# A Convex Combination Method of Subset Selection Criterion in a Linear Regression

Satish Bhat

*Department of Statistics, Yuvaraja's College, University of Mysore, Mysuru, Karnataka, India-570005*

**Abstract: In multiple linear regression analysis, problem multicollinearity and outliers are the two serious issues. Due to these problems, the ordinary least squares(OLS) method does not yield good estimates to the regression parameters. When such a situation arise, there are various methods to estimate the optimum estimates to the regression parameters and one such is called subset selection methods. Several authors have been suggested various modes of subset selection methods based on OLS methods, which are sensitive to multicollinearity, heteroscadesticity, non-normal error distribution, etc. In order to overcome the problem of these factors, here we suggest a new method of obtaining the good estimates to the regression parameters by selecting a good subset model, and we called it as Weighted Method of Subset Selection Criteria. It is obtained by shrinking the *$GS_p$* and *$MGS_p$* criterions using convex combination concepts. The performance of the suggested criteria is evaluated empirically and compared with some of the existing methods of subset selection criterions. Empirical results indicate that the proposed criteria performs better in presence of both multicollinearity and outliers.**

**Key Words: Multiple linear regression, multicollinearity, outliers, subset selection, ridge estimator.**

## 1. INTRODUCTION

Consider the general form of the multiple linear regression model

$$y = X\beta + \varepsilon \ , \qquad (1)$$

Where (1) is called the full model, $X$ is data matrix of order ( $n \times p$ ) with ( $p-1$ ) predictors, which are non stochastic, $y$ is a $(n \times 1)$ vector of response variable, $\beta$ is a ( $p \times 1$ ) vector of unknown regression coefficients and $\varepsilon$ is $(n \times 1)$ vector of random errors, such that E( $\varepsilon$ )=0 and E $(\varepsilon\varepsilon') = \sigma^2 I$.

For the full model(1.1), it is known that, the ordinary least square estimator for $\beta$ is $\hat{\beta}_{ols} = (X'X)^{-1} X'y$ and the vector of predicted values is $\hat{y}_p = X\hat{\beta}_{ols} = Hy$, where $H$ is the hat matrix or prediction matrix.

### 1.1 SUBSET MODEL

Here we partition the data matrix $X$ as $X = [X_t : X_\upsilon]$ and the vector of regression coefficients $\beta$ as $\beta' = [\beta'_t : \beta'_\upsilon]$ such that the equation (1.1) become

$$y = X_t \beta_t + X_\upsilon \beta_\upsilon + \varepsilon \qquad (2)$$

Where the matrix $X_t$ is of order $n \times k$ with ones' in the first column, and the matrix $X_\upsilon$ is of order $n \times (p-k)$. The vectors $\beta'_t$ and $\beta'_\upsilon$ of regression coefficients are of orders $(k \times 1)$ and $(p-k) \times 1$ respectively.

Consider the subset model based on $k-1$ predictors as

$$y = X_t \beta_t + \varepsilon \qquad (3)$$

Then the ordinary least squares estimator for the subset model based on the full model for $\beta_t$ is $\hat{\beta}_t = (X'_t X_t)^{-1} X'_t y$ and the vector $\hat{y}_k$ of predicted values is $\hat{y}_k = X_t \hat{\beta}_t = H_1 y$, where $H_1$ is the hat matrix based on the subset model.

Now we briefly introduce the various methods of subset selection, which are defined in the literature by various authors. Viz., Cp Criterion,(Mallows, 1973), Sp criterion( Kashid and Kulkarni, 2002), etc.

#### 1.1.1 THE *Cp* - CRITERION

It is one of the most wiedly used subset selection criterion due to Mallows(1973) in the method of regression analysis. It is obtained by shrinking the OLS estimator. The Mallows Cp statistic is defined by

$$C_p = \frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - N + 2(p+1) \quad (4)$$

It is well-known that OS estimator is very sensitive to outliers, the linear dependence between any two predictors or the multicollinearity or violation of the normality on the error variable(Huber, 1991). Since Cp criterion is based on OLS estimator and thereby the *Cp* statistic fails to select the correct subset model and there by the estimates of regression coefficients are then generally far away from the true parameter value. Note that a small value of $C_p$ means, the model is relatively precise

### 1.1.2 THE *Sp* - CRITERION

Outliers in the data may lead to wrong selection of subset and hence Kashid and Kulkarni,2002 motivated to rectify this problem and came with an innovative idea and suggested $S_p$ criterion, which is based on the fitted values of $y_i$, $i$ =1,2,..., $n$, using M-Estimator of the full model and the subset model respectively. They define the Sp statistic as

$$S_p = \frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - (p - 2k) \quad (5)$$

Where $p$ and $k$ are the number of predictors in full model and the subset model respectively. The estimators $\hat{y}_{ip}$ and $\hat{y}_{ik}$ the vector are the predicted values of $y_i$ based on the full model. The unknown variance $\sigma^2$ is estimated by its suitable estimate based on the full model as

$\hat{\sigma} = 1.4826\, median\,|\, e_i - median(e_i)\,|$ where, $e_i$ is the $i^{th}$ residual. (6)

### 1.1.3 THE Rp- CRITERION

When the data suffers from multicollinearity, the two criterions discussed above fail to select the correct subset model(Dorugade and Kashid, 2010 ), and therefore a new criteria called Rp criteria, which was due to Dorugade and Kashid, (2010 ). They have proposed this criteria based on the ordinary ridge regression(ORR) estimator(Hoerl and Kennard, 1970a,b), when multicollinearity is present in the data. It is defined as under

$$R_p = \frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - tr(H'_R H_R) + tr(H'_{Rt} H_{Rt}) + k \quad (7)$$

Where $H_R = X(X'X + rI)^{-1} X'$, and $H_{Rt} = X_t (X'_t X_t + r_t I)^{-1} X'_t$ and $\sigma^2$ is the unknown error variance, and it is estimated by it's suitable estimate using the ORR estimator, defined by

$$\hat{\sigma}^2 = (y - X\hat{\beta}_{ORR})'(y - X\hat{\beta}_{ORR})/(n-p) \quad (8)$$

The Sp and Rp statistics are equivalent to Mallows $C_p$ criterion when the OLS estimator of $\beta$ is used.

Note that $\sigma^2$ the above Criterions $C_p$ and $S_p$ would fail to select subset model when data suffers from multicollinearity whereas the Rp criterion fails when outliers are present (Jadhav, adn Khashid, 2014) in the data. Now we discuss the criteria in the following subsection which is one of the better methods of subset selection when simultaneous presence of multicollinearity and outliers in the data.

### 1.1.4 THE *GS<sub>p</sub>* CRITERION

When data suffer from both multicollinearity and outliers together, Jadhav and Kashid(2014) suggested a new subset selection criterion called the Gsp criterion, which is based upon the Jackknifed ridge-M estimator(JRM), and is defined as follows.

$$GS_p = \frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - tr\{(H - H_{Rt})'(H - H_{Rt}) + k \quad (9)$$

Where $\sigma^2$ is the unknown error variance estimated by its suitable estimate using the full model based on JRM estimator and tr(.), denote the trace of the matrix. The matrices $H$ and $H_{Rt}$ are the prediction matrices based on JRM estimator.

### 1.1.5 THE *MGS<sub>P</sub>* CRITERION

In presence of multicollinearity and the violation of normality on the error variable, Huber(1981) pointed out that $C_p$ statistic fails to select correct subset model and thereby one used to get poor estimates for the regression parameters. The subset selection criterion $S_p$ is also based on OLS estimator like $C_p$ criterion, and hence $S_p$ criteria is also fail to sect correct subset model when multicollinearity is present in the data. Thus in order to overcome the problem of both multicollinearity and outliers, Modified Generalised Subset Selection (MGSp) Criteria (Satish, 2024), *GS<sub>p</sub>* Criterion (Jadhav and Kashid, 2014) etc. are available in the literature.

The, Modified Generalised Subset Selection (*MGS<sub>p</sub>*) Criteria is due to Satish (2024), and is slightly superior to the existing estimators which are defined above. Modified Generalised Subset Selection (*MGS<sub>p</sub>*) Criteria is also obtained by shrinking the JRM estimator in order to overcome the problem of multicollinearity and outliers are present in the data and it is defined as under.

$$GS_p = \frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - tr(H'H_{Rt}) + k \qquad (10)$$

Where the matrices $H$ and $H_{Rt}$ are the prediction matrices, based on JRM estimator.

## 2. PRPOSED SUBSET SELECTION (WEIGHTED SUBSET SELECTION) CRITERION

Here we propose a new method of subset selection criterion, which is obtained by shrinking the GSp and MGSP criterions using convex combination approach. It is defined as follows

$$\boldsymbol{WMGS}_p = \alpha\left\{\frac{\sum_{i=1}^{n} (\hat{y}_{ip} - \hat{y}_{ik})^2}{\sigma^2} - tr[(H - H_{Rt})'(H - H_{Rt})]\right\} + (1-\alpha)\left\{\frac{\sum_{i=1}^{n} \left(\hat{y}_{ip} - \hat{y}_{ik}\right)^2}{\sigma^2} - tr(H'H_{Rt})]\right\} + k$$

Where $\alpha$ is chosen such that $\alpha = 1/\sqrt{\max(\lambda_j})$ of the matrix $X'X$ .

It is observed that from the empirical study that, when both multicollinearity and outliers are present in the data, the suggested estimator will select the correct subset model as compared to that of the existing estimators which are considered under this study.

Subset Selection Procedure based on *WMGS<sub>p</sub>* Statistic

Step 1. Compute the value of *WMGS<sub>p</sub>* statistic fro all possible subset models

Step2. Select the subset of minimum size, for which the value of the *WMGS<sub>p</sub>* statistic is close to k, the number of predictors in the subset model.

In the following section, we study the performance of the subset selection criterions such as $C_p$, $S_p$, $R_p$, $GS_p$, $MGS_p$, and $WMGS_p$ using empirical data.

## 3. SIMULATION STUDY

A simulation study is carried out to illustrate the performance of the suggested subset selection criterion $WMGS_p$ and some of the existing criterions. The simulation study is performed in three steps

Step 1. The performance of the estimators which are considered under study through numerical example for all combinations of presence and absence of multicollinearity and outliers

Step 2. Evaluation of correct subset selection ability of the criterions

Step 3. Performances of different choices of the estimator of $\sigma^2$, through numerical example.

For the computation of M-estimator, Huber's robust criterion function with tuning parameter t=1.345 is used.

To evaluate the performance of the existing and the suggested methods of subset selection criterions in different situations viz., multicollinearity etc., first we generate a random data matrix $X$ of sige ( $n \times p$ ) using the relation

$$x_{ij} = (1-\rho^2)^{1/2}\xi_{ij} + \rho\xi_{i(l+1)}$$

where, $i = 1,2,...,n;\ j = 1,2,...,l$

where $\xi_{ij}$ 's are independent standard normal pseudo random numbers, $\rho$ is fixed such that $\rho^2$ is the degree of correlation between any two predictors. Let $l = 4$, and $\rho = 0.999$ and $n = 30$ are considered to generate observations on the response variable $y$ using the following regression model

$$y_i = 10 + 3x_{i1} + 2x_{i2} + 0x_{i3} + 0x_{i4} + \varepsilon_i,$$

$i = 1,2,...,30$ and $\varepsilon_i \sim N(0,\ 0.25)$.

The simulated results are given in Table-1 through Table-3. The single outlier observation is introduced in the response variable $y$ corresponding to the maximum residual value. The true value of the variable $y_{27} = 12.0021$ is changed to 240.0420. To identify the sternness of the multicollinearity, the variance inflation factor(VIF) values are used(Marquardt, 1970, Montgomery, Peck and Vinning, 2003). For the data which we have obtained, the VIF's for each term are 21.7.4324, ..., 312.5734. These VIF values indicate the presence of strong linear dependence between the predictors. We have computed and presented the results in table 1 through table 3 for various subset selection criterion statistic which are used in this study for all possible subset models. For the above subset model, the correct subset model contains the predictor variables are $X_1$ and $X_2$. Analysis shows that the suggested criterion *WMGS<sub>p</sub>* selects the correct subset model{ $X_1, X_2$ }, when both multicollinearity and outliers are present in the data. On the other hand, the condition for selection for the subset model by other existing criterions fail to select the correct subset model or in other words they are little more deviated from the selection of the correct subset model { $X_1, X_2$ }.

Table 1. The results of $C_p$, $S_p$, $R_p$, $GS_p$, $MGS_p$ and $WMGS_p$ statistic for all subsets with multicollinearity but without an outlier.

| predictors in the model | $C_p$ | $S_p$ | $R_p$ | $JRMGS_p$ | $LRMGS_p$ | $JRMMGS_p$ | $LRMMGS_p$ | $WMGS_p$ |
|---|---|---|---|---|---|---|---|---|
| X1 | -0.9183 | -0.9218 | -0.9144 | -0.919 | 1.0666 | 1.0648 | 1.0808 | 1.0427 |
| X2 | 1.0766 | **2.0584** | 3.0423 | 3.0637 | 3.0471 | 3.0048 | 5 | -0.8158 |
| X3 | -0.6157 | -0.6316 | -0.6218 | 1.1134 | 1.0703 | 1.1195 | 1.3665 | 1.3444 |
| X4 | 1.3653 | 3.0436 | 3.0132 | 3.2061 | 3.3435 | 5 | 1.8843 | 1.8806 |
| X1X2 | **1.8907** | **1.8844** | **2.9359** | **2.9571** | **2.9868** | **2.9647** | **2.9685** | **3.0135** |
| X1X3 | 4.0427 | 4.0407 | **3.0522** | 4.0648 | 5 | 2.2455 | 1.8362 | 2.2136 |
| X1X4 | **2.0114** | 5.4123 | 1.2098 | 9.2603 | 10.3102 | 27.0987 | 33.8142 | 17.744 |
| X2X3 | 17.1154 | 3.2784 | 5.3209 | 12.9760 | 7.0025 | 6.982 | 7.0469 | 7.0819 |
| X2X4 | 14.0449 | 5.3702 | 8.4 | 8.0898 | 9.807 | 9.3927 | 17.077 | 16.9696 |
| X3X4 | 10.4982 | 9.0622 | 15.8046 | 3.4113 | 2.9989 | 3.3834 | 3.2833 | 2.4558 |
| X1X2X3 | 59.394 | 94.1313 | 12.4422 | 42.2006 | 45.9691 | 75.668 | 73.198 | 40.226 |
| X1X2X4 | 8.4345 | 12.8543 | 8.0156 | 8.0293 | 8.0739 | 8.1207 | 16.0179 | 17.3478 |
| X1X3X4 | 10.3689 | 10.1814 | 11.935 | 11.6093 | 20.049 | 19.9302 | 13.7834 | 12.1393 |
| X2X3X4 | 19.7906 | 6.0206 | 6.0341 | 6.0788 | 6.1255 | 13.023 | 14.3529 | 14.3764 |
| X1X2X3X4 | 5.0186 | 5.0394 | 5.0632 | 5.0542 | 5.0954 | 5.0097 | 5.0008 | 5.0006 |

Table 2. The results of $C_p$, $S_p$, $R_p$, $GS_p$, $MGS_p$ and $WMGS_p$ statistic for all subsets with multicollinearity but without one outlier

| predictors in the model | $C_p$ | $S_p$ | $R_p$ | $JRMGS_p$ | $LRMGS_p$ | $JRMMGS_p$ | $LRMMGS_p$ | $WMGS_p$ |
|---|---|---|---|---|---|---|---|---|
| X1 | 10.9823 | 11.0092 | 10.9181 | 10.9551 | 7.9666 | 4.5775 | 9.2919 | 4.1209 |
| X2 | 4.7415 | 10.9217 | 6.0433 | 6.5637 | 3.5359 | 4.0448 | 5 | 29.408 |
| X3 | 29.4737 | 2.0942 | 29.3461 | 17.9701 | 9.4995 | 22.8578 | 8.1197 | 10.8407 |
| X4 | 25.5834 | 9.9641 | 12.1397 | 4.597 | 5.3132 | 5 | 7.8358 | 7.8354 |
| X1X2 | 5.3317 | **3.2393** | **2.8644** | **3.2840** | **3.2819** | **3.0835** | **3.0811** | **3.0698** |
| X1X3 | 10.0605 | 10.0577 | 10.0777 | 10.0686 | 5.0000 | 25.794 | 25.8425 | 25.6754 |
| X1X4 | **3.0549** | **3.0603** | **2.9805** | 13.2923 | 90.133 | 215129 | 32.1293 | 16.9793 |
| X2X3 | 23.6596 | 20.7242 | 98.6809 | 24.4935 | 11.8703 | 11.5707 | 11.8305 | 11.8641 |
| X2X4 | 746.5037 | 438.1506 | 320.5805 | 220.6711 | 519.5617 | 795.1106 | 40.3247 | 56.5453 |
| X3X4 | 492.1544 | 233.6417 | 589.8519 | 29.4571 | 29.5198 | 29.3192 | 29.4023 | 306006 |
| X1X2X3 | 179.9085 | 132.9257 | 901.3616 | 215.1316 | 213.2132 | 169.7966 | 23.6598 | 20.7244 |
| X1X2X4 | 98.6823 | 24.4937 | 15.4044 | 15.081 | 15.3632 | 15.3996 | 7.4655 | 4.3819 |
| X1X3X4 | 32.0621 | 22.7163 | 51.9613 | 79.5521 | 40.7517 | 56.5515 | 49.2207 | 23.3694 |
| X2X3X4 | 58.9909 | 13.3967 | 13.0734 | 13.3555 | 13.3919 | 74.6523 | 43.8167 | 32.0594 |
| X1X2X3X4 | 5.0084 | 5.0047 | 5.0043 | 5.0002 | 5.0002 | 5.0001 | 5.0001 | 5.0001 |

Table 3. The results of $C_p$, $S_p$, $R_p$, $GS_p$, $MGS_p$ and $WMGS_p$ statistic for all subsets with multicollinearity but without Two outliers.

| predictors in the model | $C_p$ | $S_p$ | $R_p$ | $JRMGS_p$ | $LRMGS_p$ | $JRMMGS_p$ | $LRMMGS_p$ | $WMGS_p$ |
|---|---|---|---|---|---|---|---|---|
| X1 | 3.8973 | 3.9094 | 3.9084 | 3.8661 | 5.5661 | 5.6251 | 3.6657 | 5.9073 |
| X2 | 3.2769 | 2.1347 | 7.3906 | 4.816 | 3.6096 | 3.0488 | 5 | 9.9499 |
| X3 | 9.9811 | 9.981 | 9.876 | 11.1506 | 11.1648 | 7.345 | 11.9848 | 6.2683 |
| X4 | 3.3226 | 12.6222 | 7.4016 | 4.381 | 3.0525 | 5 | 25.0172 | 25.0255 |
| X1X2 | 4.9307 | 4.7111 | 1.4228 | 4.908 | 4.9674 | 3.0601 | 2.8781 | 2.9884 |
| X1X3 | 2.9719 | 2.6317 | 2.8131 | 2.8833 | 5.0873 | 12.4319 | 12.5145 | 12.5053 |
| X1X4 | 12.3303 | 3.1508 | 99.4556 | 94.7926 | 13.6548 | 80.8713 | 196.492 | 151.1141 |
| X2X3 | 24.5065 | 46.5347 | 48.4438 | 65.7852 | 26.4317 | 26.5169 | 26.4641 | 26.3331 |
| X2X4 | 89.5384 | 322.4569 | 342.8249 | 27.7359 | 301.9215 | 776.2565 | 751.8262 | 93.8792 |
| X3X4 | 179.3066 | 185.6802 | 259.0643 | 16.3246 | 16.3989 | 16.3915 | 16.2223 | 31.8927 |
| X1X2X3 | 1002.5346 | 895.0866 | 18.4254 | 809.0531 | 196.5196 | 151.3753 | 245.0908 | 465.3652 |
| X1X2X4 | 481.06.3 | 65.7865 | 29.5102 | 29.5915 | 29.5479 | 29.4007 | 94.0524 | 32.6174 |
| X1X3X4 | 34.3692 | 31.7214 | 30.2366 | 77.6673 | 75.7668 | 93.2571 | 17.9362 | 18.5737 |
| X2X3X4 | 259.1209 | 276.5048 | 327.5861 | 227.5425 | 127.3954 | 91.0448 | 323.9101 | 342.9685 |
| X1X2X3X4 | 5.0002 | 5.0002 | 5.0002 | 5.0001 | 5.0001 | 5.0000 | 5.0000 | 5.0000 |

## 4. DISCUSSION AND CONCLUSION

It is observed from the simulation study that, the assumed linear regression model(see eqn.) , the correct subset model contains the predictors are $X_1$ and $X_2$, i.e., the subset model{ $X_1, X_2$ }. It can be observed from the Table -1 that when multicollinearity is considered, the suggested subset selection criterion $WMGS_p$ selects the correct subset model along with the other existing methods $R_p$, $GS_p$, $MGS_p$. But if we observe carefully, the statistic value of $WMGS_p$ is so closer to *k*, the number of predictors in the subset model as compared to that of the criteria $GS_p$ , and $MGS_p$; and further note that the criteria *Cp*, $R_p$ and *Sp* do not select the correct subset model and these will select more than one subset models. Table-2 and Table- 3 results are obtained when both multicollinearity and outliers were introduced in the model, and we observe from these simulation results that the suggested criterion $WMGS_p$ yield better results than any other subset selection criteria in selecting the correct subset model{ $X_1, X_2$ }in all the three cases viz., multicollinearity, multicollinearity with one outlier etc. Thus from the simulation study it is observed that the proposed method of subset selection behaves better and it is comparable. However there is always a scope for research to verify the performance of the suggested estimator under different conditions such as various degree of multicollinearity viz., low, moderate, etc., large sample size, outliers, error variances, error distributions etc., it may yield poor results. Yet, under the high degree of multicollinearity, the performance of the proposed method is satisfactory as compared to that of the other criteria which are considered under study.

## REFERENCE

[1] D. Brickes, and Y. Dodge, *Alternative Methods of Regression*, John Wiley & Sons, New York, 1993, Ch.8, pp. 173-182.

[2] N. R. Draper, and H.Smith, *Applied Regression Analysis*, Third Edn., John Wiley, New York., 1998, ch. 10,

[3] A.V. Dorugade, and D.N. Kashid, Variable selection in Linear Regression Based Ridge Estimator, Journal of Statistical Computation and Simulation, vol. 80(11), pp.1211-1224, Dec.2009

[4] A.E. Hoerl, and R.W. Kennard, Ridge regression: Biased Estimation for nonorthogonal Problems, Technometrics,12(1), pp. 55-67, Feb. 1970.

[5] N.H. Jadhav, and D.N. Kashid, and S.R. Kulkarni, Subset selection in multiple Linear regression the presence of outlier and multicollinearity. Statistical methodology 19, pp.44-59, July, 2014.

[6] N.H. Jadhav, and D.N. Kashid, A Jackknifed ridge M-estimator fro regression model with multicollinearity and Outliers. Journal of Statistical Theory and Practice, vol. 5(4) pp. 659-673, Feb. 2012.

[7] D.N. Kashid, and S.R. Kulkarni, More general Criterion for Subset Selection in Multiple Linear regression, Communications in Statistics-Theory and Methods, vol. 31, pp. 795-811, Feb.2002.

[8] Kashid, D.N. and Kulkarni, S.R.(2002), Subset Selection in Multiple regression, with heavy tailed Error Distribution, J. Stat. Comp. And Simulation, 73,pp. 791-805, Feb. 2002.

[9] C.L. Mallows, Some Comments on Cp. Technometrics, 15(4), pp. 661-675, Nov.1973.

[10] D.C. Montgomery, E. A. Peck, G.G. Vinning, G.G. *Introduction to Linear Regression Analysis*, Third Edn., John Wiley and Sons, New York, 2006, Ch. 3, pp. 133-155.

[9] Satish Bhat, Robust method of Subset selection in Regression Analysis: A Substitute, International journal of all Education and Scientific Methods, vol.12(7), July-2024.