremely flexible Constrained Least Squares Method (CLS) [23, 33, 28, 25]. These approaches share the use of a priori information presented as assumptions about the norm of the sought-for (true) vector, the existence of the region in which it is located, the possibility of excluding the components corresponding to small eigenvalues of the matrix $X'X$ in PCR, a priori distribution of the parameter, etc. However, the abundance of methods and studies on their development indicates that each of them has its own area of suitable applications, and there is no superior method. This well-known fact is based primarily on numerous studies by simulation and analysis of the estimators under different conditions [29, 9, 21, 38, 14, 20].

To illustrate this situation, we shall consider the possibilities and limitations of some classical methods, namely, Ridge Regression, Principal Component Regression (PCR), James-Stein Regression, and Constrained Least Squares (CLS).

The first and the second methods can be effective in the presence of multicollinearity while they are nearly useless otherwise [38]. If b_s is the OLS estimate calculated for the given sample X, Y, then the inequality ${b_s}'b_s > \beta'\beta$ is a necessary condition for their applicability. The validity of this inequality in the literature is usually based on the fact that the mathematical expectation of the left side is greater than $\beta'\beta$. However, this does not mean that the inequality will be true in a specific case. Moreover, under certain conditions, the probability that the inequality is true could be close to 0.5. Therefore, this property must be based on a priori information. This is exactly the view that is substantiated in [6]. However, for Ridge Regression the mentioned a priori information is not sufficient, and a certain parameter must be set. There are dozens of ways to do that [5]. Whichever way it is done, the specified parameter depends on the sought-for vector β and the unknown σ, which creates serious difficulties. The main one is the estimation of the real mean squares error (MSE).

For PCR it is necessary to choose the number of components, which can be vital for the regression coefficient estimating problem. As far back as 1982, in [21] it was indicated that components with low variance can be just as important as components having large variance. The same warning is contained in [32, 14]. In addition, as the study of shrinkage estimators using the Monte Carlo method has shown [38], PCR is an unstable estimator. In a regression model with three variables, the sum of squares of the preset positive factors was constant, but the values of these factors varied. Thus, depending on these values and the standard deviation, PCR could give an 8 times smaller or a 15 times larger MSE than the variance of the OLS estimator. In these circumstances, evaluation of the actual MSE is a hard problem.

The James-Stein estimator does not require any parameter setting, always has a lower MSE than the OLS estimator variance, and allows analytical evaluation of its value (see, e.g., [26]). These are its undoubted advantages. However, this

estimator has significant limitations as well. The first is that this method only works if $k \geq 3$; therefore, it cannot be used to solve such important problems as direct measurements processing, estimation of a line slope, or numerous two-dimensional problems. The second and very significant drawback of the James-Stein estimator is that it does not fully correct the distortion of the regression coefficients caused by multicollinearity, and, in particular, incorrect signs of these coefficients [1]. Finally, the same work presents the James-Stein estimator in the following form:

$$b_{JS} = \left[1 - \frac{(k-2)(n-k-1)(1-R^2)}{n(n-k+2)R^2}\right] b, \tag{5}$$

where R^2 is the coefficient of determination.

To obtain useful results in real life applications the regression model should be adequate, and the significance of the coefficient of determination is the essential condition of its adequacy. In many cases this leads to the fact that b_{JS} differs little from b. Indeed, if, for example, $R^2 = 0.8$, $k = 5$, and, as recommended in the literature, $n/k = 15$, we get $b_{JS} = 0.9904b$. Accordingly, the decrease in MSE and the usefulness of the approach will be very small.

The Constrained Least Squares Method is a universal method applicable to the problems of any dimension and any conditionality – orthogonal, multicollinear, intermediate. However, this method also has limitations. First, it cannot be used with constraints that have the form of strict inequalities. Second, even if the inequalities are not strict, the method does not allow using a priori relationships for the desired parameters regarding an OLS estimate, such as the inequality $b_s{}'b_s > \beta'\beta$. In such cases, it gives the OLS estimate. Finally, the distribution of the estimate obtained by this method may be unfavourable for the meaningful interpretation of the results. For example, if it is known that in the one-dimensional regression the true coefficient is positive, then in all cases where $b_s < 0$ we derive the CLS estimate $b_s = 0$.

All of the above confirms that the development and improvement of estimation methods is still very relevant.

This article presents a method of parameter estimation that is based on an unusual estimator that we called a Dual estimator. This approach has been partially described in the preprint [13]. Here we present an in-depth study of the method, and simultaneously change the name of the Quasi-estimator for Dual estimator, believing that the latter is more appropriate. We shall show that for a linear regression model we can obtain two alternative estimators by subtracting or adding a certain vector to the OLS estimator vector. One of them strictly dominates the OLS estimator (2). Moreover, under the normality assumptions this estimator is unbiased, and consistent, and has significantly smaller variance than the OLS estimator. Furthermore, it is robust with respect to deviations from initial preconditions, relative to both the distribution and the properties of the variance-covariance matrix of the error ϵ. Use of a priori information is a universal way to choose the better alternative. An important difference of the proposed method from the known ones is the possibility of using strict inequalities as a priori information. In particular, if it is known that the L2-norm

of the OLS estimate calculated for the given sample of X, Y exceeds the same norm of a vector of true coefficients, one can choose a better alternative without additional parameters. When a priori constraints in the form of non-strict linear inequalities are used, the method under consideration has a smaller mean squared error, averaged on a priori interval, than the CLS approach, and a more favourable distribution of the resulting estimates. Finally, a priori information on two possible parameter values is successfully used for experimental confirmation of one of the two competing theories, which is illustrated by a verification of the General Theory of Relativity based upon astronomical data.

2. The Dual estimator and its properties

Consider two non-linear estimators, non-homogeneous in Y:

$$b_1 = b + c\sqrt{e'e}\, q, \tag{6}$$

$$b_2 = b - c\sqrt{e'e}\, q, \tag{7}$$

where b is an OLS-estimate (2), e is the $n \times 1$ vector of the known regression residuals:

$$e = Y - Xb, \tag{8}$$

q is an arbitrary, normalized ($q'q = 1$), $k \times 1$ vector, and c is a constant, as yet unknown, that we will subsequently define.

The estimators (6) and (7) differ only in the sign of the additive correction to the vector of the OLS-estimator.

Let us define the random error of the OLS estimator as

$$\delta = b - \beta \tag{9}$$

and stipulate that, in every application, we choose, out of the two estimators (6) and (7), the one whose correction sign is equal to $sign(-q'\delta)$, *where the sign-function is defined by the rule:* $sign(x) = 1$, *if* $x \geq 0$, *and* $sign(x) = -1$, *if* $x < 0$. Then we derive the following Dual estimator:

$$\tilde{b} = b - sign(q'\delta)\, c\, \sqrt{e'e}\, q. \tag{10}$$

The term "Dual estimator" is used because the last expression (10) involves a discrete random variable $sign(q'\delta)$, which receives just two values: $+1$ or -1. Let us define the constant c in such a way that the average square of the distance between the Dual estimator (10) and β is minimized for any q. The aforementioned mean square of the distance, for $\tilde{b}$, is equal to:

$$\widetilde{L}^2 = E((\tilde{b} - \beta)'(\tilde{b} - \beta)). \tag{11}$$

Substituting the value $\tilde{b}$ from (10) and b from (9) into (11), taking the derivative with respect to c, and then equating the result to zero, we obtain the following:

$$\tilde{c} = \arg\min_{c \in \mathbb{R}^1} \widetilde{L}^2 = E(\sqrt{(q'\delta)^2 e'e}\,)/E(e'e). \tag{12}$$

Now the Dual estimator with the minimal $\widetilde{L}^2$ will have the following form:

$$\tilde{b} = b - sign(q'\delta)\,\tilde{c}\,\sqrt{e'e}\,q, \tag{13}$$

where $\tilde{c}$ is taken from (12). Note that when $n = k$ the residual sum of squares $e'e = 0$ and, as follows from (13), $\tilde{b} = b$. Thus, hereafter we shall restrict ourselves to the case $n > k$. The aforementioned is already sufficient to prove the following:

Proposition 1. *Let $E(\epsilon) = 0, cov(\epsilon) = \sigma^2 I_n$, $n > k$, and q be an arbitrary, normalized $k \times 1$ vector. Then $\widetilde{L}^2 < L^2$, for the Dual estimator (13).*

Proof. Substitute (12) into (13), and the result obtained into (11). After transformations, writing $sign(q'\delta)\,q'\delta = |q'\delta| = \sqrt{(q'\delta)^2}$ and using (9), we derive:

$$\widetilde{L}^2 = L^2 - [E(\sqrt{(q'\delta)^2 e'e})]^2/E(e'e) \tag{14}$$

Since the value subtracted from L^2 is positive, the proposition is proven. □

Up to this point, we did not consider the distribution of the error ϵ. Assume now that ϵ is normally distributed, i.e. that $\epsilon \sim N(0, \sigma)$, maintaining all our previous stipulations. In that case we can, first of all, determine what the vector q should be in order to minimize the value $\widetilde{L}^2$ from (14).

Proposition 2. *Let $\epsilon \sim N(0, \sigma^2 I_n), n > k$. Then the minimum of $\widetilde{L}^2$ is achieved when the vector q from (13) is equal to the normalized eigenvector z_1 ($z_1'z_1 = 1$) which corresponds to the maximal eigenvalue of the inverse matrix $(X'X)^{-1}$: $z_1 = \arg\min_{q \in \mathbb{S}^k} \widetilde{L}^2$, where the unit sphere of order k is denoted as $\mathbb{S}^k$.*

Proof. Let us consider (14) and, first, prove the independence of the quadratic forms $(q'\delta)^2$ and $e'e$. From (1), (2), (9) we obtain

$$\delta = (X'X)^{-1}X'\epsilon. \tag{15}$$

Further, using (2), (8), (14) and, taking into account that $q'\delta$ is a scalar, let us represent these forms as $\epsilon' T \epsilon$, and $\epsilon' B\,\epsilon$, where the matrices T and B are equal, correspondingly, to:

$$T = X(X'X)^{-1}qq'(X'X)^{-1}X', \tag{16}$$

$$B = I_n - X(X'X)^{-1}X' \tag{17}$$

By direct verification we ascertain that $TB = BT = 0$. This is a necessary and sufficient condition for independence of the quadratic forms from the normal random variables ϵ being considered [36].

It follows from the independence that the numerator of the second term in the right hand side of equation (14) can be represented as a product:

$$\left[E(\sqrt{(q'\delta)^2 e'e})\right]^2 = \left[E(\sqrt{(q'\delta)^2})\right]^2 \left[E(\sqrt{e'e})\right]^2 = \left[E(|\,q'\delta\,|)\right]^2 \left[E(\sqrt{e'e})\right]^2.$$

In order to minimize $\widetilde{L}^2$, one needs to maximize the first multiplicand of the above product. The scalar $q'\delta$ is a normal random variable with expectation zero.

The modulus of such a variable has a half-normal distribution. Its mathematical expectation is given by [10]:

$$E(|q'\delta|) = \sqrt{\frac{2}{\pi}D(q'\delta)} \Rightarrow [E(|q'\delta|)]^2 = \frac{2}{\pi}D\,(q'\delta),$$

where D denotes its variance. It is well known [36] that the variance is

$$D(q'\delta) = \sigma^2(q'(X'X)^{-1}q)$$

where, as designated above, σ is the standard deviation of ϵ.

To finish the proof of the proposition, we are left to find out what should be the value of the normalized vector q in order to maximize the value of the obtained scalar. The answer is given by the Rayleigh-Ritz theorem [18], according to which this maximum is equal to the maximal eigenvalue λ_1 of the matrix $(X'X)^{-1}$ and is obtained at the normalized eigenvector of this matrix corresponding to such a maximal eigenvalue. The proposition is proven. □

Now the Dual estimator gets the following form:

$$\tilde{b} = b - sign(z_1'\delta)\cdot\tilde{c}\sqrt{e'e}\cdot z_1. \tag{18}$$

Normality of ϵ and the above result allow us to define concretely the expression $\tilde{c}$ in (18) and, eventually, derive the final form for the optimal Dual estimator.

Proposition 3. *Let $\epsilon \sim N(0, \sigma^2 I_n), n > k$. Then the Dual estimator (18) has the form:*

$$\tilde{b}_o = b - sign(z_1'\delta)\sqrt{\frac{\lambda_1}{\pi}}\,\frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+2)/2)}\sqrt{e'e}\,z_1, \tag{19}$$

where $\Gamma(x)$ is the gamma-function, and λ_1, z_1, as noted above, are the maximal eigenvalue of the matrix $(X'X)^{-1}$ and the normalized eigenvector corresponding to it, respectively.

Proof. Let us find $\tilde{c}$ from (12) and substitute the derived result into (18). First we find the numerator in (12). As is shown in the proof of Proposition 2, the random variables $q'\delta$ and $e'e$ are independent. Hence, this numerator is equal to $E(|q'\delta|)E(\sqrt{e'e})$. It is also established in the proof of Proposition 2 that $E(|q'\delta|) = \sigma\sqrt{\frac{2}{\pi}q'(X'X)^{-1}q}$. If $q = z_1$, we obtain $E(|q'\delta|) = \sigma\sqrt{\frac{2}{\pi}\lambda_1}$. Since $e'e/\sigma^2 \sim \chi^2_{n-k}$ the mathematical expectation of $\sqrt{e'e}$ is known [8] and is equal to $\sigma\sqrt{2}\,\frac{\Gamma((n-k+1)/2)}{\Gamma((n-k)/2)}$. Substituting the mathematical expectation $e'e$ into the denominator in expression (12), specifically, $E(e'e) = (n-k)\sigma^2$, and using the property $\Gamma(x+1) = x\Gamma(x)$ of the gamma-function [2], we obtain $\Gamma((n-k+2)/2) = ((n-k)/2)\Gamma((n-k)/2)$ and

$$\tilde{c} = \sqrt{\frac{\lambda_1}{\pi}}\,\frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+2)/2)} \tag{20}$$

We are left to substitute this expression for $\tilde{c}$ in equation (18), and Proposition 3 has been proven. □

Corollary 1. *Let some components of the vector q be equal to zero. That is, we only improve regression coefficients with numbers from* $\mathbf{m}$ *which is an arbitrary subset of the set* $1, 2, \ldots, k$. *Let* $(X'X)^{-1}_{mm}$ *be a part of the matrix* $(X'X)^{-1}$ *corresponding to the subset* $\mathbf{m}$, b_m *be a subvector of the OLS-vector b corresponding to the subset* $\mathbf{m}$, *and* λ_m, z_m *be the maximal eigenvalue and its eigenvector of the matrix* $(X'X)^{-1}_{mm}$. *Then the optimal Dual estimator is*

$$\tilde{b}_{mo} = b_m - sign(z'_m \delta_m) \sqrt{\frac{\lambda_m}{\pi}} \frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+2)/2)} \sqrt{e'e}\, z_m. \tag{21}$$

The considered Corollary gives us a convenient tool in cases when there is no necessity to correct all regression coefficients. An obvious example is a situation when variances of these coefficients differ markedly from each other. In this case, it is reasonable to improve the coefficients with large variance. A second example is an event when in the multivariate regression there exists a priori information relative to only one coefficient. This case will be discussed in Section 4.

Now let us consider the ratio of the $\widetilde{L}_o^2$ for the optimal Dual estimator to the L^2 of the OLS-estimator. Possessing the definitions and the results (4), (14) and using the expression for $E(\sqrt{(z'_1\delta)^2 e'e})$ from the proofs of Propositions 2, and 3, we can compute the mentioned ratio as follows:

$$\widetilde{L}_o^2 / L^2 = 1 - \frac{n-k}{\pi} \frac{\Gamma^2((n-k+1)/2)}{\Gamma^2((n-k+2)/2)} \frac{\lambda_1}{\sum_{i=1}^k \lambda_i}. \tag{22}$$

Using the ratio from [2]:

$$\Gamma(\alpha+p)/\Gamma(\alpha+h) = \alpha^{p-h}\left(1 + \frac{1}{2\alpha}(p-h)(p+h-1)\right) + O(1/\alpha^2),$$

we can obtain an approximate but more obvious expression:

$$if\ (n-k)^2/4 \gg 1$$

$$\widetilde{L}_o^2 / L^2 \cong 1 - \frac{2}{\pi}\left(1 - \frac{0.25}{n-k}\right)^2 \frac{\lambda_1}{\sum_{i=1}^k \lambda_i}. \tag{23}$$

As can be seen from (22) and (23), the relative gain for the Dual estimator depends mainly on the distribution of the eigenvalues λ_i of the matrix $(X'X)^{-1}$ and can be quite substantial. Thus, in the one-dimensional case, when $\lambda_1 = \sum_{i=1}^k \lambda_i$, the ratio $\widetilde{L}_o^2 / L^2$ is close to 0.4, i.e. the Dual estimator has $\widetilde{L}_o^2$ smaller than the OLS-estimator L^2 by the ratio of 2.5. This fact can have a great significance, in particular, when processing direct measurements.

In the case of multicollinearity, when $\lambda_1 \cong \sum_{i=1}^k \lambda_i$, we obtain approximately the same result.

When processing orthogonal data, the effect will naturally be smaller and will substantially depend on the number of variables k. But if we only correct some of the coefficients, as is shown in Corollary 1, the situation improves.

Let us establish some additional properties of the optimal Dual estimator (19). When the property under consideration holds for the original Dual estimator (10), with an arbitrary vector q, we shall mention this in our remarks.

Proposition 4. *Let $\epsilon \sim N(0, \sigma^2 I_n), n > k$. Then the optimal Dual estimator $\tilde{b}_o$ from (19) is unbiased, i.e. $E(\tilde{b}_o) = \beta$.*

Proof. Let us compute the mathematical expectations of both sides of equation (19), taking into account the independence of $z_1'\delta$ from $e'e$ proved above and therefore of their functions:

$$E(\tilde{b}_o) = E(b) - E(sign(z_1'\delta))E(\sqrt{e'e})\sqrt{\frac{\lambda_1}{\pi}}\frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+1)/2)}z_1.$$

Using (15) and the defining property of the eigenvector, we obtain:

$$z_1'\delta \sim N(0, \sqrt{\lambda_1}\,\sigma), \tag{24}$$

that is, the random variable $z_1'\delta$ has normal distribution with expectation zero and variance $\lambda_1\sigma^2$. The expectation of the function $sign(z_1'\delta)$ is equal to:

$$E(sign(z_1'\delta)) = \int_{-\infty}^{\infty} sign(x)f(x)\,dx,$$

where $f(x)$, the distribution density of the normally distributed variable in (24), is continuous, symmetrical and bounded above.

In view of our definition of the sign-function ($sign(x) = 1$ when $x \geq 0$ and $sign(x) = -1$ when $x < 0$), the given integral is equal to zero due to the symmetry of the integrand function for all $x \neq 0$ and its boundedness at the point $x = 0$.

From this, since $E(sign(z_1'\delta)) = 0$ and, because of the unbiasedness of the OLS-estimator $E(b) = \beta$, we get $E(\tilde{b}_o) = \beta$. □

Remark 1. We shall obtain the same result for any continuous, symmetric, and bounded above distribution of the error ϵ, for which the mathematical expectation exists.

Remark 2. The unbiasedness property, under conditions of Remark 1, applies also to the estimator (10).

Proposition 5. *Let $\epsilon \sim N(0, \sigma^2 I_n), n > k$. Then the variance-covariance matrix for the optimal Dual estimator (19) is equal to:*

$$Q = E((\tilde{b}_o - E\tilde{b}_o)(\tilde{b}_o - E\tilde{b}_o)') = \sigma^2\left[(X'X)^{-1} - (n-k)\frac{\lambda_1}{\pi}\frac{\Gamma^2((n-k+1)/2)}{\Gamma^2((n-k+2)/2)}z_1z_1'\right]. \tag{25}$$

Proof. Proceeding from the definition of the variance-covariance matrix given on the left side of (25), and taking into account the unbiasedness of the estimator $\tilde{b}_o$, and also the relations (9), (12), (19), after some transformations, we obtain:

$$Q = E(\delta\delta' - \tilde{c}\, sign(z_1'\delta) z_1 \delta' \sqrt{\epsilon' B \epsilon} - \tilde{c}\, sign(z_1'\delta)\delta z_1' \sqrt{\epsilon' B \epsilon} + \tilde{c}^2 \epsilon' B \epsilon\, z_1 z_1') \quad (26)$$

Using (15), the well known expansion $(X'X)^{-1} = \sum_{i=1}^{k} \lambda_i z_i z_i'$ and a property of the scalar $\epsilon' X z_1$, after certain transformations we derive:

$$\begin{aligned} Q = E\Bigg[& \delta\delta' - 2\tilde{c} sign(z'_1\delta) z'_1 \delta \sqrt{e'e}\, z_1 z'_1 + \tilde{c}^2 e'e\, z_1 z' \\ & + \tilde{c} sign(z'_1\delta)\sqrt{e'e}\left(\sum_{i=2}^{k} \lambda_i z'_i X' \epsilon z_1 z'_i + \sum_{i=2}^{k} \lambda_i z_i X' \epsilon z_i z'_1\right)\Bigg]. \end{aligned} \quad (27)$$

The expectation of the fourth summand is equal to zero. This follows from the pairwise and, therefore, mutual independence [36] of the three random scalars $z_1'\delta$, $\epsilon' X z_i$, $\sqrt{\epsilon' B \epsilon}$ and the zero expectation of two of them. Substituting the earlier obtained results into the first three summands of expression (27) we derive (25). □

Let us present also the following facts regarding the variance-covariance matrix Q. Denote

$$\breve{Q} = Q/\sigma^2. \quad (28)$$

We ascertain by checking that the maximal eigenvalue of the matrix $\breve{Q}$ is equal to:

$$\lambda_{\breve{Q}} = \lambda_1 \left(1 - \frac{n-k}{\pi} \frac{\Gamma^2((n-k+1)/2)}{\Gamma^2((n-k+2)/2)}\right), \quad (29)$$

and that all other eigenvalues are equal to the corresponding eigenvalues of the matrix $(X'X)^{-1}$. In addition, obviously, the matrix $\breve{Q}$ is positively defined.

Let us establish yet another property of the optimal Dual estimator (19).

Proposition 6. *If the OLS-estimator is consistent in the mean–square sense, then the optimal Dual estimator (19) is also consistent in the same sense.*

Proof. Suppose that $\lim_{n\to\infty}(X'X)_n^{-1} = 0$, i.e. that the OLS-estimator is quadratic mean consistent [36]. Consider the variance-covariance matrix (25) for the optimal Dual estimator $\tilde{b}_o$ as a function of n.

We substitute $(X'X)_n^{-1} z_{1n}$ instead of $\lambda_{1n} z_{1n}$ and take out $(X'X)_n^{-1}$:

$$Q_n = \sigma^2 (X'X)_n^{-1} \left(I_k - \frac{n-k}{\pi} \frac{\Gamma^2((n-k+1)/2)}{\Gamma^2((n-k+2)/2)} z_{1n} z'_{1n}\right).$$

Using the relation for the gamma-function first used in the derivation of (23), we obtain:

$$\lim_{n\to\infty} \left((n-k)\frac{\Gamma^2((n-k+1)/2)}{\Gamma^2((n-k+2)/2)}\right) = 2.$$

Taking into account that $z_{1n} z'_{1n}$ is bounded, since the eigenvector z_{1n} is normalized, we finally derive $\lim_{n\to\infty} Q_n = 0$. □

Let us now consider questions related to confidence intervals for the optimal Dual estimator (19). Taking into account its unbiasedness, we represent the estimator in the following form: $\tilde{b}_o = \beta + \tilde{\delta}$, where, in view of (9), (19), and (20),

$$\tilde{\delta} = \delta - sign(z_1'\delta)\tilde{c}z_1\sqrt{e'e}. \tag{30}$$

Let us consider the central moments of the *i-th* component $\tilde{\delta}_i$ of the vector $\tilde{\delta}$. The odd moments are equal to zero, as shown earlier. The second moment is equal to the corresponding diagonal element of the matrix Q from (25). Let us evaluate the kurtosis of $\tilde{\delta}_i$. If its value is only slightly different from three, the distribution of the $\tilde{\delta}_i$ can be considered as approximately normal with zero expectation and the variance shown above. To obtain an upper bound of the kurtosis, we use the fact that the greatest distortion of the initial normal distribution of δ_i takes place under the greatest improvement of the estimator, i.e. when the value of the $\widetilde{L}_o^2/L^2$ from (22) is minimal. We can see from (22) and (23) that this index decreases as the ratio of the largest eigenvalue λ_1 to the sum of all eigenvalues increases. This ratio is maximal and equals 1 in the one-dimensional case. Thus, it is sufficient to consider the one-dimensional case to find the maximum of the kurtosis. Without loss of generality, let us assume that the X is an n-dimensional vector of ones and take into account that in this case $z_1 = 1$ and $\lambda_1 = 1/n$. Then, using the standard procedure for kurtosis calculation as well as the known relationships for moments about the origin of the half-normal and chi-squares distributions [8], we derive from (30) the kurtosis Ku:

$$Ku = \frac{\frac{3}{n^2} - 2\tilde{c}^2\frac{n-1}{n} - 16\frac{\tilde{c}^3}{\sqrt{\pi n}}\frac{\Gamma(1.5+(n-1)/2)}{\Gamma((n-1)/2)} + 4\tilde{c}^4\frac{\Gamma(2+(n-1)/2)}{\Gamma((n-1)/2)}}{\left(1/n - (n-1)\,\tilde{c}^2\right)^2}, \tag{31}$$

where $\tilde{c}$ is calculated by (20).

One can compute using (31) that for $n < 6$, $Ku < 4.017$, and for $n \geq 6$, $Ku < 4$. Moreover, Monte-Carlo simulation shows that, in particular, the quantile of the probability density function of $\tilde{\delta}_1$ at the significance level of 0.05 equals 1.923, i.e. it is smaller than the one for normal distribution. Hence, it is reasonable to present approximately

$$\tilde{b}_o \cong N(\beta,\, Q). \tag{32}$$

And now we can extend known results from the theory of regression analysis [32, 36] to the optimal Dual estimator (19). In particular, the individual confidence interval for the *j-th* component of the vector of the Dual estimator is computed with the help of the expression:

$$\tilde{b}_o\,(j) \pm t(n-k, 1-\alpha/2)\sqrt{\breve{Q}_{j,j}}\,s, \tag{33}$$

where $t(n-k, 1-\alpha/2)$ is the $1-\alpha/2$ point of the Student distribution with $n-k$ degrees of freedom, $\breve{Q}_{j,j}$ is the corresponding diagonal element of the matrix $\breve{Q}$ from (28), and s is the standard deviation estimate:

$$s = \sqrt{e'e/(n-k)}. \tag{34}$$

TABLE 1
One–dimensional problem

	Theoretical value $\widetilde{L}_o^2/L^2$	Experimental value $\widetilde{L}^2/L^2$
Normal distribution N(0, 1)	0.39772	0.39472
Uniform distribution in interval $[-2, 2]$	–	0.41328
Mixture N(0, 1)–80% and N(0, 10)–20%	–	0.25025
Exponential autocorrelation $\sigma = 1, q = 0.3$	–	0.65405

TABLE 2
Two–dimensional problem

	Theoretical value $\widetilde{L}_o^2/L^2$	Experimental value $\widetilde{L}^2/L^2$
Normal distribution N(0, 1)	0.40319	0.40535
Uniform distribution in interval $[-2, 2]$	–	0.41470
Mixture N(0, 1)–80% and N(0, 10)–20%	–	0.39195
Exponential autocorrelation $\sigma = 1, q = 0.3$	–	0.51469

Finishing our consideration of the basic properties of the optimal Dual estimator (19), let us answer the following important question: how does the efficiency index (22) change when the original assumptions are violated, namely, if the error distribution ϵ is symmetrical, but differs from the normal one, and the error covariance matrix is not the identity matrix. An answer to this question was obtained with the help of statistical modeling using Statistics Toolbox from the Matlab package. Errors with the normal distribution $N(0, 1)$, uniformly distributed in the interval $[-2, 2]$, a mixture of the normal distributions $N(0, 1)$–80% and $N(0, 10)$–20%, and also the errors representing a time series with the exponential autocorrelation function $R(\tau) = \sigma^2 e^{-q|\tau|}$ with σ^2=1 and q=0.3 were modeled. The number of tests was $10,000$. As data, the vector of ones $X_{10,1}$ has been used (i.e. the one-dimensional problem of direct measurements $Y = \beta + \epsilon$ was modeled), as well as the matrix $X_{10,2}$ with a significant linear contingency of its columns (i.e. a two-dimensional problem $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$ was modeled under conditions of multicollinearity). The degree of multicollinearity is characterized by the ratio of the maximal eigenvalue of the matrix $(X'X)^{-1}$ to the minimal one, and is equal to 448.8. The data for the two-dimensional problem were centered. We have taken $\beta = 2$ in the first case and $\beta = \left[\begin{smallmatrix} 14 \\ 6 \end{smallmatrix}\right]$ in the second one. For each test, computing b by (2) and knowing β we can find: $\tilde{b}_o$ by (19), $(b - \beta)'(b - \beta)$, and $(\tilde{b}_o - \beta)'(\tilde{b}_o - \beta)$. The averages of the last two values provides a close approximation to the required L^2 and $\widetilde{L}_o^2$. The modeling results are further presented in the Tables 1 and 2. One can see from Tables 1 and 2 that the optimal Dual estimator shows a high degree of robustness, i.e. the index $\widetilde{L}^2/L^2$ stability in the presence of deviations from the original assumptions regarding the error ϵ. A visible reduction in efficiency is observed only in the one-dimensional problem with the autocorrelated errors ϵ, which is explained by the absence of centered data in the one-dimensional case. On the other hand, in the one-dimensional case, and with very heavy characteristics of the mixture of the distributions, one observes an increase in efficiency, i.e. in this case the

robustness of the Dual estimator is increased. For the two-dimensional problem, almost the same pattern is typical, but with smaller scatter for the efficiency criterion $\widetilde{L}^2/L^2$.

Now we shall consider the use of various kinds of a priori information to choose the better estimate.

Let us formulate the general principle of the use of the extra information, keeping in mind that, as it is defined above in the introduction, b_s is the OLS estimate computed for the given sample (X, Y). To choose a better estimate one must know the sign of the scalar $q'(b_s - \beta)$ in the general case, or ${z_1}'(b_s - \beta)$ for the optimal vector. The first product is known. If there exists a constraint directly on the scalar $q'\beta$ or on the vector β, one can derive the constraint on the difference. The simplest rule is as follows. If for a specified constraint the difference can be only positive or, on the contrary, only negative, one has the correct estimate (note, that there are also the more complicated rules as it will be shown later). If it is impossible to recognize the sign of the difference, the estimate remains as the OLS estimate b_s.

3. The use of a priori information in the form of strict inequalities

As follows from previously obtained outcomes, a priori information in the form of $q'(b_s - \beta) > 0$ or $q'(b_s - \beta) < 0$ is naturally used in the proposed method. The method provides the maximal effect, corresponding to (22) or, approximately, (23). Besides, one can use the indirect extra information relative to β. Let us see how that works in several examples.

The first example relates to a one-dimensional problem. Suppose we know that $a_1 < \beta < a_2$ and b_s is calculated. Then, taking into account that in this case $q = z_1 = 1$, we obtain $q'(b_s - \beta) > 0$ when $b_s - a2 > 0$ or $q'(b_s - \beta) < 0$ when $b_s - a1 < 0$ as well as the required estimates from (19). Further, our estimate remains as b_s when the above inequalities are not satisfied.

In the second example, we consider the situation when a priori information in some sense is qualitative. Suppose that we have a two-dimensional regression, and the external information is the fact that $\beta_1 > \beta_2$ where β_1, β_2 are components of the vector β. Let the signs of corresponding components of the vector q be positive and negative respectively. Then for all samples of X, Y in which the inequality $q'b_s < 0$ is true we obtain $q'(b_s - \beta) < 0$. Under opposite signs of components of the vector q we obtain $q'(b_s - \beta) > 0$.

In the third example we turn to the non-linear case and show how a priori information in the form of $\beta'\beta \leq {b_s}'b_s$ can be used to obtain the sign of $q'(b_s - \beta)$. In what follows, let us rename $\tilde{b}$ to b_q and b to b_s in order to emphasize that we are dealing with the sample but not with the general population. Let us show that with the given a priori information the estimator is of the form:

$$b_q = b_s - sign(q'b_s)\sqrt{\frac{\lambda_1}{\pi}}\,\frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+2)/2)}\sqrt{e'e}\;q. \tag{35}$$

Indeed, one can obtain from (13), replacing b by b_s and $\tilde{b}$ by b_q: ${b_q}'b_q = {b_s}'b_s - 2q'b_s sign(q'\delta)\tilde{c}\sqrt{e'e} + \tilde{c}^2 e'e$. If ${b_q}'b_q < {b_s}'b_s$, $sign(q'\delta) = sign(q'b_s)$.

But in some samples we can derive that ${b_s}'b_s < {b_q}'b_q$. This situation usually arises when the vector b_s is close to the vector β. It is obvious that our estimate should be b_s in such a case. Thus, the result of (35) can be strengthened in the following manner:

$$b_{qf} = \begin{cases} b_q, \ if \ \ {b_q}'b_q \le {b_s}'b_s \\ b_s, \ if \ \ {b_q}'b_q > {b_s}'b_s \end{cases} \tag{36}$$

An application of the rule (35) will now be demonstrated on the problem reviewed in Hoerl's first publication dedicated to ridge regression [16]. This problem is characterized by multicollinearity as seen from the the eigenvalues of the correlation matrix, which are [0.0137 0.0985 2.8878]. To exclude the intercept we have centered the data. Then the vector of remaining coefficients given by A. Hoerl is $\beta = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$. The standard deviation and coefficient of determination respectively are 0.915 and 0.956. Let us use the estimator (35) replacing the vector q by the optimal vector z_1. This vector is equal to $z_1 = \begin{pmatrix} -0.7520 \\ 0.6539 \\ 0.0834 \end{pmatrix}$. Let us find the OLS estimate using (2): $b_s = \begin{pmatrix} 8.1837 \\ -4.2694 \\ 4.9864 \end{pmatrix}$. One can see that this estimate is strongly distorted by comparison with the true coefficients. The estimate b_q calculated by (35) is equal to: $b_q = \begin{pmatrix} 4.4576 \\ -1.0292 \\ 5.3996 \end{pmatrix}$. For this estimate, the index $\widetilde{L}_o^2/L^2$ equals 0.246 while the theoretical index (22) equals 0.4733, and it is easy to see that the estimate b_q is better than the OLS estimate. Unfortunately, the sign of the second coefficient remains incorrect, but the direction of its change is accurate.

In summary, let us compare the capabilities of the proposed method with the possibilities of the classical techniques. As is known, strict linear and non-linear inequalities cannot be used in the regression analysis under a priori parameter constraints, and only non-strict inequalities are available [25]. Regarding shrinkage (biased) estimators, their features have been briefly discussed in the Introduction section. To this we can add that in the literature the application of only one type of strict inequalities is mentioned, namely the inequality relative to norms examined above. However, in this case the specified inequality is a preliminary condition which is not sufficient to obtain the estimate. The exception is the James-Stein estimator, but it has restrictions considered above in the Introduction. In particular, in our third example the James-Stein estimate differs very little from the OLS estimate b_s, i.e., it is useless.

4. The use of a priori information in the form of linear non-strict inequalities

We shall begin with the one-dimensional problem and consider two cases: $y = x\beta + \epsilon$ and $y = \beta_0 + x\beta + \epsilon$, where, in the first case, x is a vector of ones, in the second case, x is an arbitrary vector with unequal elements, and β_0, β are

scalars. After centering of the second equation, β_0 disappears. Assume that the true coefficient is constrained by the linear non-strict inequality $a_1 \le \beta \le a_2$. This extra information is also used in the Constrained Least Squares technique. Of course, to find a better alternative of the Dual estimator we could apply an approach considered in the first example of the previous Section. But the estimation algorithm has to be more sophisticated, if we aim to compete with the Constrained Least Squares method. Indeed, if, for example, the mentioned better alternative is less than a_1, it is reasonable to take a_1 as the estimate. Next we give the estimation algorithm which takes into account a number of similar situations, and then we investigate its properties. But we stress in advance the following. A priori inequalities considered in this section do not include the OLS estimator b. On the other hand, we will study the behaviour of the proposed estimator in the general population. Therefore in the given section we express our estimator through b. Thus, we present the estimator b_q in the following form:

$$b_q = \begin{cases} b, & if\ a_1 \le b \le a_2 \\ b+p, & if\ a_1 - p\ < b < a_1 \\ b-p, & if\ a_2 < b < a_2 + p \\ a_1, & if\ b < a_1 - p \\ a_2, & if\ b > a_2 + p \end{cases}, \tag{37}$$

where p is obtained from (19), (34) and is equal to

$$p = \sqrt{\frac{n-1}{x'x\,\pi}}\ \frac{\Gamma(n/2)}{\Gamma((n+1)/2)}\, s. \tag{38}$$

Setting $p = 0$ in (37) we obtain the obvious rule for Constrained Least Squares (CLS) in the one-dimensional case:

$$b_{cls} = \begin{cases} b_s, & if\ a_1 \le b \le a_2 \\ a_1, & if\ b < a_1 \\ a_2, & if\ b > a_1 \end{cases} \tag{39}$$

It is known that $b \sim N(\beta, \hat{\sigma})$, where $\hat{\sigma}$ is calculated as

$$\hat{\sigma} = \sigma / \sqrt{x'x}. \tag{40}$$

Now it is possible to establish the following.

Proposition 7. *Suppose in one-dimensional linear regression, a priori information in the form $a_1 \le \beta \le a_2$ is symmetric about β, that is $(a_1 + a_2)/2 = \beta$. Suppose also that $a_2 - a_1 < \vartheta_\alpha \hat{\sigma}$, where ϑ_α is the $(1-\alpha)100\%$ quantile of the standard normal distribution, $\alpha > 0$ and $p < a_2 - a_1$. Then for an arbitrary $\hat{\sigma}$ the Mean Squares Error (MSE) of the estimator b_q (37) is less than the MSE of the estimator b_{cls} (39).*

Proof. Estimators b_q and b_{cls} are functions of $b \sim N(\beta, \hat{\sigma})$. Hence, the MSE of our estimators are:

$$MSE_q = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \int_{-\infty}^{\infty} (b_q - \beta)^2 e^{-(b-\beta)^2/2\hat{\sigma}^2}\, db$$

$$MSE_{cls} = \frac{1}{\hat{\sigma}\sqrt{2\pi}} \int_{-\infty}^{\infty} (b_{cls} - \beta)^2 e^{-(b-\beta)^2/2\hat{\sigma}^2} db$$

One can see from (37), (39) that the integrands of the derived expressions differ from each other only in the intervals $(a_1 - p, a_1)$ and $(a_2, a2 + p)$. In these intervals b_{cls} is equal to a_1 and a_2 respectively, i.e. the values of b_{cls} are on the bounds of a priori interval. On the other hand, under accepted assumptions, the values of b_q always lie within this interval. Therefore, $(b_q - \beta)^2 < (b_{cls} - \beta)^2$ together with the corresponding integrands. □

Proposition 8. *Suppose in one-dimensional linear regression a priori information in the form $a_1 \le \beta \le a_2$ is asymmetric about β. Then under conditions $a_2 - a_1 < \vartheta_\alpha \hat{\sigma}$, where ϑ_α is the $(1-\alpha)100\%$ quantile of the standard normal distribution, $\alpha > 0$, $p < a_2 - a_1$, $a_1 < \beta - p$, $a_2 > \beta + p$, then for an arbitrary $\hat{\sigma}$ the Mean Squares Error (MSE) of the estimator b_q (37) is less than the MSE of the estimator b_{cls} (39).*

Proof. We can prove this proposition in the same way as the previous proposition. □

Remark 3. We have proven the given propositions for any value of p which is within the interval $a_2 - a_1$. It is obvious, that the propositions are fair, also for a value of p, calculated by (38), that lies inside the specified interval.

The last result shows that the algorithm (37) based on the Dual estimator has a definite potential advantage over the Constrained Least Squares approach. However, we should establish the quantitative characteristics of the method in order to evaluate its usefulness in practical application. For this purpose, we provide the following result

Proposition 9. *Suppose in one-dimensional regression the sought-for coefficient β is subject to the linear non-strict inequality $a_1 \le \beta \le a_2$, p is a positive constant such that $p < a_2 - a_1$, and $\hat{\sigma}$ is defined by (40). Then the Mean Squares Error of the estimator (37) can be found from the equations*

$$MSE_q = V_q + (E_q - \beta)^2, \tag{41}$$

where E_q, the mathematical expectation of b_q, and its variance V_q are defined as follows:

$$\begin{aligned} E_q = {} & \frac{a_1 + a_2}{2} + \frac{a_1 - \beta - p}{2} \, erf\left(\frac{a_1 - \beta - p}{\hat{\sigma}\sqrt{2}}\right) \\ & - \frac{a_2 - \beta + p}{2} \, erf\left(\frac{a_2 - \beta + p}{\hat{\sigma}\sqrt{2}}\right) + \frac{p}{2}\left[erf\left(\frac{a_1 - \beta}{\hat{\sigma}\sqrt{2}}\right)\right. \\ & \left. + erf\left(\frac{a_2 - \beta}{\hat{\sigma}\sqrt{2}}\right)\right] + \frac{\hat{\sigma}}{\sqrt{2\pi}}\left(e^{-\frac{(a_1-\beta-p)^2}{2\hat{\sigma}^2}} - e^{-\frac{(a_2-\beta+p)^2}{2\hat{\sigma}^2}}\right), \end{aligned} \tag{42}$$

$$
\begin{aligned}
V_q &= \frac{1}{2}[(a_1 - E_q)^2 + (a_2 - E_q)^2] \\
&+ \sqrt{\frac{2}{\pi}}\,\hat{\sigma}\left[\left(E_q - \frac{1}{2}(a_2 - p + \beta)\right) e^{-\frac{(a_2+p-\beta)^2}{2\hat{\sigma}^2}}\right. \\
&\left. - \left(E_q - \frac{1}{2}(a_1 + p + \beta)\right) e^{-\frac{(a_1-p-\beta)^2}{2\hat{\sigma}^2}} - p\left(e^{-\frac{(a_1-\beta)^2}{2\hat{\sigma}^2}} + e^{-\frac{(a_2-\beta)^2}{2\hat{\sigma}^2}}\right)\right] \\
&+ \frac{1}{2}\left[((a_1 - E_q)^2 - (E_q - p - \beta)^2 - \hat{\sigma}^2)\, erf\left(\frac{a_1 - p - \beta}{\sqrt{2}\,\hat{\sigma}}\right)\right. \\
&+ ((E_q + p - \beta)^2 + \hat{\sigma}^2 - (a_2 - E_q)^2)\, erf\left(\frac{a_2 + p - \beta}{\sqrt{2}\,\hat{\sigma}}\right) \\
&+ ((E_q - p - \beta)^2 - (E_q - \beta)^2)\, erf\left(\frac{a_1 - \beta}{\sqrt{2}\,\hat{\sigma}}\right) + ((E_q - \beta)^2 \\
&\left. - (E_q + p - \beta)^2)\, erf\left(\frac{a_2 - \beta}{\sqrt{2}\,\hat{\sigma}}\right)\right].
\end{aligned} \tag{43}
$$

Proof. Here we give the sketch of a proof to avoid cumbersome mathematical operations. Let us use well-known relationships for the mathematical expectation and variance of a function of random variables [10]. In our case, the piecewise function has the form (37). The domain of integration is divided into five intervals corresponding to the definition of the function. □

Corollary 2. *If we set $p = 0$ in (42) and (43), we will get the MSE for Constrained Least Squares (39) in the one-dimensional case:*

$$
MSE_{cls} = V_{cls} + (E_{cls} - \beta)^2, \tag{44}
$$

$$
\begin{aligned}
E_{cls} = \frac{a_1 + a_2}{2} + \frac{a_1 - \beta}{2}\, erf\left(\frac{a_1 - \beta}{\hat{\sigma}\sqrt{2}}\right) - \frac{a_2 - \beta}{2}\, erf\left(\frac{a_2 - \beta}{\hat{\sigma}\sqrt{2}}\right) \\
+ \frac{\hat{\sigma}}{\sqrt{2\pi}}\left(e^{-\frac{(a_1-\beta)^2}{2\hat{\sigma}^2}} - e^{-\frac{(a_2-\beta)^2}{2\hat{\sigma}^2}}\right),
\end{aligned} \tag{45}
$$

$$
\begin{aligned}
V_{cls} &= \frac{1}{2}\left((a_1 - E_{cls})^2 + (a_2 - E_{cls})^2\right) \\
&+ \sqrt{\frac{2}{\pi}}\hat{\sigma}\left(\left(E_{cls} - \frac{a_2 + \beta}{2}\right)e^{-\frac{(a_2-\beta)^2}{2\hat{\sigma}^2}} - \left(E_{cls} - \frac{a_1 + \beta}{2}\right)e^{-\frac{(a_1-\beta)^2}{2\hat{\sigma}^2}}\right) \\
&+ \frac{1}{2}\left((a_1 - E_{cls})^2 - (E_{cls} - \beta)^2 - \hat{\sigma}^2\right) erf\left(\frac{a_1 - \beta}{\hat{\sigma}\sqrt{2}}\right) \\
&- \frac{1}{2}\left((a_2 - E_{cls})^2 - (E_{cls} - \beta)^2 - \hat{\sigma}^2\right) erf\left(\frac{a_2 - \beta}{\hat{\sigma}\sqrt{2}}\right).
\end{aligned} \tag{46}
$$

The formulas (41)–(46) allow us to calculate and compare MSE_q and MSE_{cls} for known a_1, a_2, and $\hat{\sigma}$ when β is given, in other words, to simulate the various situations. But in solving actual problems, the coefficient β is unknown, and therefore we must suppose that its value is arbitrary within the interval $[a_1, a_2]$.

TABLE 3
Outcome of the numerical investigation of the one-dimensional task

$\frac{a_2-a_1}{\hat{\sigma}}$	0.6	0.9	1.2	1.5	1.8	2.1	2.4	2.7	3.0	3.3
MI_{cls}/L^2	0.092	0.18	0.27	0.36	0.44	0.50	0.56	0.61	0.64	0.68
MI_q/L^2	0.071	0.13	0.19	0.27	0.34	0.41	0.48	0.53	0.57	0.61

In this situation it is perfectly natural to use the next two integrals as criteria of the quality of estimation :

$$MI_q = \frac{1}{a_2 - a_1} \int_{a_1}^{a_2} MSE_q \, d\beta, \tag{47}$$

$$MI_{cls} = \frac{1}{a_2 - a_1} \int_{a_1}^{a_2} MSE_{cls} \, d\beta. \tag{48}$$

Now, to obtain the all-round tool for investigation of the properties of the algorithms (37), and (39) we should concretize the calculation of p in (42), and (43). Obviously, for this purpose it is necessary to use the formula (38). But the Proposition 9 stipulates that must be $p < a_2 - a_1$, and this condition can not be fulfilled for a small difference $a_2 - a_1$. It is possible to solve this problem in the obvious way, if to find the optimal value of p, having used (41)–(48):

$$p_o = \arg\min_{p \in u}(MI_q/MI_{cls}) \tag{49}$$

where u is the interval $[0, a_2 - a_1]$ when $a_2 - a_1 < p$. On such an interval one should use p_o instead of p from (38).

Let us examine the numerical performance of the estimators in question, using equations (40)–(49). Doing so, we consider that there is a good reason to normalize the functions (47), and (48) by the variance of the OLS estimator L^2. Without reducing the generality of results, we will model the case when x is the vector of ones. Here it turns out that the final result depends only on the ratio $(a_2 - a_1)/\hat{\sigma}$. Varying this ratio and setting $s = \sigma$ we obtain the results presented in Table 3. This table represents the results for those ratios $(a_2 - a_1)/\hat{\sigma}$ that are interesting in a practice. But the calculation in the range $0 < (a_2 - a_1)/\hat{\sigma} < \infty$ shows that for the accepted index of the quality of the estimation, algorithm (37) is uniformly better than algorithm (39).

At the end of this section, we will discuss one of the possibilities of using the Dual estimator in multidimensional cases for which a priori information in the form $a_1 \leq \beta_i \leq a_2$ exists. We should note that this is the most available constraint in many applications, because, if we have a system of a priori inequalities for several coefficients, then a complicated question of its consistency arises. The technique based on Corollary 1 of Proposition 3 will be applied. Let us denote the OLS estimate of β_i as b_i, and the estimate for this coefficient obtained using rule (37) as b_{qi}. We obtain the value of p_i from the formula (21). In this case,

TABLE 4
Outcome of the simulation for Hald's problem

$\frac{a_2-a_1}{\hat{\sigma}}$	1	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
MI_{cls}	0.23	0.29	0.34	0.40	0.45	0.49	0.53	0.56	0.59	0.62	0.64
MI_q	0.18	0.22	0.27	0.33	0.37	0.42	0.46	0.50	0.53	0.56	0.59

TABLE 5
Outcome of the simulation for Hoerl's problem

$\frac{a_2-a_1}{\hat{\sigma}}$	1	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
MI_{cls}	0.33	0.38	0.43	0.48	0.52	0.55	0.59	0.62	0.64	0.67	0.69
MI_q	0.29	0.33	0.37	0.42	0.46	0.50	0.53	0.56	0.59	0.62	0.64

$z_p = 1$, and the λ_p is equal to the i-th diagonal element of the inverse matrix, $\lambda_p = (X'X)^{-1}_{ii}$, and therefore

$$p_i = \sqrt{\frac{(X'X)^{-1}_{ii}}{\pi}} \frac{\Gamma((n-k+1)/2)}{\Gamma((n-k+2)/2)} \sqrt{e'e}. \tag{50}$$

Having found b_{qi}, we can obtain the remaining coefficients using the known expression for the LS estimator with restrictions in the form of equalities [36]:

$$\breve{b} = b + (X'X)^{-1}H'(H(X'X)^{-1}H')^{-1}(b_{qi} - Hb), \tag{51}$$

in which H is a row vector of size k that contains 1 in the *i-th* position and zeroes in all the remaining positions. The comparison of the MSI_q of the proposed technique with the MSI_{cls} of the CLS estimator was performed by Monte-Carlo simulation using the lsqlin program of the Matlab package on the two problems. The first is the A. Hald problem, with four variables, presented in [6], and the second one is the A. Hoerl problem considered in Section 3. In both cases we used a priori constraints for the second coefficient, and $\hat{\sigma}$ was calculated using the formula

$$\hat{\sigma} = \sqrt{(X'X)^{-1}_{2,2}} s. \tag{52}$$

In the lsqlin program, the following parameters were used: $A = [0\,1\,0\,0;\ 0\,{-1}\,0\,0]$, and $c = \left[\begin{smallmatrix} a_2 \\ -a_1 \end{smallmatrix}\right]$ for the first example, and $A = [0\,1\,0;\ 0\,{-1}\,0]$ with the same c for the second example. The number of tests for each case was 1500. The results for the above mentioned problems are presented in Tables 4 and 5 respectively.

As one can see from the tables, in both cases the proposed method is preferable.

5. Confirmation of one of two competing theories with the help of an experiment

As an example, let us consider the widely known experiment conducted by astronomers Dyson, F.W., Eddington, A.S., and Davidson, C.R. in the year of 1919 (cf. [7]). The purpose of the experiment was to determine the deflection of a ray

of light in the gravitational field of the sun. Here they considered three possibilities: the deflection is absent; the deflection conforms to Newton's theory and is equal to $0.87''$; the deflection conforms to the General Theory of Relativity of A. Einstein and is equal to $1.75''$. The experiment was performed in three countries by different groups. As the outcome of the experiment, there were obtained three independent values: $1.98 \pm 0.12, 1.61 \pm 0.3$, and 0.93. For the first two values, after the symbol $\pm$, the probable error is shown, which is equal, as it is known, to 0.6745σ. For the third value, the authors did not give the value of the probable error. Noting that it is too big, they discarded this value. Subsequently, this fact led to doubts and prolonged discussions, and only a double checking, conducted in 1979, showed that the error in the third measurement was indeed big (cf. [24]). We will process these data without discarding the third measurement 0.93 and assuming that its probable error is big and equals to 0.6, i.e. five times greater than for the first measurement. Note also that this figure is consistent with the results of the analysis carried out later in 1979. Doing so, we use the model (1), where $X = \binom{1}{1}{1}, Y = \binom{1.98}{1.61}{0.93}$, but with the difference that we have $\mathrm{cov}(\epsilon) = V$, where V is the non-identity diagonal matrix. Using known probable errors we execute the obvious calculations and obtain the following matrix of variances:

$$V = \begin{pmatrix} 0.03165 & 0 & 0 \\ 0 & 0.19782 & 0 \\ 0 & 0 & 0.79129 \end{pmatrix}.$$

Let us transform the initial variables X, Y, ϵ as follows: $X_t = V^{-0,5}X, Y_t = V^{-0,5}Y, \epsilon_t = V^{-0,5}\epsilon$. And now, as is well known (see e.g. [6]), the model

$$Y_t = X_t\beta + \epsilon_t, \tag{53}$$

where

$$X_t = \begin{pmatrix} 5.675 \\ 2.248 \\ 1.124 \end{pmatrix}, \quad Y_t = \begin{pmatrix} 11.237 \\ 3.620 \\ 1.046 \end{pmatrix},$$

has uncorrelated errors and completely corresponds to model (1). Then we obtain, using (2), the OLS estimate $b_s = 1.897$. Next, since the possible values of β are 0.87 or 1.75, it is easy to see that $\delta = b_s - \beta$ is positive. Considering that for the one-dimensional task $z_1 = 1$, we obtain $sign(z_1'\delta) = 1$. Now we obtain from (19) $\tilde{b}_o = 1.788$, which is close to 1.75. In addition, taking into account the importance of the problem being studied, let us accept a significance level $\alpha = 0.02$. Then, in accordance with (33), under the normality assumption, we derive the confidence interval of $\tilde{b}_o$ as [1.030 2.545]. We see that it only covers one of the possible theoretical values, namely 1.75. Thus, under the assumptions accepted, there are strong reasons to accept the outcome obtained with the Dual estimator as the final estimate and to confirm the General Theory of Relativity.

6. Conclusion

The article has demonstrated that for the linear regression model we can obtain two alternative estimators by subtracting or adding a certain vector to the Ordinary Least Squares Estimator (OLSE) vector. One of them strictly dominates the latter. Moreover, if errors are normally distributed, this estimator is unbiased, and consistent, and has significantly smaller variance than the OLSE. In general, the use of a priori information is a natural way to choose a better alternative. The article discusses several types, traditional and nontraditional, of this information, and shows the advantage of the proposed method over the biased estimators and the Constrained Least Squares technique. Also, it was demonstrated that in the proposed method a priori information on two possible values of the parameter being sought can be successfully used, which is characteristic for the problem of confirming one of the two alternative theories. The example of verification of the General Theory of Relativity based on astronomical data is examined. It should be noted that there are many other types of constraints, both with respect to the parameters being estimated and with respect to the regression response, which can be used for the purpose of selecting a better alternative. However, as can be seen from the article's material, for each kind of a priori information a rule (algorithm) of its efficient use must be found, which is not trivial and deserves special consideration.

Acknowledgments

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

References

[1] Aivazyan, S.A., Yenyukov, I.S., Meshalkin, L.D. (1985). Applied Statistics. Study of Relationships. Finansy i statistika, Moskow (in Russian). MR0803501

[2] Bateman, H., Erdelyi, A. (1953). Higher Transcendental Functions, Vol. 1, 2, Mc Grow-Hill.

[3] Blaker, H. (1999). A Class of Shrinkage Estimators in Linear Regression. The Canadian Journal of Statistics, 27, 207–220. MR1703631

[4] Carlin, B.P., Louis, T.A. (2008). Bayesian Methods for Data Analysis, 3rd Ed. MR2442364

[5] Dorugade, A.V., Kashid, D.N. (2010). Alternative Method for Choosing Ridge Parameter for Regression. Applied Mathematical Sciences, 4(9), 447–456. MR2580639

[6] Draper, N.R., Smith, H. (1998). Applied Regression Analysis, 3rd Ed., New York, J. Wiley and Sons, Inc. MR1614335

[7] Dyson, F.W., Eddington, A.S., Davidson, C.R. (1920). A Determination of the Deflection of Light by the Sun's Gravitational Field, from Observations Made at the Total Eclipse of May 29. Mem. R. Astron. Soc., 220, 291–333.

[8] Forbes, C., Evans, M. et al. (2011). Statistical Distributions, 4th Ed., Wiley and Sons, Inc. MR2964192
[9] Gibbons, D.G. (1981). A Simulation Study of Some Ridge Estimators. J. Amer. Statist. Assoc., 76, 131–139.
[10] Gnedenko, B. (1978). The Theory of Probability. Translated from Russian, MIR, Moscow. MR0676302
[11] Goldstein, M., Wooff, D. (2007). Bayes Linear Statistics. Theory and Methods, John Wiley and Sons. MR2335584
[12] Gordinsky, A. et al. (2000). A New Approach to Statistic Processing of Steam Parameter Measurements in the Steam Turbine Path to Diagnose Its Condition. Proc. of the International Joint Power Generation Conference, Miami Beach, FL, July 23–26, 1–5.
[13] Gordinsky, A. (2010). Quasi-Estimation as a Basis for Two-Stage Solving of Regression Problem, http://arxiv.org/abs/1010.0959.
[14] Hadi, A.S., Ling, R.F. (1998). Some Cautionary Notes on the Use of Principal Components Regression. The American Statistical Association, 52(1).
[15] Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed., Springer-Verlag. MR2722294
[16] Hoerl, A.F. (1962). Application of Ridge Analysis to Regression Problems. Chemical Engineering Progress, 58, 54–59.
[17] Hoerl, A.E., Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 42(1).
[18] Horn, R.A., Johnson, C.R. (1986). Matrix Analysis, Cambridge. MR0832183
[19] James, W., Stein, C. (1961). Estimation with Quadratic Loss. Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I, pp. 361–379, Univ. California Press, Berkeley, Calif. MR0133191
[20] Jiang, Y.H., Smith, P.L. (2002). Understanding and Interpreting Regression Parameter Estimates in Given Contexts: A Monte Carlo Study of Characteristics of Regression and Structural Coefficients, Effect Size R Squared and Significance Level of Predictors. Annual Meeting of the American Educational Research Association (New Orleans, LA, April 1–5, 2002), 25 p.
[21] Jolliffe, I.T. (1982). A Note on the Use of Principal Components in Regression. Royal Statistical Society.
[22] Jolliffe, I.T. (2002). Principal Component Analysis: Springer Series in Statistics, 2nd Ed., Springer, NY. MR2036084
[23] Judge, G.G., Takayama, T. (1966). Inequality Restrictions in Regression Analysis. Journal of the American Statistical Association, 61(313), 166–181. MR0193713
[24] Kennefick, D. (2007). Not Only Because of Theory: Dyson, Eddington and the Competing Myths of the 1919 Eclipse Expedition. 2007, http://arxiv.org/abs/0709.0685

[25] Knopov, P.S., Korkhin, A.S. (2012). Regression Analysis Under a Priori Parameter Restrictions. Springer, New York, Dordrecht, Heidelberg, London. MR3014923
[26] Li, T.F., Bhoj, S. (1988). A Modified James-Stein Estimator with Application to Multiple Regression Analysis. Scandinavian Journal of Statistics, 15(1), 33–37. MR0967955
[27] McDonald, G.C. (2009). Ridge Regression. John Wiley and Sons, WIREs Comp. Stat., 1, 93–100.
[28] Meada, J.L., Renau, R.A. (2010). Least Squares Problems with Inequality Constraints as Quadratic Constraints. Linear Algebra and its Applications, 432(8), 1936–1949 MR2599833
[29] Van Nostrand, R.C. (1980). A Critique of Some Ridge Regression Methods (1980): Comment Author(s): Source. Journal of the American Statistical Association, 75(369), 92–94. MR0568580
[30] Petersen, J.H. et al. (2010). Correcting a Statistical Artifact in the Estimation of the Hubble Constant Based on Type IA Supernovae Results in a Change in Estimate of 1.2
[31] Rao, C.R. (1972). Linear Statistical Inference and Its Applications, 2nd Ed., Wiley, New-York. MR0346957
[32] Rao, C.R., Toutenburg, H. (1999). Linear Models: Least Squares and Alternatives, Springer Series in Statistics. MR1707290
[33] Rothenberg, T.J. (1973). Efficient Estimation with a Priori Information. New Haven and London, Yale University Press. MR0326904
[34] Samaniego, F.J. (2010). A Comparison of the Bayesian and Frequentist Approaches to Estimation. Springer Series in Statistics, New York, Dordrecht, Heidelberg, London. MR2664350
[35] Schafer, C.M., Stark, P.B. (2003). Using What We Know: Inference with Physical Constraints. Statistical Problems in Particle Physics, Astrophysics, and Cosmology, SLAC, September 8–11.
[36] Seber, G.A.F. (1977). Linear Regression Analysis, John Wiley and Sons. MR0436482
[37] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society B, 58, 267–288. MR1379242
[38] Yoshioka, S. (1986). Multicollinearity and Avoidance in Regression Analysis. Bihaviormetrika, 19, 103–120.