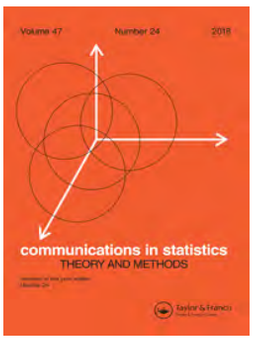

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Model selection in linear regression using paired bootstrap

Fazli Rabbi, Salahuddin Khan, Alamgir Khalil, Wali Khan Mashwani, Muhammad Shafiq, Pınar Göktaş & Yuksel.Akay Unvan

To cite this article: Fazli Rabbi, Salahuddin Khan, Alamgir Khalil, Wali Khan Mashwani, Muhammad Shafiq, Pınar Göktaş & Yuksel.Akay Unvan (2020): Model selection in linear regression using paired bootstrap, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2020.1725829

To link to this article: https://doi.org/10.1080/03610926.2020.1725829

Published online: 10 Feb 2020.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2020.1725829

Model selection in linear regression using paired bootstrap

Fazli Rabbi[a], Salahuddin Khan[b], Alamgir Khalil[a], Wali Khan Mashwani[c], Muhammad Shafiq[c], Pınar Göktaş[d], and Yuksel.Akay Unvan[e]

[a]Department of Statistics, University of Peshawar, Peshawar, Pakistan; [b]CECOS University of IT and Emerging Sciences, Hayatabad, Pakistan; [c]Institute of Numerical Sciences, Kohat University of Science& Technology, Kohat, Pakistan; [d]Department of Strategy Development, Muğla Sıtkı Koçman University, Muğla, Turkey; [e]Ankara Yildirim Beyazit University, Ankara, Turkey

ABSTRACT

Model selection is an important and challenging problem in statistics. The model selection is inevitable in a large number of applications including life sciences, social sciences, business, or economics. In this article, we propose a resampling-based information criterion called paired bootstrap criterion (PBC) for model selection. The proposed criterion is based on minimizing the conditional expected prediction loss for selecting the best subset of variables. We estimate the conditional expected prediction loss by using the out-of-bag (OOB) bootstrap approach. Other classical criteria for model selection such as AIC, BIC are also presented for comparison purpose. We demonstrate that the proposed paired bootstrap model selection criterion is effective in selecting accurate models via real and simulated data examples. The results confirm the satisfactory behavior of the proposed model selection criterion to select parsimonious models that fit the data well. We apply the proposed methodology to a real data example.

ARTICLE HISTORY
Received 7 December 2019
Accepted 30 January 2020

KEYWORDS
Residual bootstrap; paired bootstrap; model selection; prediction loss; out-of-bag bootstrap; OOB error

1. Introduction

Regression analysis is the most generally used procedure to demonstrate the relationship between a response variable and a set of predictors. When performing a linear regression on a set of observations, usually p predictor variables are available for predicting a response variable $\boldsymbol{y}$, and one has the desire to select the best subset of these predictor variables. This selected model may contain all possible p explanatory variables or may contain only a subset p_α where $\alpha \in A$ and A is the set of all possible models being examined. Working with the largest number of explanatory variables that explains the most variability in the observations does not automatically produce the best model. We should instead use a systematic process for model selection to determine which model best explains the data. Model selection is a basic issue in statistics which helps to identify the set of significant predictors which explain the response variable well.

Several model selection procedures have been suggested for the least squares linear regression model. The most widely used selection procedures are forward, backward,

CONTACT Wali Khan Mashwani walikhan@kust.edu.pk Institute of Numerical Sciences, Kohat University of Science & Technology, House 189 /D3, Street 17, Phase 1, Peshawar 25000, Pakistan.

stepwise, and best subsets regression. Selection criteria for these procedures are often based on R^2, adjusted-R^2, F test statistics (F-to-enter and F-to-remove), Mallow's C_p criterion (Mallows 1973), and the final prediction error (FPE) (Akaike 1970; Shibata 1984). Unfortunately, all of these selection criteria are biased and are, therefore, not recommended for variable selection by researchers, for example, (see Breiman 1995; Davison and Hinkley 1997; Miller 1990; Shao 1993; Wisnowski et al. 2003; Zhang 1992). Direct minimization of these criteria leads to models that have too many significant variables, suggesting that the dimension of the active variable set ($< \boldsymbol{p}$) is too large. Shao (1993, 1996) and Breiman (1995) proposed different resampling procedures to address the limitations of the traditional methods for least-squares subset model selection. These authors used the resampling procedures such as the bootstrap and crossvalidation to estimate the prediction error. A model having a minimum value for prediction error is considered as the correct one. Some other good overviews based on the resampling techniques to model selection are Sauerbrei (1999), Sauerbrei, Boulesteix, and Binder (2011), Lee, Babu, and Rao (2012), Babu (2011), Arlot (2009), De Bin et al. (2016).

Shao (1996) bootstrap procedure in its original form is an n-out-of-n bootstrap, the first n refers to the number of observations to take out as a bootstrap sample and the second n refers to the number of original observations. Shao (1996) procedure is asymptotically equivalent to the Akaike Information Criterion (AIC) (Akaike 1974), Mallows C_p criterion, and leave-one-out crossvalidation selection technique. These all tools share the same property of being asymptotically inconsistent. The bootstrap selection technique is inconsistent in the sense that the probability of selecting the optimal subset of variables does not converge to 1 as$n \to \infty$. To obtain asymptotic consistency, Shao (1996) treats the issue through an m-out-of-n bootstrap for an appropriately chosen $m < n$ (where m refers to the bootstrap sample and n refers to number of original observations).

The Shao (1996) bootstrap procedure for model selection is strongly depends on bootstrap sample m. So, the key strength driving this research is to improve the Shao (1996) criterion which is less dependent on m. We pursue the investigation in Shao (1996) and make some refinements, by utilizing the concept of out-of-bag (OOB) bootstrap. The OOB observations are those which are not a 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 (Breiman 2001; Zhang, Zhang, and Zhang 2010). We believe that our proposal will provide a consistent procedure to be used for model selection in linear regression problems.

This article is organized as follows. Section 2 considers the linear relationship between x and y, bootstrapping in the regression model and the two distinctive methods for generating bootstrap samples: residuals bootstrapping and pairs bootstrapping. Section 3 discusses the Bootstrap estimate of the expected prediction loss. Section 4 illustrates the existing bootstrap model selection criterion. Section 5 presents the proposed paired bootstrap criterion for model selection. Section 6 discusses our simulation results. Section 7 demonstrates the data example. Finally, Section 8 summarizes our conclusion.

2. Linear regression model

Suppose that we have a vector of n responses $y = (y_1, y_2, \ldots, y_n)^T$. Also, we have p explanatory variables for each observed response contained in a vector X_i. Let X be an

$n \times p$ matrix with full rank, and let β be a vector of p unknown regression parameters. Then the linear regression model between Y and X is

$$y_i = X^T\beta + \varepsilon \tag{1}$$

where ε is an n-dimensional vector of location zero and scale one errors. Moreover, X and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ are independent of each other.

2.1. Bootstrapping in regression model

The bootstrap procedures can be easily extended to linear regression models. There are many articles and books available describing the procedure and its application. In particular, applying the bootstrap to regression models is covered in Freedman (1981), Bunke and Droge (1984), and Shao (1996). Two different approaches are used for generating the bootstrap sample observations in linear regression models, including residual bootstrap (Efron 1979) and paired bootstrap (Efron 1982). We present brief details of these procedures in the following subsections.

2.1.1. Residual bootstrapping

Let $\hat{Y} = X^T\hat{\beta}$ is the fitted values and $\hat{\beta}$ is the least squares regression coefficients. Suppose $e_i = y_i - \hat{y}_i$ is the ith residual calculated from an original sample. Generate bootstrap observations y_i^* by using $y_i^* = \hat{y}_i + e_i^*$ for $i = 1, 2, ..., n$ where e_i^* are the bootstrap residuals selected from e_i. The residual bootstrap samples are $\{(x_i, y_i^*),$ where $i = 1, 2, ..., n\}$. The bootstrap estimate of $\hat{\beta}$ is given by

$$\hat{\beta}^* = (X^TX)^{-1}X^Ty^*$$

where $Y^* = (y_1^*, y_2^*, ..., y_n^*)$. The residual bootstrap is generally used when the explanatory variables x_i are deterministic. In this case, they are assumed to be fixed and nonrandom, and so the only variability in y_i is attributed to the bootstrapped errors e_i^*.

2.1.2. Paired bootstrap

In the paired bootstrap, we produce the pairs (response, explanatory variable) bootstrap samples by sampling n observations from $\{(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)\}$ with replacement and having equal selection probability. Then, the bootstrap sample is (y_i^*, x_i^*) for $i = 1, 2, ..., n$. The bootstrap estimate of $\hat{\beta}$ is given by

$$\hat{\beta}^* = (X^{*T}X^*)^{-1}X^{*T}y^*$$

where $y^* = (y_1^*, y_2^*, ..., y_n^*)$ and $X^* = (x_1^*, x_2^*, ..., x_n^*)$. A paired bootstrap is often used when the explanatory variables x_i are considered to be random, although the method can also be used when x_i are deterministic.

3. Bootstrap estimate of the expected prediction loss

Suppose that we have a response vector $y = (y_1, y_2, ..., y_n)^T$ and X be an $n \times p$ matrix. Let α denote any subset of size p_α from $\{1, 2, ..., p\}$, β_α is the subvector of β, and let X_α denote

the $n \times p_\alpha$ matrix that contains n observations (rows) and only the p_α explanatory variables (columns). Let $x_{\alpha i}^T$ denote the ith row vector of the matrix X_α. Then model α is given by

$$y_i = x_{\alpha i}^T \beta_\alpha + \varepsilon_{\alpha i}, \qquad i = 1, 2, ..., n \tag{2}$$

where $\varepsilon_{\alpha i}$'s are mean-zero and scale one errors. Moreover, X_α and $\varepsilon_\alpha = (\varepsilon_{\alpha 1}, \varepsilon_{\alpha 2}, ..., \varepsilon_{\alpha n})^T$ are independent of each other.

To fit model (2), the least squares procedure is used. The least squares estimate of β_α is

$$\hat{\beta}_\alpha = (X_\alpha^T X_\alpha)^{-1} X_\alpha^T y$$

Note that model (2) is said to be correct model if $E(y_i/x_{i\alpha}) = x_{i\alpha}^T \beta_\alpha$, i.e., β_α contains all non-zero components of β. However, if a model with parameter β_α is not a correct model, then $E(y_i/x_{i\alpha}) \neq x_{i\alpha}^T \beta_\alpha$, since $E(\hat{\beta}_\alpha)$ will not be the same as the non-zero components of β. We can measure the dissimilarity of the model α and the full model by the loss which is given by

$$l(\alpha) = \frac{1}{n} \sum_{i=1}^{n} (x_i^T \beta - x_{i\alpha}^T \hat{\beta}_\alpha)^2 \tag{3}$$

Suppose, we have n future responses z_i that are independent of the past responses, y_i but with the same explanatory variables X_i for $i = 1, 2, ..., n$. Then the average conditional expected prediction loss (EPL) is

$$L(\alpha) = E\left[\frac{1}{n} \sum_{i=1}^{n} (z_i - x_{i\alpha}^T \hat{\beta}_\alpha)^2 | Y, X\right] \tag{4}$$

$$L(\alpha) = E\left[\frac{1}{n} \sum_{i=1}^{n} [(z_i - x_i^T \beta) + (x_i^T \beta - x_{i\alpha}^T \hat{\beta}_\alpha)]^2\right]$$

$$L(\alpha) = \sigma^2 + l(\alpha) \tag{5}$$

where $var(z_i/x_i) = \sigma^2$.

Initially, the bootstrap estimate of the Expected Prediction Loss (EPL) is derived by Efron (1982, 1983) using n-out-of-n bootstrap procedure. The suggested bootstrap estimate of $L(\alpha)$ in (4) is given by

$$L^*(\alpha) = \frac{\|Y - X_\alpha \hat{\beta}_\alpha\|^2}{n} + e_n^*(\alpha) \tag{6}$$

where $e_n^*(\alpha)$ is the bootstrap estimate of expected excess error for model α given by

$$e_n^*(\alpha) = E_*\left[\frac{\|Y - X_\alpha \hat{\beta}_\alpha^*\|^2}{n} - \frac{\|Y^* - X_\alpha^* \hat{\beta}_\alpha^*\|^2}{n}\right] \tag{7}$$

where E_* is the expectation with respect to the bootstrap sample and $\hat{\beta}_\alpha^*$ is the bootstrap estimator of $\hat{\beta}_\alpha$. Almost this estimator $L_n^*(\alpha)$ is unbiased, but a straightforward n-out-of-n bootstrap is asymptotically inconsistent for regression models (Shao 1996). A simple modification by Shao (1996) to an m-out-of-n selection procedure rectified this consistency condition.

4. Existing bootstrap model selection criterion

In this section, we discuss the existing model selection procedure based on expected prediction loss. Consider a vector of n responses $y_i = (y_1, y_2, ..., y_n)^T$ and the design matrix $X = (x_1, x_2, ..., x_n)^T$.

Shao (1996) estimated the average conditional expected prediction loss [defined in (4)] by using an m-out-of-n bootstrap. In bootstrapping pairs, obtaining a consistent estimate is a simple matter of using m pairs of observations (y_i, x_i) for $i = 1, 2, ..., m$ selected from the full set of n observations. The m-out-of-n bootstrap estimate of $\hat{\beta}_\alpha$ based on the model α is given by

$$\hat{\beta}^*_{\alpha,m} = \left[\sum_{i=1}^{m} x^*{}_{i\alpha} x^{*T}_{i\alpha}\right]^{-1} \sum_{i=1}^{m} x_{i\alpha} y^*{}_{i\alpha} \tag{8}$$

The corresponding bootstrap estimate of the expected prediction loss proposed by Shao (1996) is given by

$$L^*_n(\alpha) = E_* \left[\frac{\|Y - X^T_\alpha \hat{\beta}^*_{\alpha,m}\|^2}{n}\right] \tag{9}$$

where E_* is the expectation with respect to the bootstrap sample and $\hat{\beta}^*_{\alpha,m}$ is the bootstrap estimator of $\hat{\beta}_\alpha$. Here, the focus is on the model $\hat{\alpha}^s_{n,m} \in$ A that minimizes $L^*_n(\alpha)$ i.e.,

$$\hat{\alpha}^s_{m,n} = \underset{\alpha \in \mathrm{A}}{argmin}\ L^*_n(\alpha) \tag{10}$$

5. The proposed model selection criterion

In this section, we present a paired bootstrap model selection criterion based on modified expected prediction loss. To estimate the modified expected prediction loss we make some refinements in Shao (1996), by utilizing the concept of out-of-bag bootstrap. Following Shao (1996), we use an m-out-of-n bootstrapping method rather than traditional methods to obtain asymptotic consistency. To estimate the modified expected prediction loss, we proceed as follows:

(i) sample rows of (y, X) independently with replacement so that total bootstrap sample is of size m ($\leq n$),
(ii) construct the estimator $\hat{\beta}^*_{\alpha,m}$ from data obtained in step (i),
(iii) calculate the modified criterion function by using the out-of-bag bootstrap expectation i.e., m observations used to obtain $\hat{\beta}^*_{\alpha,m}$, are not included when calculating $L^{**}_n(\alpha)$,
(iv) repeat the steps (i) to (iii) K independent times and then estimate the modified expected prediction loss by

$$L^{**}_n(\alpha) = E_* \left[\frac{\|Y_{[-m]} - X^T_{\alpha[-m]} \hat{\beta}^*_{\alpha,m}\|^2}{n-m}\right] \tag{11}$$

where E_* denotes expectation with respect to the bootstrap distribution and m is the number of distinct observations in the bootstrap sample, and $[-m]$ denotes the m observations are excluded when calculating $L_n^{**}(\alpha)$. As in Müller and Welsh (2005, 2009), we suggest to take the bootstrap sample size m in between 0.25 n to 0.50 n for moderate n i.e., 50 to 200, but for large n, m can be smaller than 0.25 n. Moreover, m satisfies the conditions given by

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

In practice, the interest lies in all of the models that make $L_n^{**}(\alpha)$ small. By using the modified bootstrap criterion function, we select a model $\hat{\alpha}_{m,n}^{f} \in$ A that minimizes $L_n^{**}(\alpha)$, i.e.,

$$\hat{\alpha}_{m,n}^{f} = \underset{\alpha \in \mathrm{A}}{\operatorname{argmin}} \; L_n^{**}(\alpha) \tag{12}$$

Here, we prefer paired bootstrapping over residual bootstrapping because the former can be used in both situations, i.e., either the explanatory variables X_i are random or deterministic whereas the later can be used only when the explanatory variables X_i are deterministic (Efron 1982).

6. Simulation study

To perform simulations, we may use a real dataset with known explanatory variables (in Simulation Setting 1) or we may generate our own hypothetical dataset with known parameter coefficients (in Simulation Setting 2). In the following subsections, the finite-sample performance of the proposed criterion is compared with existing model selection procedures via MC simulation and real dataset.

6.1. Simulation setting 1

To compare the finite-sample performance of the proposed bootstrap model selection criterion with the existent procedure suggested by Shao (1996), the classical AIC and the BIC (Schwarz 1978), we use the solid waste data of Gunst and Mason (1980), as used in Shao (1993, 1996, 1997); Wu (2001), Wisnowski et al. (2003), Müller and Welsh (2005), and Salibian-Barrera and Van Aelst (2008) in the context of model selection. Consider the following model with $\boldsymbol{p} = 5$ predictors and sample size $n = 40$,

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i, \qquad i = 1, 2, \ldots, 40 \tag{13}$$

where ε_i are iid standard normal errors. The first component of each X_i is 1 and the values of other components of X_i are taken from the solid waste data example of Gunst and Mason (1980). Following Shao (1996), we generate bootstrap samples from the model given by Equation (9). We apply the two model selection procedures to choose a model from a pre-specified list. To show a better performance for any model selection procedure, the sample size n must be increased if the ratio of a component of β over standard deviation σ is too small (i.e., < 2) (Shao 1996). The estimated selection probabilities, for the existing bootstrap estimator $\hat{\alpha}_{m,n}^{s}$[defined in Equation (10)] and the proposed bootstrap estimator $\hat{\alpha}_{m,n}^{f}$[defined in Equation (12)] are computed for various m

Table 1. Selection probabilities of $\hat{\alpha}^{s}_{m,n}$ and $\hat{\alpha}^{f}_{m,n}$ based on simulation setting 1.

Trueβ	Model	$\hat{\alpha}^{s}_{15,40}$	$\hat{\alpha}^{f}_{15,40}$	$\hat{\alpha}^{s}_{20,40}$	$\hat{\alpha}^{f}_{20,40}$	$\hat{\alpha}^{s}_{25,40}$	$\hat{\alpha}^{f}_{25,40}$	$\hat{\alpha}^{s}_{30,40}$	$\hat{\alpha}^{f}_{30,40}$	$\hat{\alpha}^{s}_{40,40}$	$\hat{\alpha}^{f}_{40,40}$	AIC	BIC
(2,0,0,4,0)	**1,4***	**0.943**	**0.972**	**0.875**	**0.943**	**0.770**	**0.903**	**0.673**	**0.864**	**0.479**	**0.799**	**0.583**	**0.835**
	1,4,5	0.010	0.006	0.024	0.014	0.042	0.023	0.054	0.036	0.084	0.046	0.106	0.046
	1,3,4	0.019	0.010	0.050	0.014	0.100	0.038	0.138	0.049	0.190	0.080	0.105	0.046
	1,2,4	0.028	0.012	0.046	0.029	0.069	0.034	0.090	0.043	0.128	0.060	0.107	0.057
	1,3,4,5	0.000	0.000	0.001	0.000	0.005	0.001	0.016	0.002	0.034	0.004	0.027	0.004
	1,2,4,5	0.000	0.000	0.001	0.000	0.004	0.000	0.009	0.002	0.022	0.002	0.027	0.009
	1,2,3,4	0.000	0.000	0.003	0.000	0.008	0.001	0.016	0.004	0.041	0.009	0.024	0.003
	1,2,3,4,5	0.000	0.000	0.000	0.000	0.002	0.000	0.004	0.000	0.022	0.000	0.021	0.000
(2,0,0,4,8)	**1,4,5***	**0.965**	**0.978**	**0.907**	**0.948**	**0.838**	**0.910**	**0.765**	**0.888**	**0.607**	**0.832**	**0.694**	**0.877**
	1,3,4,5	0.013	0.007	0.043	0.019	0.077	0.041	0.119	0.052	0.199	0.080	0.124	0.054
	1,2,4,5	0.022	0.015	0.048	0.031	0.071	0.045	0.094	0.055	0.135	0.073	0.135	0.063
	1,2,3,4,5	0.000	0.000	0.002	0.002	0.014	0.004	0.022	0.005	0.059	0.015	0.047	0.006
(2,9,0,4,8)	1,4,5	0.013	0.022	0.002	0.012	0.000	0.000	0.000	0.007	0.000	0.003	0.000	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
	1,3,4,5	0.001	0.003	0.004	0.005	0.004	0.005	0.002	0.004	0.002	0.006	0.000	0.001
	1,2,4,5*	**0.976**	**0.966**	**0.956**	**0.966**	**0.916**	**0.942**	**0.872**	**0.928**	**0.778**	**0.902**	**0.827**	**0.934**
	1,2,3,4,5	0.009	0.007	0.038	0.017	0.080	0.044	0.126	0.061	0.220	0.089	0.173	0.065
(2, 4, 6, 8, 9)	1,3,4,5	0.071	0.097	0.015	0.032	0.008	0.018	0.003	0.013	0.002	0.012	0.000	0.001
	1,2,4,5	0.010	0.020	0.000	0.003	0.001	0.003	0.000	0.001	0.000	0.000	0.000	0.000
	1,2,3,5	0.011	0.014	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	1,2,3,4,5*	**0.908**	**0.869**	**0.985**	**0.964**	**0.991**	**0.979**	**0.997**	**0.986**	**0.998**	**0.988**	**1.000**	**0.999**

Note: (*) denote the optimal model.

using $L = 1000$ Monte Carlo (MC) simulations with bootstrap replications of $K = 100$, are tabulated in Table 1.

The results in Table 1 can be summarized as follows:

- The modified bootstrap selection procedure outperforms the existing bootstrap selection procedure, the AIC and BIC. For example, for $\beta = (2, 0, 0, 4, 0)$ we see that $\hat{\alpha}^{f}_{15,40}$ selects the optimal model 97.2% ($\text{sd}_{0.972} = 0.005$), $\hat{\alpha}^{s}_{15,40}$ selects the optimal model 94.3% ($\text{sd}_{0.943}$=0.007), the AIC selects the optimal model 58.3% ($\text{sd}_{0.583} = 0.016$) and the BIC selects the optimal model 83.5% ($\text{sd}_{0.835}$=0.012).
- The modified bootstrap selection procedure clearly improves for smaller m. For example, for $\beta = (2, 0, 0, 4, 8)$, we see that $\hat{\alpha}^{f}_{40,40}$ selects the optimal model, 83.2% of the time, which is much lesser than the 97.8% by using $\hat{\alpha}^{f}_{15,40}$.
- Our modified criterion $\hat{\alpha}^{f}_{m,n}$ is less dependent on a bootstrap sample of size m as compared to the existing procedure $\hat{\alpha}^{s}_{m,n}$.
- If the optimal model is the full model, then the existing bootstrap model selection procedure outperforms our modified bootstrap model selection procedure.

6.2. Simulation setting 2

To evaluate the performance of the proposed criterion on simulated data, the following regression model with $p = 5$ and sample size $n = 60$ is considered

$$y_i = x_i^T \beta + \varepsilon_i, i = 1, 2, ..., n \tag{14}$$

where ε_i is generated from standard normal distribution, the regression variables are generated from $\text{N}(0, 1)$, and adding an intercept column of 1's to produce design matrix

Table 2. Selection Probabilities of $\hat{\alpha}^{s}_{m,n}$ and $\hat{\alpha}^{f}_{m,n}$ based on simulation setting 2.

True β	Model	$\hat{\alpha}^{s}_{16,60}$	$\hat{\alpha}^{f}_{16,60}$	$\hat{\alpha}^{s}_{24,60}$	$\hat{\alpha}^{f}_{24,60}$	$\hat{\alpha}^{s}_{32,60}$	$\hat{\alpha}^{f}_{32,60}$	$\hat{\alpha}^{s}_{40,60}$	$\hat{\alpha}^{f}_{40,60}$	$\hat{\alpha}^{s}_{60,60}$	$\hat{\alpha}^{f}_{60,60}$	AIC	BIC
(1,0,0,1,0)	**1,4***	**0.893**	**0.951**	**0.730**	**0.889**	**0.586**	**0.836**	**0.480**	**0.789**	**0.314**	**0.739**	**0.587**	**0.853**
	1,4,5	0.037	0.020	0.091	0.042	0.124	0.055	0.136	0.072	0.161	0.084	0.086	0.045
	1,3,4	0.038	0.015	0.081	0.037	0.115	0.052	0.127	0.065	0.145	0.076	0.100	0.042
	1,2,4	0.030	0.014	0.079	0.030	0.106	0.046	0.132	0.060	0.143	0.077	0.136	0.048
	1,3,4,5	0.001	0.000	0.006	0.001	0.019	0.004	0.034	0.004	0.064	0.007	0.022	0.001
	1,2,4,5	0.000	0.000	0.006	0.000	0.022	0.005	0.040	0.007	0.068	0.009	0.034	0.005
	1,2,3,4	0.001	0.000	0.005	0.001	0.021	0.002	0.036	0.003	0.069	0.005	0.025	0.005
	1,2,3,4,5	0.000	0.000	0.002	0.000	0.007	0.000	0.015	0.000	0.036	0.003	0.010	0.001
(1,0,0,1,1)	**1,4,5***	**0.948**	**0.976**	**0.833**	**0.936**	**0.722**	**0.895**	**0.635**	**0.866**	**0.478**	**0.827**	**0.672**	**0.902**
	1,3,4,5	0.028	0.012	0.084	0.034	0.131	0.055	0.162	0.067	0.209	0.081	0.120	0.044
	1,2,4,5	0.024	0.012	0.080	0.030	0.125	0.050	0.157	0.063	0.215	0.085	0.171	0.048
	1,2,3,4,5	0.000	0.000	0.003	0.000	0.022	0.000	0.046	0.004	0.098	0.007	0.037	0.006
(1,1,0,1,1)	**1,2,4,5***	**0.976**	**0.988**	**0.916**	**0.965**	**0.861**	**0.942**	**0.799**	**0.927**	**0.697**	**0.911**	**0.842**	**0.949**
	1,2,3,4,5	0.024	0.012	0.084	0.035	0.139	0.058	0.201	0.073	0.303	0.089	0.158	0.051
(1,1,1,1,1)	**1,2,3,4,5***	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Note: (*) denote the optimal model.

X. To generate the response variables y_i, we use Equation (14). The estimated selection probabilities for the existing bootstrap estimator $\hat{\alpha}^{s}_{m,n}$ and our proposed bootstrap estimator $\hat{\alpha}^{f}_{m,n}$ are calculated for $m = 16$, 24, 32, 40, and 60, using $L = 1000$ Monte–Carlo (MC) simulations with bootstrap replications of $K = 100$ and are tabulated in Table 2.

The simulation results presented in Table 2, confirm the satisfactory behavior of our modified bootstrap model selection criterion. For $m \approx 0.25n$, the modified bootstrap criterion selects the optimal models with high probability. Moreover, it is obvious from the results that the modified model selection criterion performs very well as compared to the existence criterion suggested by Shao (1996), the AIC and BIC for $m < 0.50n$.

The estimated selection probabilities based on Table 2 are plotted in Figures 1 and 2. The four different models are:

- M_1 shows that the optimal model has one non-zero predictor, i.e., $\beta_1 = (1, 0, 0, 1, 0)$,
- M_2 shows that the optimal model has two non-zero predictors, i.e., $\beta_2 = (1, 0, 0, 1, 1)$,
- M_3 indicates that the model has three non-zero predictors, i.e., $\beta_3 = (1, 1, 0, 1, 1)$, and
- M_4 indicates that the optimal model is the full model, i.e., $\beta_4 = (1, 1, 1, 1, 1)$.

Furthermore, F shows the selection probabilities plotted for our modified criterion $\hat{\alpha}^{f}_{m,n}$ and S indicates the selection probabilities plotted for Shao (1996) criterion $\hat{\alpha}^{s}_{m,n}$.

In Figure 1, the estimated selection probabilities are plotted against M_1, M_2, M_3, and M_4 for $m = 16$, 24, 32, and 40, whereas in Figure 2, the selection probabilities are plotted against m values for M_1, M_2, and M_3.

From Figures 1 and 2, we observe that:

- for $m \approx 0.25n$, the modified bootstrap criterion selects the optimal models with high probability,
- if bootstrap sample size m is less than 50% of the original sample size n, i.e., $m < 0.50n$ then our modified bootstrap criterion outperforms the existence criterion, the AIC, and BIC,

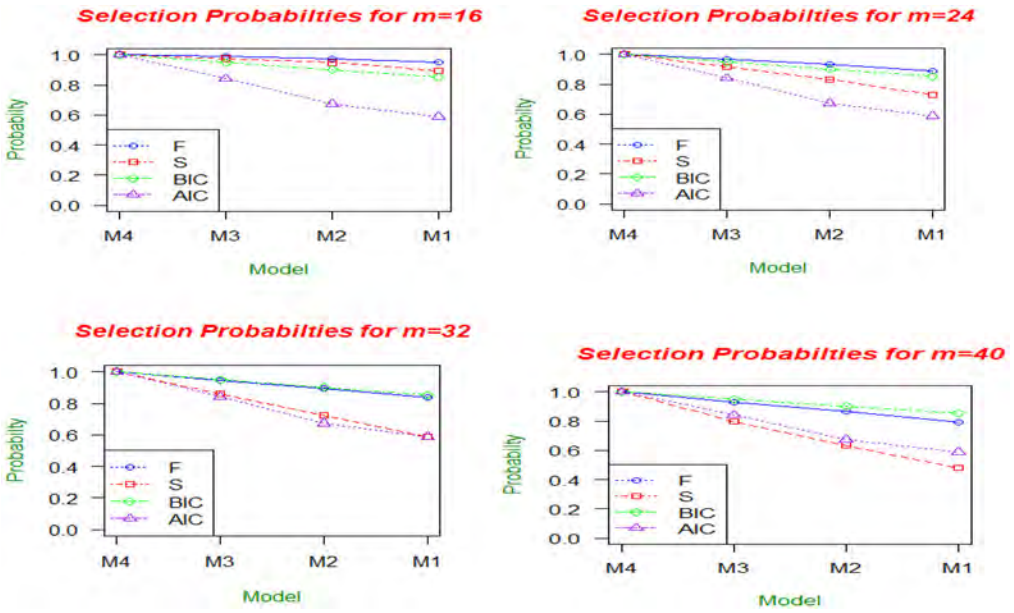

Figure 1. The selection probabilities for various m plotted against different models.

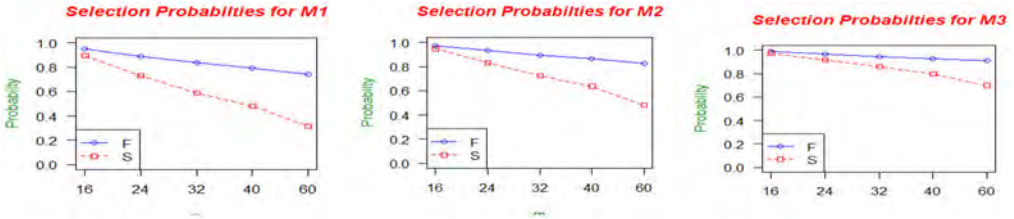

Figure 2. The selection probabilities for various models plotted against different values of m.

- if bootstrap sample size m is nearly half of the original sample size n ($m \approx 0.50n$), then the performance of our modified criterion and the BIC is almost the same, whereas the performance of Shao (1996) criterion is similar to the AIC,
- for $m > 0.50n$, then the performance of the BIC is better than our modified criterion.
- with the substantial increase in the value of m, the estimated selection probabilities may decline,
- all selection criteria select the full optimal model with probability 1,
- our modified criterion is more stable and less dependent on m as compared to the existing criterion.

7. Real data example (body density data)

In this section, we analyze the body density data of Johnson (1996). This dataset consists of thirteen explanatory variables. The response variable is the Body fat observed on

Table 3. Selected best model for the body density data using a range of model.

Selection criterion	Selected variables
$\hat{\alpha}^{f}_{m,n}$	weight, neck, and abdomen
$\hat{\alpha}^{s}_{m,n}$	neck, abdomen, and hip
BIC	weight, abdomen, and hip
AIC	weight, neck, abdomen, and forearm

$n = 128$ individuals. The explanatory variables are age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist. A summary of selected best models is presented in Table 3.

Table 3 presents a summary of selected best models. We calculate $\hat{\alpha}^{f}_{m,n}$ and $\hat{\alpha}^{s}_{m,n}$ with the same specifications as in the simulation study using $m = 35 \approx 0.27\ n$. According to our criterion, the variables included in the final selected model are weight, neck, and abdomen.

8. Conclusion

We proposed a paired bootstrap criterion (PBC) for model selection in linear regression. The criterion is a modification to the bootstrap model selection method proposed by Shao (1996). The results of our study reveal that the performance of the bootstrap model selection procedure is improved when using the OOB error. The simulations study confirms the satisfactory behavior of the modified bootstrap model selection criterion for finite samples to select parsimonious models that fit the data well. The paired bootstrap criterion results in a consistent model selection in the sense that the probability of selecting the optimal model can be improved as n increases. Moreover, there is an indication that our paired bootstrap criterion is less dependent on m than the existing approach. In conclusion, our proposed criterion is superior to the existing criterion suggested by Shao (1996), the AIC and the BIC.

References

Akaike, H. 1970. Statistical predictor identification. *Annals of the Institute of Statistical Mathematics* 22 (1):203–17.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6):716–23.

Arlot, S. 2009. Model selection by resampling penalization. *Electronic Journal of Statistics* 3: 557–624. doi:10.1214/08-EJS196.

Babu, G. J. 2011. Resampling methods for model fitting and model selection. *Journal of Biopharmaceutical Statistics* 21 (6):1177–86. doi:10.1080/10543406.2011.607749.

Breiman, L. 1995. Better subset regression using the nonnegative garrote. *Technometrics* 37 (4): 373–84. doi:10.1080/00401706.1995.10484371.

Breiman, L. 2001. Random forests. *Machine Learning* 45 (1):5–32. doi:10.1023/A:1010933404324.

Bunke, O., and B. Droge. 1984. Bootstrap and cross-validation estimates of the prediction error for linear regression models. *The Annals of Statistics* 12 (4):1400–24. doi:10.1214/aos/1176346800.

Davison, A. C., and D. V. Hinkley. 1997. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.

De Bin, R., S. Janitza, W. Sauerbrei, and A. L. Boulesteix. 2016. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 72 (1):272–80. doi:10.1111/biom.12381.

Efron, B. 1979. Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review* 21 (4):460–80. doi:10.1137/1021092.

Efron, B. 1982. *The jackknife, the bootstrap and other resampling plans*. Philadelphia, PA: SIAM.

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 78 (382):316–31. doi:10.1080/01621459.1983.10477973.

Freedman, D. A. 1981. Bootstrapping regression models. *The Annals of Statistics* 9 (6):1218–28. doi:10.1214/aos/1176345638.

Gunst, G. P., and R. L. Mason. 1980. *Regression analysis and its applications*. New York, NY: Marcel Dekker.

Johnson, R. W. 1996. Fitting Percentage of Body Fat to Simple Body Measurements. *Journal of Statistics Education* 4 (1). doi:10.1080/10691898.1996.11910505.

Lee, H., G. J. Babu, and C. Rao. 2012. A jackknife type approach to statistical model selection. *Journal of Statistical Planning and Inference* 142 (1):301–11. doi:10.1016/j.jspi.2011.07.017.

Mallows, C. L. 1973. Some comments on Cp. *Technometrics* 15 (4):661–75. doi:10.2307/1267380.

Miller, A.J. 1990. *Subset selection in regression*. London: Chapman & Hall.

Müller, S., and A. Welsh. 2005. Outlier robust model selection in linear regression. *Journal of the American Statistical Association* 100 (472):1297–310. doi:10.1198/016214505000000529.

Müller, S., and A. Welsh. 2009. Robust model selection in generalized linear models. *Statistica Sinica* 19:1155–70.

Salibian-Barrera, M., and S. Van Aelst. 2008. Robust model selection using fast and robust bootstrap. *Computational Statistics & Data Analysis* 52 (12):5121–35. doi:10.1016/j.csda.2008.05.007.

Sauerbrei, W. 1999. The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 48:313–29. doi:10.1111/1467-9876.00155.

Sauerbrei, W., A.-L. Boulesteix, and H. Binder. 2011. Stability investigations of multivariable regression models derived from low-and high-dimensional data. *Journal of Biopharmaceutical Statistics* 21 (6):1206–31. doi:10.1080/10543406.2011.629890.

Schwarz, G. 1978. Estimating the dimension of a model. *The Annals of Statistics* 6 (2):461–4. doi:10.1214/aos/1176344136.

Shao, J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* 88 (422):486–94. doi:10.1080/01621459.1993.10476299.

Shao, J. 1996. Bootstrap model selection. *Journal of the American Statistical Association* 91 (434):655–65. doi:10.1080/01621459.1996.10476934.

Shao, J. 1997. An asymptotic theory for linear model selection. *Statistica Sinica* 7:221–42.

Shibata, R. 1984. Approximate efficiency of a selection procedure for the number of regression variables. *Biometrika* 71 (1):43–9. doi:10.1093/biomet/71.1.43.

Wisnowski, J. W., J. R. Simpson, D. C. Montgomery, and G. C. Runger. 2003. Resampling methods for variable selection in robust regression. *Computational Statistics & Data Analysis* 43 (3):341–55. doi:10.1016/S0167-9473(02)00235-9.

Wu, Y. 2001. An M-estimation-based model selection criterion with a data-oriented penalty. *Journal of Statistical Computation and Simulation* 70 (1):71–87.

Zhang, G.-Y., C.-X. Zhang, and J.-S. Zhang. 2010. Out-of-bag estimation of the optimal hyperparameter in subbag ensemble method. *Communications in Statistics - Simulation and Computation* 39 (10):1877–92. doi:10.1080/03610918.2010.521277.

Zhang, P. 1992. On the distributional properties of model selection criteria. *Journal of the American Statistical Association* 87 (419):732–7. doi:10.1080/01621459.1992.10475275.