scientific reports

Check for updates

OPEN

Robust model selection using the out-of-bag bootstrap in linear regression

Fazli Rabbi[1], Alamgir Khalil[1], Ilyas Khan[2]✉, Muqrin A. Almuqrin[3]✉, Umair Khalil[4] & Mulugeta Andualem[5]✉

Outlying observations have a large influence on the linear model selection process. In this article, we present a novel approach to robust model selection in linear regression to accommodate the situations where outliers are present in the data. The model selection criterion is based on two components, the robust conditional expected prediction loss, and a robust goodness-of-fit with a penalty term. We estimate the conditional expected prediction loss by using the out-of-bag stratified bootstrap approach. In the presence of outliers, the stratified bootstrap ensures that we obtain bootstrap samples that are similar to the original sample data. Furthermore, to control the undue effect of outliers, we use the robust MM-estimator and a bounded loss function in the proposed criterion. Specifically, we observe that instead of minimizing the penalized loss function or the conditional expected prediction loss separately, it is better to minimize them simultaneously. The simulation and real-data based studies confirm the consistent and satisfactory behavior of our bootstrap model selection procedure in the presence of response outliers and covariate outliers.

A variety of models are used in statistical modeling. Often the focus is to identify the single best model, which describes the data well while being parsimonious. The model selection procedure involves fitting a set of competing models and then selecting the best model by comparing the values of their goodness-of-fit statistics, their prediction loss, or both of these two. Several studies on model selection procedures have concluded that these methods depend on maximum likelihood-type or least squares approaches[1–6] and are possibly affected by the presence of outlying observations in the data. Robust model selection methods aim to work well in situations when some of the observations are outliers and/or the error distribution is not normal. Several robust model selection procedures have been proposed in the literature. To cope with these problems in model selection, different approaches are proposed. Some of them are based on robust modifications of well-known standard criteria such as Akaike information criterion or Mallows' C_p criterion, or on various resampling techniques, like bootstrap or cross-validation[7–18]. The main objective of this research work is to propose a modified version of[19] for model selection in the presence of outliers.

Suppose that we have a column vector of $\boldsymbol{n}$ responses $Y = (\mathbf{y_1}, \mathbf{y_2}, ..., \mathbf{y_n})^{\mathbf{T}}$, and X is an $\mathbf{n} \times \mathbf{p}$ design matrix. Let $\boldsymbol{\alpha}$ denote any subset of size $\mathbf{p_\alpha}$ from $\{\mathbf{1}, \mathbf{2}, \ldots, \mathbf{p}\}$, and let $\mathbf{X_\alpha}$ is an $\mathbf{n} \times \mathbf{p_\alpha}$ matrix. Let $\mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}$ denote the $\mathbf{i}$th row vector of the design matrix $\mathbf{X_\alpha}$. Then the linear regression model corresponding to model α is given by

$$\mathbf{y_i} = \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\boldsymbol{\beta}_\alpha + \varepsilon_{\alpha\mathbf{i}}, \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, \ldots, \mathbf{n} \tag{1}$$

where X_α and $\boldsymbol{\varepsilon}_{\alpha i} = (\varepsilon_{\alpha\mathbf{1}}, \varepsilon_{\alpha\mathbf{2}}, ..., \varepsilon_{\alpha\mathbf{n}})^{\mathbf{T}}$ are independent, and the errors $\boldsymbol{\varepsilon}_{\alpha\mathbf{i}}$ are assumed to have location 0 and scale 1, $\boldsymbol{\beta}_\alpha$ is an unknown $\mathbf{p}_\alpha$-vector of regression coefficients. Let A represent a collection of candidate models. The interest here is to select a model $\boldsymbol{\alpha}$ from A based on the specified properties of the corresponding fit. To fit the linear regression model, the MM-estimator of[20] was adopted, which combines excellent robustness properties along with high efficiency in the absence of outliers in the data. In model selection, three aspects are generally considered i.e., specifying an estimator, fitting models by using the specified estimator and finally, the fitted models are compared. Furthermore, for each of the models $\boldsymbol{\alpha}$ the approach can be extended by considering

[1]Department of Statistics, University of Peshawar, Peshawar, Pakistan. [2]Department of Mathematics, College of Science Al-Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia. [3]Department of Mathematics, College of Science in Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia. [4]Department of Statistics Abdul Wali, Khan University, Mardan, Pakistan. [5]Department of Mathematics, Bonga University, Bonga, Ethiopia. ✉email: i.said@mu.edu.sa; m.almuqrin@mu.edu.sa; mulugetaandualem4@gmail.com

various types of estimators like LS-estimator, M-estimator, and MM-estimator etc. The models are indexed by $\boldsymbol{\alpha} \in$ A and estimate $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$ by the estimator $\hat{\beta}_{\boldsymbol{\alpha}}$.

The following two minimal requirements for a good model are discussed by[19]:

(i) it has the capability to fit the sample data $\boldsymbol{y}$ and $\boldsymbol{X}$ reasonably well, and
(ii) it has the ability to predict future observations with great accuracy.

The ability of a model to fit the sample data $\boldsymbol{y}$ and $\boldsymbol{X}$ is measured by applying a penalized loss function and the expected prediction loss is used to measure the ability to predict future observations. It has been found in the literature that bootstrapping a robust estimator encounters some difficulties in the presence of outliers. For robust regression, an $\boldsymbol{m}$-out-of-$\boldsymbol{n}$ paired bootstrap approach is proposed by[12]. Their study findings revealed that implementing the bootstrap procedure directly to a data set containing outliers, generally, fails due to two reasons: (1) The use of $\rho(x) = x^2$, which is non-robust against outliers, and (2) the bootstrap samples may contain a high proportion of outliers as compared to the original data set. Müller and Welsh[19] addressed both of these issues by using stratified bootstrap with appropriate choice of $\rho(.)$ in the presence of outliers. Their approach ensured that one can obtain bootstrap samples that are similar to the sample data. According to their approach, bootstrap samples are constructed in such a manner that the residuals distribution for each bootstrap sample will reflect the relatively same residuals distribution observed in the original data. Their strategy seems to solve the issue well in practice.

Our objective in this paper is to pursue the investigation in[19] and make some refinements, by utilizing the concept of out-of-bag bootstrap to develop a robust model selection criterion which deals with outliers and heavy tailed error distributions. The out-of-bag (OOB) observations are those which are not part of the bootstrap sample. These OOB observations can be used for estimating the prediction error, yielding the so-called OOB error. This type of error is often claimed to be an unbiased estimator for the true error rate[21,22].

The rest of the paper is organized as follows: We discuss the existing robust model selection criteria in "Robust model selection criteria" section. Section "The proposed robust model selection criterion" describes a proposed robust model selection criterion. We show the performance of our modified robust criterion via simulation studies in "Simulation studies" section. We present a data examplein "Data example (Stack loss data)" section and conclude with a short discussion in "Conclusion" section.

Robust model selection criteria

In this section, we discuss the existing robust model selection criteria based on robust expected prediction loss. Consider a vector of $\boldsymbol{n}$ responses $\mathbf{y_i} = (\mathbf{y_1}, \mathbf{y_2},..., \mathbf{y_n})^{\mathbf{T}}$ and the design matrix $\mathbf{X} = (\mathbf{x_1}, \mathbf{x_2},..., \mathbf{x_n})^{\mathbf{T}}$. The conditional expected prediction loss of a model α for a given non- negative loss function $\rho(.)$ is calculated by

$$\mathbf{M^{PE}}(\alpha) = \frac{\boldsymbol{\sigma^2}}{\mathbf{n}} \mathbf{E}\left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{z_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}_{\boldsymbol{\alpha}}}{\boldsymbol{\sigma}}\right\} \Big| \mathbf{y}, \mathbf{X}\right] \tag{2}$$

where $\hat{\beta}_{\boldsymbol{\alpha}}$ is the estimator of $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$, $\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2},..., \mathbf{z_n})^{\mathbf{T}}$ is a vector of future responses at X, independent of y, and σ is the measure of spread for a given data. Initially, this type of prediction loss was introduced by[5] as a model selection criterion by using a loss function $\rho(x) = \frac{x^2}{2}$ in the least squares regression.

To select a model α from a set A,[19] proposed the following criterion function

$$\mathbf{M_n}(\alpha) = \frac{\boldsymbol{\sigma^2}}{\mathbf{n}}\left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}_{\boldsymbol{\alpha}}}{\boldsymbol{\sigma}}\right\} + \delta(\mathbf{n})\mathbf{p}_{\boldsymbol{\alpha}} + \mathbf{E}\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{z_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}_{\boldsymbol{\alpha}}}{\boldsymbol{\sigma}}\right\} \Big| \mathbf{y}, \mathbf{X}\right] \tag{3}$$

Following[5,19] estimated the unknown distribution of the data by usingan m-out-of-n stratified bootstrap procedure, whereas the penalized in-sample term in (3) is estimated directly. The estimated selection criteria functions are given by

$$\mathbf{M_{m,n}^{PE}}(\alpha) = \frac{\hat{\sigma}^2}{\mathbf{n}} \mathbf{E}_* \left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}^*_{\alpha,\mathbf{m}}}{\hat{\sigma}}\right\}\right] \tag{4}$$

$$\mathbf{M_{m,n}^{PPE}}(\alpha) = \frac{\hat{\sigma}^2}{\mathbf{n}}\left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}_{\boldsymbol{\alpha}}}{\hat{\sigma}}\right\} + \delta(\mathbf{n})\mathbf{p}_{\boldsymbol{\alpha}} + \mathbf{E}_* \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}_{\alpha\mathbf{i}}^{\mathbf{T}}\hat{\beta}^*_{\alpha,\mathbf{m}}}{\hat{\sigma}}\right\}\right] \tag{5}$$

where $\hat{\beta}^*_{\alpha,\mathbf{m}}$ is the bootstrap estimate of $\hat{\beta}_{\boldsymbol{\alpha}}$, E_* denotes expectation with respect to the bootstrap distribution and $\boldsymbol{m}$ is the number of distinct observations in the bootstrap sample which satisfies the conditions given by

$$m \to \infty \text{ and } \frac{m}{\sqrt{n}} \to 0 \text{ as } n \to \infty.$$

The criterion function in (4) was modified by[18] using the following steps:

(i) calculate and order the residuals,
(ii) set the number of strata S at between 3 and 8 depending on the sample size n,
(iii) set stratum boundaries of the residuals,

(iv) allocate observations into different strata so that observations in the extreme tail are kept in lower or upper tail strata and other strata comprising the remaining observations,
(v) sample rows of (y,X) independently with replacement from each stratum so that total bootstrap sample of size is m ($\leq n$),
(vi) construct the estimator $\hat{\beta}^*_{\alpha,m}$ from data obtained in step (v),
(vii) calculate the criterion function $\mathbf{M}^{\mathbf{PE}*}_{\mathbf{m,n}}(\alpha)$ from n-m observations i.e., $\boldsymbol{m}$ observations used to obtain $\hat{\beta}^*_{\alpha,\mathbf{m}}$ are not included when calculating $\mathbf{M}^{\mathbf{PE}*}_{\mathbf{m,n}}(\alpha)$,
(viii) repeat the steps (vi) and (vii) $\boldsymbol{K}$ independent times and then estimate the modified robust expected prediction loss by

$$\mathbf{M}^{\mathbf{PE}*}_{\mathbf{m,n}}(\alpha) = \frac{\hat{\sigma}^2}{\mathbf{n}}\left[\mathbf{E}_* \sum_{\mathbf{i=1}}^{\mathbf{n-m}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_{i[-m]}} - \mathbf{x}^{\mathbf{T}}_{\alpha\mathbf{i[-m]}}\hat{\beta}^*_{\alpha,\mathbf{m}}}{\hat{\sigma}}\right\}\right] \tag{6}$$

where $\hat{\beta}^*_{\alpha,\mathbf{m}}$ is the bootstrap estimate of $\hat{\beta}_{\boldsymbol{\alpha}}$, E_* denotes expectation with respect to the bootstrap distribution and $\boldsymbol{m}$ is the number of distinct observations in the bootstrap sample used to obtain $\hat{\beta}^*_{\alpha,m}$ and [-m] means that the $\boldsymbol{m}$ observations are excluded from total observations when calculating $M^{PE*}_{m,n}(\alpha)$. Here the focus is on the model $\alpha \in A$ that minimizes $M^{PE}_{m,n}(\alpha)$, $M^{PPE}_{m,n}(\alpha)$ or $\mathbf{M}^{\mathbf{PE}*}_{\mathbf{m,n}}(\alpha)$ i.e.

$$\overline{\alpha}_{\mathbf{m,n}} = \underset{\alpha\in \mathrm{A}}{\mathbf{arg\ min}}\, \mathbf{M}^{\mathbf{PE}}_{\mathbf{m,n}}(\alpha) \tag{7}$$

$$\tilde{\alpha}_{\mathbf{m,n}} = \underset{\alpha\in \mathrm{A}}{\mathbf{arg\ min}}\, \mathbf{M}^{\mathbf{PPE}}_{\mathbf{m,n}}(\alpha) \tag{8}$$

$$\hat{\alpha}_{\mathbf{m,n}} = \underset{\alpha\in \mathrm{A}}{\arg\min}\, \mathbf{M}^{\mathbf{PE}*}_{\mathbf{m,n}}(\alpha) \tag{9}$$

The proposed robust model selection criterion

In this section, we propose a robust model selection procedure based on two components, a robust penalized loss function, and a modified robust expected prediction loss.

We estimate the penalized in-sample term in the criterion function by

$$\mathbf{M}^{\mathbf{P}}_{\mathbf{n}}(\alpha) = \frac{\hat{\sigma}^2}{\mathbf{n}}\left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}^{\mathbf{T}}_{\alpha\mathbf{i}}\widehat{\boldsymbol{\beta}_{\alpha}}}{\hat{\sigma}}\right\} + \delta(\mathbf{n})\mathbf{p}_{\boldsymbol{\alpha}}\right] \tag{10}$$

where $\delta(\mathbf{n})$ denotes a function of sample size $\boldsymbol{n}$. The two restrictions on function $\delta(\mathbf{n})$ are that $\delta(\mathbf{n}) \to \infty$ and $\frac{\delta(\mathbf{n})}{\mathbf{n}} \to 0$ as $n \to \infty$. The two restrictions on δ (n) are imposed to penalize complexity, which expresses a preference for smaller and simpler models. These conditions are satisfied by the choice $\delta(\mathbf{n}) = \log(n)$. We combine (6) and (10) to estimate the robust criterion function by

$$\mathbf{M}^{\mathbf{PPE}*}_{\mathbf{m,n}}(\alpha) = \frac{\hat{\sigma}^2}{\mathbf{n}}\left[\sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_i} - \mathbf{x}^{\mathbf{T}}_{\alpha\mathbf{i}}\hat{\beta}_{\alpha}}{\hat{\sigma}}\right\} + \delta(\mathbf{n})\mathbf{p}_{\boldsymbol{\alpha}} + \mathbf{E}_* \sum_{\mathbf{i=1}}^{\mathbf{n-m}} \boldsymbol{\rho}\left\{\frac{\mathbf{y_{i[-m]}} - \mathbf{x}^{\mathbf{T}}_{\alpha\mathbf{i[-m]}}\hat{\beta}^*_{\alpha,\mathbf{m}}}{\hat{\sigma}}\right\}\right] \tag{11}$$

where $\hat{\beta}^*_{\alpha,\mathbf{m}}$ is the bootstrap estimate of $\hat{\beta}_{\boldsymbol{\alpha}}$, E_* denotes expectation with respect to the bootstrap distribution and $\boldsymbol{m}$ is the number of distinct observations in the bootstrap sample. An important issue is "how large should the number of bootstrap replications K in our proposed criterion. There is no hard and fast rule for the number of bootstrap replications. However, for estimation of standard error, it is usually in the range of 25–250[23]. The first term in criterion function (11) measures the relationship between the observed sample data $\boldsymbol{y}$ and $\boldsymbol{X}$; the second term penalizes complexity (i.e., preference for smaller models), while the ability to predict future observations is measured by the last term. To use (11), we have to specify $\rho(.)$ and σ. The robustness viewpoint is adopted for the purpose of fitting the core of the data and predicting core observations, rather than fitting and predicting the tails having atypical observations. So a bounded ρ function is selected. Here, trimming is preferred, so that for sufficiently large |x| the $\rho(x)$ function is constant.

As in[11,14,18,19], the simplest ρ function is given by

$$\rho(\mathbf{x}) = \mathbf{min}(\mathbf{x^2}, \mathbf{b^2}) \tag{12}$$

which is quadratic near the origin and becomes constant when it is away from the origin. As in[19], we use $\boldsymbol{b}$ = 2. To measure spread σ, we use the full model α_f, because for residuals spread, a large model can produce a valid measure. For simplicity, we measure σ by the median absolute deviation (MAD) from the median multiplied by 1.483 and is given by

$$\hat{\sigma} = 1.483 \underset{1\leq i\leq n}{med}\left|e_i - \underset{1\leq j\leq n}{med}(e_j)\right|$$

where $\mathbf{e_i} = \mathbf{y_i} - \mathbf{x}^{\mathbf{T}}_{\boldsymbol{\alpha}_\mathbf{f}\mathbf{i}}\hat{\beta}_{\boldsymbol{\alpha}_\mathbf{f}}$, $\mathbf{e_j} = \mathbf{y_j} - \mathbf{}^{\mathbf{T}}_{\boldsymbol{\alpha}_\mathbf{f}\mathbf{j}}\hat{\beta}_{\boldsymbol{\alpha}_\mathbf{f}}$ and $\hat{\beta}_{\alpha}$ is the estimator for $\boldsymbol{\beta}_{\boldsymbol{\alpha}}$.

Among the models being considered, we select a model $\alpha \in A$ that minimizes $\mathbf{M}^{\mathbf{PPE}*}_{\mathbf{m,n}}(\alpha)$, i.e.

True β^T	Model	m = 15				m = 20				m = 25				
		$\overline{\alpha}_{15,40}$	$\hat{\alpha}_{15,40}$	$\tilde{\alpha}_{15,40}$	$\hat{\alpha}^*_{15,40}$	$\overline{\alpha}_{20,40}$	$\hat{\alpha}_{20,40}$	$\tilde{\alpha}_{20,40}$	$\hat{\alpha}^*_{20,40}$	$\overline{\alpha}_{25,40}$	$\hat{\alpha}_{25,40}$	$\tilde{\alpha}_{25,40}$	$\hat{\alpha}^*_{25,40}$	$\breve{\alpha}_{40}$
(2,0,0,4,0)	**1,4***	**0.943**	**0.972**	**0.928**	**0.958**	**0.875**	**0.943**	**0.876**	**0.925**	**0.770**	**0.903**	**0.814**	**0.896**	**0.835**
	1,4,5	0.010	0.006	0.013	0.009	0.024	0.014	0.030	0.018	0.042	0.023	0.038	0.027	0.046
	1,3,4	0.019	0.010	0.021	0.011	0.050	0.014	0.044	0.023	0.100	0.038	0.075	0.039	0.046
	1,2,4	0.028	0.012	0.036	0.022	0.046	0.029	0.046	0.032	0.069	0.034	0.060	0.036	0.057
	1,3,4,5	0.000	0.000	0.001	0.000	0.001	0.000	0.001	0.001	0.005	0.001	0.004	0.001	0.004
	1,2,4,5	0.000	0.000	0.000	0.000	0.001	0.000	0.001	0.001	0.004	0.000	0.005	0.000	0.009
	1,2,3,4	0.000	0.000	0.001	0.000	0.003	0.000	0.002	0.000	0.008	0.001	0.004	0.001	0.003
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.000
(2,0,0,4,8)	**1,4,5***	**0.965**	**0.978**	**0.955**	**0.972**	**0.907**	**0.948**	**0.914**	**0.939**	**0.838**	**0.910**	**0.866**	**0.914**	**0.877**
	1,3,4,5	0.013	0.007	0.021	0.010	0.043	0.019	0.039	0.025	0.077	0.041	0.063	0.040	0.054
	1,2,4,5	0.022	0.015	0.024	0.018	0.048	0.031	0.045	0.035	0.071	0.045	0.060	0.043	0.063
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.002	0.002	0.002	0.001	0.014	0.004	0.011	0.003	0.006
(2,9,0,4,8)	1,4,5	0.013	0.022	0.005	0.007	0.002	0.012	0.001	0.003	0.000	0.000	0.000	0.003	0.000
	1,2,5	0.001	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,3,4,5	0.001	0.003	0.001	0.001	0.004	0.005	0.001	0.003	0.004	0.005	0.001	0.003	0.001
	1,2,4,5*	**0.976**	**0.966**	**0.979**	**0.984**	**0.956**	**0.966**	**0.958**	**0.971**	**0.916**	**0.942**	**0.930**	**0.949**	**0.934**
	1,2,3,4,5	0.009	0.007	0.015	0.008	0.038	0.017	0.040	0.023	0.080	0.044	0.069	0.045	0.065
(2,9,6,4,8)	1,3,4,5	0.071	0.097	0.027	0.049	0.015	0.032	0.006	0.012	0.008	0.018	0.002	0.008	0.001
	1,2,4,5	0.010	0.020	0.006	0.009	0.000	0.003	0.000	0.000	0.001	0.003	0.000	0.001	0.000
	1,2,3,5	0.011	0.014	0.000	0.003	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3,4,5*	**0.908**	**0.869**	**0.967**	**0.939**	**0.985**	**0.964**	**0.994**	**0.988**	**0.991**	**0.979**	**0.998**	**0.991**	**0.999**

Table 1. Estimated selection probabilities of $\overline{\alpha}_{m,n}$, $\tilde{\alpha}_{m,n}$, $\hat{\alpha}_{\mathbf{m,n}}$ and $\hat{\alpha}^*_{m,n}$ based on the least squares estimator and $\rho(x^2)$. The (*) indicates the optimal model. Significant values are in bold.

$$\hat{\alpha}^*_{\mathbf{m,n}} = \arg\min_{\alpha \in \mathrm{A}} \mathbf{M}^{\mathbf{PPE}*}_{\mathbf{m,n}}(\alpha) \tag{13}$$

The optimal $\boldsymbol{m}$ depends on the true model. As in[14,19], one should use $\boldsymbol{n/4} \leq \boldsymbol{m} \leq \boldsymbol{n/2}$ for moderate $\boldsymbol{n}$ ($50 \leq \boldsymbol{n} \leq 200$). If $\boldsymbol{n}$ is small, $\boldsymbol{m}$ is small and the parameter estimators do not converge for some bootstrap samples; but if $\boldsymbol{n}$ is large, $\boldsymbol{m}$ may be smaller than a fourth of $\boldsymbol{n}$. Choosing the number of strata $\mathbf{S}$ at between 3 and 8, depending on the sample size $\boldsymbol{n}$[24].

The penalized loss function in the proposed criterion function, given in (10), is just like a robust version of AIC proposed by[25,26]. But the main difference is due to the ρ function and the estimator in our criterion. The penalized in-sample term in (11) is similar to the robust version of[3]. Furthermore, for $\rho(x) = (x^2)$, the penalized in-sample term was reduced to[3] criterion.

Simulation studies

To assess and compare the finite sample performance of our proposed method with the existent model selection methods, we carried out two simulation studies, that is, one for contamination free dataset in a simulation setting 1 and the other for the contaminated data set in a simulation setting 2.

Simulation setting 1. The finite-sample performance of our proposed criterion is compared with existing model selection procedures via real dataset and simulated data set.

The Gunst and Mason data. To compare the finite sample performance of our proposed method with the existent model selection methods through the real dataset, we use the following regression mode

$$\mathbf{Y_i} = \beta_0\mathbf{X_{i0}} + \beta_1\mathbf{X_{i1}} + \beta_2\mathbf{X_{i2}} + \beta_3\mathbf{X_{i3}} + \beta_4\mathbf{X_{i4}} + \mathbf{u_i}, \quad \mathbf{i} = \mathbf{1}, \mathbf{2}, ..., \mathbf{40}$$

where u_i are iid standard normal errors;X_0 is the column of 1's; and the values of X_1, X_2, X_3 and X_4 are taken from the solid waste data of[27], as in[5,12,13,18,19]. We compare the estimator $\hat{\alpha}^*_{m,n}$ [expressed in (13)], with $\overline{\alpha}_{m,n}$ [expressed in (7)], $\tilde{\alpha}_{m,n}$ [expressed in (8)], $\hat{\alpha}_{\mathbf{m,n}}$ [expressed in (9)] and robust BIC $\breve{\alpha}_n$ [expressed in (14)],

$$\breve{\alpha}_{\mathbf{n}} = \arg\min_{\alpha \in \mathrm{A}} \mathbf{M}^{\mathbf{P}}_{\mathbf{n}}(\alpha) \tag{14}$$

In the zero contamination case, the least squares estimator is used to fit the regression models. The penalty term $\delta(n) = \log(n)$ is used in all simulations. The estimated selection probabilities for $\hat{\alpha}^*_{m,n}$,$\overline{\alpha}_{m,n}$,$\tilde{\alpha}_{m,n}$ and $\breve{\alpha}_n$ based on the LS estimator and $\rho(\mathbf{x}) = \mathbf{x^2}$ are mentioned in Table 1, whereas the estimated selection probabilities based

True β^T	Model	m = 15				m = 20				m = 25				
		$\overline{\alpha}_{15,40}$	$\hat{\alpha}_{15,40}$	$\tilde{\alpha}_{15,40}$	$\hat{\alpha}^*_{15,40}$	$\overline{\alpha}_{20,40}$	$\hat{\alpha}_{20,40}$	$\tilde{\alpha}_{20,40}$	$\hat{\alpha}^*_{20,40}$	$\overline{\alpha}_{25,40}$	$\hat{\alpha}_{25,40}$	$\tilde{\alpha}_{25,40}$	$\hat{\alpha}^*_{25,40}$	$\breve{\alpha}_{40}$
(2,0,0,4,0)	**1,4***	**0.897**	**0.971**	**0.879**	**0.929**	**0.800**	**0.927**	**0.832**	**0.902**	**0.672**	**0.846**	**0.781**	**0.885**	**0.839**
	1,4,5	0.028	0.010	0.028	0.021	0.053	0.021	0.045	0.028	0.077	0.046	0.056	0.032	0.043
	1,3,4	0.029	0.010	0.029	0.018	0.071	0.018	0.057	0.027	0.116	0.054	0.076	0.037	0.047
	1,2,4	0.042	0.009	0.042	0.029	0.060	0.031	0.054	0.039	0.094	0.046	0.065	0.039	0.055
	1,3,4,5	0.001	0.000	0.001	0.001	0.006	0.001	0.004	0.001	0.013	0.003	0.007	0.002	0.004
	1,2,4,5	0.002	0.000	0.002	0.002	0.005	0.002	0.006	0.002	0.011	0.002	0.009	0.004	0.009
	1,2,3,4	0.001	0.000	0.001	0.000	0.005	0.000	0.002	0.001	0.012	0.003	0.006	0.001	0.003
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000
(2,0,0,4,8)	**1,4,5***	**0.934**	**0.981**	**0.918**	**0.957**	**0.860**	**0.951**	**0.875**	**0.933**	**0.781**	**0.893**	**0.835**	**0.909**	**0.878**
	1,3,4,5	0.023	0.010	0.032	0.017	0.065	0.021	0.057	0.029	0.098	0.052	0.075	0.040	0.055
	1,2,4,5	0.043	0.009	0.050	0.026	0.068	0.028	0.061	0.038	0.100	0.049	0.077	0.050	0.061
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.007	0.000	0.007	0.000	0.021	0.006	0.013	0.001	0.006
(2,9,0,4,8)	1,4,5	0.000	0.005	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.001	0.002
	1,3,4,5	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.000	0.001	0.001	0.001	0.001
	1,2,4,5*	**0.979**	**0.989**	**0.972**	**0982**	**0.932**	**0.972**	**0.933**	**0.968**	**0.875**	**0.939**	**0.910**	**0.951**	**0.933**
	1,2,3,4,5	0.021	0.006	0.027	0.017	0.068	0.027	0.066	0.031	0.125	0.059	0.089	0.047	0.064
(2,9,6,4,8)	1,3,4,5	0.008	0.036	0.001	0.007	0.002	0.010	0.001	0.005	0.000	0.005	0.001	0.001	0.001
	1,2,4,5	0.001	0.004	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
	1,2,3,5	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3,4,5*	**0.991**	**0.959**	**0.999**	**0.992**	**0.998**	**0.989**	**0.999**	**0.995**	**1.000**	**0.994**	**0.999**	**0.999**	**0.999**

Table 2. Estimated selection probabilities of $\overline{\alpha}_{m,n}$, $\tilde{\alpha}_{m,n}$, $\hat{\alpha}_{\mathbf{m,n}}$ and $\hat{\alpha}^*_{m,n}$ based on the least squares estimator and $\rho(x) = \min(x^2, b^2)$. The results are based on 1000 Monte Carlo simulations and K = 100 bootstrap replications. Significant values are in bold.

on the LS estimator and $\rho(\mathbf{x}) = \mathbf{min}(\mathbf{x^2}, \mathbf{b^2})$ are given in Table 2. The results given in Tables 1 and 2 are based on $\boldsymbol{L} = 1000$ simulations and $\boldsymbol{K} = 100$ bootstrap samples for $\boldsymbol{m} = 15, 20, 25$.

The simulation resultspresented in Tables 1 and 2 are summarized as follows:

- The performance of the modified model selection procedure using the least squares estimator is comparable to the existing methods $\overline{\alpha}_{m,n}$, $\tilde{\alpha}_{m,n}$, $\hat{\alpha}_{\mathbf{m,n}}$ and the BIC($\breve{\alpha}_n$).
- The proposed selection criterion outperforms the existent procedures in both cases, i.e., either using the squared loss function $\rho(x) = x^2$ or the robust loss function $\rho(x) = \min(x^2, b^2)$.
- For the full model, if bootstrap sample size $\boldsymbol{m}$ increases, the estimated selection probabilities also increase. For example, in the case of $\boldsymbol{m} = 15$, the correct percent is 93.9%, whereas, for $\boldsymbol{m} = 25$, the correct percent is 99.1%.
- Moreover, model selection based on the robust loss function is superior to the squared loss function. For instance, in the case when the optimal model has all the predictors, then the modified model selection procedure $\hat{\alpha}^*_{15,40}$ using the squared loss function selects the optimal model 93.9% of the time, which is less than the 99.2% obtained by using a robust loss function.
- Furthermore, the modified selection criterion $\hat{\alpha}^*_{m,n}$ is less dependent on bootstrap sample size $\boldsymbol{m}$ as compared to the existent criteria $\overline{\alpha}_{m,n}$ and $\tilde{\alpha}_{m,n}$.

Simulated data and model selection consistency. To show model selection consistency and performance of the proposed criterion on simulated data, the following regression model with $\boldsymbol{p} = 5$ is considered.

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, 2, ..., n \tag{15}$$

where ε_i is generated from standard normal distribution, the regression variables are generated from N(0, 1), and added an intercept column of 1's to produce the design matrix X. To generate the response variable y_i, we use Eq. (15).

The true data generating models are:

- $\beta_1 = (1, 0, 0, 1, 0)$, i.e. the model had one nonzero variable,
- $\beta_2 = (1, 0, 0, 1, 1)$, i.e. the model had two nonzero variables,
- $\beta_3 = (1, 1, 0, 1, 1)$, i.e. the model had three nonzero variables and

The estimated selection probabilities for $\hat{\alpha}^*_{m,n}$, $\tilde{\alpha}_{m,n}$, $\hat{\alpha}_{m,n}$ and $\overline{\alpha}_{m,n}$ are calculated for $\boldsymbol{m} = 24$ and $n = 40,80,120,160$, based on $\boldsymbol{L} = 1000$ simulations with bootstrap replications of $\boldsymbol{K} = 50$ and are tabulated in Table 3.

True	Model	$m = 24, n = 40$				$m = 24, n = 80$				$m = 24, n = 120$				$m = 24, n = 160$			
		$\overline{\alpha}_{24,40}$	$\hat{\alpha}_{24,40}$	$\tilde{\alpha}_{24,40}$	$\hat{\alpha}^*_{24,40}$	$\overline{\alpha}_{24,80}$	$\hat{\alpha}_{24,80}$	$\tilde{\alpha}_{24,80}$	$\hat{\alpha}^*_{24,80}$	$\overline{\alpha}_{24,120}$	$\hat{\alpha}_{24,120}$	$\tilde{\alpha}_{24,120}$	$\hat{\alpha}^*_{24,120}$	$\overline{\alpha}_{24,160}$	$\hat{\alpha}_{24,160}$	$\tilde{\alpha}_{24,160}$	$\hat{\alpha}^*_{24,160}$
(1,0,0,1,0)	1	0.000	0.002	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,4*	**0.405**	**0.731**	**0.672**	**0.785**	**0.752**	**0.886**	**0.851**	**0.897**	**0.893**	**0.953**	**0.923**	**0.948**	**0.953**	**0.978**	**0.954**	**0.969**
	1,4,5	0.158	0.088	0.104	0.071	0.065	0.034	0.039	0.030	0.038	0.016	0.031	0.017	0.012	0.007	0.012	0.010
	1,3,4	0.132	0.073	0.080	0.063	0.084	0.036	0.052	0.032	0.036	0.016	0.026	0.018	0.017	0.008	0.015	0.010
	1,2,4	0.172	0.087	0.102	0.069	0.071	0.040	0.046	0.036	0.029	0.014	0.018	0.015	0.016	0.007	0.019	0.011
	1,3,4,5	0.041	0.006	0.016	0.004	0.003	0.001	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,4,5	0.048	0.008	0.014	0.004	0.015	0.002	0.009	0.003	0.002	0.000	0.001	0.001	0.001	0.000	0.000	0.000
	1,2,3,4	0.025	0.004	0.010	0.002	0.010	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000
	1,2,3,4,5	0.019	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
(1,0,0,1,1)	1,5	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,4	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,4,5*	**0.566**	**0.821**	**0.784**	**0.865**	**0.828**	**0.939**	**0.896**	**0.938**	**0.930**	**0.971**	**0.950**	**0.969**	**0.975**	**0.990**	**0.968**	**0.984**
	1,3,4,5	0.175	0.080	0.092	0.064	0.079	0.025	0.049	0.026	0.035	0.014	0.027	0.016	0.011	0.007	0.016	0.007
	1,2,4,5	0.206	0.093	0.114	0.066	0.085	0.035	0.053	0.036	0.033	0.014	0.022	0.014	0.013	0.003	0.016	0.009
	1,2,3,4,5	0.053	0.004	0.010	0.003	0.008	0.001	0.002	0.000	0.002	0.001	0.001	0.001	0.001	0.000	0.000	0.000
(1,1,0,1,1)	**1,2,4,5***	**0.779**	**0.921**	**0.894**	**0.938**	**0.922**	**0.971**	**0.953**	**0.973**	**0.965**	**0.983**	**0.975**	**0.982**	**0.991**	**0.994**	**0.987**	**0.991**
	1,2,3,4,5	0.221	0.079	0.106	0.062	0.078	0.029	0.047	0.027	0.035	0.017	0.025	0.018	0.009	0.006	0.013	0.009

Table 3. Selection probabilities of $\hat{\alpha}^*_{m,n}$, $\tilde{\alpha}_{m,n}$,$\hat{\alpha}_{m,n}$ and $\overline{\alpha}_{m,n}$ based on LS-estimator and $\rho(x) = \min(x^2, b^2)$. The results are based on $L = 1000$ MC simulations and $K = 50$ bootstrap replications. Significant values are in bold.

From the simulation results presented in Table 3, we see that our proposed criterion is comparatively consistent procedure for model selection in linear regression problems.

Simulation setting 2. *Simulated data from uniform distribution.* In this subsection, the finite-sample performance of our modified criterion is compared with existing model selection procedures in the presence of outliers. The sample data is generated from the following model

$$y_i = 2 + 2x_{i1} + 0x_{i2} + \in_i \quad i = 1, 2, ..., 64$$

where the design matrix X has columns generated as uniform on [− 1, 1]. The following six different error distributions are considered:

i. $\in_1$ is [3/8] outliers (i.e., [5/8]from a standard normal and [3/8] from a normal with $\mu = 30 - 2 - 2x_1$ and $\sigma = 1$);
ii. $\in_2$ is [1/4] outliers (i.e., [3/4] from a standard normal and [1/4] from a normal with $\mu = 30 - 2 - 2x_1$ and $\sigma = 1$);
iii. $\in_3$ is [1/8] outliers (i.e., [7/8] from a standard normal and [1/8] from a normal with $\mu = 30 - 2 - 2x_1$ and $\sigma = 1$);
iv. $\in_4$, the Gaussian distribution with $\mu = 0$ and $\sigma = 1$;
v. $\in_5$, the Cauchy distribution;
vi. $\in_6$, the slash distribution(i.e. $\in_6 \sim Z/U$ where $Z \sim N(0,1)$ and $U \sim U(0,1)$)

In Table 4, the following possible models are considered:

- Model (1) means, a model with intercept only;
- Model (1, 2) means a model having intercept and X_1;
- Model (1, 3) means a model having intercept and X_2;
- Model (1, 2, 3) means the full model.

Following[19], the MM-estimator of[20] is used to fit the robust regression models. For this purpose, the rlm ( ) function in **R** is used for estimating the regression parameters. Furthermore, the LS estimates are computed for comparison with MM-estimates. As mentioned by[28], when the proportion of extreme observations in some of the bootstrap samples is higher than in the original sample, then the bootstrap distribution may provide a very poor estimator of the distribution of the MM-estimates. To deal with this numerical instability, we use the stratified bootstrap with equal-sized strata. In this approach, bootstrap samples are constructed so that the distribution of the residuals in each bootstrap sample reflects the one observed in the original data set. The selection probabilities based on $\boldsymbol{L} = 1000$ simulations with bootstrap replications of $\mathbf{K} = 100$ are given in Table 4.

Errors	Model	MM-estimator								LS-estimator			
		Simple bootstrap				Stratified bootstrap				Simple bootstrap			
		$\overline{\alpha}_{24,64}$	$\hat{\alpha}_{24,64}$	$\tilde{\alpha}_{24,64}$	$\hat{\alpha}^{*}_{24,64}$	$\overline{\alpha}^{S8}_{24,64}$	$\hat{\alpha}^{S8}_{24,64}$	$\tilde{\alpha}^{S8}_{24,64}$	$\hat{\alpha}^{*S8}_{24,64}$	$\overline{\alpha}_{24,64}$	$\hat{\alpha}_{24,64}$	$\tilde{\alpha}_{24,64}$	$\hat{\alpha}^{*}_{24,64}$
$\in_1$	1	0.368	0.392	0.267	0.257	0.000	0.000	0.001	0.001	0.756	1.000	1.000	1.000
	1,3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.244	0.000	0.000	0.000
	1,2*	**0.290**	**0.362**	**0.722**	**0.738**	**0.916**	**0.968**	**0.997**	**0.997**	**0.000**	**0.000**	**0.000**	**0.000**
	1,2,3	0.342	0.246	0.011	0.005	0.084	0.032	0.002	0.002	0.000	0.000	0.000	0.000
$\in_2$	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.999	1.000	1.000	1.000
	1,3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2*	**0.935**	**0.968**	**0.983**	**0.986**	**0.886**	**0.958**	**0.978**	**0.984**	**0.001**	**0.000**	**0.000**	**0.000**
	1,2,3	0.065	0.032	0.017	0.014	0.0114	0.042	0.022	0.016	0.000	0.000	0.000	0.000
$\in_3$	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000	1.000	1.000
	1,3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2*	**0.894**	**0.952**	**0.947**	**0.963**	**0.869**	**0.946**	**0.943**	**0.960**	**0.000**	**0.000**	**0.000**	**0.000**
	1,2,3	0.106	0.048	0.053	0.037	0.131	0.054	0.057	0.040	0.000	0.000	0.000	0.000
$\in_4$	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2*	**0.891**	**0.961**	**0.933**	**0.956**	**0.859**	**0.939**	**0.919**	**0.948**	**0.869**	**0.949**	**0.929**	**0.958**
	1,2,3	0.109	0.039	0.067	0.044	0.141	0.061	0.081	0.052	0.131	0.051	0.071	0.042
$\in_5$	1	0.008	0.016	0.012	0.018	0.005	0.013	0.010	0.013	0.770	0.823	0.841	0.866
	1,3	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.000
	1,2*	**0.954**	**0.974**	**0.961**	**0.973**	**0.929**	**0.966**	**0.952**	**0.969**	**0.227**	**0.175**	**0.159**	**0.134**
	1,2,3	0.038	0.010	0.027	0.009	0.065	0.021	0.038	0.018	0.002	0.001	0.000	0.000
$\in_6$	1	0.062	0.139	0.081	0.124	0.042	0.093	0.067	0.099	0.828	0.882	0.909	0.921
	1,3	0.004	0.005	0.004	0.003	0.008	0.005	0.004	0.004	0.005	0.002	0.001	0.001
	1,2*	**0.885**	**0.842**	**0.882**	**0.855**	**0.879**	**0.878**	**0.892**	**0.875**	**0.164**	**0.116**	**0.090**	**0.078**
	1,2,3	0.049	0.014	0.033	0.018	0.071	0.024	0.037	0.022	0.003	0.000	0.000	0.000

Table 4. Estimated selection probabilities of $\overline{\alpha}_{24,64}$, $\hat{\alpha}_{24,64}$, $\tilde{\alpha}_{24,64}$ and $\hat{\alpha}^{*}_{24,64}$ based on MM-estimator and LS-estimator. The outputs are based on 1000 MCsimulations and K = 100 bootstrap replications. The $\rho(x) = \min(x^2, 2^2)$ is used for all selection criteria. Significant values are in bold.

From the simulation results presented in Table 4, it is clear that the modified selection procedure using the robust $\rho(.)$ function and MM-estimator is robust in the presence of highly contaminated data. For example, the percent correct is 73.8% for un-stratified bootstrap, whereas the percent correct is 99.7% for stratified bootstrap under the contaminated normal situation $\in_1$. For all error distributions, the modified robust criterion outperforms the existence criteria. The simulation studies suggest that when errors are non-normal, then using robust regression is superior to using LS, but in the case of normal errors, robust regression is inferior to LS. Furthermore, in the presence of outliers and heavy-tailed error distributions, the modified robust criterion using MM-estimator outperforms LS-estimator by a large margin. For example, under $\in_5$ error distribution, for MM-estimator the percent correct is 96.9%, whereas, the percent correct is 13.4% for LS-estimator. These results demonstrate that the modified robust procedure has good robustness characteristics with contaminated normal and heavy-tailed distributions, whereas the LS procedure performs very poorly in both cases. This clearly proves the lack of robustness of the LS procedure in the presence of outliers and heavy-tailed distributions. An excellent amount of improvement is obtained in the bootstrap model selection procedure by using the combined criterion as observed in the above simulation study.

Modified solid waste data of Gunst and Mason. To evaluate the performance of our proposed robust model selection method, we modified the Gunst and Mason data by planting 10% and 20% outliers. The response vector is generated as

$$\mathbf{Y_i} = \beta_0\mathbf{X_{i0}} + \beta_1\mathbf{X_{i1}} + \beta_2\mathbf{X_{i2}} + \beta_3\mathbf{X_{i3}} + \beta_4\mathbf{X_{i4}} + \varepsilon_i, \quad i = 1, 2, ..., 40$$

where X_0 is the column of 1's; and the values of X_1, X_2, X_3 and X_4 are taken from the solid waste data of[26]. To create high-leverage points, we replace the first four to eight observations of each regressor variable value by 20. The true generating model has two non-zero predictors, i.e. $\boldsymbol{\beta}^{\mathbf{T}} = (\mathbf{2}, \mathbf{0}, \mathbf{0}, \mathbf{4}, \mathbf{8})$ and we choose the following five different error distributions to represent various deviations from normality:

i. $\in_1$ is 10% wild (i.e., 90% from a standard normal and 10% from a normal with, $\mu = 0$ and $\sigma = 0.7$);
ii. $\in_2$ is 20% wild (i.e., 80% from a standard normal and 20% from a normal with ,$\mu = 0$ and $\sigma = 5$) ;
iii. $\in_3$.is t(3) (i.e., t-distribution with 3 degrees of freedom);
iv. $\in_4$, is standard normal;

Errors	True β^T	Model	10% X-outliers				20% X-outliers			
			$\overline{\alpha}_{16,40}$	$\hat{\alpha}_{16,40}$	$\tilde{\alpha}_{16,40}$	$\hat{\alpha}^*_{16,40}$	$\overline{\alpha}_{16,40}$	$\hat{\alpha}_{16,40}$	$\tilde{\alpha}_{16,40}$	$\hat{\alpha}^*_{16,40}$
$\in_1$	**(2,0,0,4,8)**	1,5	0.002	0.004	0.002	0.002	0.008	0.034	0.007	0.018
		1,4,5*	**0.912**	**0.966**	**0.902**	**0.940**	**0.823**	**0.883**	**0.824**	**0.876**
		1,3,5	0.005	0.006	0.008	0.008	0.034	0.035	0.036	0.033
		1,2,5	0.002	0.004	0.003	0.003	0.011	0.016	0.013	0.015
		1,3,4,5	0.036	0.008	0.034	0.019	0.051	0.018	0.052	0.027
		1,2,4,5	0.041	0.012	0.045	0.026	0.062	0.014	0.048	0.026
		1,2,3,4,5	0.002	0.000	0.006	0.002	0.011	0.000	0.020	0.005
$\in_2$	**(2,0,0,4,8)**	1,5	0.008	0.032	0.017	0.027	0.018	0.093	0.029	0.056
		1,4,5*	**0.606**	**0.793**	**0.684**	**0.781**	**0.531**	**0.714**	**0.642**	**0.717**
		1,3,5	0.021	0.030	0.033	0.033	0.034	0.039	0.038	0.046
		1,2,5	0.013	0.028	0.016	0.021	0.019	0.017	0.021	0.022
		1,3,4,5	0.091	0.027	0.072	0.041	0.110	0.039	0.085	0.046
		1,2,4,5	0.255	0.089	0.170	0.94	0.282	0.098	0.179	0.109
		1,2,3,4,5	0.006	0.001	0.008	0.003	0.004	0.000	0.001	0.001
$\in_3$	**(2,0,0,4,8)**	1,5	0.004	0.021	0.003	0.007	0.026	0.077	0.027	0.055
		1,4	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.001
		1,4,5*	**0.887**	**0.929**	**0.886**	**0.920**	**0.807**	**0.815**	**0.817**	**0.824**
		1,3,5	0.015	0.022	0.019	0.021	0.052	0.053	0.057	0.052
		1,2,5	0.010	0.012	0.013	0.017	0.032	0.028	0.031	0.030
		1,3,4,5	0.045	0.012	0.042	0.018	0.042	0.013	0.030	0.017
		1,2,4,5	0.036	0.0.03	0.032	0.014	0.039	0.012	0.034	0.018
		1,2,3,4,5	0.003	0.001	0.005	0.003	0.002	0.001	0.003	0.003
$\in_4$	**(2,0,0,4,8)**	1,5	0.001	0.004	0.001	0.002	0.005	0.019	0.008	0.015
		1,4,5*	**0.921**	**0.965**	**0.913**	**0.948**	**0.863**	**0.904**	**0.859**	**0.897**
		1,3,5	0.005	0.007	0.003	0.005	0.029	0.035	0.032	0.032
		1,2,5	0.003	0.005	0.005	0.004	0.012	0.012	0.009	0.010
		1,3,4,5	0.039	0.011	0.034	0.021	0.046	0.017	0.040	0.022
		1,2,4,5	0.029	0.008	0.039	0.019	0.038	0.011	0.039	0.020
		1,2,3,4,5	0.002	0.000	0.005	0.000	0.007	0.001	0.013	0.003
$\in_5$	**(2,0,0,4,8)**	1,5	0.045	0114	0.050	0.081	0.162	0.274	0.162	0.251
		1.4	0.001	0.002	0.001	0.001	0.007	0.018	0.009	0.011
		1,4,5*	**0.839**	**0.820**	**0.842**	**0.843**	**0.670**	**0.611**	**0.692**	**0.637**
		1,3,5	0.046	0.034	0.044	0.035	0.055	0.041	0.056	0.046
		1,2,5	0.020	0.022	0.019	0.022	0.045	0.042	0.037	0.033
		1,3,4,5	0.026	0.004	0.017	0.009	0.024	0.009	0.016	0.008
		1,2,4,5	0.018	0.003	0.020	0.008	0.0028	0.005	0.019	0.013
		1,2,3,5	0.002	0.000	0.003	0.000	0.007	0.000	0.007	0.000
		1,2,3,4,5	0.003	0.001	0.002	0.001	0.002	0.000	0.002	0.001

Table 5. Estimated selection probabilities of $\overline{\alpha}_{m,n}$, $\tilde{\alpha}_{m,n}$, $\hat{\alpha}_{\mathbf{m,n}}$ and $\hat{\alpha}^*_{m,n}$ based on MM estimator. The results are based on 1000 Monte Carlo simulations and K = 50 bootstrap replications. Significant values are in bold.

v. $\in_5$, is Cauchy distribution with location = 0 and scale = 1.

The selection probabilities of $\overline{\alpha}_{m,n}$, $\hat{\alpha}_{m,n}$, $\tilde{\alpha}_{m,n}$ and $\hat{\alpha}^*_{m,n}$ on the basis of stratified bootstrap with the MM estimator are computed. The selection probabilities based on $\boldsymbol{L} = 1000$ simulations with bootstrap replications of $\mathbf{K} = 50$ are given in Table 5.

Table 5 demonstrates the simulation results with 10% and 20% of outliers in the covariates and five different error distributions as discussed in the simulation setting. If we look at the results, we see that the performance of our robust procedure is very good for $\in_4$ amongst all error distributions while it does not perform very well for $\in_5$ in the presence of x-outliers. The selection probabilities for error distribution $\in_3$ are similar to that of $\in_4$. Furthermore, the selection probabilities are good for distribution $\in_1$ (10% symmetric wild case) as compared to contamination type $\in_2$ (20% symmetric wild case). Overall, the selection probabilities for each of the criteria decrease when the percentage of both x- and y-outliers goes up. Moreover, selection probabilities in the presence of response outliers and covariate outliers, the performance of our proposed model selection criterion based on MM-estimation is comparable to the existing criteria even when the contamination level changes from i.e., 10% to 20%.

Selected variables	$\bar{\alpha}_{10,21}$	$\hat{\alpha}_{10,21}$	$\tilde{\alpha}_{10,21}$	$\hat{\alpha}^{*}_{10,21}$	AIC	BIC
X_1	2.97	3.50	4.42	4.82	4.61	4.71
X_2	3.02	3.59	4.58	5.00	2.19	2.28
X_3	3.48	4.20	5.32	5.86	3.21	3.31
X_1, X_2	**1.70**	**2.55**	**3.23**	**3.87**	1.55	1.70
X_1, X_3	2.38	3.42	4.67	5.45	2.69	2.84
X_2, X_3	2.09	3.49	3.72	4.77	1.47	1.62
X_1, X_2, X_3	1.81	3.05	3.90	5.28	**1.34**	**1.54**

Table 6. Selected best model for the stack loss data using a range of model selection procedures. Significant values are in bold.

Data example (Stack loss data)

In this section, we analyze the Stack loss data presented by[29]. This dataset consists of three explanatory variables, and it contains four outliers,namely observations 1, 3, 4, and 21.The response is the Stack loss (y) observed on $n = 21$ observations. The explanatory variables are theFlow of cooling air (X_1), Cooling Water Temperature (X_2), and Concentration of acid (X_3).We applied our robust method $\hat{\alpha}^{*}_{m,n}$, the existing methods, and the traditional methods on the data. Table 5 presents a summary of selected best models.

Table 6 shows the classical methodsselect the full model, whereas robust criteria agreed with the importance of the two variables, X_1 and X_2.The best model according to our criterion includes X_1 and X_2.

Conclusion

In this article, we have presented a novel procedure for robust model selection in linear regression. The criterion is a modification to the bootstrap model selection method based on robust estimator proposed by[19]. The simulation results reveal that the performance of model selection is improved when using the OOB error in the present studies. Moreover, the undue effect of outliers is controlled by using both a robust MM-estimator and a bounded loss function in the proposed criterion. The proposed model selection criterion can maintain their robust properties in the presence of response outliers and covariate outliers. The proposed criterion is compared with other robust model selection criteria described in previous literature.

We observed that in the presence of outliers and heavy-tailed error distributions, the MM-estimator outperformed the least squares estimator by a large margin. This clearly proved the lack of robustness of the least squares procedure in the presence of outliers and in heavy-tailed distributions. Furthermore, when errors are non-normal, then robust regression is found superior to least squares, but in the case of normal errors, robust regression is found inferior to least squares.

From simulation-based and real-data based results, we conclude that our modified robust model selection procedure is consistent and works well in situations where outliers are present in the data. As observed in our simulation study, an excellent amount of improvement is gained by minimizing the combined criterion, rather than minimizing the penalized loss function or the modified conditional expected prediction loss separately. Furthermore, our robust model selection criterion will perform better when the data generating model is small.

Received: 29 January 2022; Accepted: 9 May 2022
Published online: 29 June 2022

References

1. Akaike, H. Statistical predictor identification. *Ann. Inst. Stat. Math.* **22**(1), 203–217 (1970).
2. Mallows, C. L. Some comments on C_p. *Technometrics* **15**(4), 661–675 (1973).
3. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**(2), 461–464 (1978).
4. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **37**(4), 373–384 (1995).
5. Shao, J. Bootstrap model selection. *J. Am. Stat. Assoc.* **91**(434), 655–665 (1996).
6. Rabbi, F. *et al.* Model selection in linear regression using paired bootstrap. *Commun. Stat. Theory Methods* **50**(7), 1629–1639 (2021).
7. Ronchetti, E. Robust model selection in regression. *Stat. Probab. Lett.* **3**(1), 21–23 (1985).
8. Ronchetti, E. & Staudte, R. G. A robust version of Mallows's C_p. *J. Am. Stat. Assoc.* **89**(426), 550–559 (1994).
9. Sommer, S. & Staudte, R. G. Robust variable selection in regression in the presence of outliers and leverage points. *Aust. J. Stat.* **37**(3), 323–336 (1995).
10. Sommer, S. & Huggins, R. M. Variables selection using the Wald test and a robust C_p. *Appl. Stat.* **45**, 15–29 (1996).
11. Ronchetti, E., Field, C. & Blanchard, W. Robust linear model selection by cross-validation. *J. Am. Stat. Assoc.* **92**(439), 1017–1023 (1997).
12. Wisnowski, J. W., Simpson, J. R., Montgomery, D. C. & Runger, G. C. Resampling methods for variable selection in robust regression. *Comput. Stat. Data Anal.* **43**(3), 341–355 (2003).
13. Salibian-Barrera, M. & Van Aelst, S. Robust model selection using fast and robust bootstrap. *Comput. Stat. Data Anal.* **52**(12), 5121–5135 (2008).
14. Müller, S. & Welsh, A. Robust model selection in generalized linear models. *Stat. Sin.* **19**, 1155–1170 (2009).
15. Tharmaratnam, K. & Claeskens, G. A comparison of robust versions of the AIC based on M-, S-and MM-estimators. *Statistics* **47**(1), 216–235 (2013).
16. Saleh, S. Robust AIC with high breakdown scale estimate. *J. Appl. Math.* https://doi.org/10.1155/2014/286414 (2014).

17. Sakate, D. & Kashid, D. A new robust model selection method in GLM with application to ecological data. *Environ. Syst. Res.* **5**(9), 1–8. https://doi.org/10.1186/s40068-016-0060-7 (2016).
18. Rabbi, F., Khan, S., Khalil, A. & Salahuddin, N. Robust linear model selection using paired bootstrap. *Indian J. Sci. Technol.* **12**(10), 1–7. https://doi.org/10.17485/ijst/2019/v12i10/142190 (2019).
19. Müller, S. & Welsh, A. Outlier robust model selection in linear regression. *J. Am. Stat. Assoc.* **100**(472), 1297–1310 (2005).
20. Yohai, V. J. High breakdown-point and high-efficiency robust estimates for regression. *Ann. Stat.* **15**, 642–656 (1987).
21. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32 (2001).
22. Zhang, G.-Y., Zhang, C.-X. & Zhang, J.-S. Out-of-bag estimation of the optimal hyperparameter in subbag ensemble method. *Commun. Stat. Simul. Comput.* **39**(10), 1877–1892 (2010).
23. Imon, A. H. M. R. & Ali, M. M. Bootstrapping regression residuals. *J. Korean Data Inf. Sci. Soc.* **16**(3), 665–682 (2005).
24. Cochran, W. G. *Sampling Technique* 3rd edn. (Wiley, New York, 1977).
25. Hampel, F. Some aspects of model choice in robust statistics. In *Proceedings of the 44th Session of the ISI, Madrid, Book*, Vol. 2 767–771 (1983).
26. Ronchetti, E. Robustness aspects of model choice. *Stat. Sin.* **7**, 327–338 (1997).
27. Gunst, R. F. & Mason, R. L. *Regression Analysis and its Applications* (Marcel Dekker, New York, 1980).
28. Salibian-Barrera, M. & Zamar, R. H. Bootstrapping robust estimates of regression. *Ann. Stat.* **30**, 556–582 (2002).
29. Brownlee, K. A. *Statistical Theory and Methodology in Science and Engineering* (Wiley, New York, 1965).

Acknowledgements

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under Project R-2022-151.

Author contributions

F.R.: Random sampling and data analysis A.K.: Formulation of the model and simulations I.K.: Results computations and discussion M.A.A.: Simulations in revision, computing results, revision U.K.: Software, coding, computing, R-code, revision M.A: Writing revision, computing results, analysis of results, data analysis.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-14398-1.

Correspondence and requests for materials should be addressed to I.K., M.A.A. or M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.