

# On Robust Estimation of Error Variance in (Highly) Robust Regression

Jan Kalina<sup>1,2</sup>, Jan Tichavský<sup>1</sup>

<sup>1</sup>*The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Praha 8, Czech Republic, kalina@cs.cas.cz*

<sup>2</sup>*The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 00 Praha 8, Czech Republic*

The linear regression model requires robust estimation of parameters, if the measured data are contaminated by outlying measurements (outliers). While a number of robust estimators (i.e. resistant to outliers) have been proposed, this paper is focused on estimating the variance of the random regression errors. We particularly focus on the least weighted squares estimator, for which we review its properties and propose new weighting schemes together with corresponding estimates for the variance of disturbances. An illustrative example revealing the idea of the estimator to down-weight individual measurements is presented. Further, two numerical simulations presented here allow to compare various estimators. They verify the theoretical results for the least weighted squares to be meaningful. MM-estimators turn out to yield the best results in the simulations in terms of both accuracy and precision. The least weighted squares (with suitable weights) remain only slightly behind in terms of the mean square error and are able to outperform the much more popular least trimmed squares estimator, especially for smaller sample sizes.

Keywords: High robustness, robust regression, outliers, variance of errors, least weighted squares, simulation.

## 1. INTRODUCTION

The simplicity and interpretability of the linear regression model has resulted in an enormous number of applications in modeling real data. The aim of linear regression is to model a continuous variable taking into account one or more independent variables (regressors). Regression can be used to predict values of the response based on such values of one or more (continuous and/or categorical) variables, which are not available. Already Carl Friedrich Gauss was using regression for equalization of networks constructed from geodesic measurements and thus he connected to the two intertwined science disciplines, namely measurement science and statistical point estimation. Later, Ronald A. Fisher or Francis Galton, who both contributed to development of regression analysis, were also passionate about measurement [1]. Regression analysis has become an important part of analyzing measurements by means of exploratory data analysis (EDA) or predictive data mining, with numerous applications not limited to calibration of measuring instruments [2], estimating missing values, or detecting mixtures of two (or more) populations or clusters of data [3].

Real measurements can be influenced by (random or systematic) errors coming from different sources (e.g. because some measurements are performed under different condi-

tions), while there are established tools for estimating the uncertainty of measurements in different situations (see e.g. [4]). We can say that real measurements often suffer from outlying values (outliers), which remain however only vaguely defined. If the linear model exists objectively, then any measurement, for which the considered linear regression model does not objectively hold, may be perceived as an outlier in the model. If the linear model represents only our approximation to reality, then any anomalous measurements not fitted well by the model play the role of outliers. In general, outliers typically appear in real data in across various disciplines, e.g. in engineering applications [5] or image analysis based on markers measured within images [6]. Outliers appear practically always in measurements of molecular genetic and metabolomic biomarkers, for which severe measurement errors are immanent to the measurement technology [7], [8]. Thus, there emerges a need for robust estimation of regression parameters.

While the least squares estimator in linear regression is notoriously known to suffer from the presence of outlying values in the data, numerous robust regression estimators have been proposed as resistant alternatives [9], including various recent tools tailor-made for very specific tasks. The concept of robust estimation is traditionally tied to resistance with respect to outliers. However, there remains a gap in the area of esti-

imating the variance of the random regression errors (i.e. error variance), as there seem no systematic comparisons among different robust estimates.

Within regression modeling, it is important to estimate the regression parameters (say  $\beta$ ) together with the variance of errors, commonly denoted as  $\sigma^2$  (if this is assumed to exist). If the measurements are contaminated by outliers, then it is necessary to estimate  $\sigma^2$  in a highly robust way. It is possible to accompany each of the commonly used robust estimators of  $\beta$  by a corresponding estimator of  $\sigma^2$ , which allows to evaluate the precision (uncertainty) of the regression fit. Estimation of  $\sigma^2$  within robust regression is crucial for robust hypothesis tests about the significance of regression parameters, for the corresponding confidence intervals for the parameters, for outlier detection or comparing the efficiency of various robust estimates. Outlier detection based on robust estimates of  $\sigma^2$  has been popularized by the seminal book [10] and has been several times used in practical applications [11].

In this paper, we work with several important classes of (possibly highly) robust regression estimators, including S-estimators, MM-estimators, or the least trimmed squares estimator. Nevertheless, we are primarily interested in the least weighted squares estimator, which turns out to have appealing statistical properties [12]. As recalled in Section 2, their corresponding estimators of  $\sigma^2$  are consistent (under specific assumptions). We are particularly interested in the least weighted squares estimator, which remains much less known compared to other available robust estimators. For the least weighted squares, we propose new weighting schemes together with corresponding estimates for the variance of the regression errors. We present an illustrative example in Section 3 and numerical simulations in Section 4 comparing the variance of the errors obtained by several highly robust estimators. Finally, Section 5 concludes the paper.

## 2. SUBJECT & METHODS

We consider the standard linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + e_i, \quad i = 1, \dots, n, \quad (1)$$

where  $n$  denotes the total number of measurements available for a continuous response  $Y = (Y_1, \dots, Y_n)^T$  and for  $p$  independent variables (regressors), which may be either random or deterministic. We will use the usual notation  $X_i = (X_{i1}, \dots, X_{ip})^T$  and  $X = (X_1^T, \dots, X_n^T)^T$ . The random errors (disturbances)  $e_1, \dots, e_n$  are assumed to be independent and identically distributed (i.i.d.) with a distribution function  $F$ , which is absolutely continuous with the corresponding probability density function  $f$ . Throughout the paper, we assume  $Ee_1^2$  to exist and denote the common variance of the errors by  $\sigma^2 := \text{var } e_1$ . The statistical concept of variability corresponds to the precision in measurement science, as it holds  $\text{var } Y_1 = \text{var } e_1 = \sigma^2$  for measurements with non-random regressors.

It seems natural to estimate  $\sigma^2$  by the same estimation principle, which is used for the very estimate of  $\beta$ . A joint estimation of  $\beta$  and  $\sigma^2$  is however conceptually difficult and it was even impossible for the M-estimators, i.e. the very first ro-

bust estimators, which naturally generalize maximum likelihood estimation principles. In fact, studentized M-estimators were proposed, which had to rely on an initial estimate of  $\sigma^2$ , which plays the role of a nuisance parameter [9]. This reveals the importance of estimating  $\sigma^2$  and we focus here only on such estimates, which are able to estimate  $\beta$  and  $\sigma^2$  jointly. They are suitable for the so-called contaminated normal distribution corresponding to mixture of normal errors with outliers [13], or for errors with a heavy tailed distribution.

This section recalls the least squares and its weighted version together with several (possibly highly) robust estimators, namely S-estimators, MM-estimators and least trimmed squares. In addition, we present new ideas for the least weighted squares estimator. We also cite references, where the consistency of these estimators of  $\beta$  and corresponding estimators of  $\sigma^2$  was proven; usually it is needed to assume that the distribution of errors is continuous and symmetric around 0, apart from additional technical assumptions. Highly robust estimators are defined as those, which may attain a high value of the breakdown point. We can say that the breakdown point, which represents a fundamental concept of robust statistics [9], is a measure of robustness of a statistical estimator of an unknown parameter. Formally, the finite-sample breakdown point evaluates the minimal fraction of data that can drive an estimator beyond all bounds when set to arbitrary values.

### 2.1. Least squares and weighted least squares

The least squares estimator with the analytical expression  $b_{LS} = (X^T X)^{-1} X^T Y$  is vulnerable to the presence of outliers in the data. To avoid further confusions, let us also recall its weighted version with given non-negative weights  $w_1, \dots, w_n$  fulfilling  $\sum_{i=1}^n w_i = 1$ . Let  $u_i(b)$  denote the residual of the  $i$ -th measurement based on a given estimate  $b = (b_0, b_1, \dots, b_p)^T$  of  $\beta$ , i.e.

$$u_i(b) = Y_i - b_0 - b_1 X_{i1} - \cdots - b_p X_{ip}, \quad i = 1, \dots, n. \quad (2)$$

While the least squares estimator is suitable for measurements with the same precision, the weighted least squares (WLS) estimator, also known as Aitken estimator or generalized least squares, represents an analogy for differently precise measurements, however with a known precision. Formally, the WLS estimator is obtained as

$$b_{WLS} = \arg \min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^n w_i u_i^2(b), \quad (3)$$

i.e. by minimization of a weighted estimate of  $\sigma^2$  over  $b$ . Equivalently, the WLS estimator is obtained as the solution of the set of normal equations

$$\sum_{i=1}^n w_i X_i (Y_i - X_i^T b) = 0. \quad (4)$$

Denoting by  $W$  the diagonal matrix with elements  $w_1, \dots, w_n$  on the main diagonal, the explicit form of  $b_{WLS}$  is

$$b_{WLS} = (X^T W X)^{-1} X^T W Y. \quad (5)$$

## 2.2. Estimation of $\sigma^2$ and outlier detection

An important direction of robust statistics was established on (ordinary) residuals (2). While outlier detection is often performed on studentized residuals instead of the ordinary ones in (2), we are not aware of robust estimators based on studentized residuals, where the latter require an estimate of  $\sigma^2$  and of the projection matrix  $H = X(X^T X)^{-1} X^T$ . An externally studentized residual (jackknife residual, studentized deleted residual) for the  $i$ -th measurement for  $i = 1, \dots, n$  uses an estimate of  $\sigma^2$  computed from all observations except for the  $i$ -th. On the other hand, an (internally) studentized residual evaluates an estimate of  $\sigma^2$  exploiting all observations. An example of a recent result is the work [14] investigating (both internal and external) studentized residuals of regression quantiles, proposed with the aim to construct regression diagnostic tools resistant to outliers.

The literature on robust regression commonly considers the following outlier detection method. If  $u_1, \dots, u_n$  are ordinary residuals of (any) robust estimate of  $\beta$ , the  $i$ -th measurement is classified  $(\mathbf{X}_i, Y_i)^T$  as an outlier if and only if

$$|u_i| > 2.5\hat{\sigma}, \quad (6)$$

where  $\hat{\sigma}^2$  denotes an estimate of  $\sigma^2$ .

The threshold 2.5 in (6) was advocated in references on outlier detection in linear regression, especially in the seminal papers published by promoters of the concept of the breakdown point [10], [15] and in numerous subsequent papers as well (see e.g. p. 74 of [16]), although other values (e.g. 2 or 3) are popular among practitioners as rules of thumb. The cut-off value 2.5 is chosen with regard to the situation with normal errors; in such case only 1.24 % of residuals are classified as outliers [11]. Still, it is necessary to admit that each robust estimator would require its own particular value of the threshold [5].

The current paper is interested in robust estimation of  $\sigma^2$ , which itself is a still overlooked topic in robust regression. It is natural to use the same robust estimation principle to estimate both  $\beta$  and  $\sigma^2$  in (1). In other words, a robust estimator of  $\sigma^2$  can be adapted to obtain a corresponding (and consistent) estimate of  $\sigma^2$ .

## 2.3. S-estimators

S-estimators for linear regression defined as

$$b_S = \arg \min_{b \in \mathbb{R}^{p+1}} s(u_1(b), \dots, u_n(b)), \quad (7)$$

where  $s$  is a scale functional fulfilling certain conditions, were investigated e.g. by [17]. Their corresponding estimator of  $\sigma^2$  is equal directly to  $s(u_1(b_S), \dots, u_n(b_S))$ . Sufficient conditions for the consistency of S-estimators for  $\beta$  and of the corresponding estimator of  $\sigma^2$  was presented also in [17], while consistency factors for S-estimators to achieve a given value of the breakdown point (or alternatively to achieve a given efficiency) for normal data were derived in [18].

## 2.4. MM-estimators

As S-estimators possess a low efficiency, MM-estimators were proposed in [19] as two-stage estimation tools, whose efficiency can be tuned independently of the robustness. The computation starts with an initial S-estimator, while the corresponding (estimated) scale (say  $\hat{s}$ ) is used (provided technical assumptions) in the second stage defining MM-estimators as

$$\frac{1}{n} \sum_{i=1}^n \psi \left( \frac{u_i(b)}{\hat{s}} \right) Y_i = 0. \quad (8)$$

While (8) itself has the form of an M-estimator (however with a specific fixed scale), the MM-estimator of  $\sigma^2$  is defined as the corresponding M-estimator of  $\sigma^2$ . The consistency of MM-estimators for both  $\beta$  and  $\sigma^2$  under specific assumptions was derived in [19].

## 2.5. Least trimmed squares

The least trimmed squares (LTS) estimator of  $\beta$  for a selected value of a trimming constant  $h$  ( $n/2 \leq h < n$ ) was proposed in [10]. Denoting residuals corresponding to a particular  $b \in \mathbb{R}^p$  as  $u_i(b)$  and order statistics of their squares as

$$u_{(1)}^2(b) \leq \dots \leq u_{(n)}^2(b), \quad (9)$$

the estimator is defined as

$$b_{LTS} = \arg \min_{b \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^h u_{(i)}^2(b). \quad (10)$$

It will be convenient to use the notation  $\alpha = (n-h)/n \in [1/2, 1)$ . Consistency of the LTS estimator and other properties were derived in [20] for random regressors under the assumption that  $Ee_1^4 \in (0, \infty)$ . The LTS estimator may attain a high robustness but cannot achieve a high efficiency, which is a consequence of the hard trimming, i.e. complete rejection of outliers. We may consider the LTS as a special case of the least weighted squares estimator, where the latter will be recalled in Section 2.7.

A consistent estimator  $\hat{\sigma}_{LTS}^2$  of  $\sigma^2$  based on the LTS regression was derived in [20]. Introducing auxiliary notation in the form

$$c_\alpha = F^{-1} \left( 1 - \frac{\alpha}{2} \right) \quad \text{and} \quad \chi_\alpha = \frac{\alpha}{\alpha - 2c_\alpha f(c_\alpha)} \quad (11)$$

and denoting residuals corresponding to the LTS estimator by  $u_i^{LTS}$  for  $i = 1, \dots, n$ , we may express Višek's estimator  $\hat{\sigma}_{LTS}^2$  as

$$\hat{\sigma}_{LTS}^2 = \frac{\chi_\alpha}{h} \sum_{i=1}^h (u_i^{LTS})^2. \quad (12)$$

The effect of  $\alpha$  on the values of  $\chi_\alpha$  is revealed in Fig.1. Let us note that a formal choice  $\alpha = 1$  yields  $\chi_\alpha = 1$ , i.e. (12) reduces to the standard estimate of  $\sigma^2$  for the least squares. We may observe in Fig.1., that values of  $\chi_\alpha$  are decreasing with an increasing  $\alpha$ .

## 2.6. Least weighted squares

The least weighted squares (LWS) estimator (see e.g. [12]) for the model (1) generalizes the LTS based on implicit weighting of individual measurements. It performs down-weighting of individual measurements through the idea to assign small (or zero) weights to potential outliers. This down-weighting may also be denoted as soft trimming to stress the contrary of the hard trimming of the LTS. The LWS estimator, which has acquired only much smaller attention compared to the popular LTS, may attain a high breakdown point, if a suitable weight function is chosen. The high breakdown point can be interpreted as a high resistance against noise or outliers in the data. The LWS estimator is at the same time robust to heteroscedasticity [12], but its primary attention is focused on estimating  $\beta$  and not on outlier detection. The idea of the LWS, i.e. the implicit weights based on ranks of residuals, was successful in a variety of recent applications including denoising gene expression measurements acquired by the microarray technology [8] or image analysis based on landmarks measured within facial images [11]. There is a good experience with implicit weighting also for multivariate robust estimation, the minimum weighted covariance determinant (MWCD) estimator proposed in [21], [22].

The definition of the LWS exploits the concept of weight function, which is defined as a function  $\psi : [0, 1] \rightarrow [0, 1]$ ; it must be continuous on  $[0, 1]$  with  $\psi(0) = 1$  and  $\psi(1) = 0$ . The weight function is assumed to have both one-sided derivatives existing in all points of  $(0, 1)$ , where the one-sided derivatives are bounded by a common constant; also, the existence of a finite left derivative in 0 and finite right derivative in point 1 is assumed [23], [12].

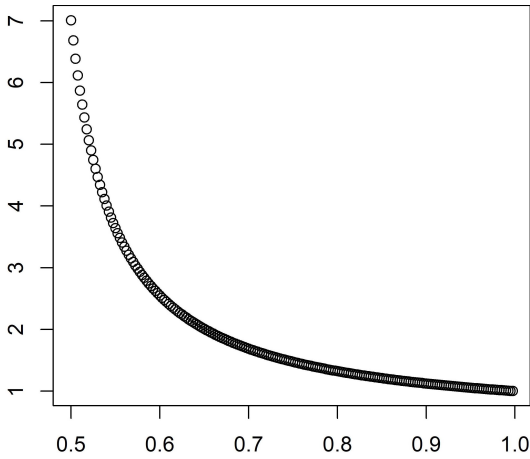


Fig. 1. Horizontal axis: values of  $\alpha \in [1/2, 1]$ . Vertical axis: values of  $\chi_\alpha$  for the LTS estimator depending on  $\alpha$ .

The LWS estimator with a given  $\psi$  is defined as

$$b_{LWS} = \arg \min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^n \psi \left( \frac{i-1/2}{n} \right) u_{(i)}^2(b). \quad (13)$$

We may understand the quantities

$$w_i = \psi \left( \frac{i-1/2}{n} \right), \quad i = 1, \dots, n, \quad (14)$$

as weights. Alternatively, we may start with choosing a fixed non-increasing sequence of non-negative weights  $w_1, \dots, w_n$  and formulate an equivalent definition of the LWS estimator of  $\beta$  in the form

$$b_{LWS} = \arg \min_{b \in \mathbb{R}^{p+1}} \sum_{i=1}^n w_i u_{(i)}^2(b). \quad (15)$$

Note that we do not need to require the weights to be standardized to the condition  $\sum_{i=1}^n w_i = 1$ .

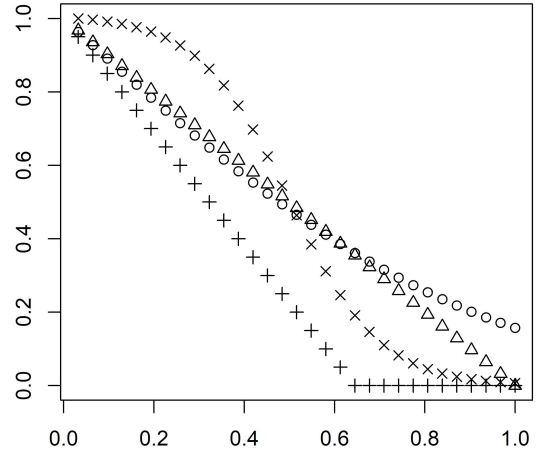


Fig. 2. Horizontal axis: values of  $t \in [0, 1]$ . Vertical axis: weight functions  $\psi(t)$  corresponding to LWS1 (triangles), LWS2 (crosses), LWS3 (circles) and LWS4 (plus signs).

If we denote the ranks of (2) by  $R_1(b), \dots, R_n(b)$ , i.e. with  $R_i(b)$  denoting the rank of  $u_i^2(b)$  among  $u_1^2(b), \dots, u_n^2(b)$ , we may express the LWS estimator as

$$b_{LWS} = \arg \min_{b \in \mathbb{R}^{p+1}} \sum_{k=1}^n \psi \left( \frac{R_k(b) - 1/2}{n} \right) u_{(k)}^2(b). \quad (16)$$

This expression is appealing from the computational perspective. For the computation of the LWS estimator, an analogy of the FAST-LTS algorithm of [24] exploiting the form (16) can be formulated in a straightforward way. This approximate algorithm was characterized as reliable based on empirical evidence [25].

The derivative of the loss function

$$l(b) = \sum_{i=1}^n \psi \left( \frac{i-1/2}{n} \right) u_{(i)}^2(b) \quad (17)$$

has the form (in our notation)

$$\frac{\partial l}{\partial b} = \sum_{i=1}^n X_i u_i(b) \psi \left( \hat{F}^{(n)}(|u_i|, b) \right), \quad (18)$$

where  $\hat{F}^{(n)}$  denotes the empirical distribution function

$$\hat{F}^{(n)}(r, b) = \frac{1}{n} \sum_{k=1}^n \mathbb{1}[|u_k(b)| < r], \quad r \in \mathbb{R}; \quad (19)$$

a detailed proof was given on p. 183 of [12]. As a consequence, the set of normal equations for the LWS estimator in (1) has the form

$$\sum_{i=1}^n X_i u_i(b) \psi(\hat{F}(|u_i|)) = 0, \quad (20)$$

which apparently differs from (4) of the WLS by permuting the weights before assigning to individual measurements.

It is meaningful to consider only weight functions which are non-increasing; only then, less reliable measurements obtain smaller weights. Four possible (fixed) weighting schemes for the LWS will be now introduced in the form of different weight functions described below; the corresponding LWS estimators will be denoted as LWS1 to LWS4.

1. LWS1. Linear weights

$$\psi(t) = 1 - t, \quad t \in [0, 1]. \quad (21)$$

2. LWS2. Weights generated by the (strictly decreasing) logistic function

$$\psi(t) = \frac{1 + \exp\{-s/2\}}{1 + \exp\{s(t - \frac{1}{2})\}}, \quad t \in [0, 1], \quad (22)$$

for a given constant  $s > 0$  responsible for the shape of the logistic curve, while we consider  $s = 10$  in all the computations.

3. LWS3. Trimmed linear weights generated for a fixed  $\tau \in [1/2, 1)$  by

$$\psi(t) = \left(1 - \frac{t}{\tau}\right) \cdot \mathbb{1}[t < \tau], \quad t \in [0, 1], \quad (23)$$

where  $\mathbb{1}[\cdot]$  denotes an indicator function. Here,  $\tau$  is in relationship with the trimming, i.e. there are  $\lfloor \tau n \rfloor$  measurements retained and the remaining measurements are ignored, where  $\lfloor x \rfloor$  denotes the integer part of  $x \in \mathbb{R}$ ; this is analogous to  $\alpha$  for the LTS.

4. LWS4. Weights generated by the so-called error function in the form

$$\psi(t) = 1 - \frac{2}{\sqrt{\pi}} \int_0^t \exp\{-x^2\} dx. \quad (24)$$

These four weight functions are shown in Fig.2. In real-life applications, the LWS was successfully considered with weighting schemes LWS1 [8] and LWS2 [11]. No other weighting scheme seems to have been used in real-life applications, i.e. neither for the LWS nor for the MWCD estimator. Vřšek [12] proved the LWS estimator to be consistent for random regressors under general assumptions on the weight function and on the joint distribution of the regressors and the weights. It is worth noting that only second moments of  $e_1$  are assumed to exist. These weak assumptions, i.e. unexpectedly weak from the point of view of the discussion on p. 10 of [20], however required a very long proof of the consistency. The

proof is valid for each of the four weighting functions considered above.

Concerning now the LWS estimator with a given weight function  $\psi$ , let us denote the corresponding residuals (i.e. after the optimal permutation of the weights) by  $u_i^{LWS}$  for  $i = 1, \dots, n$ . Denoting

$$\gamma = \int \xi(r) dr \quad \text{and} \quad \xi(r) = \psi(F(|r|)) r^2 f(r), \quad (25)$$

let us express the consistent estimator of  $\sigma^2$  based on the LWS by [23] in the form

$$\hat{\sigma}_{LWS}^2 = \frac{1}{n\gamma} \sum_{i=1}^n w_i (u_i^{LWS})^2. \quad (26)$$

We now propose to consider a straightforward approximation of the integral in formula (25). We choose a large  $A > 0$  and a large  $N > 0$  in order to consider the integral over  $N$  subintervals of the interval  $[-A, A]$ . We now approximate  $\gamma$  by

$$\gamma_A = \frac{2A}{N} \sum_{k=1}^N \xi\left(\frac{A}{N}(2k - N - 1)\right). \quad (27)$$

For real data, the variance (26) can be estimated using (27) in a straightforward way, while it holds  $\gamma_A \rightarrow \gamma$  for  $n \rightarrow \infty$  for all weight functions of above. We computed values of  $\gamma_A$  (and also of  $1/\gamma_A$ , useful within (27)), for different choices of weights for the LWS, and the results are presented in Table 2. There, we choose  $A = 10$  and  $N = 10000$ . Because different weight functions have remarkably different values of  $\int_0^1 \psi(t) dt$  and thus corresponding values of  $\sum_{i=1}^n w_i$  (for a particular fixed  $n$ ), which is not standardized (as stated already above), values of  $\gamma_A$  are not directly comparable among different weight functions. Thus, for example, LWS4 has a much larger value of  $\gamma_A$  compared to other versions of the LWS, because it also assigns larger (unnecessarily large) weights to potential outliers.

2.7. The simplest regression model

The simplest special case of the regression model (1), known as the location model, has the form

$$Y_i = \mu + e_i, \quad i = 1, \dots, n, \quad (28)$$

where  $\mu \in \mathbb{R}$ . This model, extremely useful to describe any simple measurement of an unknown constant  $\mu$ , considers  $X$  from (1) to be equal to  $(1, \dots, 1)^T$ . Various simple tools are available for (28), which cannot be extended to (1); this is true e.g. for outlier detection based on the interquantile range [9]. Still, the least squares estimator remains to have a zero breakdown point in (28). i.e. to be extremely vulnerable to severe outliers. The LWS estimator, which remains meaningful for (28) as well, has been successfully applied in [11], [8].

The LWS estimator of  $\mu$  in (28) computed with fixed magnitudes of the weights  $w_1, \dots, w_n$  fulfilling  $\sum_{i=1}^n w_i = 1$  is equal to the weighted mean of the data  $Y_1, \dots, Y_n$  with such

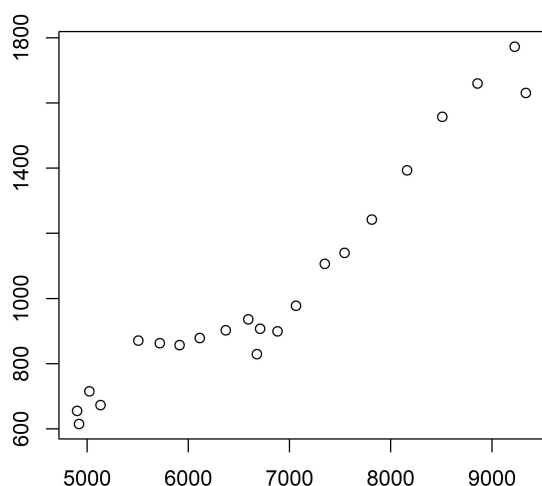


Fig. 3. Raw data of Section 3. Horizontal line: measurements of drug D. Vertical line: measurements of protein P.

permutation of the weights, which has the smallest weighted variance

$$S_w^2(\mathbf{Y}) = \sum_{i=1}^n w_i \left( Y_i - \sum_{j=1}^n w_j Y_j \right)^2; \quad (29)$$

see [11] for a proof. We give an important remark here that the estimate of  $\text{var} e_1 = \sigma^2$  of above, i.e. based on (26) with (27), holds also in (28), so Table 2 remains valid in (28) as well.

### 3. ILLUSTRATIVE EXAMPLE

To illustrate the performance of robust regression estimators, we present the following illustrative artificial example. Let us say that the level of drug D in the blood of patients with hypothyroidism is a biomarker of the excretion level of protein P in the body. In other words, there is a hypothesis that the amount of protein P in the body is proportional to the level of drug D in the blood. Measurements of both these (continuous) variables were performed over a random sample of  $n = 22$  patients, so the measurements  $D_1, \dots, D_n$  of the level of drug D and measurements  $P_1, \dots, P_n$  of the level of protein P are available. We consider the linear regression model To illustrate the performance of robust regression estimators, we present the following illustrative artificial example. Let us say that the level of drug D in the blood of patients with hypothyroidism is a biomarker of the excretion level of protein P in the body. In other words, there is a hypothesis that the amount of protein P in the body is proportional to the level of drug D in the blood. Measurements of both these (continuous) variables were performed over a random sample of  $n = 22$  patients, so the measurements  $D_1, \dots, D_n$  of the level of drug D and measurements  $P_1, \dots, P_n$  of the level of protein P are available. We consider the linear regression model

$$P_i = \beta_0 + \beta_1 D_i + e_i, \quad i = 1, \dots, n. \quad (30)$$

The raw data are shown in Fig.3. We also consider a contaminated version of the data shown in Fig.4, corresponding to the situation when measurement errors, methodological errors or errors of other types and sources lead to modification of two observations with indexes 4 and 17.

The results of various estimators of  $(\beta_0, \beta_1)^T$  are presented in Table 1. We can see that the least squares estimates are highly sensitive to the contamination. While different robust estimates give different results, we illustrate the interpretability of the LWS. We namely present the plot of the weights determined by the LWS3 estimator, i.e. with the trimmed linear weights, which completely ignore (about) 25 % of the measurements, in our case exactly 5 measurements. Fig.5. for the raw data and in Fig.6. for the contaminated data. Only the weights corresponding to the two measurements with index 4 and 17 are shown as full circles. These measurements are well explained by the linear model in the raw dataset, but their weights are reduced to 0 if their values are replaced by outlying values. This shows that the LWS3 method is indeed able to recognize them as being contaminated, while the values of the weights of the remaining measurements are modified only slightly after the contamination. Thus, the LWS3 estimator is able to rely on the same measurements here, which are reliable and well explained by the linear model.

### 4. SIMULATIONS

We present two simulation studies denoted as A and B, which were performed with the aim to investigate various estimates of  $\sigma^2$  based on the least squares (LS) as well as on several robust estimators. Particularly, we compute S-estimators with breakdown point 0.5, MM-estimators with breakdown point 0.5 and efficiency 0.95, the LTS (with two different choices of  $\alpha$ ), and the LWS with four different weighting schemes of above. In both simulations, these estimators of  $\beta$  and corresponding estimators of  $\sigma^2$  are consistent, as the consistency assumptions are fulfilled for each of

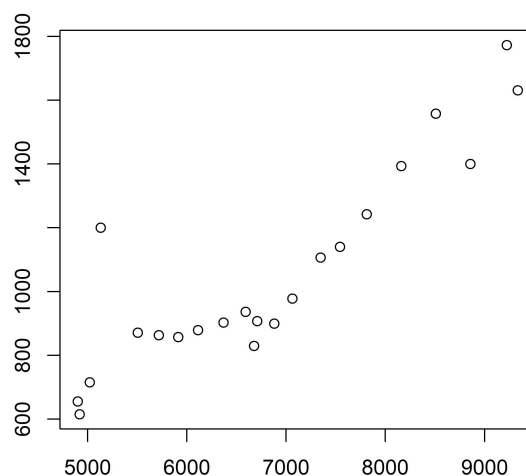


Fig. 4. Contaminated data of Section 3. Horizontal line: measurements of drug D. Vertical line: measurements of protein P.

Table 1. Points estimates of  $\beta$  in the example of Section 3 over raw as well as contaminated data.

Estimator	Raw data		Contaminated data	
	Intercept	Slope	Intercept	Slope
LS	-582	0.239	-331	0.204
LTS ( $\alpha = 0.55$ )	-371	0.203	-359	0.201
LTS ( $\alpha = 0.80$ )	-515	0.228	-423	0.212
S-estimator	-542	0.206	-387	0.207
MM-estimator	-567	0.237	-486	0.222
LWS1	-465	0.221	-396	0.209
LWS2	-280	0.189	-282	0.189
LWS3 ( $\tau = 0.75$ )	-244	0.183	-251	0.185
LWS4	-519	0.230	-344	0.203

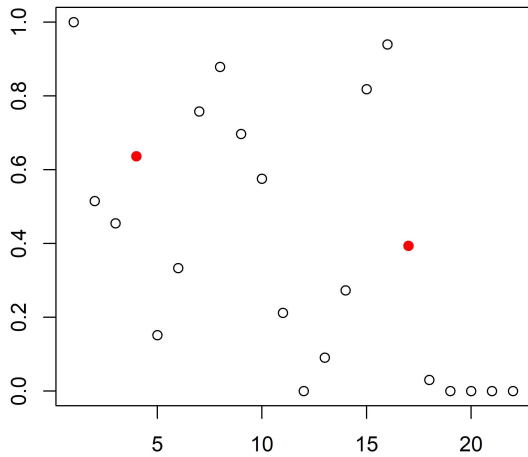


Fig. 5. Weights of the LWS3 estimator for raw data of Section 3, shown in Fig.3. Horizontal line: index  $1, \dots, n$ . Vertical line: values of the weight function.

the estimators. For the computations, we use R software [26] with additional packages (robustbase, NORMT3).

In simulation A, we consider a simple setup with  $p = 1$  and  $n = 30$  or  $n = 60$ . The regressor is randomly generated from uniform distribution  $U(0, 10)$ , i.e. over the interval  $(0, 10)$ , and we use fixed values  $\beta = (2, 1)^T$  and  $\sigma^2 = 1$ . In this fixed design, we generate  $J = 10000$  datasets, where the errors are generated as i.i.d. with  $e_i \sim N(0, \sigma^2)$ . In each dataset, we replace 10 % of measurements by outliers. First, a constant  $m$  from  $U(0, 10)$  is generated, and we randomly select 10 % of the measurements. The regressor in these selected data points is replaced by  $m + v$ , where  $v \sim N(0, 0.25)$  are i.i.d. normally distributed. The response in these data points is obtained according to (1) with i.i.d. errors generated from the uniform distribution over  $(-6, -3) \cup (3, 6)$ . Such design allows to compare the ability of different regression estimators to estimate  $\sigma^2$ , as we evaluate quartiles of the estimates together with (empirical) mean square error

$$MSE = \frac{1}{J} \sum_{j=1}^J (\hat{\sigma}_j^2 - \sigma^2)^2; \quad (31)$$

the value of MSE is small (i.e. close to 0) for estimators which are accurate and at the same time precise in estimating  $\sigma^2$ . In other words, the values of MSE give only a combination of accuracy and precision, which can be separated from each other when interpreting the presented quartiles.

The top part of Table 3. presents the results of simulation A. There, we evaluate averaged values of estimates of  $\sigma^2$  over all datasets (i.e. also over all designs). Particularly, the averaged first (Q1), second (Q2) and third (Q3) quartiles are reported in the table. The least squares are extremely misled by the contamination by outliers in terms of both accuracy and precision. The best values of MSE are shown in bold-face. These are acquired by MM-estimators, while the LWS estimator stays only slightly behind. It is quite important to choose suitable weights for the LWS; if zero weights are assigned to outliers, which is the case of LWS2 and LWS3, the results are superior to those of the LTS (independently on choosing its parameter  $\alpha$ ). We also observe that the estimation of  $\sigma^2$  is improved with an increasing  $n$  in terms of MSE.

It is remarkable that the quartiles obtained by S-estimators are identical and the same is true for the three quartiles of MM-estimators, i.e. the methods are able to estimate  $\sigma^2$  precisely with an extreme reduction of the influence of the outliers. Of course, this precision holds only for the reported quartiles, but still this phenomenon has not been (to the best of our knowledge) presented. We perceive it as an appealing property for analyzing real measurements, especially for MM-estimates, which are much more accurate compared to S-estimates.

Simulation B considers a more complex dataset with  $p = 5$  and again  $n = 30$  or  $n = 60$ . Let us introduce the notation  $\mathcal{I}_r$  for a unit matrix of size  $r \times r$  and  $\mathbb{1}_r$  for the  $r$ -dimensional vector  $(1, \dots, 1)^T$ , where  $r \in \mathbb{N}$ . The regressor is constructed as  $(X_1, \dots, X_5)^T \sim N(0, \Sigma)$  with  $\Sigma = (1 - c)\mathcal{I}_5 + c\mathbb{1}_5\mathbb{1}_5^T$  and  $c = 0.3$ . We use fixed values  $\beta = \mathbb{1}_6$  and  $\sigma^2 = 1$ . In this

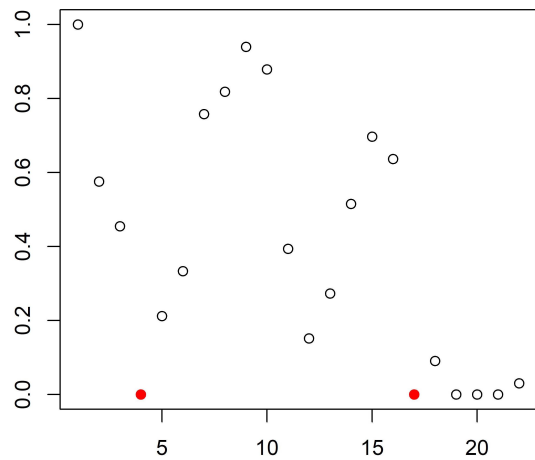


Fig. 6. Weights of the LWS3 estimator for contaminated data of Section 3, shown in Fig.4. Horizontal line: index  $1, \dots, n$ . Vertical line: values of the weight function.

Table 2. Values of  $\gamma_A$ , i.e. approximations to  $\gamma$ , together with values of  $1/\gamma_A$  for different choices of weights for the LWS estimator. The approximation is independent on  $n$ .

Choice of weights	LWS1	LWS2	LWS3 ( $\tau = 0.90$ )	LWS3 ( $\tau = 0.75$ )	LWS4
Value of $\gamma_A$ (27)	0.091	0.026	0.035	0.006	0.798
Value of $1/\gamma_a$	11.01	37.75	28.44	173.32	1.25

Table 3. Performance of different regression estimators in the simulations A and B. For each estimator, three quartiles of a consistent estimate of  $\sigma^2$  are presented. In the final column, the mean square error (31) of the estimate is reported.

Regression estimator	Estimate of $\sigma^2$				Estimate of $\sigma^2$			
	Q1	Q2	Q3	MSE	Q1	Q2	Q3	MSE
	Simulation A ( $n = 30$ )				Simulation A ( $n = 60$ )			
LS	2.29	4.46	6.20	16.2	2.97	4.18	4.85	10.5
LTS ( $\alpha = 0.55$ )	0.87	0.96	1.01	0.08	1.25	1.28	1.31	0.05
LTS ( $\alpha = 0.80$ )	0.96	1.02	1.06	0.09	1.11	1.16	1.20	0.05
S-estimator	1.48	1.48	1.48	0.25	1.43	1.43	1.43	0.19
MM-estimator	0.82	0.82	0.82	<b>0.05</b>	0.96	0.96	0.96	<b>0.03</b>
LWS1	1.20	1.28	1.32	0.09	1.24	1.27	1.33	0.07
LWS2	1.25	1.30	1.32	0.08	1.13	1.18	1.22	0.06
LWS3 ( $\tau = 0.75$ )	1.20	1.27	1.28	0.07	1.14	1.17	1.22	0.06
LWS4	1.18	1.30	1.34	0.10	1.27	1.29	1.33	0.08
	Simulation B ( $n = 30$ )				Simulation B ( $n = 60$ )			
LS	2.51	3.29	3.80	5.20	2.31	2.84	3.16	3.97
LTS ( $\alpha = 0.55$ )	1.28	1.32	1.36	0.11	1.23	1.26	1.30	0.08
LTS ( $\alpha = 0.80$ )	1.27	1.34	1.36	0.12	1.22	1.28	1.34	0.10
S-estimator	1.60	1.60	1.60	0.37	1.52	1.52	1.52	0.27
MM-estimator	1.29	1.29	1.29	<b>0.09</b>	1.25	1.25	1.25	<b>0.07</b>
LWS1	1.22	1.38	1.42	0.12	1.24	1.32	1.35	0.11
LWS2	1.22	1.27	1.36	0.10	1.16	1.20	1.31	0.10
LWS3 ( $\tau = 0.75$ )	1.29	1.31	1.36	0.10	1.17	1.22	1.27	0.10
LWS4	1.21	1.39	1.44	0.13	1.33	1.37	1.40	0.12

fixed design, we generate  $J = 10000$  datasets, where the errors are generated as i.i.d. with  $e_i \sim N(0, \sigma^2)$ . In each dataset, we replace 10 % of measurements by outliers. First,  $m_1$  and  $m_2$  are generated as independent from  $U(0, 3)$ , and we randomly select 10 % of the measurements. The regressor in these selected data points is replaced by  $m + v$ , where  $m = (m_1, m_2, 5, 5, 5)^T$  and  $v \sim N(0, 0.25)$ . The response in these data points is obtained according to (1) with i.i.d. errors generated from the uniform distribution over  $(-6, -3) \cup (3, 6)$ . In this design, the outliers always have the form of leverage points (in contrary to simulation A), i.e. are outlying also in terms of the values of the regressors.

The results of simulation B are presented in the bottom part of Table 3. The results again reveal the least squares to be very unsuitable due to the contamination, while the MM-estimator turns out to be the most successful in terms of MSE, leaving behind the best versions of the LWS, i.e. LWS2 and LWS3. Estimates of  $\sigma^2$  obtained by MM- and S-estimators have the

highest precision, while the (inefficient) S-estimates have a very low accuracy. Again, the resulting values of MSE are improved for  $n = 60$  compared to  $n = 30$ . However, we can say that these most suitable versions of the LWS (LWS2 and LWS3) are able to come closer to those of MM-estimators and also to overcome the LTS especially for  $n = 30$ , while they do not become so much improved for  $n = 60$ .

## 5. CONCLUSIONS

This paper is devoted to estimating the variance  $\sigma^2$  of random errors for several highly robust regression estimators. The paper presents an overview of consistent estimates of  $\sigma^2$  corresponding to highly robust regression estimators and recalls possible applications of the estimated  $\sigma^2$ . We are particularly interested in the LWS estimator, for which we formulate a possible approximation (27) applicable to real data. We performed two numerical studies on simulated datasets,



while the best results in both of them are obtained with MM-estimators, which are also theoretically known to combine high efficiency with high robustness. The results confirm that meaningfulness of the proposal (27). Particularly, the LWS estimator (and especially with weights functions suppressing outliers) performs reasonably well, especially for  $n = 30$  compared to  $n = 60$ . This is an interesting novel argument in favor of the LWS estimator, which has some other appealing properties recalled in Section 2.7. The LTS and S-estimators, which are theoretically known to be rather inefficient [19], perform much weaker compared to MM-estimators or LWS. Such conclusions go much beyond the recommendations of an automated meta-learning approach of [25] for finding the most suitable robust regression estimator for a given dataset.

#### ACKNOWLEDGEMENT

The work was supported by the Czech Science Foundation projects 19-05704S and 17-07384S. The authors are grateful to Lubomír Soukup for discussion.

#### REFERENCES

- [1] Hald A. (2006). *A history of parametric statistical inference from Bernoulli to Fisher, 1713 to 1935*. New York: Springer.
- [2] Dietrich, C.F. (1991). *Uncertainty, calibration and probability. The statistics of scientific and industrial measurement*. 2nd edn. Boca Raton: Taylor & Francis.
- [3] Neykov, N., Filzmoser, P., Dimova, R., & Neytchev, P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, 52, 299–308.
- [4] White, G.H. (2008). Basics of estimating measurement uncertainty. *Clinical Biochemist Reviews*, 29, S53–S60.
- [5] Hekimoglu, S., Erenoglu, R.C., & Kalina, J. (2009). Outlier detection by means of robust regression estimators for use in engineering science. *Journal of Zhejiang University Science A*, 10, 909–921.
- [6] Luukinen J.M., Aalto D., Malinen, J., Niikuni, N., Saunavaara, J., Jääsaari, P., Ojalampi, A., Parkkola, R., Soukka, T., & Happonen, R.P. (2018). A novel marker based method to teeth alignment in MRI. *Measurement Science Review*, 18, 79–85.
- [7] Kumar, N., Hoque, M.A., Shahjaman, M., Islam, S.M.S., Mollah, M.N.H. (2017). Metabolomic biomarker identification in presence of outliers and missing values. *BioMed Research International*, 2017, Article 2437608.
- [8] Kalina, J. (2018). A robust pre-processing of BeadChip microarray images. *Biocybernetics and Biomedical Engineering*, 38, 556–563.
- [9] Jurečková, J., Picek, J., & Schindler, M. (2019). *Robust statistical methods with R*. 2nd edn. Boca Raton: CRC Press.
- [10] Rousseeuw, P.J. & Leroy, A.M. (1987). *Robust regression and outlier detection*. New York: Wiley.
- [11] Kalina, J. (2012). Implicitly weighted methods in robust image analysis. *Journal of Mathematical Imaging and Vision*, 44, 449–462.
- [12] Vříšek, J.Á. (2011). Consistency of the least weighted squares under heteroscedasticity. *Kybernetika*, 47, 179–206.
- [13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., & Stahel, W.A. (1986). *Robust statistics. The approach based on influence functions*. New York: Wiley.
- [14] Ranganai, E. (2016). On studentized residuals in the quantile regression framework. *SpringerPlus*, 5, Article 1231.
- [15] Davies, P.L. & Kovac A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics*, 29, 1–65.
- [16] Rousseeuw, P.J. & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1, 73–79.
- [17] Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *Annals of Statistics*, 18, 1651–1675.
- [18] Riani, M., Cerioli, A., & Torti, F. (2014). On consistency factors and efficiency of robust S-estimators. *Test*, 23, 356–387.
- [19] Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15, 642–656.
- [20] Vříšek, J.Á. (2006). The least trimmed squares. Part I: Consistency. *Kybernetika*, 42, 1–36.
- [21] Roelant, E., Van Aelst, S., & Willems, G. (2009). The minimum weighted covariance determinant estimator. *Metrika*, 70, 177–204.
- [22] Kalina, J. (2019). Common multivariate estimators of location and scatter capture the symmetry of the underlying distribution. *Communications in Statistics—Simulation and Computation*. In press.
- [23] Vříšek, J.Á. (2010). Robust error-term-scale estimate. *IMS Collections*, 7, 254–267.
- [24] Rousseeuw, P.J. & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12, 29–45.
- [25] Kalina, J. & Neoral, A. (2019). A robustified metalearning procedure for regression estimators. In *Conference Proceedings, The 13th International Days of Statistics and Economics MSED 2019*. Slaný: Melandrium, 617–726.
- [26] R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.

Received September 10, 2019.

Accepted January 25, 2020.