Acta Universitatis Carolinae. Mathematica et Physica

TomĂĄĹĄ Jurczyk

Ridge least weighted squares

Acta Universitatis Carolinae. Mathematica et Physica, Vol. 52 (2011), No. 1, 15--26

Persistent URL: http://dml.cz/dmlcz/143664

Ridge Least Weighted Squares

TOMÁŠ JURCZYK

Praha

Received May 10, 2010
Revised August 10, 2010

Multicollinearity and outlier presence are classical problems of data within the linear regression framework. We are going to present a proposal of a new method which can be a potential candidate for robust ridge regression as well as a robust detector of multicollinearity. This proposal arises as a logical combination of principles used in the ridge regression and in the least weighted squares estimate. We will also show the properties of the new method.

1. Notation and goals

Let us set up notation first. Let $\mathscr{N}$ denote the set of all positive integers, $\mathscr{R}$ the real line. All vectors are supposed to be column ones.

Throughout the paper we will be investigating regression methods. We consider the linear regression model

$$Y_i = X_i'\beta^0 + e_i = \sum_{j=1}^{p} X_{ij}\beta_j^0 + e_i,\ i = 1, 2, \ldots, n,$$

where vector $Y = (Y_1, \ldots, Y_n)'$ is the response variable, X_{ij} is an element of the design matrix $X = (X_{ij})_{i=1,j=1}^{n,p}$, which has the full rank. X_i denotes the i-th row of X, and $e_i, i = 1, \ldots, n$ are error terms, which are random variables with $Ee_i = 0$. For any $\beta \in \mathscr{R}^p$, $r_i(\beta) = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j$ denotes the i-th residual and $r_{(h)}^2(\beta)$ stands for the h-th order statistic among the squared residuals, i.e., we have $r_{(1)}^2(\beta) \leq r_{(2)}^2(\beta) \leq \ldots \leq \leq r_{(n)}^2(\beta)$.

The purpose of the paper is to find an estimate of vector parameter β^0 to simultaneously handle multicollinearity and contamination (both problems and their conse-

MFF UK, KPMS, Sokolovská 83, 186 75 Praha 8 – Karlín

This work was supported by grant LC06024 and grant GAČR 402/09/0557.

Key words and phrases. Multicollinearity, robust ridge regression, least weighted squares

E-mail address: jurczyk@karlin.mff.cuni.cz

quences are explained in part 2 of the paper). It will be recalled that methods routinely used to solve one of these problems are no longer suitable when the second problem also arises in the data. Moreover, we will try to describe the substance of this issue.

Many procedures have been suggested for this situation – mainly multistep procedures where one of the steps uses the ridge weighted least squares estimation (given in definition 4); such proposals could be seen for example in [5], [7] or [10]. We try to find our method in a different way: we want to combine two methods directly into a one-step procedure. We are going to use the ridge regression (proposed by [2]), which is a classical method for dealing with multicollinearity, and the least weighted squares (first presented in [11]), which can be used for outlier detection.

2. Consequences of multicollinearity and outlier presence

In order to show our idea of finding a suitable estimate, we first have to understand the consequences of both problems separately, as well as the principle of why the ridge regression and the least weighted squares do their jobs. We will start with problems of classical least squares method (LS).

2.1 Multicollinearity

Least squares method is a simple and widely used method. Unfortunately, there exist many situations in which this method is clearly not suitable. One of these problematic situations can be a presence of multicollinearity.

Multicollinearity is a situation in which the regressors are nearly linear dependent. In this situation the normal equations for the LS estimate do not have a stable solution, the LS estimate has a large expected value of its length, and components of the estimate may have a large variation (for more details see [15]). If we imagine the loss function (function which is being minimized) of the LS estimate ($\sum_{i=1}^{n} r_i^2(\beta)$), we will see the problem immediately. We have a nearly multiple solution of the normal equations (caused by dependence of the regressors) – such a solution forms a linear subspace in $\mathscr{R}^p$, so the graph of the loss function of the LS estimate is "almost" flat (see also [3]) in certain direction(s). Therefore we get a large expected length and a large variance of the estimate.

One of the methods which is recommended and used instead of LS in case of the multicollinearity presence is the ridge regression estimator.

Definition 1 *Let $\delta > 0$, then*

$$\hat{\beta}^{(RR,n,\delta)} = \operatorname*{argmin}_{\beta \in \mathscr{R}^p} \left(\sum_{i=1}^{n} r_i^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2 \right)$$

is called the Ridge Regression (RR) estimate of parameter β^0.

RR estimate can be computed as $(X'X+\delta I)^{-1}X'Y$, thus avoiding problems with inversion of matrix $X'X$ (which is ill-conditioned under the multicollinearity presence) and also ensures the stability of the solutions. It is known that the RR estimate is biased but, at the same time, it has a smaller mean square error (for small values of δ) than the LS estimate (the proof can, for example, be found in [15]).

LS and RR loss functions are different in penalization for large values of β. Just because of this penalization, the RR estimator avoids estimates with large length (this is also visible from lemma 5 with $w = (1, 1, \ldots, 1)'$) and therefore the variation of the estimate is also reduced. The loss function $\sum_{i=1}^{n} r_i^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2$ is not so flat as $\sum_{i=1}^{n} r_i^2(\beta)$.

2.2 Contamination problem

Contamination is a problem of the data with the presence of other observations (outliers) which do not follow the regression model, and typically have large values. It is known that already one outlier far away from the model will move the minimum of the loss function of the LS estimate in direction of its influence. It is caused by the fact that all residuals have the same importance.

Dealing with contaminated data is one of the tasks of robust statistics. The main goal of the robust statistics is searching for models which would work for majority of data. There exist many different robust methods which are used to reveal contaminating observations. We are going to present one typical representative of such robust methods called *least weighted squares*. This estimator was proposed in [11]. We choose this estimator because of its nice properties (see [6], [11], [13], etc.) and also because it is a direct generalization of another well-known and widely used *least trimmed squares* (LTS) estimator (firstly mentioned in [8]).

Definition 2 *For any $n \in \mathcal{N}$, let $1 = w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$ be some weights. Then*

$$\hat{\beta}^{(LWS,n,w)} = \underset{\beta \in \mathcal{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i r_{(i)}^2(\beta)$$

is called the Least Weighted Squares (LWS) estimator.

We can see that the robust aspect is ensured here by weighting. The largest residual gets the smallest weight. Please notice that a single weight is not directly related to a specific observation. The LWS estimator assigns the weights to the observations "by itself".

It remains to say how this method can be used for outlier detection. With the special choice of weights $w_{n-h+1} = \ldots = w_n = 0$, we can take for outliers those observations to which these zero weights are assigned. The idea is following: if the assigned weight for the observation is 0, the residual of this observation will not affect the value of the loss function (regardless of how large the respective residual is).

2.3 Multicollinearity and outliers together

We are interested now in a special type of data, in which the majority of the data suffers from multicollinearity and follow the regression model, while the rest of the data represents contamination.

The ridge regression is useless on this type of data because it is not robust. This follows from the fact that each residual has the same importance; hence already one large outlier (one potential large residual) considerably affects the estimate.

We expect the revealing of all contaminating observations from a good robust method, and consequently the revealing of the true structure of the data. This task is important, because dependence of regressors is an essential feature of the data. We could try LWS (as a classical representative of robust methods) on this type of data. Unfortunately, according to paper [3], the LWS method is not suitable either. The robust regression methods based on residual weighting fail in detection of outliers (although they are "built" for this purpose) with the increasing rate of multicollinearity. Simply, we have nearly dependence of the regressors in the majority of the data, so there are nearly some degrees of freedom which are filled by additional outliers. In other words – the LWS method prefers the weights assignment which assigns large weights to some outliers, as compared with the assignment in which all outliers have zero weights. For more details see [3].

3. Ridge least weighted squares

Now, we are going to show the promised new method which should be able to cope with both presented problems.

Let us again recall the three methods we have already presented in this paper. The RR is derived from the classical LS by addition of penalization. The LWS is designed as a weighted version of LS. If we wrote down minimization problems or loss functions of all presented estimators, we would reach a possible candidate for our method immediately.

$$\begin{array}{ccc} \text{Least Squares} & & \text{Ridge Regression} \\ \sum_{i=1}^{n} r_i^2(\beta) & \rightarrow & \sum_{i=1}^{n} r_i^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2 \\ \downarrow & & \downarrow \\ \text{Least Weighted Squares} & & \text{Ridge Least Weighted Squares} \\ \sum_{i=1}^{n} w_i r_{(i)}^2(\beta) & \rightarrow & \sum_{i=1}^{n} w_i r_{(i)}^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2 \end{array}$$

We can see that ridge least weighted squares (as we call this new method) is a logical combination of both principles – penalization for large β in case of multicollinearity and weighting against outliers.

The ridge least weighted squares should solve the problem of multicollinearity for LWS in the same way as RR does for the LS estimate (see again lemma 5). It

makes the loss function less flat than that of LWS. Therefore, the influence of outliers (expected to be located far away from origin) will be reduced. We should define new estimator precisely:

Definition 3 *Let $\delta > 0$ and $1 = w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$ be some weights, then the solution of the extremal problem*

$$\hat{\beta}^{(RLWS,n,w,\delta)} = \underset{\beta \in \mathscr{R}^p}{\operatorname{argmin}} \left(\sum_{i=1}^{n} w_i r_{(i)}^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2 \right) \tag{1}$$

is called the Ridge Least Weigthed Squares (RLWS) estimator.

3.1 Existence of RLWS estimate

To show the existence of the solution in (1), let us first recall a slightly simpler estimate and its properties.

Definition 4 *Let $\delta > 0$, $w = (w_1, w_2, \ldots, w_n)'$ be nonnegative weights, then we define the Ridge Weighted Least Squares Estimator (RWLS) as*

$$\hat{\beta}^{(RWLS,n,w,\delta)} = \underset{\beta \in \mathscr{R}^p}{\operatorname{argmin}} \left(\sum_{i=1}^{n} w_i r_i^2(\beta) + \delta \sum_{j=1}^{p} \beta_j^2 \right). \tag{2}$$

This weighted version of the ridge regression estimator is also called the *Weighted Ridge* in literature (see [1]). We can rewrite (2) in matrix notation:

$$\hat{\beta}^{(RWLS,n,w,\delta)} = \underset{\beta \in \mathscr{R}^p}{\operatorname{argmin}} \left((Y - X\beta)' W (Y - X\beta) + \delta \beta' \beta \right), \tag{3}$$

where $W = diag\{w_1, w_2, \ldots, w_n\}$.

Lemma 1 *The solution of the normal equations*

$$X'WY = X'WX\beta + \delta\beta \tag{4}$$

is also the solution of the minimization expressed in (2). Therefore

$$\hat{\beta}^{(RWLS,n,w,\delta)} = (X'WX + \delta I)^{-1} X'WY.$$

Proof: Let $b \in \mathscr{R}^p$ be a solution of (4), which means $X'W(Y - Xb) - \delta b = 0$. We are going to show that the loss function in (3) is, for any $\beta \in \mathscr{R}^p$, greater than or equal to the value of the loss function for b. We have

$$\begin{aligned}
&(Y - X\beta)' W (Y - X\beta) + \delta \beta' \beta \\
&= [(Y - Xb) + (Xb - X\beta)]' W [(Y - Xb) + (Xb - X\beta)] + \delta [b - (b - \beta)]' [b - (b - \beta)] \\
&= (Y - Xb)' W (Y - Xb) + (b - \beta)' X' W X (b - \beta) + [(Y - Xb)' WX - \delta b'](b - \beta) \\
&\quad + (b - \beta)' [X' W (Y - Xb) - \delta b] + \delta b' b + \delta (b - \beta)' (b - \beta) \\
&= (Y - Xb)' W (Y - Xb) + \delta b' b + (b - \beta)' (X' WX + \delta I)(b - \beta) \\
&\geq (Y - Xb)' W (Y - Xb) + \delta b' b.
\end{aligned}$$

For $\delta > 0$, matrix $X'WX + \delta I$ is regular and positive definite, therefore the last inequality becomes an equality if and only if $b = \beta$. □

Theorem 1 *For any $\delta > 0$, $1 = w_1 \geq w_2 \geq \ldots \geq w_n \geq 0$ and arbitrary observations $\{Y_i, X_i\}_{i=1}^n$ the solution of (1) always exists.*

Proof: We have fixed $\delta > 0$, w, design matrix $X = (X_1, X_2, \ldots, X_n)'$ and response variable $Y = (Y_1, Y_2, \ldots, Y_n)'$. Denote by $W = diag\{w_1, w_2, \ldots, w_n\}$ the weight matrix. For a given permutation π of indices $\{1, 2, \ldots, n\}$, denote $Y(\pi)$ and $X(\pi)$ the vector and the matrix obtained as the corresponding permutation of vector Y coordinates and of matrix X rows, respectively. For data $(Y(\pi), X(\pi))$, w and δ, we are able to compute the RWLS estimate

$$\hat{\beta}^{(RWLS,n,w,\delta)}(\pi) = [X'(\pi)WX(\pi) + \delta I]^{-1}X'(\pi)WY(\pi).$$

According to Lemma 1, $\hat{\beta}^{(RWLS,n,w,\delta)}(\pi)$ minimizes

$$\sum_{i=1}^{n} w_i\Big(Y_i(\pi) - X_i'(\pi)\beta\Big)^2 + \delta\sum_{j=1}^{p}\beta_j^2$$

over β. Compute $\hat{\beta}^{(RWLS,n,w,\delta)}(\pi)$ for all permutations and select that permutation, say π_{min}, for which

$$\sum_{i=1}^{n} w_i\Big(Y_i(\pi) - X_i'(\pi)\hat{\beta}^{(RWLS,n,w,\delta)}(\pi)\Big)^2 + \delta\sum_{j=1}^{p}\Big(\hat{\beta}_j^{(RWLS,n,w,\delta)}(\pi)\Big)^2$$

is minimal. For any other permutation of indices $\tilde{\pi}$ we have

$$\begin{aligned}
&\sum_{i=1}^{n} w_i\Big(Y_i(\pi_{min}) - X_i'(\pi_{min})\hat{\beta}^{(RWLS,n,w,\delta)}(\pi_{min})\Big)^2 \\
&\qquad +\delta\sum_{j=1}^{p}\Big(\hat{\beta}_j^{(RWLS,n,w,\delta)}(\pi_{min})\Big)^2 \qquad (5)\\
\leq &\sum_{i=1}^{n} w_i\Big(Y_i(\tilde{\pi}) - X_i'(\tilde{\pi})\hat{\beta}^{(RWLS,n,w,\delta)}(\tilde{\pi})\Big)^2 + \delta\sum_{j=1}^{p}\Big(\hat{\beta}_j^{(RWLS,n,w,\delta)}(\tilde{\pi})\Big)^2 \\
&= \min_{\beta\in\mathscr{R}^p}\left(\sum_{i=1}^{n} w_i\Big(Y_i(\tilde{\pi}) - X_i'(\tilde{\pi})\beta\Big)^2 + \delta\sum_{j=1}^{p}\beta_j^2\right)
\end{aligned}$$

The only difference between RLWS and RWLS estimations is implied by the way of their assigning the weights to observations. In more detail – the RWLS assignment of the weights is fixed, while the RLWS method chooses one of the assignments by itself. Therefore, if we knew the permutation (say π^*) chosen by the RLWS method just for β which minimizes the RLWS loss function, we would have $\hat{\beta}^{(RLWS,n,w,\delta)} = \hat{\beta}^{(RWLS,n,w,\delta)}(\pi^*)$. Together with inequality (5), we get that the value of the loss function for $\hat{\beta}^{(RWLS,n,w,\delta)}(\pi_{min})$ is less than or equal to the value of the loss function for $\hat{\beta}^{(RLWS,n,w,\delta)}$. But if we look at the RLWS minimization and realize that weights $w_1, \ldots, w_n$ are non-increasing, we see that, for each β, the loss function of RLWS follows the rule "the larger the residual, the smaller the weight". This is clearly the best possible (minimizing) assignment of the weights (for any β including $\hat{\beta}^{(RLWS,n,w,\delta)}$); we thus arrive to $\pi^* = \pi_{min}$. □

Remark 1 *The proof of theorem 1 shows the way how to find the RLWS estimate. Instead of searching for the estimate through*

$$\min_{\beta}\ \min_{\pi}\ \left(\sum_{i=1}^{n} w_i\Big(Y_i(\pi) - X_i'(\pi)\beta\Big)^2 + \delta\sum_{j=1}^{p}\beta_j^2\right)$$

(i.e., in fact RLWS minimization), we find the estimate by minimizing

$$\min_{\pi}\ \min_{\beta}\ \left(\sum_{i=1}^{n} w_i\Big(Y_i(\pi) - X_i'(\pi)\beta\Big)^2 + \delta\sum_{j=1}^{p}\beta_j^2\right)$$

(i.e., the procedure used in Theorem 1). So there exist π_{min} *and* $\hat{\beta}^{(RLWS,n,w,\delta)}$ *such that* $\hat{\beta}^{(RLWS,n,w,\delta)}$ *is the solution of the same normal equations as for* $\hat{\beta}^{(RWLS,n,w,\delta)}(\pi_{min})$*, i. e.* $X'(\pi_{min})W(Y(\pi_{min}) - X(\pi_{min})\beta) - \delta\beta = 0$*. Let us also emphasize that the inversion of the matrix* $X'(\pi_{min})WX(\pi_{min}) + \delta I$ *for* $\delta > 0$ *always exists because of its positive definiteness.*

According to the previous remark, the only possible non-uniqueness of the RLWS solution can appear when the permutation π_{min} is not unique. Let us discuss possible situations:

1) We have weights with $w_i = w_j$ for some $i \neq j$: then the permutations which have the pair i and j and the swapped pair j and i at the same positions give the same value of the loss function.

2) In another situation there may be $r_i^2(\hat{\beta}^{(RLWS,n,w,\delta)}) = r_j^2(\hat{\beta}^{(RLWS,n,w,\delta)})$ for some $i \neq j$: then the permutation with swapped π_i and π_j gives the same value of the loss function as permutation $\pi_{min} = (\pi_1, \pi_2, \ldots, \pi_n)'$.

Even if there one (or both) of situation described in 1) or 2) arises, the RLWS estimate is the same for all minimizing permutations mentioned in 1) or 2). Therefore cases 1) and 2) are not problematic and the estimate is unique. The only situation when the estimate is not unique is the following:

3) For fixed δ there exist $\beta^1 \neq \beta^2$ in which we have global minimum of RLWS loss function. Let 1) and 2) not hold. Then two different permutations $\pi^1_{min} = (\pi^1_1, \ldots, \pi^1_n)'$ and $\pi^2_{min} = (\pi^2_1, \ldots, \pi^2_n)'$ must exist such that

$$\sum_{i=1}^{n} w_i r^2_{\pi^1_i}(\beta^1) + \delta(\beta^1)'\beta^1 = \sum_{i=1}^{n} w_i r^2_{\pi^2_i}(\beta^2) + \delta(\beta^2)'\beta^2. \tag{6}$$

If we rewrite (6), we get

$$\sum_{i=1}^{n} w_i\left([e_{\pi^1_i} - X'_{\pi^1_i}(\beta^1 - \beta^0)]^2 - [e_{\pi^2_i} - X'_{\pi^2_i}(\beta^2 - \beta^0)]^2\right) = \delta\left((\beta^2)'\beta^2 - (\beta^1)'\beta^1\right).$$

So, if the error term e_i, $i = 1, \ldots, n$ is a continuous random variable then the occurence of situation 3) has probability 0.

Combinations of π_{min} non-uniqueness of types 1), 2) and 3) may of course occur, but only case 3) can cause non-uniqueness of the estimate.

To complete the picture of all solutions of (1), let us show the shape of the loss function of RLWS in more detail. For each permutation $\pi^k = (\pi^k_1, \ldots, \pi^k_n)'$ $k = = 1, \ldots, n!$, let us define functions $f_k(\beta) = \sum_{i=1}^{n} w_i r^2_{\pi^k_i}(\beta) + \delta\sum_{j=1}^{p}\beta_j^2$. Each f_k is continuous and strictly convex as a quadratic function of β (the strict convexity is

caused by the δ term). The loss function of RLWS is then $l_{RLWS}(\beta) = \min_k f_k(\beta)$. Therefore, $l_{RLWS}(\beta)$ is continuous and the parameter space is divided into parts on which it is strictly convex. This implies that $l_{RLWS}(\beta)$ can have several local minima. If there is a multiple solution of the RLWS method then the solutions are in different parts of the parameter space and there is only a finite number of them (no more than $n!$), furthermore, if we change the value of δ, the number of solutions will change (unless they are all exactly in the same distance from the origin, which is also highly improbable). Let us stress again that there cannot appear a situation in which the solutions would lie on a line (or in a linear subspace) like it appears for LS (LWS) when X ($X(\pi_{min})$) does not have the full rank (see, for example, [3]). (π_{min} is analogous to π_{min} from Theorem 1 for LWS).

4. Properties of RLWS

We have already considered the existence and uniqueness of the estimate. The shape of the loss function was also briefly mentioned. Now we are ready to show some basic properties of the RLWS estimate.

It is immediately visible that the estimate is biased because of $\delta > 0$. It inherits this property from the ridge regression estimator (which is one of the special cases of RLWS).

From the robust point of view, the RLWS estimate has the same properties concerning the breakdown point as the LWS (or LTS) estimate. It is so because the δ term does not force the estimate to be unbounded in any way. The breakdown point is determined by the number of zero weights in w. For example, the maximal possible breakdown point for LWS $(\lfloor (n-p)/2 \rfloor + 1)/n$ (which is at the same time maximal for any regression equivariant estimator – for proof see [9]) can also be attained by RLWS – we have to choose $\lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ zero weights to reach it.

Now, we are going to investigate equivariance properties of RLWS. Let us denote by $\hat{\beta}^{RLWS}(\{X_i', Y_i\}_{i=1}^n)$ the RLWS estimate obtained from data $Y = (Y_1, \ldots, Y_n)'$ and X where X_i is the i-th row vector.

Lemma 2 *The RLWS estimate is scale equivariant, i.e.,* $\hat{\beta}^{RLWS}(\{X_i', cY_i\}_{i=1}^n) = c\hat{\beta}^{RLWS}(\{X_i', Y_i\}_{i=1}^n)$ *for any constant* c.

Proof: Denote by $Y_{(i)}$ and $X_{(i)}'$ the observation which gives $r_{(i)}^2(\beta)$.

$$\hat{\beta}^{RLWS}(\{X_i', cY_i\}_{i=1}^n) = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} w_i \left(cY_{(i)} - X_{(i)}'\beta\right)^2 + \delta\beta'\beta$$

$$= \operatorname*{argmin}_{\beta} \; c^2 \left(\sum_{i=1}^{n} w_i \left(Y_{(i)} - X_{(i)}' \frac{\beta}{c}\right)^2 + \delta \frac{\beta'\beta}{c^2} \right)$$

$$= \underset{\beta=c\beta^*}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}\beta^*\right)^2 + \delta(\beta^*)'\beta^* = c\hat{\beta}^{RLWS}(\{X'_i, Y_i\}_{i=1}^n).$$

□

Lemma 3 *The RLWS estimate is not regression equivariant, i.e., there exists vector* v *such that* $\hat{\beta}^{RLWS}(\{X'_i, Y_i + X'_i v\}_{i=1}^n) \neq \hat{\beta}^{RLWS}(\{X'_i, Y_i\}_{i=1}^n) + v$.

Proof: $\hat{\beta}^{RLWS}(\{X'_i, Y_i + X'_i v\}_{i=1}^n) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i [Y_{(i)} - X'_{(i)}(v - \beta)]^2 + \delta\beta'\beta$

$$= \underset{\beta=\beta^*+v}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}\beta^*\right)^2 + \delta(\beta^* + v)'(\beta^* + v). \tag{7}$$

In order to have regression equivariance, the term (7) should be in the form

$$\underset{\beta=\beta^*+v}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}\beta^*\right)^2 + \delta(\beta^*)'(\beta^*).$$

□

We will also state the result for affine equivariance. Let us recall the definition first. We say that an estimator T is affine equivariant if $T(\{X'_i A, Y_i\}_{i=1}^n) = A^{-1}T(\{X'_i, Y_i\}_{i=1}^n)$ for any nonsingular square matrix A.

Lemma 4 *The RLWS estimate is not affine equivariant. Nevertheless RLWS estimate is equivariant with respect to transformations of the type* $A^{-1} = A'$.

Proof: $\hat{\beta}^{RLWS}(\{X'_i A, Y_i\}_{i=1}^n) = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}A\beta\right)^2 + \delta\beta'\beta$

$$= \underset{\beta=A^{-1}\beta^*}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}\beta^*\right)^2 + \delta(\beta^*)'(A^{-1})'A^{-1}\beta^*,$$

which is not generaly equal to

$$\underset{\beta=A^{-1}\beta^*}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(Y_{(i)} - X'_{(i)}\beta^*\right)^2 + \delta(\beta^*)'\beta^*.$$

For A satisfying $(A^{-1})'A^{-1} = I$ (the same condition as $A^{-1} = A'$) the equivariance holds. □

Previous results are implied by the penalization for β far from the origin. The estimate cannot be equivariant with respect to transformations not preserving the distance structure. On the other hand, for example rotation transformations preserve equivariace (lemma 4).

As the last issue we will discuss consistency.

Lemma 5 *Let* $\hat{\beta}^{(LWS,n,w)}$ *be the solution of LWS and* $\hat{\beta}^{(RLWS,n,w,\delta)}$ *be the solution of RLWS minimization for the same dataset.*
Then $\|\hat{\beta}^{(RLWS,n,w,\delta)}\| \leq \|\hat{\beta}^{(LWS,n,w)}\|$ *for all* $\delta > 0$.

Proof: The minimization forms of LWS and RLWS (definition 2 and 3) imply

$$\sum_{i=1}^{n} w_i r_{(i)}^2\left(\hat{\beta}^{(LWS,n,w)}\right) \leq \sum_{i=1}^{n} w_i r_{(i)}^2\left(\hat{\beta}^{(RLWS,n,w,\delta)}\right) \tag{8}$$

and

$$\sum_{i=1}^{n} w_i r_{(i)}^2\left(\hat{\beta}^{(LWS,n,w)}\right) + \delta \sum_{j=1}^{p} \left(\hat{\beta}_j^{(LWS,n,w)}\right)^2$$

$$\geq \sum_{i=1}^{n} w_i r_{(i)}^2\left(\hat{\beta}^{(RLWS,n,w,\delta)}\right) + \delta \sum_{j=1}^{p} \left(\hat{\beta}_j^{(RLWS,n,w,\delta)}\right)^2. \tag{9}$$

From inequalities (8) and (9) it directly follows

$$\delta \sum_{j=1}^{p} \left(\hat{\beta}_j^{(RLWS,n,w,\delta)}\right)^2 \leq \delta \sum_{j=1}^{p} \left(\hat{\beta}_j^{(LWS,n,w)}\right)^2;$$

hence $||\hat{\beta}^{(RLWS,n,w,\delta)}|| \leq ||\hat{\beta}^{(LWS,n,w)}||$. □

Lemma 5 together with the result in [13] (Lemma 2), which implies that LWS estimate is bounded in probability, gives that RLWS is bounded in probability as well (for the conditions see Lemma 2 in [13]). This is the first step to show that RLWS (although it is biased) is also weakly consistent (under the same conditions as for the LWS estimate). Due to the limits on the scope of this paper, let us only hint that with the increasing n it becomes more important to reduce the part $\sum_{i=1}^{n} w_i r_{(i)}^2$ than $\delta \sum_{j=1}^{p} \beta_j^2$ (which is negligible in comparison with $\sum_{i=1}^{n} w_i r_{(i)}^2$). The $\sqrt{n}$-consistency of RLWS can be proven as well.

Theorem 2 *Let all conditions of Lemma 2 in* [14] *hold, then RLWS estimate is* $\sqrt{n}$*-consistent.*

Sketch of the proof: Unfortunately, we did not build sufficient notation in this paper and also we do not have enough space for the whole proof, so again only the idea. To prove the $\sqrt{n}$-consistency of RLWS we will use all results from [14]. In this setup, X_i is a random vector. An outline of the proof of the $\sqrt{n}$-consistency of the LWS estimate from [14]: At first we derive the normal equations for LWS (denoted as $\boldsymbol{N}E_{Y,X,n}^{LWS}(\beta)$) using the empirical distribution function of absolute values of the residuals. Then we are working with $\frac{1}{\sqrt{n}}\boldsymbol{N}E_{Y,X,n}^{LWS}(\beta) = 0$. Using weak consistency of LWS, closeness of the empirical and theoretical distribution functions, and the conditions of the theorem, we arrive at equality

$$A(n,X,e)\,\sqrt{n}(\hat{\beta}^{(LWS,n,w)} - \beta^0) + R(\hat{\beta}^{(LWS,n,w)}, n, X, e) = P(\hat{\beta}^{(LWS,n,w)}, n, X, e)$$

where term $R(\hat{\beta}^{(LWS,n,w)}, n, X, e)$ is $o_p(1)$, $P(\hat{\beta}^{(LWS,n,w)}, n, X, e)$ is $\mathcal{O}_p(1)$ and $A(n,X,e)$ converges in probability to a regular matrix.

Because the RLWS estimate is also weakly consistent and the normal equations for RLWS (using empirical distribution function – not derived in this paper) are

$\boldsymbol{NE}^{RLWS}_{Y,X,n,\delta}(\beta) = \boldsymbol{NE}^{LWS}_{Y,X,n}(\beta) - \delta\beta$, we can repeat all steps of Lemma 2 in [14] and get equation

$$A(n,X,e)\,\sqrt{n}(\hat{\beta}^{(RLWS,n,w,\delta)} - \beta^0) - \frac{1}{\sqrt{n}}\delta\hat{\beta}^{(RLWS,n,w,\delta)}$$
$$+R(\hat{\beta}^{(RLWS,n,w,\delta)}, n, X, e) = P(n,X,e,\hat{\beta}^{(RLWS,n,w,\delta)}).$$

Terms A, P and R have the same properties as above. This is in fact the end of the proof, because $\frac{1}{\sqrt{n}}\delta\hat{\beta}^{(RLWS,n,w,\delta)}$ is $o_p(1)$. The exact proof will be available in upcoming paper [4]. □

5. Discussion

The aim of this short paper is to show a new estimate which seems to be a reasonable candidate for a robust version of the ridge regression. Another role of this estimate can also be a robust detector of multicollinearity. It is known that already one additional observation may hide or create multicollinearity for classical methods of multicollinearity detection (such as the condition number or Pearson's correlation coefficient). The idea of utilization RLWS as robust diagnostics of multicollinearity is simple: We have the RLWS estimate as well as the weights assignment of this estimate. So, if RLWS works well, we drop off observations which are identified (by their weights) as outliers and use the classical multicollinearity diagnostics on the rest of the (noncontaminated) data.

It is also important to know the way of obtaining this estimate because it tells us that our proposal is reasonable. Basic properties such as existence, uniqueness, equivariance and consistency is discussed. The next step will be to investigate the performance of the estimate on real and also simulated data as well as to make a comparison with other methods. This way is open because RLWS can be simply computed using the same type of algorithm as the LWS (more details about algorithm can be found in [12]).

References

[1] Askin, R. G., Montgomery, D. C.: *Augmented robust estimators*. Technometrics, **22** (1980), 333–341.

[2] Hoerl, A. E., Kennard, R. W.: *Ridge regression: biased estimation for nonorthogonal problems*. Technometrics **12** (1970), 55–68.

[3] Jurczyk, T.: *Outlier detection under multicollinearity*. Preprint.

[4] Jurczyk, T.: *Consistency and $\sqrt{n}$-consistency of the ridge least weighted squares.* Preprint.

[5] Lawrence, K. D., Arthur, J. L.: Robust Regression: Analysis and Application. Marcel Dekker, New York, 1990.

[6] Mašíček, L.: *Diagnostics and Sensitivity of Robust Models*. Disertation thesis at Charles University in Prague, 2004.

[7] Midi, H., Zahari, M.: *A simulation study on ridge regression estimators in the presence of outliers and multicollinearity*. Jurnal Teknologi **47** (2007), 59–74.

[8] Rousseeuw, P. J.: *Least median of squares regression*. J. Amer. Statist. Assoc. **79** (1984), 871–880.

[9] Rousseeuw, P. J., Leroy, A. M.: Robust Regression and Outlier Detection. John Wiley & Sons, New York, 1987.
[10] Simpson, J. R., Montgomery, D. C.: *A biased-robust technique for the combined outlier-multicollinearity problem*. J. Stat. Comput. Simul. **56** (1996), 1–22.
[11] Víšek, J. Á.: *Regression with high breakdown point*. Robust 2000 (eds. Jaromír Antoch & Gejza Dohnal, published by Union of Czech Mathematicians and Physicists), Matfyzpress, Prague (2001), 324–356.
[12] Víšek, J. Á.: *Consistency of the Instrumental Weighted Variables*. Ann. Inst. Statist. Math., **61(3)** (2009), 543–578.
[13] Víšek, J. Á.: *Consistency of the least weighted squares under heteroscedasticity*. Submitted to Kybernetika.
[14] Víšek, J. Á.: $\sqrt{n}$*-consistency of the least weighted squares under heteroscedasticity*. Preprint.
[15] Zvára, K.: Regresní analýza. Academia, Praha, 1989.