Communications for Statistical Applications and Methods
2019, Vol. 26, No. 3, 273–293

https://doi.org/10.29220/CSAM.2019.26.3.273
Print ISSN 2287-7843 / Online ISSN 2383-4757

Least absolute deviation estimator based consistent model selection in regression

K. S. Shende[1,a], D. N. Kashid[a]

[a]Department of Statistics, Shivaji University, India

Abstract

We consider the problem of model selection in multiple linear regression with outliers and non-normal error distributions. In this article, the robust model selection criterion is proposed based on the robust estimation method with the least absolute deviation (LAD). The proposed criterion is shown to be consistent. We suggest proposed criterion based algorithms that are suitable for a large number of predictors in the model. These algorithms select only relevant predictor variables with probability one for large sample sizes. An exhaustive simulation study shows that the criterion performs well. However, the proposed criterion is applied to a real data set to examine its applicability. The simulation results show the proficiency of algorithms in the presence of outliers, non-normal distribution, and multicollinearity.

Keywords: linear regression, model selection, consistency, robustness, sequential algorithm

1. Introduction

The primary goal of regression analysis is to evolve a useful model to accurately predict the response variable for the given values of predictors. Consider the following general multiple linear regression model

$$y = X\beta + \varepsilon, \tag{1.1}$$

where y is $n \times 1$ vector of observed values of the response variable, X is $n \times k$ full rank matrix of $(k-1)$ predictor variables with ones in the first column, and β is corresponding $k \times 1$ vector of an unknown regression coefficients. The ε is $n \times 1$ vector of independent errors, and has the same distribution function F.

While developing the model, it is necessary to find out the unknown regression coefficients by using the appropriate method. The eminent ordinary least squares (OLS) estimator is obtained by minimizing the residual sum of squares. The OLS estimator is easy to compute and satisfies many properties. Nevertheless, the OLS method is not resistant to inconvenient observations in y space (known as outliers) and departs from the normality assumption of error in real data. The least absolute deviation (LAD) furnishes a useful and plausible alternative, resistant estimator. The LAD has many applications in Econometric and other studies. The resistant LAD estimator is obtained by minimizing the sum of absolute residuals. Dielman (2005) presented a rich literature review on LAD regression. LAD estimator has asymptotic $N(\beta, \tau^2(X'X)^{-1})$ distribution, $\tau = 1/\{2f(m)\}$, and $f(m)$ is the probability density of error evaluated at the median. The τ^2/n is a variance of the sample median

[1] Corresponding author: Department of Statistics, Shivaji University, Kolhapur, India (MS).
E-mail: shendeks.stat99@gmail.com

Published 31 May 2019 / journal homepage: http://csam.or.kr

of error. It is assumed that $F(0) = 1/2$ and $f(0) > 0$. The LAD estimator is useful for the existence of outliers and the non-normal error distribution problem.

Problems such as increase in complexity, prediction error, and economical aspects arise due to the addition of irrelevant predictor variables in the regression model. Such problems can be handled by a decisive aspect known as the model selection or variable selection procedure. Model selection has recently attracted significant attention in statistical research. The selection of a less complex model is essential. The model selection criteria are represented in the following form

Lack of Fit + Model Complexity.

Hence, the model selection can be done by trading off the lack of fit against model complexity. Many model selection methods have been proposed in the literature to choose a parsimonious model in multiple linear regression. Rao *et al.* (2001) given an extensive literature review on model selection. Most methods are based on OLS such as Mallows's C_p (Mallows, 1973). For zero bias, the expected value of C_p is p; therefore, Mallows's C_p selects the model for which C_p close to p. The C_p plot is a useful tool to graphically represent Mallows's C_p. There are alternative graphical methods available to select predictor variables. Siniksaran (2008) recently suggested an alternative plot with some advantages using a geometric approach. Gilmour (1995) modified Mallows's C_p because the expected value of Mallows's C_p of a model which includes all relevant predictor variables is not equal to p when the mean squared error (MSE) is used as an estimate of σ^2. Other methods like Akaike information criterion (AIC) (Akaike, 1973) and Bayesian information criterion (BIC) (Schwarz, 1978), are also available in the literature. Yamashita *et al.* (2007) studied stepwise AIC as well as other stepwise methods such as partial F, partial correlation and semi-partial correlation for variable selection in multiple linear regression that showed certain advantages of stepwise AIC.

The above methods are based on OLS or likelihood and are vulnerable to outliers. Researchers have proposed various robust variable selection methods to deal with outliers such as robust AIC (RAIC) (Ronchetti, 1985), robust BIC (RBIC) (Machado, 1993), RC_p (Ronchetti and Staudte, 1994), $C_p(d)$ (Kim and Hwang, 2000), S_p (Kashid and Kulkarni, 2002), and Tharmaratnam and Claeskens (2013) compared AIC based on different robust estimators. The model selection criteria $C_p, RC_p, C_p(d)$ and AIC are inconsistent; therefore, the probability of selection of only relevant predictor variables is less than one for large sample size. Methods like BIC and GIC-LR are consistent model selection methods that select only relevant predictor variables with probability one for a large sample size (Rao *et al.*, 2001). The BIC and GIC-LR methods are based on likelihood function and ordinary least squares (OLS) estimator respectively; however, these perform poorly in existence of outliers or departures from the normality assumption. BIC or GIC-LR methods existing in the literature are therefore consistent but not robust. To overcome this drawback, we have proposed a consistent and robust model selection criterion based on LAD estimator.

The remaining article is organized as follows. In Section 2, we propose a new variable selection criterion. We also studied its theoretical properties. In Section 3, the performance of the proposed criterion is studied through simulation and real data. The algorithms for model selection are explained in Section 4 with simulation and body fat real dataset. The article ends with some discussions of the results in Section 5.

2. Proposed method

The model (1.1) can be rewritten as

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{2.1}$$

where X and β are partitioned so that X_1 is a matrix of $(p-1)$ predictor variables with ones in the first column, and β_1 is a $p \times 1$ vector of associated regression coefficients including intercept. X_2 is a matrix of $(k-p)$ predictor variables, and β_2 is a $(k-p)\times 1$ vector of associated regression coefficients. Consider the test for regression coefficient with the null hypothesis $H_0 : \beta_2 = 0$. Under the null hypothesis, the reduced model is

$$y = X_1\beta_1 + \varepsilon. \tag{2.2}$$

Consider $\hat{y}_f$ and $\hat{y}_r$ are the predicted values of y based on full model and reduced model respectively. The predicted values are obtained using the LAD estimator of the respective models. We propose a criterion based on these fitted values of y and model complexity. It is defined as

$$CR_p = \frac{|y-\hat{y}_r|'\mathbf{1} - |y-\hat{y}_f|'\mathbf{1}}{\frac{\tau}{2}\left(1+\frac{k-p}{n-k+p}\right)} + C_n(p). \tag{2.3}$$

The first term $D_p = [|y-\hat{y}_r|'\mathbf{1} - |y-\hat{y}_f|'\mathbf{1}]/[(\tau/2)\{1+(k-p)/(n-k+p)\}]$ represents the lack of fit and is non-negative, $\mathbf{1}$ is the n-dimensional column vector of ones, and τ is a scale parameter that can be replaced by a suitable estimator based on a full model. The D_p is a scaled likelihood test statistic and scaled by the quantity $(1+(k-p)/(n-k+p))$. This statistic is accurate for moderate sample size as compared to likelihood test statistic (Birkes and Dodge, 1993). For $n \to \infty, D_p$ and likelihood test statistic are equivalent. $Dp = 0$ for the full model and is minimum among all possible subsets; therefore, if we select a model that has a minimum Dp, then the full model is always selected. Hence, the 'minimum D_p' criterion does not select the parsimonious model which explain data with few predictor variables and has better prediction ability. To make a consistent criterion, consider the model complexity $C_n(p)$ is an increasing function of model dimension (p) that often depends on sample size (n). Generally, the model dimension considered as the model complexity, but this complexity measure does not make a consistent criterion. To overcome this problem, we consider the function of the sample size and model dimension as a complexity measure. The model having small complexity will be the best model as long as discrepancy measure (D_p) is also small. The CR_p criterion selects the model which has a small CR_p value among all possible models. The established theoretical results of the CR_p are given below:

Proposition 1. *Under the null hypothesis* $H_0, E(CR_p) = (k-p) + C_n(p)$.

Proof: The proposed criterion is

$$CR_p = \frac{|y-\hat{y}_r|'\mathbf{1} - |y-\hat{y}_f|'\mathbf{1}}{\frac{\tau}{2}\left(1+\frac{k-p}{n-k+p}\right)} + C_n(p).$$

Under the null hypothesis, D_p approximately follows χ^2 distribution with $k-p$ degree of freedom (Birkes and Dodge, 1993). The expected value of CR_p is

$$E(CR_p) = (k-p) + C_n(p).$$

Hence, the proof. □

Alternatively, for large n the proposed criterion can be written as

$$CR_{\text{alt}_p} = \frac{\tau}{2}\left(|y-\hat{y}_r|'\mathbf{1} - |y-\hat{y}_f|'\mathbf{1}\right) + C_n(p). \tag{2.4}$$

The performance of both criteria expressed in (2.3) and (2.4) will be same for large n. Consider, α^1 be the subset of $\{1, 2, \ldots, k-1\}$, and α^0 represents intercept. Let the selected model denoted by M_α, $\alpha = \alpha^1 \cup \alpha^0$, and α_0 represents a set of all necessary predictor variables. The selected model belongs to one of the following classes:

- Optimal Model: $\mathcal{M}_o = M_o = \{M_\alpha : \alpha = \alpha_0\}$
- Class of correct models: $\mathcal{M}_c = \{M_\alpha : \alpha \supseteq \alpha_0\}$
- Class of wrong models: $\mathcal{M}_w = \{M_\alpha : \alpha \not\supseteq \alpha_0\}$

Let $CR_{p_{\alpha^*}}$ and $CR_{p_{\alpha^{**}}}$ denotes the values of criterion corresponding to any correct model $M_{\alpha^*} \in \mathcal{M}_c$ and wrong model $M_{\alpha^{**}} \in \mathcal{M}_w$ with dimension p_{α^*} and $p_{\alpha^{**}}$ respectively. The $\hat{y}_c$ and $\hat{y}_w$ are vectors of fitted values of the respective correct model and wrong model. Under mild conditions, the Theorem 1 exhibits the consistency property of the proposed criterion for fixed k.

Condition 1. *For any $M_{\alpha^*} \in \mathcal{M}_c$ and $M_{\alpha^{**}} \in \mathcal{M}_w$,* $\liminf_{n\to\infty} \left(\frac{|y-\hat{y}_w|'\mathbf{1}}{n} - \frac{|y-\hat{y}_c|'\mathbf{1}}{n} \right) > 0.$

It is expected that the average of absolute residuals of the wrong model is greater than any correct model. Thus, the difference $(|y - \hat{y}_w|'\mathbf{1}/n - |y - \hat{y}_c|'\mathbf{1}/n)$ is positive, large, and Condition 1 is reasonably true.

Condition 2. *$C_n(p) = o(n)$ and $C_n(p) \to \infty$ as $n \to \infty$.*

The Condition 2 is required to prove the following consistency property.

Theorem 1. (Consistency Property) *Assume that above conditions are satisfied. Then*

$$\lim_{n\to\infty} \Pr(M_\alpha = M_o) = 1.$$

Proof: From the definition of criterion,

$$\begin{aligned}
CR_{p_{\alpha^{**}}} - CR_{p_{\alpha^*}} &= \frac{|y - \hat{y}_w|'\mathbf{1} - |y - \hat{y}_f|'\mathbf{1}}{\frac{\tau}{2}\left(1 + \frac{k-p_{\alpha^{**}}}{n-k+p_{\alpha^{**}}}\right)} - \frac{|y - \hat{y}_c|'\mathbf{1} - |y - \hat{y}_f|'\mathbf{1}}{\frac{\tau}{2}\left(1 + \frac{k-p_{\alpha^*}}{n-k+p_{\alpha^*}}\right)} + C_n(p_{\alpha^{**}}) - C_n(p_{\alpha^*}) \\
&= \frac{2}{\tau}\left(\left(1 - \frac{k-p_{\alpha^{**}}}{n}\right)|y - \hat{y}_w|'\mathbf{1} - |y - \hat{y}_c|'\mathbf{1}\right) + \frac{2(k-p_{\alpha^*})}{\tau n}|y - \hat{y}_c|'\mathbf{1} \\
&\quad + \frac{2(p_{\alpha^*} - p_{\alpha^{**}})}{\tau n}|y - \hat{y}_f|'\mathbf{1} + C_n(p_{\alpha^{**}}) - C_n(p_{\alpha^*}) \\
&= \frac{2}{\tau}\left(\left(1 - \frac{k-p_{\alpha^{**}}}{n}\right)|y - \hat{y}_w|'\mathbf{1} - |y - \hat{y}_c|'\mathbf{1}\right) + \xi_1 + \xi_2 + C_n(p_{\alpha^{**}}) - C_n(p_{\alpha^*}), \qquad (2.5)
\end{aligned}$$

where

$$\xi_1 = \frac{2(k-p_{\alpha^*})}{\tau n}|y - \hat{y}_c|'\mathbf{1} \quad \text{and} \quad \xi_2 = \frac{2(p_{\alpha^*} - p_{\alpha^{**}})}{\tau n}|y - \hat{y}_f|'\mathbf{1}.$$

For any selected model M_α,

$$|y - \hat{y}_r|'\mathbf{1} - |y - \hat{y}_f|'\mathbf{1} \le |X\hat{\beta} - X\beta|'\mathbf{1} + |X_\alpha\hat{\beta}_\alpha - X_\alpha\beta_\alpha|'\mathbf{1} + |X\beta - X_\alpha\beta_\alpha|'\mathbf{1}. \qquad (2.6)$$

Whenever, $M_\alpha \in \mathcal{M}_c$, $X\beta = X_\alpha\beta_\alpha$ and by consistency and asymptotic normality property (Dielman, 2005) we have $|X_\alpha\hat{\beta}_\alpha - X_\alpha\beta_\alpha|'\mathbf{1} = O_p(1)$, $|y - \hat{y}_c|'\mathbf{1} = O_p(1)$, $|y - \hat{y}_f|'\mathbf{1} = O_p(1)$ and consequently, $\xi_1 = \xi_2 = o_p(1)$. Hence,

$$\begin{aligned}
&\liminf_{n\to\infty} \Pr\left(CR_{p_{\alpha^{**}}} - CR_{p_{\alpha^*}} > 0\right) \\
&= \liminf_{n\to\infty} \Pr\left(\frac{2}{\tau}\left(\left(1 - \frac{k - p_{\alpha^{**}}}{n}\right)|y - \hat{y}_w|'\mathbf{1} - |y - \hat{y}_c|'\mathbf{1}\right) + o_p(1) + C_n(p_{\alpha^{**}}) - C_n(p_{\alpha^*}) > 0\right) \\
&\geq \Pr\left(\liminf_{n\to\infty} \frac{2}{\tau}\left(\left(1 - \frac{k - p_{\alpha^{**}}}{n}\right)|y - \hat{y}_w|'\mathbf{1} - |y - \hat{y}_c|'\mathbf{1}\right) + o_p(1) + o_p(n) > 0\right) \\
&= 1. \qquad (2.7)
\end{aligned}$$

Now, to complete the proof, it is sufficient to show that CR_p selects the optimal model with probability one among the class of correct models. Consider $D_{p_{\alpha_o}}$ and $D_{p_{\alpha^*}}$ are values of D_p corresponding to the optimal and correct model respectively. Under Condition 2, we have

$$\begin{aligned}
\lim_{n\to\infty} \Pr(M_\alpha = M_o) &= \lim_{n\to\infty} \Pr\left(CR_{p_{\alpha_o}} \leq CR_{p_{\alpha^*}}\right) \\
&= \lim_{n\to\infty} \Pr\left(D_{p_{\alpha_o}} - D_{p_{\alpha^*}} < \infty\right) \\
&= \lim_{n\to\infty} \Pr\left(\chi^2_{p_{\alpha^*} - p_{\alpha_o}} < \infty\right) \\
&= 1. \qquad (2.8)
\end{aligned}$$

Hence, the CR_p selects only all relevant predictor variables with probability one for large n. □

2.1. Choice of τ

The CR_p requires the estimation of an unknown scale parameter τ. Birkes and Dodge (1993) have the given estimator $\hat{\tau}_1$ of τ, and recommended to use only non-zero residuals to improve the performance. Dielman (2006) examined the performance of the likelihood ratio (LR) test, the Wald test and the Lagrange multiplier (LM) test for the testing hypothesis regarding the regression coefficient in the LAD regression. He considered four different estimators $\hat{\tau}_2$, $\hat{\tau}_3$, $\hat{\tau}_4$, and $\hat{\tau}_5$ of τ for a comparative study of these significance tests as well as showed that these types of estimators are performed well. In this study, we considered the following five existing estimators of τ to calculate CR_p.

$$\hat{\tau}_1 = \frac{\sqrt{m}\,(r_{(k_2)} - r_{(k_1)})}{4}, \quad k_1 = \left[\frac{m+1}{2} - \sqrt{m}\right],\ k_2 = \left[\frac{m+1}{2} + \sqrt{m}\right], \text{ and } m = \sum_{i=1}^{n} I_{(r_i \neq 0)},$$

$$\hat{\tau}_2 = \frac{\sqrt{m}\,(r_{(m-k_1-1)} - r_{(k_1)})}{z_{\frac{\alpha}{2}}}, \quad k_1 = \left[\frac{m+1}{2} - z_{\frac{\alpha}{2}}\sqrt{\frac{m}{4}}\right],\ m = \sum_{i=1}^{n} I_{(r_i \neq 0)}, \text{ and } \alpha = 0.05,$$

$$\hat{\tau}_3 = \frac{\sqrt{m}\,(r_{(m-k_1-1)} - r_{(k_1)})}{z_{\frac{\alpha}{2}}}, \quad k_1 = \left[\frac{m+1}{2} - z_{\frac{\alpha}{2}}\sqrt{\frac{m}{4}}\right],\ m = n, \text{ and } \alpha = 0.05,$$

$$\hat{\tau}_4 = \frac{\sqrt{m}\,(r_{(m-k_1-1)} - r_{(k_1)})}{t_{\frac{\alpha}{2}}}, \quad k_1 = \left[\frac{m+1}{2} - t_{\frac{\alpha}{2}}\sqrt{\frac{m}{4}}\right],\ m = \sum_{i=1}^{n} I_{(r_i \neq 0)}, \text{ and } \alpha = 0.05,$$

$$\hat{\tau}_5 = \frac{\sqrt{m}\,(r_{(m-k_1-1)} - r_{(k_1)})}{t_{\frac{\alpha}{2}}}, \quad k_1 = \left[\frac{m+1}{2} - t_{\frac{\alpha}{2}}\sqrt{\frac{m}{4}}\right],\ m = n, \text{ and } \alpha = 0.05.$$

Table 1: Penalty functions

Sr. No.	Penalty function $C_n(p)$
1	$P_1 = 2p$
2	$P_2 = 3p$
3	$P_3 = 2p\log(p)$
4	$P_4 = p\log(n)$
5	$P_5 = p(\log(n) + 1)$
6	$P_6 = p\sqrt{n}$
7	$P_7 = p(\sqrt{n} + 2)$

Here, $r_{(\cdot)}$ denotes ordered residuals of full model, and $[\cdot]$ denotes nearest positive integer. Only nonzero residuals are considered to estimate $\hat{\tau}_1$, $\hat{\tau}_2$, and $\hat{\tau}_4$; however, all n residuals are considered to estimate $\hat{\tau}_3$ and $\hat{\tau}_5$. An exhaustive simulation compares the performance of these estimators in the next section.

3. Performance of CR_p

In this section, an extensive simulation study checked the superiority of the proposed criterion. Also, the real-life data analysis showed an applicability of the criterion.

3.1. Simulation study

In this simulation study, we considered seven different penalties (Table 1). The four penalties P_4–P_7 satisfy Condition 2, and remaining penalties are the functions of p only and do not satisfy Condition 2.

The independent predictor variables X_j, $j = 1, 2, \ldots, (k-1)$ and random errors are generated from $N(0, 1)$ distribution. The outliers are introduced artificially in the data by multiplying 20 to response variable y corresponding to maximum absolute residuals. The simulation has been done for different sample sizes $n = 30, 50, 70, 100, 200$ and two different models are described below:

- Model-I: $\beta = (5, 2, 3, 4, 0, 0)$
- Model-II: $\beta = (5, 2, 3, 4, 2, 0, 0)$

In both these models, the response variable y is generated using (1.1). The performance of the proposed method is studied in terms of the percentage of an optimal model selection. The percentage of an optimal model selection in 1,000 runs are recorded in Table 2 and Table 3. It shows that CR_p performs well in cases of clean data as well as outliers; however, outliers drastically affect AIC and BIC. RBIC performs uniformly better than RAIC. The performance of CR_p criterion with P_3–P_7 over RBIC is remarkable. The penalties P_3–P_7 select an optimal model with a large percentage as compared to other penalties. It is observed that $\hat{\tau}_1$, $\hat{\tau}_2$, $\hat{\tau}_4$ performs better than $\hat{\tau}_3$ and $\hat{\tau}_5$. Hence, the consideration of only non-zero residuals to estimate τ results in a good percentage for a small as well as large sample size. $\hat{\tau}_4$ performs better compared to others for small sample sizes; however, $\hat{\tau}_2$ and $\hat{\tau}_4$ perform equally for large sample sizes. Thus, $\hat{\tau}_4$ performs well in cases of small as well as large sample sizes. For further study, we consider $\hat{\tau}_4$ as an estimator of τ. CR_p criterion with all penalties performs well as the sample size increases. The simulation study confirms the consistency property of CR_p criterion for P_4–P_7 penalties.

Table 2: Percentage of optimal model selection (Model-I)

n	No. of Outliers	τ	CR_p							AIC	BIC	RAIC	RBIC
			P_1	P_2	P_3	P_4	P_5	P_6	P_7				
30	0	$\hat{\tau}_1$	74.0	86.1	95.7	88.8	94.5	96.7	98.6	64.6	84.0	54.7	92.0
		$\hat{\tau}_2$	87.1	95.0	98.8	96.4	98.3	99.2	99.5				
		$\hat{\tau}_3$	66.9	78.6	90.4	81.8	88.2	91.8	95.7				
		$\hat{\tau}_4$	91.8	97.6	99.5	98.0	99.4	99.4	98.8				
		$\hat{\tau}_5$	65.3	77.7	89.5	81.1	87.1	91.0	95.3				
	1	$\hat{\tau}_1$	65.7	79.4	93.0	83.0	90.5	94.9	98.2	13.9	12.5	43.5	84.8
		$\hat{\tau}_2$	80.4	91.1	98.6	93.9	97.9	98.8	99.2				
		$\hat{\tau}_3$	57.1	70.5	85.1	73.8	81.3	86.8	92.6				
		$\hat{\tau}_4$	87.2	96.1	99.7	97.2	99.2	99.6	98.3				
		$\hat{\tau}_5$	55.6	69.2	84.3	72.9	80.0	86.0	91.5				
	2	$\hat{\tau}_1$	64.2	77.6	90.9	80.5	88.8	92.6	96.5	3.3	2.0	43.3	80.8
		$\hat{\tau}_2$	78.5	89.2	97.1	92.0	95.8	97.6	97.9				
		$\hat{\tau}_3$	54.6	69.4	84.7	73.6	80.6	86.4	92.3				
		$\hat{\tau}_4$	85.5	93.3	98.4	94.6	97.8	98.4	97.5				
		$\hat{\tau}_5$	53.7	68.3	83.5	72.0	79.4	86.0	91.8				
	3	$\hat{\tau}_1$	62.0	77.2	90.4	80.1	86.8	91.9	96.0	2.4	0.7	40.3	78.7
		$\hat{\tau}_2$	77.7	87.9	95.8	90.4	94.5	96.1	98.2				
		$\hat{\tau}_2$	54.3	66.8	80.0	70.3	77.0	82.6	88.2				
		$\hat{\tau}_4$	83.4	92.8	98.1	94.8	97.2	98.2	97.1				
		$\hat{\tau}_5$	53.0	65.4	79.1	68.9	76.6	81.4	87.8				
50	0	$\hat{\tau}_1$	73.6	85.7	95.7	91.9	95.0	98.4	99.3	64.5	87.8	57.7	94.1
		$\hat{\tau}_2$	89.0	95.5	99.0	98.1	99.0	99.9	99.9				
		$\hat{\tau}_3$	66.4	79.9	91.5	88.0	91.2	95.8	98.1				
		$\hat{\tau}_4$	88.3	95.3	98.8	98.0	98.7	99.9	99.9				
		$\hat{\tau}_5$	77.7	88.3	96.2	93.6	96.2	98.8	99.2				
	1	$\hat{\tau}_1$	68.9	83.7	93.7	89.5	93.6	98.2	99.4	23.6	22.8	57.1	90.8
		$\hat{\tau}_2$	85.9	94.4	99.2	97.8	99.0	99.7	99.9				
		$\hat{\tau}_3$	62.9	76.9	90.0	85.7	89.8	94.7	97.5				
		$\hat{\tau}_4$	85.3	94.1	99.1	97.8	99.0	99.6	99.9				
		$\hat{\tau}_5$	74.5	85.3	95.1	91.1	94.9	98.6	99.3				
	2	$\hat{\tau}_1$	68.6	80.9	93.1	88.1	92.9	97.4	99.0	7.8	4.5	52.9	88.3
		$\hat{\tau}_2$	84.4	93.1	98.7	97.1	98.6	99.5	99.8				
		$\hat{\tau}_3$	64.0	76.3	89.6	83.1	88.9	94.7	96.8				
		$\hat{\tau}_4$	83.6	92.7	98.6	96.5	98.6	99.5	99.7				
		$\hat{\tau}_5$	72.7	84.7	94.6	90.1	94.1	98.0	98.9				
	3	$\hat{\tau}_1$	67.7	80.9	92.3	86.8	92.2	97.1	98.8	6.1	1.5	50.3	86.4
		$\hat{\tau}_2$	83.4	92.8	98.2	96.4	98.1	99.4	99.9				
		$\hat{\tau}_3$	60.8	74.6	86.8	81.1	86.0	93.8	97.0				
		$\hat{\tau}_4$	82.9	92.7	98.1	96.1	97.9	99.4	99.9				
		$\hat{\tau}_5$	71.6	83.6	93.6	88.8	92.9	97.4	99.0				
70	0	$\hat{\tau}_1$	74.1	85.9	96.1	94.8	96.5	98.7	99.5	65.7	90.7	62.3	95.3
		$\hat{\tau}_2$	89.0	96.0	98.7	98.1	98.9	100.0	100.0				
		$\hat{\tau}_3$	78.8	90.8	96.5	95.4	96.9	99.2	99.8				
		$\hat{\tau}_4$	88.4	95.7	98.6	98.0	98.8	100.0	100.0				
		$\hat{\tau}_5$	77.9	90.3	96.4	95.4	96.7	99.2	99.8				
	1	$\hat{\tau}_1$	72.3	85.4	94.2	92.1	94.7	98.6	99.4	36.6	33.4	61.1	93.9
		$\hat{\tau}_2$	88.2	94.0	98.5	97.4	98.7	99.9	99.9				
		$\hat{\tau}_3$	77.7	88.7	95.8	94.0	96.1	99.4	99.7				
		$\hat{\tau}_4$	87.9	93.6	98.2	97.4	98.6	99.9	99.9				
		$\hat{\tau}_5$	77.0	88.1	95.4	94.0	95.8	99.4	99.7				
	2	$\hat{\tau}_1$	69.8	83.1	95.0	91.8	95.5	99.0	99.7	16.4	9.8	56.9	92.6
		$\hat{\tau}_2$	84.4	94.8	98.7	97.9	99.3	100.0	100.0				
		$\hat{\tau}_3$	74.6	87.3	96.4	94.2	96.8	99.5	99.9				
		$\hat{\tau}_4$	83.9	94.4	98.7	97.7	99.0	100.0	100.0				
		$\hat{\tau}_5$	74.3	86.7	96.0	94.0	96.7	99.5	99.8				

Continued...

n	No. of Outliers	τ	CR_p							AIC	BIC	RAIC	RBIC
			P_1	P_2	P_3	P_4	P_5	P_6	P_7				
100	3	$\hat{\tau}_1$	69.9	81.8	93.4	90.1	94.0	98.5	99.1	10.4	4.1	54.7	90.8
		$\hat{\tau}_2$	83.7	92.5	98.2	97.1	98.3	99.6	99.8				
		$\hat{\tau}_3$	73.8	85.9	94.7	93.0	95.7	99.1	99.4				
		$\hat{\tau}_4$	83.1	92.4	98.2	97.0	98.3	99.6	99.8				
		$\hat{\tau}_5$	73.3	85.5	94.6	92.6	95.4	99.1	99.3				
	0	$\hat{\tau}_1$	72.1	85.3	95.1	94.1	96.4	99.5	99.8	67.8	92.6	66.5	95.7
		$\hat{\tau}_2$	88.8	96.3	99.1	98.8	99.4	100.0	100.0				
		$\hat{\tau}_3$	78.8	89.2	96.9	96.4	97.4	99.7	100.0				
		$\hat{\tau}_4$	88.6	96.2	99.1	98.8	99.4	100.0	100.0				
		$\hat{\tau}_5$	78.4	89.1	96.9	96.4	97.3	99.7	100.0				
	1	$\hat{\tau}_1$	69.5	83.4	93.4	92.2	95.9	99.7	99.9	47.1	43.9	62.6	94.2
		$\hat{\tau}_2$	88.1	95.0	99.7	99.4	99.8	100.0	100.0				
		$\hat{\tau}_3$	75.7	87.7	96.1	95.1	97.8	99.7	99.9				
		$\hat{\tau}_4$	87.9	94.6	99.7	99.4	99.7	100.0	100.0				
		$\hat{\tau}_5$	75.4	87.5	95.9	94.9	97.5	99.7	99.9				
	2	$\hat{\tau}_1$	67.1	81.0	93.6	91.2	95.4	99.2	99.5	30.6	17.8	58.2	93.3
		$\hat{\tau}_2$	85.9	95.0	98.5	98.1	99.0	99.9	100.0				
		$\hat{\tau}_3$	74.4	86.4	96.1	95.1	96.9	99.3	99.7				
		$\hat{\tau}_4$	85.7	95.0	98.4	98.1	99.0	99.9	100.0				
		$\hat{\tau}_5$	73.9	86.1	95.7	95.0	96.9	99.3	99.7				
	3	$\hat{\tau}_1$	69.0	82.1	94.1	93.4	96.0	99.1	99.8	22.5	8.5	58.2	94.2
		$\hat{\tau}_2$	88.1	95.0	98.8	98.2	99.1	99.9	100.0				
		$\hat{\tau}_3$	74.6	87.2	95.7	94.9	97.1	99.5	99.8				
		$\hat{\tau}_4$	87.7	95.0	98.8	98.2	99.1	99.9	100.0				
		$\hat{\tau}_5$	74.5	87.0	95.6	94.7	96.9	99.5	99.8				
200	0	$\hat{\tau}_1$	69.7	83.3	94.0	94.6	97.1	100.0	100.0	67.0	94.6	66.0	97.0
		$\hat{\tau}_2$	88.5	96.0	99.7	99.8	99.8	100.0	100.0				
		$\hat{\tau}_3$	82.3	91.7	98.0	98.7	99.3	100.0	100.0				
		$\hat{\tau}_4$	88.4	95.9	99.7	99.7	99.8	100.0	100.0				
		$\hat{\tau}_5$	81.9	91.7	98.0	98.6	99.3	100.0	100.0				
	1	$\hat{\tau}_1$	70.9	83.4	95.3	95.9	97.5	100.0	100.0	65.4	72.9	66.5	97.5
		$\hat{\tau}_2$	90.3	96.6	99.3	99.7	99.8	100.0	100.0				
		$\hat{\tau}_3$	82.2	92.7	98.2	98.6	99.4	100.0	100.0				
		$\hat{\tau}_4$	90.1	96.6	99.3	99.7	99.8	100.0	100.0				
		$\hat{\tau}_5$	82.1	92.6	98.2	98.6	99.3	100.0	100.0				
	2	$\hat{\tau}_1$	67.7	81.1	93.7	94.4	97.0	99.9	100.0	62.9	54.4	67.5	96.4
		$\hat{\tau}_2$	87.3	95.6	99.3	99.3	99.7	100.0	100.0				
		$\hat{\tau}_3$	80.0	90.3	98.3	98.5	98.9	100.0	100.0				
		$\hat{\tau}_4$	87.0	95.4	99.2	99.3	99.7	100.0	100.0				
		$\hat{\tau}_5$	79.6	90.1	98.3	98.5	98.9	100.0	100.0				
	3	$\hat{\tau}_1$	69.5	82.3	93.4	94.2	96.5	100.0	100.0	52.1	37.6	67.3	96.2
		$\hat{\tau}_2$	86.8	95.3	99.5	99.6	99.9	100.0	100.0				
		$\hat{\tau}_3$	81.2	91.3	97.9	98.3	99.3	100.0	100.0				
		$\hat{\tau}_4$	86.8	95.1	99.5	99.6	99.9	100.0	100.0				
		$\hat{\tau}_5$	81.0	91.2	97.9	98.3	99.3	100.0	100.0				

AIC = like Akaike information criteria; BIC = Bayesian information criteria; RAIC = robust AIC; RBIC = robust BIC.

3.2. Real data (Hald cement data)

The performance of the proposed criterion is examined with real-life data. This section analyze a Hald cement dataset (Ronchetti and Staudte, 1994). Hald cement data has 13 observations on the heat evolved in calories per gram of cement (y) and four ingredients in the mixture: tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium aluminoferrite (X_3) and dicalcium silicate (X_4). Many researchers have considered this data for model selection problem and suggested X_1, X_2 predictor

Table 3: Percentage of optimal model selection (Model-II)

n	No. of Outliers	τ	CR_p							AIC	BIC	RAIC	RBIC
			P_1	P_2	P_3	P_4	P_5	P_6	P_7				
30	0	$\hat{\tau}_1$	78.0	88.2	97.4	90.7	95.2	97.5	99.2	61.1	77.7	48.9	89.2
		$\hat{\tau}_2$	89.2	95.7	99.2	97.2	99.0	99.0	96.0				
		$\hat{\tau}_3$	54.3	66.9	85.4	70.8	78.6	85.4	91.0				
		$\hat{\tau}_4$	87.9	95.3	99.3	96.7	98.9	99.0	97.0				
		$\hat{\tau}_5$	52.6	65.3	84.4	68.9	77.7	84.5	90.7				
	1	$\hat{\tau}_1$	71.5	81.8	95.2	86.2	92.1	95.5	97.0	9.7	8.8	44.9	84.4
		$\hat{\tau}_2$	83.8	92.9	97.6	94.7	96.9	97.2	94.9				
		$\hat{\tau}_3$	50.2	62.0	77.7	64.8	71.7	78.0	85.7				
		$\hat{\tau}_4$	82.6	92.3	97.7	94.1	96.8	97.3	95.8				
		$\hat{\tau}_5$	48.3	61.1	76.6	63.9	70.4	76.8	84.9				
	2	$\hat{\tau}_1$	70.2	82.6	94.6	85.1	91.6	94.8	97.6	1.9	1.2	39.9	82.0
		$\hat{\tau}_2$	83.7	92.6	97.8	94.2	96.5	97.3	95.2				
		$\hat{\tau}_3$	46.4	60.5	77.5	63.9	71.6	77.9	84.7				
		$\hat{\tau}_4$	82.5	91.7	97.8	93.7	95.9	97.3	96.4				
		$\hat{\tau}_5$	45.3	59.2	76.2	63.0	70.5	76.4	84.0				
	3	$\hat{\tau}_1$	65.0	79.9	93.6	83.9	90.3	93.5	97.5	1.0	0.2	39.8	76.6
		$\hat{\tau}_2$	80.3	91.2	97.4	93.7	96.3	97.2	94.4				
		$\hat{\tau}_3$	43.5	56.9	74.1	60.2	68.9	74.3	83.1				
		$\hat{\tau}_4$	78.6	90.0	97.3	92.7	96.1	96.9	94.8				
		$\hat{\tau}_5$	42.1	55.8	73.3	58.9	67.5	73.3	82.4				
50	0	$\hat{\tau}_1$	74.8	88.2	97.0	93.1	96.3	98.8	99.7	66.0	87.7	57.9	93.7
		$\hat{\tau}_2$	86.5	94.0	99.3	97.0	98.7	99.8	99.9				
		$\hat{\tau}_3$	62.0	73.9	88.0	80.8	84.9	93.0	96.1				
		$\hat{\tau}_4$	90.2	96.4	99.5	98.7	99.4	100.0	100.0				
		$\hat{\tau}_5$	73.2	83.7	94.6	89.8	94.0	97.2	98.3				
	1	$\hat{\tau}_1$	72.7	84.6	95.6	90.6	94.4	97.9	99.1	18.7	16.0	53.6	88.5
		$\hat{\tau}_2$	82.6	92.1	98.2	95.7	97.9	99.2	99.8				
		$\hat{\tau}_3$	57.7	69.9	87.1	79.9	85.2	92.4	95.1				
		$\hat{\tau}_4$	87.6	94.8	99.0	97.5	98.7	99.7	99.9				
		$\hat{\tau}_5$	69.7	81.2	94.3	88.6	92.5	96.4	97.9				
	2	$\hat{\tau}_1$	70.2	84.3	95.6	91.2	94.3	96.9	98.8	5.0	3.2	51.6	88.8
		$\hat{\tau}_2$	82.2	91.6	97.6	95.4	97.4	98.8	99.6				
		$\hat{\tau}_3$	55.6	68.2	86.9	77.6	84.6	92.1	94.4				
		$\hat{\tau}_4$	87.2	94.1	98.9	96.9	98.3	99.7	100.0				
		$\hat{\tau}_5$	67.4	81.4	92.6	88.1	91.9	95.6	97.8				
	3	$\hat{\tau}_1$	68.6	80.4	95.0	88.4	93.3	97.6	98.9	2.1	0.7	47.3	86.5
		$\hat{\tau}_2$	78.9	90.8	97.9	95.3	97.4	99.1	99.8				
		$\hat{\tau}_3$	53.9	66.0	84.2	74.2	81.9	89.2	94.1				
		$\hat{\tau}_4$	83.5	93.8	98.9	97.1	98.4	99.8	100.0				
		$\hat{\tau}_5$	65.6	76.8	92.1	84.0	90.0	96.5	98.1				
70	0	$\hat{\tau}_1$	70.4	83.1	96.7	92.6	96.5	99.6	99.8	67.0	89.1	60.0	95.6
		$\hat{\tau}_2$	87.8	96.0	99.7	99.0	99.7	100.0	100.0				
		$\hat{\tau}_3$	73.4	85.4	95.8	92.2	95.4	99.0	99.8				
		$\hat{\tau}_4$	87.2	95.8	99.7	98.8	99.7	99.9	100.0				
		$\hat{\tau}_5$	72.8	85.2	95.4	91.9	95.2	99.0	99.8				
	1	$\hat{\tau}_1$	70.1	82.5	95.0	91.1	94.4	99.3	99.6	26.0	26.2	57.8	90.9
		$\hat{\tau}_2$	86.8	95.0	99.5	98.4	99.2	100.0	100.0				
		$\hat{\tau}_3$	72.8	83.6	95.8	91.9	95.2	99.2	99.9				
		$\hat{\tau}_4$	86.5	94.6	99.2	98.2	99.2	100.0	100.0				
		$\hat{\tau}_5$	72.1	83.2	95.4	91.7	95.0	99.1	99.9				
	2	$\hat{\tau}_1$	68.1	81.8	94.9	91.0	94.7	99.0	99.6	9.4	5.2	53.1	92.1
		$\hat{\tau}_2$	86.8	95.0	99.5	98.6	99.4	100.0	100.0				
		$\hat{\tau}_3$	68.4	83.7	95.6	91.8	95.4	98.7	99.3				
		$\hat{\tau}_4$	86.1	94.5	99.5	98.4	99.4	100.0	100.0				
		$\hat{\tau}_5$	67.9	83.0	95.5	91.6	95.1	98.7	99.2				

Continued...

n	No. of Outliers	τ	CR_p							AIC	BIC	RAIC	RBIC
			P_1	P_2	P_3	P_4	P_5	P_6	P_7				
100	3	$\hat{\tau}_1$	66.7	80.7	93.5	89.5	93.1	98.1	99.1	4.6	0.8	55.1	89.9
		$\hat{\tau}_2$	84.5	93.7	98.4	97.4	98.4	99.9	100.0				
		$\hat{\tau}_3$	70.1	82.4	93.7	89.7	93.1	98.4	98.9				
		$\hat{\tau}_4$	84.0	92.9	98.4	97.4	98.3	99.9	100.0				
		$\hat{\tau}_5$	69.7	82.0	93.5	89.5	92.8	98.3	98.9				
	0	$\hat{\tau}_1$	71.9	83.6	96.4	94.1	96.6	99.8	99.8	68.2	93.1	63.7	95.3
		$\hat{\tau}_2$	85.7	94.7	99.4	98.9	99.5	100.0	100.0				
		$\hat{\tau}_3$	73.3	84.7	95.8	94.1	96.2	99.6	99.8				
		$\hat{\tau}_4$	88.9	96.1	99.7	99.4	99.8	100.0	100.0				
		$\hat{\tau}_5$	73.2	84.4	95.6	93.8	96.0	99.6	99.8				
	1	$\hat{\tau}_1$	72.8	85.3	97.0	95.8	97.5	99.8	99.8	38.4	37.7	62.7	95.7
		$\hat{\tau}_2$	88.9	96.3	99.7	99.2	99.7	99.9	100.0				
		$\hat{\tau}_3$	75.5	87.3	97.5	95.5	97.7	99.5	99.9				
		$\hat{\tau}_4$	90.3	97.1	99.7	99.5	99.8	100.0	100.0				
		$\hat{\tau}_5$	75.2	87.0	97.5	95.5	97.6	99.5	99.9				
	2	$\hat{\tau}_1$	71.3	83.1	95.4	93.5	95.9	99.8	99.9	18.8	11.1	63.6	94.5
		$\hat{\tau}_2$	85.1	93.8	99.3	98.9	99.6	100.0	100.0				
		$\hat{\tau}_3$	72.4	84.8	96.0	93.9	96.5	99.5	99.7				
		$\hat{\tau}_4$	87.5	95.5	99.7	99.3	99.7	100.0	100.0				
		$\hat{\tau}_5$	72.0	84.2	96.0	93.8	96.5	99.5	99.7				
	3	$\hat{\tau}_1$	68.9	82.1	94.9	91.9	95.1	99.2	99.7	11.3	3.5	59.6	92.2
		$\hat{\tau}_2$	84.1	93.6	98.5	97.9	98.8	99.9	99.9				
		$\hat{\tau}_3$	70.3	82.8	94.9	92.1	95.4	99.3	99.6				
		$\hat{\tau}_4$	87.1	95.0	99.0	98.6	99.1	99.9	100.0				
		$\hat{\tau}_5$	69.9	82.5	94.6	91.9	95.2	99.3	99.6				
200	0	$\hat{\tau}_1$	69.6	83.1	95.9	95.6	97.2	99.6	99.9	69.7	94.2	66.1	96.8
		$\hat{\tau}_2$	88.5	96.2	99.3	99.2	99.3	100.0	100.0				
		$\hat{\tau}_3$	79.7	89.4	97.9	97.7	98.7	100.0	100.0				
		$\hat{\tau}_4$	88.4	96.1	99.3	99.1	99.3	100.0	100.0				
		$\hat{\tau}_5$	79.5	89.2	97.9	97.6	98.7	100.0	100.0				
	1	$\hat{\tau}_1$	72.1	85.9	96.7	96.6	97.9	99.9	99.9	60.4	59.4	67.8	97.3
		$\hat{\tau}_2$	91.6	96.9	99.6	99.6	99.8	100.0	100.0				
		$\hat{\tau}_3$	81.3	91.6	98.4	98.3	98.7	100.0	100.0				
		$\hat{\tau}_4$	91.5	96.9	99.6	99.6	99.8	100.0	100.0				
		$\hat{\tau}_5$	81.0	91.6	98.4	98.3	98.7	100.0	100.0				
	2	$\hat{\tau}_1$	71.7	84.0	95.4	95.1	97.5	99.9	99.9	48.8	39.2	66.9	95.5
		$\hat{\tau}_2$	89.3	96.1	99.4	99.3	99.9	100.0	100.0				
		$\hat{\tau}_3$	82.4	91.4	98.0	98.0	98.8	99.9	100.0				
		$\hat{\tau}_4$	89.2	96.0	99.3	99.3	99.9	100.0	100.0				
		$\hat{\tau}_5$	82.3	91.3	98.0	97.9	98.7	99.9	100.0				
	3	$\hat{\tau}_1$	72.3	84.4	96.6	96.2	97.9	100.0	100.0	41.9	19.5	64.1	96.7
		$\hat{\tau}_2$	90.2	96.9	99.9	99.8	100.0	100.0	100.0				
		$\hat{\tau}_3$	80.2	90.7	98.2	98.1	99.4	100.0	100.0				
		$\hat{\tau}_4$	89.7	96.9	99.9	99.7	100.0	100.0	100.0				
		$\hat{\tau}_5$	80.1	90.7	98.2	98.1	99.4	100.0	100.0				

AIC = like Akaike information criteria; BIC = Bayesian information criteria; RAIC = robust AIC; RBIC = robust BIC.

variables for Hald data. The 6^{th} observation has maximum absolute residual, to introduce an outlier by replacing the 6^{th} observation to 200 (Ronchetti and Staudte, 1994; Kashid and Kulkarni, 2002). The values of CR_p, AIC, BIC, RAIC, and RBIC for all possible subsets are recorded for original and outlier data in Tables 4 and 5. It is observed that the presence of outliers do not affect the value of CR_p. The CR_p criterion with all penalties selects X_1, X_2 variables for clean data as well as outlier data. The AIC criterion selects X_1, X_2, X_4 variables in clean data, and selects X_1, X_4 variables in the case of

Table 4: Hald Cement data (original)

Sr. No.	Submodel	CR_p							AIC	BIC	RAIC	RBIC
		P_1	P_2	P_3	P_4	P_5	P_6	P_7				
1	X_1	22.3272	24.3272	21.0997	23.4571	25.4571	25.5383	29.5383	102.4119	104.1067	12.2956	7.1850
2	X_2	17.3189	19.3189	16.0915	18.4488	20.4488	20.5300	24.5300	98.0704	99.7652	16.0992	9.1288
3	X_3	27.1334	29.1334	25.9060	28.2633	30.2633	30.3445	34.3445	107.9598	109.6547	**10.8062**	**6.7160**
4	X_4	17.4910	19.4910	16.2636	18.6209	20.6209	20.7021	24.7021	97.7440	99.4389	12.4251	7.4164
5	X_1, X_2	**6.9781**	**9.9781**	**7.5698**	**8.6730**	**11.6730**	**11.7948**	**17.7948**	64.3124	**66.5722**	15.5191	9.1283
6	X_1, X_3	26.1143	29.1143	26.7060	27.8091	30.8091	30.9309	36.9309	104.0091	106.2689	11.2779	7.7057
7	X_1, X_4	7.0266	10.0266	7.6183	8.7215	11.7215	11.8433	17.8433	67.6341	69.8939	12.1404	8.1164
8	X_2, X_3	14.2547	17.2547	14.8464	15.9496	18.9496	19.0714	25.0714	89.9295	92.1893	18.9980	10.7771
9	X_2, X_4	20.6127	23.6127	21.2043	22.3075	25.3075	25.4293	31.4293	99.5217	101.7815	16.9889	10.1520
10	X_3, X_4	9.7554	12.7554	10.3470	11.4502	14.4502	14.5720	20.5720	78.7450	81.0048	29.2977	14.4117
11	X_1, X_2, X_3	8.0552	12.0552	11.1455	10.3150	14.3150	14.4774	22.4774	63.9036	66.7283	15.3061	9.9429
12	X_1, X_2, X_4	8.2061	12.2061	11.2965	10.4659	14.4659	14.6283	22.6283	**63.8663**	66.6910	14.5849	9.6446
13	X_1, X_3, X_4	8.3405	12.3405	11.4308	10.6003	14.6003	14.7627	22.7627	64.6200	67.4447	11.8174	8.7352
14	X_2, X_3, X_4	8.9205	12.9205	12.0109	11.1803	15.1803	15.3427	23.3427	69.4683	72.2930	18.9425	10.9302
15	X_1, X_2, X_3, X_4	10.0000	15.0000	16.0944	12.8247	17.8247	18.0278	28.0278	65.8367	69.2264	18.6140	11.9214

AIC = like Akaike information criteria; BIC = Bayesian information criteria; RAIC = robust AIC; RBIC = robust BIC.

Table 5: Hald Cement data (with outlier, $y_6 = 200$)

Sr. No.	Submodel	CR_p							AIC	BIC	RAIC	RBIC
		P_1	P_2	P_3	P_4	P_5	P_6	P_7				
1	X_1	22.3272	24.3272	21.0997	23.4571	25.4571	25.5383	29.5383	129.1893	130.8842	**25.6844**	**14.0531**
2	X_2	17.3189	19.3189	16.0915	18.4488	20.4488	20.5300	24.5300	129.1579	130.8527	48.6441	25.1826
3	X_3	27.1334	29.1334	25.906	28.2633	30.2633	30.3445	34.3445	130.8619	132.5567	26.0959	14.0554
4	X_4	17.4910	19.4910	16.2636	18.6209	20.6209	20.7021	24.7021	128.9758	**130.6706**	35.3433	18.6954
5	X_1, X_2	**6.9781**	**9.9781**	**7.5698**	**8.6730**	**11.6730**	**11.7948**	**17.7948**	128.5246	130.7844	120.7875	61.7610
6	X_1, X_3	26.1143	29.1143	26.7060	27.8091	30.8091	30.9309	36.9309	131.0793	133.3391	28.2995	15.8754
7	X_1, X_4	7.0266	10.0266	7.6183	8.7215	11.7215	11.8433	17.8433	**128.4488**	130.7086	95.9106	49.7983
8	X_2, X_3	14.2547	17.2547	14.8464	15.9496	18.9496	19.0714	25.0714	129.7412	132.0010	66.9138	34.7348
9	X_2, X_4	20.6127	23.6127	21.2043	22.3075	25.3075	25.4293	31.4293	130.9744	133.2342	48.8001	25.7217
10	X_3, X_4	9.7554	12.7554	10.3470	11.4502	14.4502	14.5720	20.5720	128.9457	131.2055	122.8992	61.2152
11	X_1, X_2, X_3	8.0552	12.0552	11.1455	10.3150	14.3150	14.4774	22.4774	130.4785	133.3033	126.1412	65.3541
12	X_1, X_2, X_4	8.2061	12.2061	11.2965	10.4659	14.4659	14.6283	22.6283	130.4350	133.2597	121.2171	62.9619
13	X_1, X_3, X_4	8.3405	12.3405	11.4308	10.6003	14.6003	14.7627	22.7627	130.4121	133.2369	103.8295	54.7412
14	X_2, X_3, X_4	8.9205	12.9205	12.0108	11.1803	15.1803	15.3427	23.3427	130.3519	133.1767	117.7132	60.3156
15	X_1, X_2, X_3, X_4	10.0000	15.0000	16.0944	12.8247	17.8247	18.0278	28.0278	132.3519	135.7416	137.6374	71.4468

AIC = like Akaike information criteria; BIC = Bayesian information criteria; RAIC = robust AIC; RBIC = robust BIC.

an outlier. However, BIC selects X_1, X_2 variables in clean data, but in case of an outlier it selects only X_4 variable. RAIC and RBIC select same variable X_3 only in clean data, and X_1 only in presence of an outlier.

The selection of a model from all possible subsets will become more complicated and time consuming as the number of predictor variables increase. For example, if $k - 1 = 30$ then it is necessary to check more than a billion subsets for model selection. So, in this situation, it is reasonable to use a kick-off (Rao and Wu, 1989) or stepwise approach.

4. Algorithms for model selection

The kick-off method is based on an OLS estimator that is not robust to outliers in the data. To overcome this problem, we have modified the kick-off approach based on the LAD estimator for variable selection. The CR_p based kick-off method is explained below.

1. Kick-off method

   1) Calculate $D_{-i} = CR_{k-i} - C_n(k)$, where CR_{k-i} is the value of criterion corresponding to predictor variables excluding the i^{th} predictor variable and $C_n(k)$ penalty function of the full model.

   2) If $D_{-i} \leq 0$ then $\beta_i = 0$, else $\beta_i \neq 0$, $i = 1, 2, \ldots, k-1$. Hence, select predictor variables for which $D_{-i} > 0$.

   Alternative sequential and stepwise algorithms are described below. Let $\mathcal{S} \subseteq \mathcal{A} = \{1, 2, 3, \ldots, k-1\}$ is an index set of selected predictor variables. The sum of absolute residuals for $\mathcal{S} = \{\ \}$ is $|y - \text{Median}(y)|'\mathbf{1}$.

2. Sequential method

   1) Consider LAD estimator $\hat{\beta}$ of the full model, and using the statistical test explained by Birkes and Dodge (1993, pp. 76–77) to test the null hypothesis, $H_0 : \beta_{\{j:|\hat{\beta}_j| \leq \text{Median}(|\hat{\beta}_j|)\}} = 0$. If the null hypothesis is rejected at $\alpha\%$ level of significance, then repeat Steps 3.1–3.3 until we get final model. If null hypothesis is not rejected, then repeat Steps 2.1-2.3.

   2) Forward direction:

      2.1) Initially, consider $\mathcal{S} = \{\ \}$ null set.

      2.2) Add a new $j^{th} \in \mathcal{F} = \mathcal{S}^c \cap \mathcal{A}$ predictor variable to the previous set if $j = \arg\max_{j' \in \mathcal{F}}(CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j'\}))$ and $\mathcal{D}_j = CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j\}) > 0$ i.e., the difference $CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j\})$ is positive and large over all unselected predictor variables ($\mathcal{F}$).

      2.3) Repeat Step 2.2 until no other variable is selected.

   3) Backward direction:

      3.1) Initially, consider $\mathcal{S} = \mathcal{A}$.

      3.2) Delete $l^{th} \in \mathcal{S}$ predictor variable if $l = \arg\max_{l' \in \mathcal{S}}(CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l'))$ and $\mathcal{D}_l = CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l) \geq 0$ i.e., $CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l)$ is non-negative and large over all selected predictor variables ($\mathcal{S}$).

      3.3) Repeat Step 3.2 until no other variable is deleted.

3. Stepwise method

   1) Initially, consider $\mathcal{S} = \{\ \}$ null set.

   2) Add a new $j^{th} \in \mathcal{F}$ predictor variable to the previous set if $j = \arg\max_{j' \in \mathcal{F}}(CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j'\}))$ and $\mathcal{D}_j = CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j\}) > 0$.

      a) If any new predictor variable is not included in the null set $\mathcal{S} = \{\}$ or a singleton set, then stop.

      b) If $|\mathcal{S}| < 2$, then repeat same Step 2, else go to the next step. $|\mathcal{S}|$ is a cardinality of set $\mathcal{S}$.

   3) Delete $l^{th} \in \mathcal{S}$ predictor variable if $l = \arg\max_{l' \in \mathcal{S}}(CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l'))$ and $\mathcal{D}_l = CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l) \geq 0$ and go to Step 2.

   4) Continue Step 2 and Step 3 until consequent $\mathcal{S}$ does not change.

4.1. Addition and deletion criteria

- Addition: The new $j^{th} \in \mathcal{F}$ predictor variable is added to the previous set $\mathcal{S}$ if $\mathcal{D}_j > 0$ and maximum.

  i.e., $CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{j\}) > 0$ and maximum

  $\iff \frac{2}{\tau n}\Big(\big(n-k+|\mathcal{S}|\big)|y-\hat{y}_{\mathcal{S}}|'\mathbf{1} - \big(n-k+|\mathcal{S}|+1\big)|y-\hat{y}_{\mathcal{S}\cup j}|'\mathbf{1}\Big) + \frac{2}{\tau n}|y-\hat{y}_f|'\mathbf{1} + C_n(|\mathcal{S}|) - C_n(|\mathcal{S}|+1)$ maximum

  $\iff \frac{2}{\tau n}\big(n-k+|\mathcal{S}|\big)\Big(|y-\hat{y}_{\mathcal{S}}|'\mathbf{1} - |y-\hat{y}_{\mathcal{S}\cup j}|'\mathbf{1} - \frac{1}{n-k+|\mathcal{S}|}|y-\hat{y}_{\mathcal{S}\cup j}|'\mathbf{1}\Big)$ maximum

  $\iff \psi_1 = \frac{2}{\tau}\Big(|y-\hat{y}_{\mathcal{S}}|'\mathbf{1} - |y-\hat{y}_{\mathcal{S}\cup j}|'\mathbf{1}\Big)$ maximum

  Here, $\hat{y}_{\mathcal{S}}$ and $\hat{y}_{\mathcal{S}\cup j}$ are vectors of fitted values obtained from set of predictor variables corresponding to sets $\mathcal{S}$ and $\mathcal{S} \cup j$ respectively. The ψ_1 follows $F_{1,n-|\mathcal{S}|-2}$ distribution (Birkes and Dodge, 1993); therefore, select X_j if ψ_1 is maximum and $\psi_1 > F_{\alpha,1,n-|\mathcal{S}|-2}$.

- Deletion: We delete predictor variable X_l from existing set $\mathcal{S}$ if $\mathcal{D}_l \geq 0$ and maximum over all selected predictor variables.

  i.e., $CR_p(\mathcal{S}) - CR_p(\mathcal{S} - l)$ maximum

  $\iff CR_p(\mathcal{S} - l) - CR_p(\mathcal{S}) \leq 0$ and minimum

  $\iff \frac{2}{\tau n}\Big(\big(n-k+|\mathcal{S}|-1\big)|y-\hat{y}_{\mathcal{S}-l}|'\mathbf{1} - \big(n-k+|\mathcal{S}|\big)|y-\hat{y}_{\mathcal{S}}|'\mathbf{1}\Big) + \frac{2}{\tau n}|y-\hat{y}_f|'\mathbf{1} + C_n(|\mathcal{S}|-1) - C_n(|\mathcal{S}|)$ minimum

  $\iff \frac{2}{\tau n}\big(n-k+|\mathcal{S}|-1\big)\Big(|y-\hat{y}_{\mathcal{S}-l}|'\mathbf{1} - |y-\hat{y}_{\mathcal{S}}|'\mathbf{1} - \frac{1}{n-k+|\mathcal{S}|-1}|y-\hat{y}_{\mathcal{S}}|'\mathbf{1}\Big)$ minimum

  $\iff \psi_2 = \frac{2}{\tau}\Big(|y-\hat{y}_{\mathcal{S}-l}|'\mathbf{1} - |y-\hat{y}_{\mathcal{S}}|'\mathbf{1}\Big)$ minimum

  The ψ_2 follows $F_{1,n-|\mathcal{S}|-1}$ distribution (Birkes and Dodge,1993) and delete X_l if ψ_2 is minimum and $\psi_2 < F_{\alpha,1,n-|\mathcal{S}|-1}$.

Alternatively, we can select X_j if $\psi_1 > \chi^2_{\alpha,1}$ and delete X_l if $\psi_2 < \chi^2_{\alpha,1}$ for large n. Thus, the minimization and distribution based addition and deletion rules are equivalent.

Corollary 1. *The kick-off, sequential and stepwise algorithms select the optimal model with probability one for a large sample size.*

Proof: The proof is given separately for kick-off algorithm and other two algorithms.

- Kick-off method: If relevant predictor variable is deleted, then the reduced model belongs to $\mathcal{M}_w$. For the full model $D_p = 0$, and the full model belongs to $\mathcal{M}_c$. By (2.7),

$$\liminf_{n\to\infty} \Pr\left(D_{-i} > 0\right) = 1 \tag{4.1}$$

Similarly, if irrelevant predictor variable deleted, then the reduced model belongs to $\mathcal{M}_c$. By Condition 2 and $|y - \hat{y}_c|'\mathbf{1} = |y - \hat{y}_f|'\mathbf{1} = O_p(1)$,

$$\begin{aligned}\liminf_{n\to\infty} \Pr(D_{-i} < 0) &= \liminf_{n\to\infty} \Pr\left(CR_{k-i} - C_n(k) < 0\right) \\ &= \liminf_{n\to\infty} \Pr\left(O_p(1) + C_n(k-1) - C_n(k) < 0\right) \\ &\geq \Pr\left(\liminf_{n\to\infty} O_p(1) + C_n(k-1) - C_n(k) < 0\right) \\ &= 1. \end{aligned} \tag{4.2}$$

Hence, the kick-off method selects only relevant predictor variables with probability one for large n.

- Stepwise and sequential method:

  - Addition: Consider $r_1 \in \mathcal{F}$ and $r_2 \in \mathcal{F}$ are indices corresponding to the relevant and irrelevant predictor variable respectively. After adding r_1 in the present set $\mathcal{S}$, the value of $|y-\hat{y}|'\mathbf{1}$ is smaller than after adding r_2 in a set $\mathcal{S}$. It is equivalent to $|y - \hat{y}_{\mathcal{S}\cup\{r_2\}}|'\mathbf{1} > |y - \hat{y}_{\mathcal{S}\cup\{r_1\}}|'\mathbf{1}$ hold $\forall r_1, r_2 \in \mathcal{F}$. Since, $\text{card}(\mathcal{S} \cup \{r_1\}) = \text{card}(\mathcal{S} \cup \{r_2\}) = s_1$ (say)

$$\begin{aligned} &CR_p(\mathcal{S} \cup \{r_2\}) - CR_p(\mathcal{S} \cup \{r_1\}) \\ &= \frac{2}{\tau}\left(\left(1 - \frac{k-s_1}{n}\right)\left|y - \hat{y}_{\mathcal{S}\cup\{r_2\}}\right|' \mathbf{1} - |y - \hat{y}_{\mathcal{S}\cup\{r_1\}}|'\mathbf{1}\right) + \frac{2(k-s_1)}{\tau n}|y - \hat{y}_{\mathcal{S}\cup\{r_1\}}|'\mathbf{1} \\ &= \frac{2}{\tau}\left(1 - \frac{k-s_1}{n}\right)\left(|y - \hat{y}_{\mathcal{S}\cup\{r_2\}}|'\mathbf{1} - |y - \hat{y}_{\mathcal{S}\cup\{r_1\}}|'\mathbf{1}\right) \end{aligned} \tag{4.3}$$

  and

$$\begin{aligned} &\liminf_{n\to\infty} \Pr\left(CR_p(\mathcal{S} \cup \{r_2\}) > CR_p(\mathcal{S} \cup \{r_1\})\right) \\ &\geq \Pr\left(\liminf_{n\to\infty} \frac{2}{\tau}\left(1 - \frac{k-s_1}{n}\right)\left(\left|y - \hat{y}_{\mathcal{S}\cup\{r_2\}}\right|' \mathbf{1} - \left|y - \hat{y}_{\mathcal{S}\cup\{r_1\}}\right|' \mathbf{1}\right) > 0\right) \\ &= 1. \\ &\implies \liminf_{n\to\infty} \Pr\left(CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{r_1\}) > CR_p(\mathcal{S}) - CR_p(\mathcal{S} \cup \{r_2\})\right) = 1. \end{aligned} \tag{4.4}$$

  - Deletion: Suppose $r_3 \in \mathcal{S}$ and $r_4 \in \mathcal{S}$ are indices corresponding to the relevant and irrelevant predictor variable respectively. If we delete r_3 and r_4 from the present set $\mathcal{S}$, then $|y - \hat{y}_{\mathcal{S}-r_3}|'\mathbf{1} > |y - \hat{y}_{\mathcal{S}-r_4}|'\mathbf{1}$ hold $\forall r_3, r_4 \in \mathcal{S}$.
    Since, $\text{card}(\mathcal{S} - r_3) = \text{card}(\mathcal{S} - r_4) = s_2$ (say)

$$\begin{aligned} CR_p(\mathcal{S} - r_3) - CR_p(\mathcal{S} - r_4) &= \frac{2}{\tau}\left(\left(1 - \frac{k-s_2}{n}\right)\left|y - \hat{y}_{\mathcal{S}-r_3}\right|' \mathbf{1} - \left|y - \hat{y}_{\mathcal{S}-r_4}\right|' \mathbf{1}\right) + \frac{2(k-s_2)}{\tau n}\left|y - \hat{y}_{\mathcal{S}-r_4}\right|' \mathbf{1} \\ &= \frac{2}{\tau}\left(1 - \frac{k-s_2}{n}\right)\left(\left|y - \hat{y}_{\mathcal{S}-r_3}\right|' \mathbf{1} - \left|y - \hat{y}_{\mathcal{S}-r_4}\right|' \mathbf{1}\right) \end{aligned} \tag{4.5}$$

Table 6: Performance of algorithms in presence of outliers

n	% of	Kick-off method				Sequential method				Stepwise method			
(k/n)	outliers	P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7
	0%	82.5	91.4	100.0	100.0	96.3	96.3	96.4	96.4	99.9	99.9	100.0	100.0
	2%	68.8	80.8	99.7	100.0	94.4	94.6	94.6	94.6	99.8	100.0	100.0	100.0
200	4%	60.0	75.2	99.4	99.8	96.3	96.5	96.7	96.7	99.6	99.8	100.0	100.0
(1/2)	6%	56.8	71.2	98.8	99.3	94.7	95.7	96.0	95.9	98.7	99.7	100.0	99.9
	8%	57.7	72.2	99.4	99.9	93.8	93.9	94.0	93.7	99.8	99.9	100.0	99.7
	10%	58.4	73.1	98.4	97.9	94.0	94.4	94.2	92.9	99.6	100.0	99.8	98.3
	0%	85.4	93.7	100.0	100.0	98.8	99.3	99.6	99.6	99.2	99.7	100.0	100.0
	2%	79.9	87.9	100.0	100.0	98.2	98.9	98.9	98.9	99.3	100.0	100.0	100.0
300	4%	75.0	86.5	100.0	100.0	97.2	98.2	98.8	98.8	98.4	99.4	100.0	100.0
(1/3)	6%	67.6	80.5	100.0	100.0	97.4	98.4	99.4	99.4	98.0	99.0	100.0	100.0
	8%	65.9	79.3	99.9	100.0	97.6	99.1	99.6	99.6	98.0	99.5	100.0	100.0
	10%	61.2	75.6	100.0	100.0	97.3	98.6	99.5	99.5	97.8	99.1	100.0	100.0
	0%	90.2	95.6	100.0	100.0	98.4	99.2	99.8	99.8	98.6	99.4	100.0	100.0
	2%	83.9	92.1	100.0	100.0	97.8	99.3	100.0	100.0	97.8	99.3	100.0	100.0
400	4%	79.1	89.7	100.0	100.0	97.0	99.2	99.9	99.9	97.1	99.3	100.0	100.0
(1/4)	6%	75.6	85.9	100.0	100.0	96.3	98.2	99.7	99.7	96.5	98.5	100.0	100.0
	8%	70.1	83.2	100.0	100.0	97.0	98.2	99.7	99.7	97.2	98.5	100.0	100.0
	10%	66.0	79.2	100.0	100.0	95.3	98.2	99.7	99.7	95.6	98.5	100.0	100.0
	0%	93.7	97.6	100.0	100.0	98.5	99.4	100.0	100.0	98.5	99.4	100.0	100.0
	2%	88.3	95.3	100.0	100.0	97.7	99.0	100.0	100.0	97.7	99.0	100.0	100.0
500	4%	88.3	94.3	100.0	100.0	98.0	99.4	100.0	100.0	98.0	99.4	100.0	100.0
(1/5)	6%	83.1	89.5	100.0	100.0	96.5	98.7	100.0	100.0	96.5	98.7	100.0	100.0
	8%	77.6	88.3	100.0	100.0	96.0	99.2	100.0	100.0	96.0	99.2	100.0	100.0
	10%	75.1	85.8	100.0	100.0	94.0	97.5	100.0	100.0	94.0	97.5	100.0	100.0

and

$$\begin{aligned}
&\liminf_{n\to\infty} \Pr\Big(CR_p(\mathcal{S}-r_3) > CR_p(\mathcal{S}-r_4)\Big)\\
&\geq \Pr\left(\liminf_{n\to\infty} \frac{2}{\tau}\left(1-\frac{k-s_2}{n}\right)\left(\left|y-\hat{y}_{\mathcal{S}-r_3}\right|' \mathbf{1} - \left|y-\hat{y}_{\mathcal{S}-r_4}\right|' \mathbf{1}\right) > 0\right)\\
&= 1.\\
&\Longrightarrow \liminf_{n\to\infty} \Pr\Big(CR_p(\mathcal{S}) - CR_p(\mathcal{S}-r_3) < CR_p(\mathcal{S}) - CR_p(\mathcal{S}-r_4)\Big) = 1. \qquad (4.6)
\end{aligned}$$

◦ *Stopping*: By Theorem 1, if the present set S is an index set corresponding to optimal model then

$$\begin{aligned}
\lim_{n\to\infty} \Pr\Big(CR_p(\mathcal{S}) - CR_p(\mathcal{S}\cup\{r_2\}) < 0\Big) = 1, \quad \forall r_2 \in \mathcal{F},\\
\lim_{n\to\infty} \Pr\Big(CR_p(\mathcal{S}) - CR_p(\mathcal{S}-r_3) < 0\Big) = 1, \quad \forall r_3 \in \mathcal{S}.
\end{aligned} \qquad (4.7)$$

By (4.4), (4.6), and (4.7), the procedure of addition of relevant predictor variable ($r_1^{th} \in \mathcal{F}$) and deletion of irrelevant predictor variable ($r_4^{th} \in \mathcal{S}$) continue until getting optimal model, and the algorithms select the optimal model with probability one for large n. □

The CR_p, AIC, BIC, RAIC, and RBIC criteria requires computing $2^{k-1}-1$ criterion values to select the optimal model; however, the kick-off method needs to check only $k-1$ criterion values. In the sequential method, we fix the forward or backward direction to minimize time by using step 1; therefore, the sequential method requires $1+\sum_{i=\max(p_{\alpha_o},k-p_{\alpha_o}-1)}^{k-1} i$ criterion values, p_{α_o} is an actual

Table 7: Performance of algorithms in presence of non-normal errors

n	Distribution	Kick-off method				Sequential method				Stepwise Method			
(k/n)	of error	P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7
200 (1/2)	$N(0,3)$	81.8	88.7	89.1	80.6	95.1	95.1	79.7	62.9	99.8	99.8	82.1	64.0
	$0.95N(0,1)+0.05N(0,3)$	81.0	91.0	100.0	100.0	95.8	95.8	95.8	95.8	99.9	99.9	99.9	99.9
	$0.9N(0,1)+0.1N(0,3)$	82.4	90.2	99.9	99.9	96.8	97.0	97.0	96.9	99.8	100.0	100.0	99.9
	t_2	83.4	91.3	94.1	89.3	95.9	95.9	88.4	76.8	100.0	100.0	90.1	77.2
	Slash	20.0	15.2	0.4	0.2	64.3	47.6	0.5	0.4	70.8	50.6	0.4	0.4
	Cauchy $(0,1)$	59.9	54.5	7.9	4.8	87.1	82.2	10.3	4.6	97.3	91.6	10.5	4.1
	Laplace $(0,1)$	82.4	91.0	99.8	99.5	96.8	96.8	96.7	96.4	100.0	100.0	99.9	99.1
300 (1/3)	$N(0,3)$	88.4	94.3	100.0	100.0	98.5	99.0	99.4	99.4	99.1	99.6	100.0	100.0
	$0.95N(0,1)+0.05N(0,3)$	86.4	92.4	100.0	100.0	98.8	99.4	99.9	99.9	98.9	99.5	100.0	100.0
	$0.9N(0,1)+0.1N(0,3)$	85.0	91.4	100.0	100.0	98.9	99.4	99.6	99.6	99.3	99.8	100.0	100.0
	t_2	88.9	95.4	100.0	100.0	99.2	99.3	99.4	99.4	99.8	99.9	100.0	100.0
	Slash	88.4	93.8	33.8	23.3	98.0	98.0	58.5	39.7	100.0	100.0	56.8	37.0
	Cauchy $(0,1)$	87.6	94.1	94.2	90.4	97.1	97.1	95.5	92.5	100.0	100.0	98.3	95.0
	Laplace $(0,1)$	89.7	95.5	100.0	100.0	99.4	99.6	99.8	99.8	99.6	99.8	100.0	100.0
400 (1/4)	$N(0,3)$	91.5	96.5	100.0	100.0	98.7	99.2	99.9	99.9	98.8	99.3	100.0	100.0
	$0.95N(0,1)+0.05N(0,3)$	90.3	95.6	100.0	100.0	98.9	99.8	100.0	100.0	98.9	99.8	100.0	100.0
	$0.9N(0,1)+0.1N(0,3)$	90.4	95.2	100.0	100.0	98.8	99.7	99.9	99.9	98.9	99.8	100.0	100.0
	t_2	91.2	95.9	100.0	100.0	99.8	100.0	100.0	100.0	99.8	100.0	100.0	100.0
	Slash	92.4	97.0	93.3	89.2	99.7	99.7	98.3	95.9	100.0	100.0	98.3	95.7
	Cauchy $(0,1)$	91.8	95.5	100.0	100.0	99.9	99.9	99.9	99.9	100.0	100.0	100.0	100.0
	Laplace $(0,1)$	92.5	96.4	100.0	100.0	99.6	99.8	100.0	100.0	99.6	99.8	100.0	100.0
500 (1/5)	$N(0,3)$	92.9	97.2	100.0	100.0	98.5	99.4	100.0	100.0	98.5	99.4	100.0	100.0
	$0.95N(0,1)+0.05N(0,3)$	92.8	97.1	100.0	100.0	98.3	99.4	100.0	100.0	98.3	99.4	100.0	100.0
	$0.9N(0,1)+0.1N(0,3)$	92.6	97.0	100.0	100.0	98.7	99.8	100.0	100.0	98.7	99.8	100.0	100.0
	t_2	93.4	97.4	100.0	100.0	99.6	99.9	100.0	100.0	99.6	99.9	100.0	100.0
	Slash	95.3	98.2	99.7	99.4	99.7	99.8	99.9	99.9	99.8	99.9	100.0	100.0
	Cauchy $(0,1)$	93.9	97.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	Laplace $(0,1)$	93.3	97.2	100.0	100.0	99.6	99.8	100.0	100.0	99.6	99.8	100.0	100.0

number of relevant predictor variables. The stepwise method might be required more than sequential algorithm steps, but not more than $2^{k-1} - 1$. Hence, the stepwise method is more time-consuming when compared to the other two algorithms.

4.2. Performance and scalability of algorithms

4.2.1. Simulation study

The performance of LAD estimator based algorithms are studied through simulation. The predictor variables X_j, $j = 1, 2, \ldots, k-1$ are generated from $N(0, \Sigma)$. Here, Σ is a symmetric positive definite matrix such that $\Sigma_{ii} = 1$, $i = 1, 2, \ldots, k-1$ and $\Sigma_{ij} = 0.25$, $i \neq j = 1, 2, \ldots, k-1$. The errors are generated from $N(0, 1)$ distribution with the response variable generated using regression coefficients $\beta = (5, 2, \ldots, 2, 0, \ldots, 0)$, where 5 is intercept, and only 10 regression coefficients are non-zero in 100 coefficients. Therefore, only 10% predictor variables are significant in the simulated data. Outliers are introduced in the data by multiplying 20 to response variable y corresponding to maximum absolute residuals. The simulation is carried out for different sample sizes $n = 200, 300, 400, 500$, and the simulation results are recorded in Tables 6–8. In each table, we record the percentage of the optimal model selection in 1,000 runs considering only P_4–P_7 penalties. It is expected that the criterion selects only the first 10 predictor variables.

Table 8: Performance of algorithms in presence of multicollinearity

n (k/n)	Σ_{ij}	Kick-off method				Sequential method				Stepwise method			
		P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7	P_4	P_5	P_6	P_7
200 (1/2)	0.00	81.1	90.2	99.8	100.0	99.7	99.9	95.1	87.6	99.8	100.0	94.9	87.6
	0.50	80.5	89.0	99.9	99.9	87.8	87.9	87.9	87.9	99.9	100.0	100.0	99.8
	0.55	80.5	90.5	100	99.6	87.0	87.3	87.3	86.3	99.6	99.9	99.8	98.6
	0.60	81.5	90.2	99.7	99.3	85.0	85.1	84.8	84	99.9	100.0	99.6	98.5
	0.65	80.8	90.2	99.1	97.9	84.2	84.2	84.1	82.5	99.9	99.9	99.6	97.4
	0.70	79.8	88.7	97.7	93.7	81.7	81.8	80.5	76.6	99.8	99.9	98.5	92.6
	0.75	81.7	90.4	89.8	80.2	81.7	81.9	75.6	67.2	99.7	100.0	91.5	78.7
	0.80	80.5	89.6	71.4	57.5	76.3	76.4	61.7	47.4	99.8	99.9	75.9	55.7
	0.90	61.0	54.5	6.0	2.7	71.9	66.3	4.0	1.6	93.3	83.2	2.6	0.8
300 (1/3)	0.00	85.8	94.8	100.0	100.0	99.1	99.7	100.0	100.0	99.1	99.7	100.0	100.0
	0.50	86.2	92.8	99.9	100.0	97.3	98.1	98.2	98.2	99.1	99.9	100.0	100.0
	0.55	86.0	91.7	100.0	100.0	96.4	96.9	97.0	97.0	99.4	99.9	100.0	100.0
	0.60	86.6	93.8	100.0	100.0	96.2	96.8	96.9	96.9	99.2	99.8	100.0	100.0
	0.65	87.8	93.9	100.0	100.0	95.7	96.0	96.2	96.2	99.5	99.8	100.0	100.0
	0.70	87.3	93.8	100.0	100.0	95.7	95.8	96.2	96.2	99.4	99.6	100.0	100.0
	0.75	88.7	94.6	99.9	99.9	95.6	96.2	96.4	96.4	99.2	99.8	100.0	100.0
	0.80	85.5	92.9	99.6	99.1	92.6	93.9	94.1	94.0	98.4	99.8	100.0	99.9
	0.90	85.1	92.0	54.2	39.7	92.7	93.5	70.6	54.7	98.8	99.6	64.6	46.6
400 (1/4)	0.00	89.8	93.7	100.0	100.0	98.5	99.4	100.0	100.0	98.5	99.4	100.0	100.0
	0.50	91.5	96.6	100.0	100.0	98.8	99.5	99.8	99.8	99.0	99.7	100.0	100.0
	0.55	88.6	94.4	100.0	100.0	98.2	98.9	99.2	99.2	99.0	99.7	100.0	100.0
	0.60	92.0	95.9	100.0	100.0	97.6	98.9	99.1	99.1	98.5	99.8	100.0	100.0
	0.65	90.5	94.1	100.0	100.0	97.1	97.9	98.3	98.3	98.8	99.6	100.0	100.0
	0.70	91.1	96.4	100.0	100.0	98.1	98.6	99.1	99.1	99.0	99.5	100.0	100.0
	0.75	89.8	96.0	100.0	100.0	98.2	98.8	99.2	99.2	99.0	99.6	100.0	100.0
	0.80	88.5	94.1	100.0	100.0	97.5	98.6	98.8	98.8	98.7	99.8	100.0	100.0
	0.90	89.5	94.8	90.2	83.3	96.8	97.7	95.7	93.3	99.0	99.9	96.9	93.1
500 (1/5)	0.00	93.2	96.2	100.0	100.0	99.1	99.6	100.0	100.0	99.1	99.6	100.0	100.0
	0.50	93.7	97.4	100.0	100.0	98.6	99.7	100.0	100.0	98.6	99.7	100.0	100.0
	0.55	94.3	97.7	100.0	100.0	98.9	99.4	99.9	99.9	99.0	99.5	100.0	100.0
	0.60	92.7	96.2	100.0	100.0	98.0	99.2	99.8	99.8	98.2	99.4	100.0	100.0
	0.65	93.1	96.7	100.0	100.0	98.4	99.4	99.6	99.6	98.8	99.8	100.0	100.0
	0.70	94.6	97.7	100.0	100.0	99.0	99.7	100.0	100.0	99.0	99.7	100.0	100.0
	0.75	92.6	96.6	100.0	100.0	98.1	99.2	99.5	99.5	98.6	99.7	100.0	100.0
	0.80	92.9	96.2	100.0	100.0	98.3	99.1	99.4	99.4	98.9	99.7	100.0	100.0
	0.90	93.0	97.2	98.7	97.8	97.9	99.1	99.3	99.0	98.5	99.7	99.9	99.5

The simulation is carried out for 2%, 4%, 6%, 8%, 10% contamination of outliers in the data, and results are given in Table 6. The kick-off method performs poorly compared to sequential and stepwise methods. It is observed that the stepwise method performs well compared to the sequential method for the large k/n ratio; however, both methods perform equally for small k/n. The kick-off method is also performs well for $k/n = 1/5$, and selects the optimal model with at least 75% accuracy.

The performance of the algorithms for non-normal error distribution has been studied using the same model described above. We presented the simulation results (Table 7) with $N(0, 3)$, $0.95N(0, 1)+$ $0.05N(0, 3)$, $0.9N(0, 1) + 0.1N(0, 3)$, t_2, Slash, Cauchy $(0, 1)$, Laplace $(0, 1)$ error distributions. The sequential and stepwise method also performs reasonably well compared to the kick-off method in this case. All these algorithms have low percentages of model selection for large k/n and Slash, Cauchy distributions. However, the percentage of optimal model selection increases as the sample size increases for Slash and Cauchy error distributions. It is observed that the performance of criterion varies with a penalty function.

Table 9: Selected predictor variables in the body fat data

Criteria	Predictor Variables														
	Age (yrs)	Weight (lbs)	Height (inches)	Adiposity index (kg/m^2)	Fat Free Weight (lbs)	Circumference measure (cm)									
						Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
CRp Criterion															
CR_{p_4}		✓			✓										
CR_{p_5}		✓			✓										
CR_{p_6}		✓			✓										
CR_{p_7}		✓			✓										
AIC		✓		✓	✓		✓	✓		✓		✓	✓	✓	✓
BIC		✓		✓	✓		✓	✓		✓				✓	
RAIC			✓									✓		✓	✓
RBIC			✓									✓		✓	✓
Kick-Off algorithm															
KO_{p_4}		✓			✓										
KO_{p_5}		✓			✓										
KO_{p_6}		✓			✓										
KO_{p_7}		✓			✓										
Sequential algorithm															
S_{p_4}		✓			✓			✓							
S_{p_5}		✓			✓			✓							
S_{p_6}		✓			✓			✓							
S_{p_7}		✓			✓			✓							
Stepwise algorithm															
ST_{p_4}		✓			✓										
ST_{p_5}		✓			✓										
ST_{p_6}		✓			✓										
ST_{p_7}		✓			✓										

We also checked proficiency of the algorithms in the presence of multicollinearity (Table 8). The capability of these methods is assessed by varying $\text{cov}(X_i, X_j) = \Sigma_{ij}$ values. The high value of Σ_{ij} reveals severe multicollinearity. The percentage of the optimal model selection using algorithms decays quickly for larger values of Σ_{ij} and k/n. The algorithms select an optimal model with high precision up to the moderate multicollinearity. However, algorithms select optimal model for small k/n with high precision.

Thus, algorithms select an optimal model with a high percentage, and the performance mostly depends on the ratio k/n. The sequential method is faster than stepwise and performs well compared to the kick-off.

4.2.2. Body fat dataset

A body fat dataset is freely available in R software and contains physical measurements of 252 males. Measuring body fat is difficult compared to measuring height and weight. The physical measurements are more informative to get the percentage of body fat. The percentage of body fat is calculated using Bronzek's equation and density, and it is considered as the response variable. Another 15 variables mentioned in Table 9 are considered as predictor variables. It is observed that the data have outliers and residuals do not follow a Normal distribution (Figure 1). Consequently, the non-resistant model selection methods will not be appropriate for this dataset. The predictor variables selected by the proposed criterion with different penalties (CR_p), AIC, BIC, RAIC, RBIC, and other algorithms (Kick-Off (KO_p), Sequential (S_p), and Stepwise (ST_p)) are indicated in Table 9. The CR_p, KO_p, ST_p selects only two predictor variables, weight and fat-free weight; however, a sequential method selects one more predictor variable abdomen circumference. However, the AIC, BIC, RAIC, and RBIC select different predictor variables. For a detailed study, we compared the prediction error of these methods.

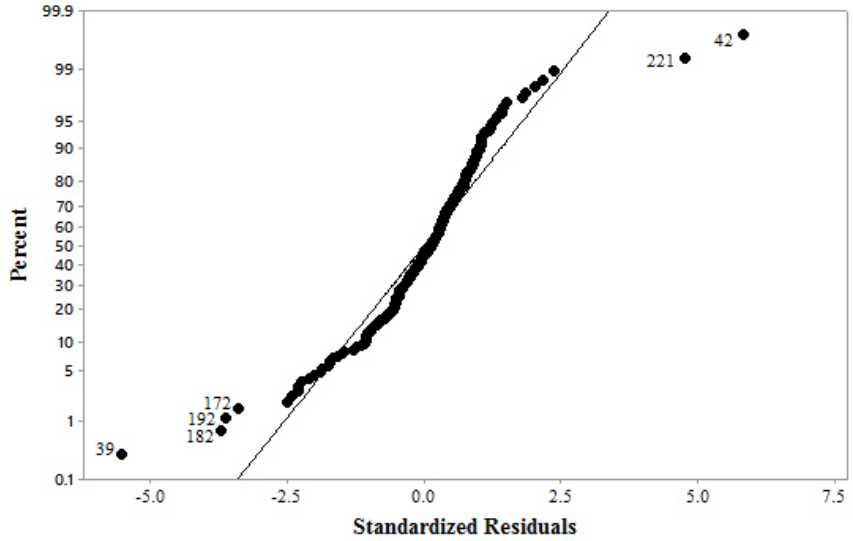

Figure 1: *Normal probability plot.*

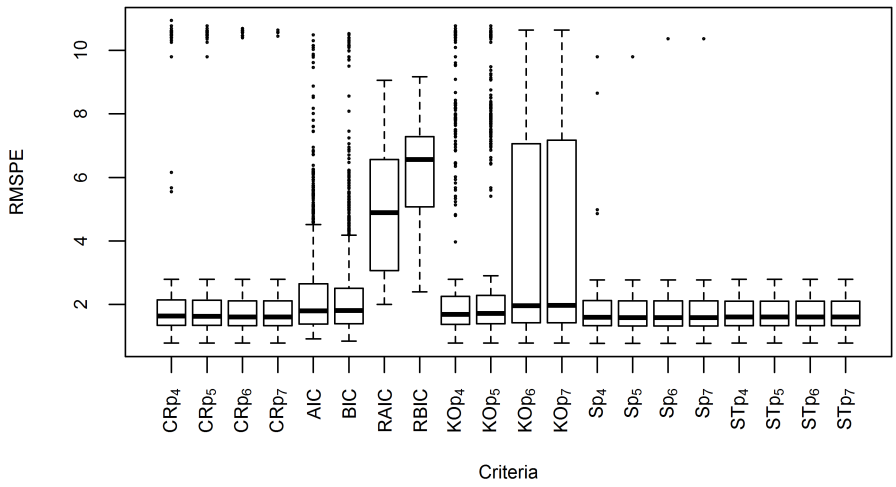

Figure 2: *Box plot of root mean square prediction error.*

The 70% (176) observations are randomly chosen to select the significant predictor variables and the remaining 30% (76) observations are used to calculate prediction error using selected predictor variables and their LAD estimator. The performance of the proposed criterion with different penalties (CR_p), AIC, BIC, RAIC, RBIC, and other algorithms (Kick-Off (KO_p), Sequential (S_p), and Stepwise

(ST_p)) has been examined by the root mean square prediction error (RMSPE) = $\sqrt{\sum_{i=1}^{76}(y_i - \hat{y}_i)^2/76}$. This procedure is repeated 1,000 times. In Figure 2, boxplots of RMSPE for different methods are plotted. It is observed that CR_p criterion, sequential and stepwise methods have a small RMSPE with low variation. However, the kick-off method with all penalties has a small RMSPE with a small variation excluding P_6 and P_7. The RMSPE of the model selected by AIC and BIC is smaller compared to RAIC and RBIC. The RMSPE of RAIC and RBIC indicate that both criteria do not select a good model for this dataset. Thus, the real-life example reveals the scalability and stability of algorithms.

5. Discussion

We have studied a robust model selection method for a class of different penalties. It is observed that the criterion with the penalty satisfying Condition 2 performs well. It is shown that the model selection criterion is consistent. The CR_p criterion is time consuming when the number of predictor variables (k) increases. LAD estimator-based algorithms will be the best option to overcome this problem. These algorithms work well for outlier data as well as the non-normality of the error term. The time required to select an optimal model for these algorithms is less than searching all possible subsets; consequently, the sequential method is preferable. Criterion based algorithms are therefore shows to have advantages such as robustness, consistency and fast.

Acknowledgement

We sincerely thank the Editor, Associate Editor and anonymous reviewers for their careful review and constructive comments which led to the significant improvement of this article.

References

Akaike H (1973). Information theory and an extension of maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267–281.

Birkes D and Dodge Y (1993). *Alternative Methods of Regression*, Wiley, New York.

Dielman TE (2005). Least absolute value regression: recent contributions, *Journal of Statistical Computation and Simulation*, **75**, 263–286.

Dielman TE (2006). Variance estimates and hypothesis tests in least absolute value regression, *Journal of Statistical Computation and Simulation*, **76**, 103–114.

Gilmour SG (1995). The interpretation of Mallows's C_p-statistic, *Journal of the Royal Statistical Society, Series D (The Statistician)*, **45**, 49–56.

Kashid DN and Kulkarni SR (2002). A more general criterion for subset selection in multiple linear regression, *Communications in Statistics - Theory and Methods*, **31**, 795–811.

Kim C and Hwang S (2000). Influence subsets on the variable selection, *Communication in Statistics-Theory and Methods*, **29**, 335–347.

Machado JAF (1993). Robust model selection and M-estimation, *Econometric Theory*, **9**, 478–493.

Mallows C (1973). Some comment on C_p, *Technometrics*, **15**, 661–675.

Rao CR and Wu Y (1989). A strong consistent procedure for model selection in a regression model, *Biometrika*, **76**, 369–374.

Rao C, Wu Y, Konishi S, *et al.* (2001). On model selection, *Lecture Notes-Monograph Series*, **38**, 1–64.

Ronchetti E (1985). Robust model selection in regression, *Statistics and Probability Letters*, **3**, 21–23.
Ronchetti E and Staudte RG (1994). A robust version of Mallows's C_p, *Journal of the American Statistical Association*, **89**, 550–559.
Schwarz G (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.
Siniksaran E (2008). A geometric interpretation of Mallows' C_p statistic and an alternative plot in variable selection, *Computational Statistics and Data Analysis*, **52**, 3459–3467.
Tharmaratnam K and Claeskens G (2013). A comparison of robust versions of the AIC based on M, S and MM-estimators, *Statistics: A Journal of Theoretical and Applied Statistics*, **47**, 216–235.
Yamashita T, Yamashita K, and Kamimura, R (2007). A stepwise AIC method for variable selection in linear regression, *Communication in Statistics-Theory and Methods*, **36**, 2395–2403.

Received November 28, 2018; Revised February 20, 2019; Accepted March 20, 2019