# Chapter 3

# Subset Selection in the Presence of Outlier and Multicollinearity

## 3.1 Introduction

In the previous chapter, we have developed a new estimation procedure for unknown regression parameters when the outliers and multicollinearity occurs simultaneously. The proposed JRM estimator performs superior as compare to the OLS, ORR, OJR and M-estimator in presence of outlier and multicollinearity in MSE sense. In this chapter, we developed a subset selection criterion based on proposed JRM estimator and the performance is evaluated through numerical examples.

In the past three decades, many robust parameter estimation methods in presence of outliers have been devised. Literature review regarding robust subset selection techniques is carried out in Chapter 1. Also, multicollinearity is a most serious and frequently encountered problem in multiple linear regression, which increases the variance of OLS estimator and hence, OLS estimator become unstable. To overcome such a problem, Hoerl and Kennard (1970) have proposed the method of ridge regression.

Many researchers have pointed out that the M-estimator is better alternative to OLS estimator in the presence of outliers (Birkes and Dodge, 1993) and ridge estimator is better in the presence of multicollinearity (Hoerl and Kennard, 1970; Hoerl, Kennard and Baldwin, 1975). Birkes and Dodge (1993), Montgomery, Peck, and Vining (2006) have given detail description of these methods in the context of parameter estimation.

In the above cited papers, subset selection method based on M-estimator and ORR estimator are good alternatives to OLS based methods in the presence of only outlier and multicollinearity respectively. However, these methods give misleading results when outlier and multicollinearity occur

simultaneously in the data. This fact is demonstrated in Section 3 through simulated data.

In this chapter, we have generalized the $S_p$ criterion called as $GS_p$ for subset selection based on JRM estimator when outliers and multicollinearity occur simultaneously in the data. Some properties of proposed statistic are studied and performance is evaluated through some numerical examples.

The rest of the chapter is organized as follows. In Section 3.2, various subset selection methods are discussed. In Section 3.3, how the problem of multicollinearity and outliers affect the existing subset selection criteria is demonstrated. In Section 3.4, we propose a new subset selection criterion $GS_p$ which combats the problem of occurrence of both outlier and multicollinearity in subset selection. Some results and equivalence of $GS_p$ with $C_p$, $S_p$ and $R_p$ is presented in Section 3.5. Section 3.6 illustrates the performance of $GS_p$ statistic in subset selection in the presence of outlier and multicollinearity for simulated as well as for real data sets. In Section 3.7, various robust estimates of $\sigma^2$ are presented and their performance is evaluated. Finally, the chapter ends with some conclusions.

## 3.2 Various Subset Selection Methods

In this section, we briefly introduced the following subset selection methods

1. $C_p$ criterion (Mallows, 1973)
2. $S_p$ criterion (Kashid and Kulkarni, 2002) when outlier present in the data
3. $R_p$ criterion (Dorugade and Kashid, 2010a) when multicollinearity present in the data.

### 3.2.1 $C_p$ Criterion

Mallows' $C_p$ statistic for subset selection is discussed in Section 1.5 of first Chapter. The performance of $C_p$ is not satisfactory in the presence of

multicollinearity or outlier in the data. In the next Section 3.3, we demonstrate the performance of $C_p$ in case of simultaneous occurrence of multicollinearity and outlier in the data.

### 3.2.2 $S_p$ Criterion

$S_p$ criterion is operationally simple to implement as compared to other robust subset selection methods and no need to modify it when we change the estimation procedure of regression coefficients. It is defined as

$$S_p = \frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\sigma^2} - (k-2p) \qquad (3.2.1)$$

where $k$ and $p$ are the number of parameters in the full and subset model respectively. $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ are the predicted values of $Y_i$ based on full model and subset model respectively. $\sigma^2$ is replaced by its suitable estimate $\hat{\sigma}^2 = 1.4826\, median|r_i - median(r_i)|$ and $r_i$ is the i[th] residual.

### 3.2.3 $R_p$ Criterion

This criterion is based on ORR estimator of regression coefficients. It is defined as

$$R_p = \frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\sigma^2} - tr\left(H_R^{'}H_R\right) + tr\left(H_{R1}^{'}H_{R1}\right) + p \quad (3.2.2)$$

where $\sigma^2$ is the unknown error variance to be replaced by its suitable estimate $\hat{\sigma}^2 = \left(Y - X\hat{\beta}_R\right)^{'}\left(Y - X\hat{\beta}_R\right)/(n-k)$ , $H_R = X\left(X^{'}X + rI\right)^{-1}X^{'}$ , $H_{R1} = X_A\left(X_A^{'}X_A + rI\right)^{-1}X_A^{'}$, $r$ is the biasing constant known as ridge constant or ridge parameter.

The above $S_p$ and $R_p$ statistics are equivalent to Mallows' $C_p$ when the OLS estimator is used.

## 3.3 Example: Effect of Outlier and Multicollinearity on Some Subset Selection Methods

In this section, we consider an example for illustrating the problem of multicollinearity and outliers in the viewpoint of subset selection. The

purpose of this section is to highlight how the simultaneous occurrence of outlier and multicollinearity affect the subset selection criteria based on various estimators like OLS estimator ($C_p$), M-estimator ($S_p$) and ridge estimator ($R_p$).

**Simulation Design**

The design proposed by McDonald and Galarneau (1975) is used to introduce multicollinearity in regressors as follows

$$x_{ij} = (1-\rho^2)^{1/2} z_{ij} + \rho z_{i(l+1)} \qquad i = 1,2,\dots,n, \qquad j = 1,2,\dots,l$$

where $z_{ij}$ are independent standard normal pseudo random numbers, $\rho^2$ is the correlation between any two regressor variables. Here, we used $n = 20$, $p = 4$ and $\rho = 0.999$. The response variable were generated using the regression model

$$Y_i = 10 + 3x_{i1} + 5x_{i2} + 0x_{i3} + 0x_{i4} + \varepsilon_i, \qquad i = 1,2,\dots 20.$$

where $\varepsilon_i$ are iid normally distributed with mean zero and variance $(0.5)^2$. The data is given in Table 3.3.1. The outlier is introduced in response variable corresponding to maximum absolute residual value. For this data, VIF values are 388.5831, 772.2531, 688.8542 and 296.6157. These VIF values indicate the severe multicollinearity in the data. We compute the values of $C_p$, $S_p$ and $R_p$ statistics for all possible subsets and results are reported in Table 3.3.2. Also, we plot these values for all possible subset models. Figure 3.3.1 shows the values of $C_p$, $S_p$ and $R_p$ versus $p$.

Table 3.3.1
Dataset

| Y | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|
| 7.3272 | -0.2614 | -0.3332 | -0.2737 | -0.2476 |
| -0.0977 | -1.2806 | -1.2601 | -1.2537 | -1.2458 |
| 11.7075 | 0.2011 | 0.1712 | 0.2823 | 0.1736 |
| 14.3274 | 0.6149 | 0.5928 | 0.5991 | 0.6173 |
| 25.9235 | 2.1414 | 2.0366 | 2.0539 | 2.1074 |
| 11.9593 | 0.1930 | 0.2185 | 0.2127 | 0.2776 |
| 8.8385 | -0.1504 | -0.1557 | -0.1142 | -0.1275 |
| 25.5449 | 1.9931 | 1.9564 | 1.9508 | 2.0669 |
| 22.5133 | 1.5604 | 1.4809 | 1.4901 | 1.4345 |
| 9.7142 | 0.0395 | -0.0204 | -0.0038 | -0.0044 |

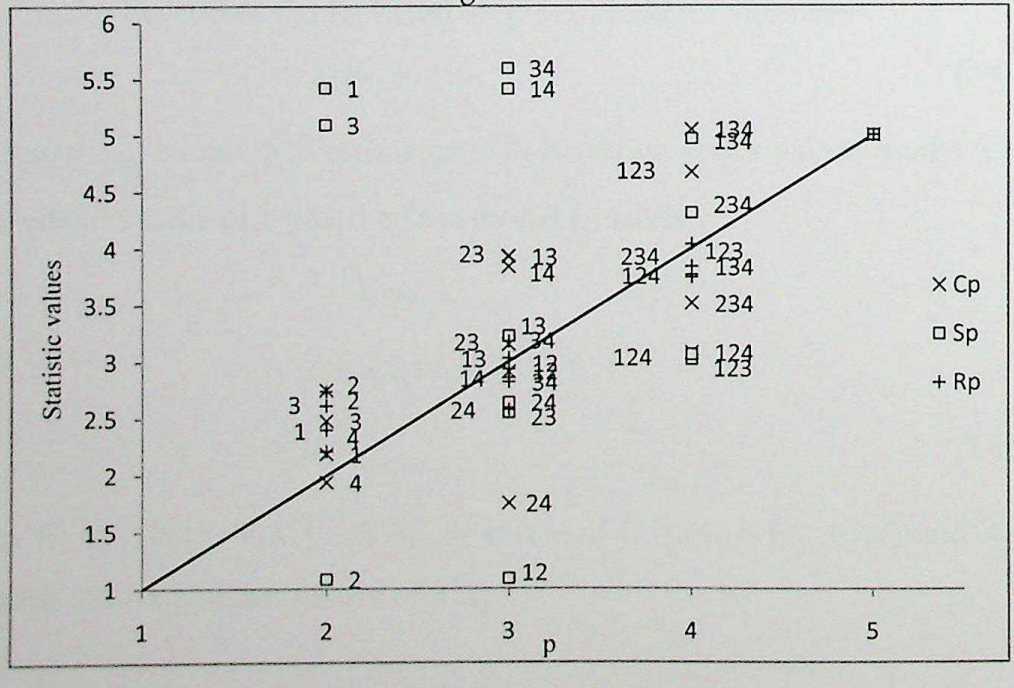
| | | | | |
|---|---|---|---|---|
| 6.8789 | -0.3658 | -0.3557 | -0.3874 | -0.4363 |
| 5.0733 | -0.6483 | -0.6797 | -0.6296 | -0.7055 |
| **30.7255** | -0.9095 | -0.9783 | -0.9203 | -0.8731 |
| 4.7199 | -0.6268 | -0.7092 | -0.6401 | -0.6227 |
| 2.9593 | -0.9802 | -0.9534 | -0.9913 | -0.9199 |
| 13.7421 | 0.3444 | 0.4100 | 0.4222 | 0.4034 |
| 4.9845 | -0.6844 | -0.5656 | -0.5384 | -0.6305 |
| 13.2898 | 0.4338 | 0.3867 | 0.4501 | 0.4035 |
| 13.4822 | 0.4670 | 0.4251 | 0.4164 | 0.5098 |
| 6.7525 | -0.3960 | -0.4657 | -0.4831 | -0.5465 |

Table 3.3.2

Values of $C_p$, $S_p$ and $R_p$ statistic for all subset models

| Regressors in subset model | $C_p$ | $S_p$ | $R_p$ |
|---|---|---|---|
| $X_1$ | 2.1903 | 5.4235 | 2.3993 |
| $X_2$ | 2.7501 | 1.0843 | 2.7413 |
| $X_3$ | 2.4824 | 5.0971 | 2.6124 |
| $X_4$ | 1.9387 | 14.6250 | 2.2086 |
| $X_1 X_2$ | 2.9131 | 1.0914 | 2.9373 |
| $X_1 X_3$ | 3.9301 | 3.2280 | 3.0330 |
| $X_1 X_4$ | 3.8417 | 5.4157 | 2.8594 |
| $X_2 X_3$ | 3.9397 | 2.5536 | 3.1816 |
| $X_2 X_4$ | 1.7539 | 2.6440 | 2.5774 |
| $X_3 X_4$ | 3.1544 | 5.5966 | 2.8219 |
| $X_1 X_2 X_3$ | 4.6727 | 3.0192 | 4.0344 |
| $X_1 X_2 X_4$ | 3.0811 | 3.0597 | 3.7427 |
| $X_1 X_3 X_4$ | 5.0529 | 4.9692 | 3.8366 |
| $X_2 X_3 X_4$ | 3.5144 | 4.3136 | 3.7644 |
| $X_1 X_2 X_3 X_4$ | 5 | 5 | 5 |

Figure 3.3.1






BARR. B. A. KHARDEKAR LIBRARY
SHIVAJI UNIVERSITY, KOLHAPUR.

From Table 3.3.2 as well as Figure 3.3.1, it seems that these three criteria fail to select proper subset $\{X_1, X_2\}$. It is clear that, when both outlier and multicollinearity are present in the data, $C_p$, $S_p$ and $R_p$ statistics select the wrong subset model. This study indicates that subset selection method based on OLS, M- and ridge estimator leads to absurd result.

### 3.4 Generalized Version of $S_p$ Statistic ($GS_p$)

Consider the multiple linear regression model as given in Eq. (1.1.2). The JRM estimator of $\beta$ based on the model (1.1.2) is denoted by $\hat{\beta}_{JRM}$. Then the predicted value of $Y$ using JRM estimator of unknown regression parameters of regression model given in Eq. (1.1.2) is

$$\begin{aligned}\hat{Y}_k &= X\hat{\beta}_{JRM} \\ &= X(I - r^2A^{-2})\left(X'WX\right)^{-1}X'WY \\ &= XR\left(X'WX\right)^{-1}X'WY \\ &= HY \end{aligned} \tag{3.4.1}$$

where $H = XR\left(X'WX\right)^{-1}X'W$, $R = (I - r^2A^{-2})$, $A = (\Lambda + rI)$ and $\Lambda$ is diagonal matrix of eigen values of $X'X$.

Consider the subset model based on $p-1$ predictor variables

$$Y = X_1\beta_1 + \varepsilon \tag{3.4.2}$$

Suppose $\hat{\beta}_{1_{JRM}}$ be the JRM estimator of $\beta_1$ based on above subset model, then the predicted value of $Y$ based on the model (3.4.2) is

$$\begin{aligned}\hat{Y}_p &= X_1\hat{\beta}_{1_{JRM}} \\ &= X_1R_1\left(X_1'W_1X_1\right)^{-1}X_1'W_1Y \\ &= H_1Y \end{aligned} \tag{3.4.3}$$

where $H_1 = X_1R_1\left(X_1'W_1X_1\right)^{-1}X_1'W_1$, $R_1 = (I - r_1^2A_1^{-2})$, $A_1 = (\Lambda_1 + r_1I)$ and $\Lambda_1$ is diagonal matrix of eigen values of $X_1'X_1$.

Based on two vectors of predicted values $\hat{Y}_k$ and $\hat{Y}_p$ of $Y$, we propose a generalized version of $S_p$ statistic. Because, Sp statistic is suitable for any type of estimators, like OLS, M-estimator. The main objective to define new criterion is that to select subset model of size $p$ ($\leq k$) such that it will give prediction 'as accurate as' that based on the full model in the presence of outliers and multicollinearity. The motivation behind the $S_p$ criterion proposed by Kashid and Kulkarni (2002) is used for proposing new criterion. It is given below.

### 3.4.1 Motivation

Suppose $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ be the $i^{th}$ predicted values of $Y$ based on $(k-1)$ and $(p-1)$ regressors respectively. The quantity $\frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\sigma^2}$ is small when $\hat{Y}_{ik}$ is close to $\hat{Y}_{ip}$ . That is, the prediction based on model with $(p-1)$ regressors is as accurate as that based on model with $(k-1)$ regressors. When the quantity $\frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\sigma^2}$ is large then $\hat{Y}_{ik}$ is far away from $\hat{Y}_{ip}$ and it implies that prediction based on the model with $(p-1)$ regressors may not be as 'accurate' as that based on the model with $(k-1)$ regressors. Using the same logic, we propose the following criterion for subset selection.

### 3.4.2 Definition of $GS_p$ Statistic

The proposed criterion is denoted by $GS_p$ and defined as

$$GS_p = \frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\sigma^2} - \left( tr(H'H) - tr(H_1'H_1) - p \right) \qquad (3.4.4)$$

where $\sigma^2$ and $\beta$ are replaced by their suitable estimates, $H$ and $H_1$ are defined at the beginning of this section.

**Remark 3.4.1** The purpose of subtracting second term from first term of Eq. (3.4.4) is simply to compare $GS_p$ with dimensions of subset model for selection of an appropriate subset model.

**Remark 3.4.2** The choice of estimator is based on nature of data and second term $(tr(H'H) - tr(H_1'H_1) - p)$ depends on choice of estimator. In case of OLS estimator, $(tr(H'H) - tr(H_1'H_1) - p) = k - 2p$ , in case of M-estimator, $(tr(H'H) - tr(H_1'H_1) - p) = k - 2p$. In case of ORR estimator, it reduces to $tr\,(H_R'H_R) - tr\,(H_{R1}'H_{R1}) - p$.

### 3.5 Some Results

In this section, we present some results to support the use of the proposed criterion to select the correct subset. Also, we have derived equivalence of proposed statistic $GS_p$ with $C_p$, $S_p$ and $R_p$ which are based on OLS estimator, M-estimator and ORR estimator respectively.

**Theorem: 3.5.1.** If predicted value is based on JRM estimator then

$$\sum_{i=1}^{n}(\hat{Y}_{ik} - \hat{Y}_{ip})^2 = RSS_p - RSS_k - 2Y'(H - H'H - H_1 + H_1'H)Y$$

**Proof**: We can write,

$$\begin{aligned}
\sum_{i=1}^{n}(\hat{Y}_{ik} - \hat{Y}_{ip})^2 &= (\hat{Y}_k - \hat{Y}_p)'(\hat{Y}_k - \hat{Y}_p) \\
&= (\hat{Y}_k - Y + Y - \hat{Y}_p)'(\hat{Y}_k - Y + Y - \hat{Y}_p) \\
&= ((Y - \hat{Y}_p) - (Y - \hat{Y}_k))'((Y - \hat{Y}_p) - (Y - \hat{Y}_k)) \\
&= (Y - \hat{Y}_p)'(Y - \hat{Y}_p) + (Y - \hat{Y}_k)'(Y - \hat{Y}_k) \\
&\quad -2(Y - \hat{Y}_p)'(Y - \hat{Y}_k) \\
&= (Y - \hat{Y}_p)'(Y - \hat{Y}_p) + (Y - \hat{Y}_k)'(Y - \hat{Y}_k) \\
&\quad -2(Y - \hat{Y}_k + \hat{Y}_k - \hat{Y}_p)'(Y - \hat{Y}_k) \\
&= RSS_p + RSS_k - 2RSS_k - 2(\hat{Y}_k - \hat{Y}_p)'(Y - \hat{Y}_k) \\
&= RSS_p - RSS_k - 2Y'(H - H'H - H_1 + H_1'H)Y \qquad (3.5.1)
\end{aligned}$$

The last term in Eq. (3.5.1) depends on estimator of $\beta$.

**Theorem: 3.5.2.** If OLS estimator of $\beta$ is used then

$$\sum_{i=1}^{n}(\hat{Y}_{ik} - \hat{Y}_{ip})^2 = RSS_p - RSS_k$$

**Proof:** From Theorem 3.5.1,

$$\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2 = RSS_p - RSS_k - 2Y'(H - H'H - H_1 + H_1'H)Y.$$

When OLS estimator is used for parameter estimation, then

$$Y'(H - H'H - H_1 + H_1'H)Y = 0.$$

Hence,

$$\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2 = RSS_p - RSS_k \qquad (3.5.2)$$

**Theorem: 3.5.3.** If selected subset model of $(p-1)$ regressors is adequate (that is $\beta_2 = 0$) then

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right] \cong \sigma^2\left[tr\,(H'H) - tr\,(H_1'H_1)\right],$$

**Proof**: From result 3.5.1,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right] = E(RSS_p) - E(RSS_k) - 2E\left[Y'(H - H'H - H_1 + H_1'H)Y\right] \qquad (3.5.3)$$

The expected values of R.H.S. of Eq. (3.5.3) are

$$E(RSS_k) = \sigma^2 tr\left(I - H - H' + H'H\right) + \beta'X'\left(I - H - H' + H'H\right)X\beta \qquad (3.5.4)$$

Similarly,

$$E(RSS_p) = \sigma^2 tr\left(I - H_1 - H_1' + H_1'H_1\right) + \beta'X'\left(I - H_1 - H_1' + H_1'H_1\right)X\beta \qquad (3.5.5)$$

Using expression for expectation of a quadratic form, we write

$$E\left[Y'(H' - H'H - H_1' + H_1'H)Y\right] = \sigma^2 tr(H' - H'H - H_1' + H_1'H) + \beta'X'(H - H'H - H_1' + H_1'H)\,X\beta \qquad (3.5.6)$$

Assuming, $H_1'H \cong H_1'H_1$, Using Eq. (3.5.4) to Eq. (3.5.6) in Eq. (3.5.3), we can write,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right] \cong \sigma^2\left[tr\left(H'H\right) - tr\left(H_1'H_1\right)\right] + \beta'X'\left(H' - H + H_1' - H_1 + H'H - H_1'H_1\right)X\beta.$$

When subset model is adequate, then biasing constant $r = r_1$ and

$$\beta'X'\left(H' - H + H_1' - H_1 + H'H - H_1'H_1\right)X\beta = 0.$$

Hence,

$$E\left[\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2\right] \cong \sigma^2[tr\left(H'H\right)-tr\left(H_1'H_1\right)]. \quad (3.5.7)$$

**Theorem: 3.5.4.** If subset model is adequate then

$$E[GS_p] \cong p.$$

**Proof**: When subset model is adequate, from result 3.5.3 we can write,

$$E\left[\frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2}\right] \cong \left[tr\left(H'H\right)-tr\left(H_1'H_1\right)\right]$$

Therefore,

$$E[GS_p] \cong \left[tr\left(H'H\right)-tr\left(H_1'H_1\right)\right]-\left[tr\left(H'H\right)-tr\left(H_1'H_1\right)-p\right]$$
$$= p. \quad (3.5.8)$$

We have mentioned that second term of proposed statistic is based on type of estimator and hence, this statistic reduces to some existing subset selection methods. Below, we discuss the relationship of $GS_p$ statistic with existing methods.

**Theorem: 3.5.5.** If $\hat{\beta}$ is OLS estimator of β, then $GS_p$ reduces to $C_p$ statistic.

**Proof**: Suppose $\hat{\beta}_{OLS}$ is OLS estimator of $\beta$. In this case, the matrix $H$ reduces to $H = X\left(X'X\right)^{-1}X'$ and $H_1 = X_1\left(X_1'X_1\right)^{-1}X_1'$. Both matrices are symmetric and idempotent.

Hence,

$$tr\left(H'H\right) = k \text{ and } tr\left(H_1'H_1\right) = p.$$

Using result 5.1, $GS_p$ can be written as

$$GS_p = \frac{RSS_p - RSS_k}{\sigma^2} - (k-2p) \quad (\text{since, } Y'(H - H'H - H_1 + H_1'H)Y = 0)$$
$$= \frac{RSS_p}{\sigma^2} - n - k - (k-2p) \qquad (\text{where } \sigma^2 \text{ is replaced by } \frac{RSS_k}{n-k})$$
$$= \frac{RSS_p}{\sigma^2} - (n-2p)$$

Therefore,

$$GS_p = C_p \quad (3.5.9)$$

Thus, $GS_p$ reduces to $C_p$ when the OLS estimator of $\beta$ is used.

**Theorem: 3.5.6.** When $\hat{\beta}$ is M-estimator of $\beta$, then $GS_p$ reduces to $S_p$.

**Proof**: Let $\hat{\beta}_M$ be M estimator of $\beta$, prediction matrix H becomes

$$H = X\left(X'WX\right)^{-1}X'W .$$

Hence,

$$tr\left(H'H\right) = k .$$

Similarly,

$$tr\left(H_1'H_1\right) = p.$$

Hence,

$$GS_p = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2} - (k-2p)$$

Therefore,

$$GS_p = S_p \tag{3.5.10}$$

Thus, $GS_p$ is equivalent to $S_p$ when the M-estimator of $\beta$ is used.

**Theorem: 3.5.7.** If $\hat{\beta}$ is ordinary ridge regression estimator of $\beta$, then $GS_p$ reduces to $R_p$ statistic.

**Proof**: Suppose, $\hat{\beta}_R$ is ordinary ridge regression estimator of $\beta$, then matrix H and $H_1$ become $H_R$ and $H_{R1}$.

Hence,

$$GS_p = \frac{\sum_{i=1}^{n}\left(\hat{Y}_{ik}-\hat{Y}_{ip}\right)^2}{\sigma^2} - \left[tr\left(H_R'H_R\right) - tr\left(H_{1R}'H_{R1}\right)\right] + p$$

$$GS_p = R_p \tag{3.5.11}$$

Thus, $GS_p$ is equivalent to $R_p$ when the ORR estimator of $\beta$ is used.

In the following section, we illustrate the performance of proposed method and compare with the existing methods.

### 3.6 Numerical Examples

To illustrate the proposed method, we considered real and simulated data and compared with the Mallow's $C_p$, $S_p$ and $R_p$ statistic. We have considered four examples and following different situations.

1. Clean data (data without outliers and multicollinearity).
2. Outlier data.
3. Multicollinearity in the data.
4. Both outlier and multicollinearity in the data.

Example 3.6.1 illustrate the performance of $C_p$, $S_p$, $R_p$ and $GS_p$ statistics when there is neither outlier nor multicollinearity present in the data. Problem of outlier is considered in Example 3.6.2. How these four criteria perform when multicollinearity is present in the data is worked out in Example 3.6.3. Finally, the performance of $C_p$, $S_p$, $R_p$ and $GS_p$ statistic is studied in Example 3.6.4 when outlier and multicollinearity simultaneously occur in the data.

Before proceeding to examples, we consider the problem given in Section 3.3. For the same simulated data, we compute the $C_p$, $S_p$, $R_p$ and $GS_p$ statistic for all possible subset models and presented in Table 3.6.1. It seems that, $GS_p$ statistic selects proper subset model $\{X_1, X_2\}$ and $C_p$, $S_p$ and $R_p$ statistic fails to do so.

Table 3.6.1
Values of $C_p$, $S_p$, $R_p$ and $GS_p$ statistic for all subset models

| Subset model | $C_p$ | $S_p$ | $R_p$ | $GS_p$ |
|---|---|---|---|---|
| $X_1$ | 2.1903 | 5.4235 | 2.3993 | 9.0449 |
| $X_2$ | 2.7501 | 1.0843 | 2.7413 | 4.2613 |
| $X_3$ | 2.4824 | 5.0971 | 2.6124 | 7.3075 |
| $X_4$ | 1.9387 | 14.6250 | 2.2086 | 19.9638 |
| **$X_1$ $X_2$** | 2.9131 | 1.0914 | 2.9373 | **2.8727** |
| $X_1$ $X_3$ | 3.9301 | 3.2280 | 3.0330 | 4.4664 |
| $X_1$ $X_4$ | 3.8417 | 5.4157 | 2.8594 | 7.5699 |
| $X_2$ $X_3$ | 3.9397 | 2.5536 | 3.1816 | 4.5578 |
| $X_2$ $X_4$ | 1.7539 | 2.6440 | 2.5774 | 5.3128 |
| $X_3$ $X_4$ | 3.1544 | 5.5966 | 2.8219 | 7.0581 |
| $X_1$ $X_2$ $X_3$ | 4.6727 | 3.0192 | 4.0344 | 3.6290 |
| $X_1$ $X_2$ $X_4$ | 3.0811 | 3.0597 | 3.7427 | 4.3270 |
| $X_1$ $X_3$ $X_4$ | 5.0529 | 4.9692 | 3.8366 | 5.4269 |
| $X_2$ $X_3$ $X_4$ | 3.5144 | 4.3136 | 3.7644 | 5.5973 |
| $X_1$ $X_2$ $X_3$ $X_4$ | 5 | 5 | 5 | 5 |

**Example: 3.6.1. Clean Data:**

In this example, we have considered the model

$$Y_i = 1 + 2X_{i1} + 3X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where, $X_{ij} \sim U(0,1)$, i = 1, 2, …, n, j = 1, 2, 3, 4 and $\varepsilon_i \sim N(0, 0.25)$. For this model, we generate the random sample of size n = 20 and compute the $C_p$, $S_p$, $R_p$ and $GS_p$ statistic for all possible subset models. These values are tabled in the Table 3.6.2.

Table 3.6.2
Values of $C_p$, $S_p$, $R_p$ and $GS_p$ for all subset Models

| Subset model | $C_p$ | $S_p$ | $R_p$ | $GS_p$ |
|---|---|---|---|---|
| $X_1$ | 109.497 | 99.970 | 113.530 | 98.700 |
| $X_2$ | 15.538 | 13.632 | 14.990 | 13.483 |
| $X_3$ | 129.343 | 117.119 | 192.420 | 115.946 |
| $X_4$ | 135.374 | 123.000 | 1074.99 | 120.726 |
| **$X_1$ $X_2$** | **2.876** | **2.733** | **2.970** | **2.725** |
| $X_1$ $X_3$ | 99.162 | 90.366 | 99.290 | 89.207 |
| $X_1$ $X_4$ | 110.870 | 101.508 | 113.790 | 100.107 |
| $X_2$ $X_3$ | 17.416 | 15.483 | 16.720 | 15.335 |
| $X_2$ $X_4$ | 17.043 | 15.272 | 16.400 | 15.125 |
| $X_3$ $X_4$ | 131.337 | 119.126 | 194.500 | 118.027 |
| $X_1$ $X_2$ $X_3$ | 3.458 | 3.318 | 3.470 | 3.337 |
| $X_1$ $X_2$ $X_4$ | 4.090 | 4.136 | 4.180 | 4.132 |
| $X_1$ $X_3$ $X_4$ | 101.135 | 92.367 | 100.620 | 91.213 |
| $X_2$ $X_3$ $X_4$ | 18.995 | 17.202 | 18.210 | 17.060 |
| $X_1$ $X_2$ $X_3$ $X_4$ | 5 | 5 | 5 | 5 |

It is clear that, these four criteria pickup the same subset model namely {$X_1$, $X_2$} for the clean data. This indicates all four methods agree upon the importance of two variables $X_1$ and $X_2$. We suggest the use of any of them for subset selection in case of clean data.

**Example: 3.6.2. Outlier Data:**

In this example, we use the simulated data in Example 3.6.1. The value corresponding to first observation of response variable ($Y_1$) is replaced by 200 to introduce outlier observation. The values of $C_p$, $S_p$, $R_p$ and $GS_p$ are computed and presented in the following table.

Table 3.6.3
Values of $C_p$, $S_p$, $R_p$ and $GS_p$ in the presence of outlier

| Subset model | $C_p$ | $S_p$ | $R_p$ | $GS_p$ |
|---|---|---|---|---|
| $X_1$ | 1.858 | 100.007 | 1.808 | 98.693 |
| $X_2$ | 1.644 | 13.751 | 1.780 | 13.430 |
| $X_3$ | 3.650 | 117.257 | 2.910 | 115.561 |
| $X_4$ | 3.107 | 123.253 | 2.227 | 120.643 |
| **$X_1$ $X_2$** | 2.157 | **2.733** | 2.481 | **2.646** |
| $X_1$ $X_3$ | 3.847 | 90.362 | 2.891 | 89.217 |
| $X_1$ $X_4$ | 3.069 | 101.487 | 2.757 | 100.220 |
| $X_2$ $X_3$ | 3.207 | 15.732 | 2.857 | 15.434 |
| $X_2$ $X_4$ | 2.936 | 15.369 | 2.762 | 15.097 |
| $X_3$ $X_4$ | 4.838 | 119.369 | 3.169 | 117.531 |
| $X_1$ $X_2$ $X_3$ | 3.989 | 3.319 | 3.672 | 3.258 |
| $X_1$ $X_2$ $X_4$ | 3.346 | 4.155 | 3.645 | 4.129 |
| $X_1$ $X_3$ $X_4$ | 4.996 | 92.362 | 3.900 | 91.428 |
| $X_2$ $X_3$ $X_4$ | 4.234 | 17.468 | 4.031 | 17.237 |
| $X_1$ $X_2$ $X_3$ $X_4$ | 5 | 5 | 5 | 5 |

From Table 3.6.3, it seems that when the data contains outlier then the values of $C_p$ and $R_p$ are drastically changed and both the criteria fail to select the proper subset model. As a contrast, we can see that $S_p$ and $GS_p$ select the same subset which is selected for clean data in Example 3.6.1. This suggests that $S_p$ and $GS_p$ performs well in case of outlier data because in both the statistics, M-estimator is used for estimation.

**Example: 3.6.3. Data with Multicollinearity:**

Here, we consider the real data given by Neter, Wasserman and Kutner (1989). The response variable is body fat (Y) and the three predictor variables are triceps skin fold thickness ($X_1$), thigh circumference ($X_2$) and mid arm circumference ($X_3$). The Condition indices are 125740083.2782, 455189.1106 and 185354.7541. The Variance inflation factor (VIF) for each regressor is 676.5, 538.6 and 99.8. This condition number and VIF values indicate the severe multicollinearity. The values of four criteria are computed and presented in the Table 3.6.4.

Table 3.6.4

Values of $C_p$, $S_p$, $R_p$ and $GS_p$ in presence of multicollinearity

| Subset Model | $C_p$ | $S_p$ | $R_p$ | $GS_p$ |
|---|---|---|---|---|
| $X_1$ | 7.3545 | 9.0873 | 6.837 | 6.3026 |
| **$X_2$** | 2.5086 | 3.2424 | **1.6930** | **2.0190** |
| $X_3$ | 63.2104 | 73.0221 | 185.392 | 13.2074 |
| $X_1$ $X_2$ | 3.9419 | 4.6679 | 2.677 | 3.1113 |
| $X_1$ $X_3$ | 3.2606 | 3.7899 | 2.92 | 3.0409 |
| $X_2$ $X_3$ | 4.1377 | 4.8936 | 3.222 | 3.6086 |
| $X_1$ $X_2$ $X_3$ | 4 | 4 | 4 | 4 |

It is clear that, $C_p$ and $S_p$ pick up more than one subset model as a proper subset and $R_p$ and $GS_p$ criteria select the same subset model with single regressor namely {$X_2$}. Therefore, $R_p$ and $GS_p$ statistics are suitable for subset selection in the presence of multicollinearity.

**Example 3.6.4: Data with Outlier and Multicollinearity**

In this example, we use the data given in Example 3.6.3 and outlier observation is introduced manually by changing response variable $Y_8$ to 50. In this data multicollinearity and outlier are present. Values of $C_p$, $S_p$, $R_p$ and $GS_p$ for all possible subset models are computed and reported in the following table.

Table 3.6.5

Values of $C_p$, $S_p$, $R_p$ and $GS_p$ in presence of outlier and multicollinearity

| Subset Model | $C_p$ | $S_p$ | $R_p$ | $GS_p$ |
|---|---|---|---|---|
| $X_1$ | 1.9513 | 9.0951 | 1.3809 | 6.1513 |
| **$X_2$** | 2.9588 | 3.2844 | 2.2652 | **1.9776** |
| $X_3$ | 10.5397 | 73.5006 | 14.0444 | 9.4328 |
| $X_1$ $X_2$ | 3.9136 | 4.6937 | 2.2959 | 3.0833 |
| $X_1$ $X_3$ | 3.8297 | 3.8122 | 2.8170 | 3.0196 |
| $X_2$ $X_3$ | 4.1875 | 4.9234 | 2.9967 | 3.5889 |
| $X_1$ $X_2$ $X_3$ | 4 | 4 | 4 | 4 |

From the above table, it is clear that the values of $C_p$, $S_p$ and $R_p$ are affected due to presence of multicollinearity and outlier. As a contrast it can be seen that $GS_p$ statistic is more stable and select the same subset which is selected in

Example 3.6.3. This indicates $GS_p$ statistic is proper subset selection method in the presence of outlier and multicollinearity. For implementing the proposed method is essential to obtain the estimator of $\sigma^2$. In the next section, we have discussed various estimators of $\sigma^2$.

### 3.7 Choice of Estimator of $\sigma^2$

In computation of $GS_p$ statistic, we use a scale parameter $\sigma^2$. Since it is unknown, we have to use its suitable estimator. Various estimators of $\sigma^2$ are available in literature. For examining the performance of $GS_p$ criterion, we use five different types of estimators of $\sigma^2$ which are based on the JRM estimator. The i$^{th}$ residual based on JRM estimator $(\hat{\beta}_{JRM})$ of $\beta$ is defined as,

$$r_i = Y_i - X_i'\hat{\beta}_{JRM}, \quad i = 1, 2, \dots, n.$$

We consider the following estimator of $\sigma^2$.

1. $\hat{\sigma}_1^2 = [1.4826\ Median\ |r_i - Median\ (r_i)|]^2$
2. $\hat{\sigma}_2^2 = [1.4826\ Median\ |r_i|]^2$
3. $\hat{\sigma}_3^2 = [\ Median\ |r_i|]^2$
4. $\hat{\sigma}_4^2 = \sum_{i=1}^{n} r_i^2 w_i^2\ /(n-k)$
5. $\hat{\sigma}_5^2 = \sum_{i=1}^{n} r_i^2 w_i^2\ /(\sum_{i=1}^{n} |r_i w_i|)$

Performance of these various estimators is illustrated by using a simulated data. Here, we have generated a random sample of size n = 20 from $N_4(0, \Sigma)$, on $X_1$, $X_2$, $X_3$, and $X_4$, where,

$$\Sigma = \begin{bmatrix} 1 & 0.531 & -0.850 & -0.531 \\ 0.531 & 1 & -0.401 & -0.972 \\ -0.850 & -0.401 & 1 & 0.288 \\ -0.531 & -0.972 & 0.288 & 1 \end{bmatrix}$$

We consider the subset model as

$$Y_i = 5 + 2X_{i1} + 4X_{i2} + \varepsilon_i, \quad i = 1, 2, \dots, 20$$

where $\varepsilon_i \sim N(0, 1)$. VIF values for this data are 25.2, 118.2, 27.4 and 134.8. We have introduced outlier artificially at $Y_6 = 200$. Values of $GS_p$ statistic for this data for these five estimators of $\sigma^2$ are presented below.

| Estimator of $\sigma^2$ | $\hat{\sigma}_1^2$ | $\hat{\sigma}_2^2$ | $\hat{\sigma}_3^2$ | $\hat{\sigma}_4^2$ | $\hat{\sigma}_5^2$ |
|---|---|---|---|---|---|
| Estimate | 1.2456 | 1.8921 | 0.7410 | 1.3130 | 1.1392 |

Using the above estimates of $\sigma^2$, $GS_p$ statistic is calculated for all possible subset models. The results are given in the following table.

Table 3.7.1
Values of $GS_p$ statistic using different estimates of $\sigma^2$

| Subset model | $\hat{\sigma}_1^2$ | $\hat{\sigma}_2^2$ | $\hat{\sigma}_3^2$ | $\hat{\sigma}_4^2$ | $\hat{\sigma}_5^2$ |
|---|---|---|---|---|---|
| $X_1$ | 8442.9658 | 5557.8458 | 14194.1878 | 8009.8015 | 9232.0755 |
| $X_2$ | 3001.1548 | 1975.3621 | 5045.9784 | 2847.1450 | 3281.7195 |
| $X_3$ | 13885.0228 | 9140.4976 | 23342.7984 | 13172.6922 | 15182.6987 |
| $X_4$ | 3839.7163 | 2527.4058 | 6455.6868 | 3642.6894 | 4198.6466 |
| **$X_1$ $X_2$** | **2.0294** | **1.6528** | **2.7803** | **1.9729** | **2.1325** |
| $X_1$ $X_3$ | 8395.5152 | 5527.2544 | 14113.1298 | 7964.8820 | 9180.0137 |
| $X_1$ $X_4$ | 642.2925 | 423.1486 | 1079.1359 | 609.3908 | 702.2307 |
| $X_2$ $X_3$ | 1033.2324 | 680.5289 | 1736.3144 | 980.2784 | 1129.7004 |
| $X_2$ $X_4$ | 2981.1376 | 1962.7376 | 5011.2246 | 2828.2377 | 3259.6804 |
| $X_3$ $X_4$ | 207.4042 | 136.8665 | 348.0146 | 196.8139 | 226.6970 |
| $X_1$ $X_2$ $X_3$ | 3.7276 | 3.4782 | 4.2247 | 3.6902 | 3.7958 |
| $X_1$ $X_2$ $X_4$ | 4.0255 | 3.6667 | 4.7408 | 3.9717 | 4.1237 |
| $X_1$ $X_3$ $X_4$ | 121.8318 | 81.2257 | 202.7764 | 115.7353 | 132.9380 |
| $X_2$ $X_3$ $X_4$ | 197.9376 | 131.3198 | 330.7342 | 187.9358 | 216.1583 |
| $X_1$ $X_2$ $X_3$ $X_4$ | 5.0000 | 5.0000 | 5.0000 | 5.0000 | 5.0000 |

From Table 3.7.1, it is clear that the $GS_p$ statistic select the same subset {$X_1$, $X_2$} for all estimators of $\sigma^2$ in the presence of multicollinearity and outlier in the data. It indicates that performance of all estimators of scale parameters of $\sigma^2$ is same for selection of subset model.

### 3.8 Conclusion

We have constructed subset model selection procedure based on JRM estimator of regression coefficients and this method works well for clean data (data without outliers and multicollinearity) or in the presence of only outlier or only multicollinearity or both outlier and multicollinearity. In this method, one can use any type of estimator (either OLS or M-estimator or JRM estimator). This is prominent feature of the proposed method. To carry out

the simulation and to study the performance of proposed method, Matlab software is used and source code is given in Appendix - II.

In the previous two chapters, we have considered the problem of outlier and multicollinearity and have proposed the new estimator and subset selection criterion based on the proposed estimator. In the next chapter, we consider the situation of violation of normality assumption of error term. We have developed a subset selection method when the distribution of error term is symmetric but non-normal.