IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, April 2026
Impact Factor: 8.2

A General Procedure for Variable Selection in Regression

Tejaswi S. Kurane[1] and Rahul H. Waliv[2]
Department of Statistics, Rajarshi Chhatrapati Shahu College, Kolhapur (MS), India[1]
Department of Statistics, Kisan Veer Mahavidyalaya, Wai, Satara (MS), India[2]
Corresponding author: tejustat@gmail.com

Abstract: *Variable selection in regression is one of the important problems in multiple linear regression. In this article, we propose robust variable selection criterion based on predicted values of full model and subset model. Proposed criterion is shown to be a consistent under certain condition. The selection ability and efficiency of proposed criterion is evaluated by the simulation study. Also evaluate the performance of proposed criterion for normal and non-normal distribution with and without outlier observation(s) in the Y-space.*

Keywords: C_p; M-estimator; bootstrap; consistency; observed L_2 efficiency.

I. INTRODUCTION

Consider the linear regression model

$$Y = X\beta + \varepsilon \qquad (1)$$

where Y is an n×1 vector of response variable, X is a matrix of order $n \times k$ with 1's in the first column, β is a $k \times 1$ vector of regression coefficient and ε is an $n \times 1$ vector of normal random errors with $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2 I$.

The motive behind fitting of regression model is predicting the future values of response when value of predictor variables is known. It is well known that, fitting a model with large number of predictor variables is neither practicable nor economical. Usually, a model based on small number of predictor variables gives more accurate prediction [1]. However, selection of relevant predictor variable is not simple task. Finding an appropriate subset of predictor variables for the model is called the variable selection or model selection in regression [2]. In the classical linear regression setup many variable selection procedures have been suggested by the researchers. Among these, Mallows' C_p [3] is most frequently used. It is defined as,

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p)$$

where RSS_p is the residual sum of squares for subset model with p parameters, n is the number of observations and σ^2 is an estimate of the error variance, which is obtained from the full model.

In literature, various subset selection procedures are available, for example the Akaike Information Criterion (AIC) [4], the Bayes Information Criterion (BIC) [5], Final Prediction Error (FPE) Method [6], the Generalized Information Criterion [7], the Modified Bootstrap Selection Method [8]. All these methods are sensitive for outlier observations. In literature there exist few robust variable selection methods likes, Ronchetti and Staudte [9] proposed robust version of C_p *(RC_p)*, Sommer and Huggins [10] suggested robust C_p *(RT_p)* based on Wald test statistic, Kim and Hwang [11] proposed an another robust version of C_p *(C_p*(k)), Kashid and Kulkarni [12] suggested more general criterion, called S_p statistic. These robust methods are based on M-estimator of regression parameters.

In this article, we propose a subset selection criterion based on average of squared difference between prediction values of full model and subset model with adding penalty term as a measure of complexity of model. The proposed method works well in the presence of outlier observation(s) in the data and also for non-normal error distributions.

DOI: 10.48175/IJARSCT-33838

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, April 2026 Impact Factor: 8.2

The remaining article is organized as: in Section 2, the problem of existing methods are demonstrated when outlier is presence in the data, Section 3, gives the notations and estimation, new variable selection criterion is proposed in Section 4, Section 5, presents or deals with performance of the proposed criterion. The article ends with some concluding remarks.

II. PROBLEM

In this section, we demonstrate the model selection ability of existing variable selection methods AIC and BIC in the presence of single outlier. The AIC is defined as

$$\text{AIC} = \; nlog\ \hat{\sigma}^2 + 2p$$

and also, BIC is defined as

$$\text{BIC} = \; nlog\ \hat{\sigma}^2 + plog(n)$$

where p is the number of parameters in the subset model with $(p-1)$ regressors, $\hat{\sigma}^2$ is MLE of σ^2, n is the number of observations.

Here, we were generated 40 observations on the predictor variables $X_1, X_2, X_3$and X_4 from U (0, 1) and ε_i (i=1, 2,…., 40) was generated from normal distribution with mean zero and unit variance. The response variable Y were generated using the regression model ,

$$Y_i = 1 + 2X_{i1} + 0X_{i2} + 0X_{i3} + 4X_{i4} + 0X_{i4} + \varepsilon_i \quad i=1, 2,\ldots., 40.$$

We introduced the one outlier by multiplying the actual Y corresponding to maximum residual by fifty. We compute the values of BIC and AIC for clean data and data with outlier. The same experiment is repeated 1000 times. The results of subset model X_1, X_4 and X_1, X_4 with some additional variables are reported in Table 1.

From Table 1, criterion AIC and BIC fail to select the correct model in the presence of outlier observation in the data. The performance of these methods is not well in the presence of outlier, because optimal model selection ability of these methods is approximately 30%. Therefore, we need to use another estimation procedure as well as new criterion. So, we have proposed new criterion.

III. NOTATIONS AND ESTIMATION

In this Section, we define full model, subset model and class of models. Consider the regression model (1)

$$Y = X\beta + \epsilon$$

Suppose $\hat{\beta}$ is M-estimator [13] of β. The prediction equation for this model is

$$\hat{Y}_{ik} = X_i'\hat{\beta} \qquad \text{where } X_i' = (X_{i0}, X_{i1}, \ldots\ldots, X_{ik-1}) \qquad (2)$$

Now model (1) can be written as

$$Y = X_\alpha\beta_\alpha + X_{k-\alpha}\beta_{k-\alpha} + \varepsilon$$

where $= (X_\alpha : X_{k-\alpha})$,$\beta = (\beta_\alpha : \beta_{k-\alpha})T, \alpha = \{\text{subset of } \{1, 2, \ldots k-1\} \cup X_0\}$.

Then model corresponding to α called subset model, it is given by

$$Y = X_\alpha\beta_\alpha + \varepsilon \qquad (3)$$

Suppose under each subset model, β_α is estimated by using any estimation procedure. The fitted regression equation for model (3) is

$$\hat{Y}_{ip} = X_i'\hat{\beta}_\alpha \qquad (4)$$

In further discussion, we called model (2) as full model and model (3) as subset model. Also, we defined class of models.

Suppose, α_N denote all necessary predictors. Following, each subset model can be associated with one of the following three categories (Sakate and Kashid [14]).

1. Optimal model $\boldsymbol{\mathcal{A}}_o = \{A_\alpha$: only all necessary predictors are present$\}$.
2. Class of correct model $\boldsymbol{\mathcal{A}}_c = \{A_\alpha$: all necessary predictor is present$\}$,

   i.e. $\boldsymbol{\mathcal{A}}_c = \{A_\alpha$: $\alpha_N \subseteq \alpha\}$.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, April 2026
Impact Factor: 8.2

3. Class of wrong model $\mathcal{A}_w = \{A_\alpha$: at least one necessary predictor is missing$\}$,
i.e. $\mathcal{A}_w = \{A_\alpha: \alpha_N \not\subseteq \alpha\}$.
Below, we defined the subset selection criterion based on fitted equations (2) and (4)

VI. PROPOSED CRITERION

We measure the efficiency of the model 'α' by averaging squared difference between prediction values of full model and subset model. It is denoted by

$$\Delta^2{}_\alpha = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2$$

Based on data, our objective is to select the subset model $\alpha \in c$, where c is set of collection of all possible subsets. The selected subset model such that $\Delta^2{}_\alpha$ is minimum for small α. In practical situation, $\Delta^2{}_\alpha$ is zero when the subset model as a full model. Therefore it is difficult to select the optimal model, so we add some another measure called as model complexity measure or penalty functions $C(n, p_\alpha)$. Using both measures, we propose the following criterion. It is denoted by G_p and defined as,

$$G_P = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_{ik} - \hat{Y}_{ip}\right)^2 + C(n, p_\alpha) \qquad (5)$$

Select the subset model corresponding to minimum value of G_P from all possible set of subset model.

V. SIMULATION STUDY

This simulation study divided into four parts. In first part, we compare the performance of G_p-statistic with AIC and BIC. The performance of G_p is evaluated through simulated data for different sample sizes in the second part. In the third part, we have obtained L_2 efficiency. In fourth part, we have explained the performance of G_p using bootstrap samples.

In the entire simulation study, we have used following penalty functions [5].

$$P_1 = 2p - k$$
$$P_2 = p\log(n)$$

where k is the number of parameter in the full model, p is the number of parameter in subset model and n is the number of observations.

The statistic G_p proposed under the assumption of normality and evaluated the performance of G_p-statistic for some non-normal error distributions. In the entire simulation study, we use normal and three non-normal distributions which are given below:

E_1= Normal(0, 25)
E_2 = 0.4Normal(0,49) + 0.6Normal(0,1)
E_3 = Cauchy(0,1)
E_4 = Slash (0, 1).
The M-estimator has obtained using Huber robust function with tuning constant $c = 1.345$.

Example : Simulated data

In this example, we use the simulated data. We were generated 40 observations of predictor variables X_i (i= 1, 2,….5) from U (0, 2). Here the errors were generated same as above Example. We were generated the response variable using the model,

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ with $\beta = (1, 5, 0, 0, 10, 0)$.

The model selection ability of proposed statistic G_p is computed for error distributions E_1, E_2, E_3 and E_4 for 1000 realizations. The outlier observations are introduced in the response variable by the same procedure discussed in Example 1. The results are reported in the Table.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, April 2026 Impact Factor: 8.2

From Table 3, it is clear that performance of the G_p is better than AIC and BIC for all used error distributions with penalty p_1 and p_2. The frequency of selecting wrong model by AIC and BIC is larger than that for optimal and correct model. The performance of G_p is better than AIC and BIC for all error distributions.

Table 1: Selection ability for 1000 simulation

Criterion Sub model	BIC	BIC+O1	AIC	AIC+O1
{1,4}*	794	352	531	308
{1,2,4}	67	19	116	46
{1,3,4}	50	16	109	54
{1,4,5}	66	18	111	48
{1,2,3,4}	11	1	30	11
{1,2,4,5}	4	4	31	13
{1,3,4,5}	7	1	40	7
{1,2,3,4,5}	1	0	12	1

o1: one outlier *: optimal model

Table 2: Selection ability of proposed criterion based on M-estimation method for 1000 simulation

Error	Model	G_p (p_1)	G_p with p_1			G_p (p_2)	G_p with p_2			BIC	AIC
			o1	o2	o3		o1	o2	o3	o1	o1
E1	{1,4}*	835	825	794	790	955	969	957	951	744	724
	{1,2,4}	40	45	47	54	15	10	12	11	0	0
	{1,3,4}	36	35	46	58	8	7	8	12	0	1
	{1,4,5}	68	64	82	67	19	12	23	21	0	2
	{1,2,3,4}	12	17	13	16	2	2	0	3	0	0
	{1,2,4,5}	4	5	7	6	0	0	0	0	0	0
	{1,3,4,5}	5	9	9	8	1	0	0	2	0	0
	{1,2,3,4,5}	0	0	2	1	0	0	0	0	0	0
	Total	1000	1000	1000	1000	1000	1000	1000	1000	744	724
E2	{1,4}*	996	996	994	992	1000	1000	1000	1000	714	727
	{1,2,4}	2	1	2	2	0	0	0	0	0	0
	{1,3,4}	0	0	2	3	0	0	0	0	0	0
	{1,4,5}	2	2	2	3	0	0	0	0	0	0
	{1,2,3,4}	0	1	0	0	0	0	0	0	0	0
	{1,2,4,5}	0	0	0	0	0	0	0	0	0	0
	{1,3,4,5}	0	0	0	0	0	0	0	0	0	0
	{1,2,3,4,5}	0	0	0	0	0	0	0	0	0	0
	Total	1000	1000	1000	1000	1000	1000	1000	1000	714	727
E3	{1,4}*	891	866	885	884	925	913	906	885	650	670
	{1,2,4}	25	26	25	10	18	14	16	9	1	0
	{1,3,4}	53	40	42	57	37	47	47	46	1	2
	{1,4,5}	11	27	14	19	3	12	12	16	0	2
	{1,2,3,4}	13	23	24	22	13	13	18	25	0	0
	{1,2,4,5}	1	3	1	1	1	0	0	0	0	0
	{1,3,4,5}	1	6	5	2	2	1	0	3	0	0

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 8, April 2026 Impact Factor: 8.2

E_4	{1,2,3,4,5}	5	9	4	5	1	0	1	16	0	0
	Total	1000	1000	1000	1000	1000	1000	1000	1000	652	674
	{1,4}*	844	826	788	781	899	890	868	853	612	640
	{1,2,4}	18	35	25	49	25	22	31	33	1	1
	{1,3,4}	17	61	69	64	45	49	55	55	1	2
	{1,4,5}	98	25	51	52	13	17	19	19	3	4
	{1,2,3,4}	0	34	41	29	14	16	20	22	0	1
	{1,2,4,5}	6	6	2	5	1	2	1	1	0	0
	{1,3,4,5}	0	2	6	11	1	0	1	1	0	0
	{1,2,3,4,5}	1	11	18	9	2	4	5	16	0	0
	Total	1000	1000	1000	1000	1000	1000	1000	1000	617	648

*The optimal model with variable x_1 and x_4 o_1=one outlier o_2=two outlier o_3=three outlier

VI. CONCLUDING REMARKS

Subset selection method introduced in this article is simple for implement and take into account goodness of fit and complexity of the model. The performance of this method is better than AIC and BIC in case of outlier data and non-normal error distribution.

REFERENCES

[1] A. J. Miller, Subset selection in regression, Chapman and Hall, 2002.
[2] D. Montgomery, E. Peck, and G. Vining, Introduction to linear regression analysis, 3rded, John Wiley and Sons Inc, New York, 2006.
[3] C. L. Mallows, Some comments on Cp, Technometrics 15 (1973), pp. 661-675.
[4] H. Akaike, Statistical predictor identification, Annals of the Institute of Statistical Mathematics, 22 (1970), pp. 203-217.
[5] G. Schwartz, Estimating the dimensions of a model, The Annals of Statistics 6 (1978), pp. 461-464.
[6] R. Shibata, Approximate efficiency of a selection procedure for the number of regression variables, Biometrika 71 (1984), pp. 43-49.
[7] C. R. Rao, and Y.Wu, A strongly consistent procedure for model selection in regression problem, Biometrika 76 (1989), pp. 369-374.
[8] J. Shao, Bootstrap model selection, Journal of the American Statistical Association, Theory and Methods 91(434) (2013), pp. 655-665.
[9] E. M. Ronchetti, and R. G. Staudte, A robust version of mallows' Cp, Journal of the American Statistical Association 89 (1994), pp. 550-559.
[10] S. Sommer and R. M. Huggins, Variable selection using the wald test and a robust Cp, Journal of the Royal Statistical Society, 45(1) (1996), pp. 5-29.
[11] C. Kim and S. Hwang, Influential subsets on the variable selection, Communication in Statistics- Theory and Methods 29(2) (2000), pp. 335-347.
[12] D. N. Kashid and S. R. Kulkarni, A more general criterion for subset selection in multiple linear regressions, Communication in Statistics-Theory & method 31(5) (2002), pp. 795-811.
[13] P. J Huber, Robust statistics, Wiley, New York, 1981.
[14] D. M. Sakate and D. N. Kashid, A deviance-based criterion for model selection in GLM, Statistics: A Journal of Theoretical and Applied Statistics (2012), pp.1-15.
[15] Y. Dodrge and D. Birkes, Alternative methods of regression, John Wiley and Sons, Inc (1973), pp. 99.
[16] H. White, Maximum likelihood estimation of misspecified models, Econometrics, 50 (1982), pp. 1-25.
[17] J. Shao, Linear model selection by cross validation, Journal of American Statistician 422 (1993) pp. 484-494.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 8, April 2026
Impact Factor: 8.2

[18] R. Shibata, An optimal autoregressive spectral estimate. Ann. Statist. 9 (1981), pp. 300-306.
[19] A. McQuarrie, R. Shumway, and C.L. Tsai, The model selection criterion AICu, Statist. Probab. Lett .34 (1997), pp. 285–292.
[20] B. M. Efron, Bootstrap Methods: Another look at the jackknife, The Annals of Statistics 7 (1979) pp. 1-26.