Chapter 4

Subset Selection Technique for Non-normal Error Terms

4.1 Introduction

In Chapter 2, we have proposed parameter estimation procedure in the presence of outlier and multicollinearity and using the same estimator, in Chapter 3, we have developed a subset selection criterion when the outlier and multicollinearity present in the data.

In this chapter, we consider the problem of non-normal distribution of error variable. The violation of assumption of normality is frequently occurring in many practical situations. The performance of OLS estimator is not reasonably well when the error distribution is non-normal. The M-estimation procedure performs well if the errors have a distribution which is not necessarily normal but close to normal. A rank based parameter estimation procedure is intended to perform well for almost any possible distribution of the errors (see Birkes and Dodge, 1993). So, we have used this estimator in this chapter. The rank based estimator has been proposed to provide more robust regression estimate of regression parameters (see Jaeckel, 1972; Hettmansperger and McKean, 1977; Hettmansperger, 1984).

In the literature, very few subset selection methods are available for non normal distribution of error terms. Usually, these methods are based on rank estimator. The N_p criterion proposed by Kashid and Kulkarni (2003) is based on rank estimator. It is defined as

$$N_p = \frac{\sum_{i=1}^{n}(\hat{Y}_{ik}-\hat{Y}_{ip})^2}{\tau^2} - (k-2p) \qquad (4.1.1)$$

where $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ denote the predicted value of Y_i based on $(k-1)$ regressors (full model) and $(p-1)$ regressors (subset model) respectively and k and p are the number of parameters of full model and subset model respectively.

In this chapter, we propose a new criterion for subset selection using rank based estimator of regression parameters when error variable is non-normally

distributed. Distribution of proposed statistic is obtained by simulation study. Under certain assumption, equivalence of proposed statistic and Mallows' C_p is proved. Also, a stepwise procedure for subset selection based on the same criterion is given.

This chapter is organized as follows. In Section 4.2, the parameter estimation method proposed by Jaeckel (1972) and Hettmansperger-McKean (1977) is discussed briefly. In Section 4.3, we have proposed subset selection method. The distribution of proposed statistic is obtained through simulation study in Section 4.4. In Section 4.5, some results related to the proposed statistic are given. The performance of proposed statistic in subset selection for simulated examples and the proper model selection ability of the method is studied in Section 4.6. A stepwise procedure based on proposed statistic with illustration is presented in Section 4.7. The chapter ends with some concluding remarks.

4.2 Regression Model and Estimation of Regression Coefficients

In this section we discuss the parameter estimation procedure for non normal error distribution. Now, we can write the multiple linear regression model given in Eq. (1.1.2) as

$$Y_i = \beta_0 + X_i^{*'}\beta^* + \varepsilon_i \qquad i = 1, 2, \ldots, n. \qquad (4.2.1)$$

where $X_i^{*'} = [X_{i1}, X_{i2}, \ldots, X_{i(k-1)}]$ is i^{th} row of a $n \times (k-1)$ matrix of independent $(k-1)$ regressor variables, β_0 and $\beta^* = (\beta_1, \beta_2, \ldots, \beta_{k-1})'$ are the unknown regression parameters, and ε_i is the i^{th} term of $n \times 1$ error vector.

We consider the following two assumptions:

(i) ε_i's are independent, identically distributed with a distribution $f_\varepsilon(.)$.

(ii) $f_\varepsilon(.)$ is absolutely continuous and symmetric about zero.

Jaeckel (1972) proposed a class of rank based estimates for estimating the regression coefficients. Hettmansperger and McKean (1977) proposed a rank based estimate of β_0. Below, we briefly discuss the estimation procedure proposed by Jaeckel (1972) and Hettmansperger-McKean (1977).

The rank based estimate $\tilde{\beta}_0$ and $\tilde{\beta}^* = (\tilde{\beta}_1, \tilde{\beta}_2, \dots, \tilde{\beta}_{k-1})'$ of unknown regression parameters are the values of β_0 and β^*, which minimize

$$\sum_{i=1}^{n} a(\varepsilon_i)\, \varepsilon_i \qquad (4.2.2)$$

where $a(\varepsilon_i)$ denotes the score function which is non-decreasing in ε_i. In the literature, various score functions are available. Wilcoxon score is most commonly used score function and its normalized form is given by

$$a(\varepsilon_i) = \sqrt{12}\left[\frac{rank(\varepsilon_i)}{(n+1)} - \frac{1}{2}\right].$$

Now we discuss estimation procedure of regression coefficients. We can write (4.2.2) using Wilcoxon scores as

$$\frac{\sqrt{12}}{(n+1)} \sum_{i=1}^{n} \left[rank\ (Y_i - \beta_0 - X_i'\beta^*) - \frac{(n+1)}{2}\right] (Y_i - \beta_0 - X_i'\beta^*) \qquad (4.2.3)$$

Note that, the value of Equation (4.2.3) is unaffected by the value of β_0 and $\frac{\sqrt{12}}{(n+1)}$. Hence, minimizing (4.2.3) with respect to β_0 and β^*, is equivalent to minimize

$$g(\beta^*) = \sum_{i=1}^{n} \left[rank\ (Y_i - X_i'\beta^*) - \frac{(n+1)}{2}\right] (Y_i - X_i'\beta^*) \qquad (4.2.4)$$

β^* can be estimated by minimizing Equation (4.2.4) using an iterative procedure. The estimator of β_0 is obtained as the median of the differences

$$(Y_i - X_i'\tilde{\beta}^*), \qquad i = 1, 2, \dots, n.$$

where $\tilde{\beta}^*$ is rank based estimator of β^* (see Birkes and Dodge, 1993).

Stepwise Algorithm for Rank Based Parameter Estimation

The iterative stepwise procedure for computing $\tilde{\beta}$ is as follows:

Step I. First obtain the ordinary least squares estimates of β^* say $\hat{\beta}^{*0} = (\hat{\beta}_1^0, \hat{\beta}_2^0, \dots, \hat{\beta}_{k-1}^0)'$.

Step II. Obtain the predicted values and residuals using OLS estimates of β^*

$$\hat{Y}_i = X_i'\hat{\beta}^{*0}$$

$$r_i = (Y_i - \hat{Y}_i), \qquad i = 1, 2, \dots, n.$$

Step III. Rank the residuals r_i and compute $U_i^0 = rank(r_i) - \frac{n+1}{2}$.

Suppose $U^0 = (U_1^0, U_2^0, \dots, U_n^0)'$.

Step IV. Compute $d = (X_c'X_c)^{-1}X_c'U^0$, where X_c is a $n \times (k-1)$ matrix with its entries are $X_{ij} - \bar{X}_j$ and $\bar{X}_j$ is a mean of the j[th] regressor.

Step V. Compute $W_i = d'X_i,\ i = 1, 2, \ldots, n.$

Step VI. Compute the ratios $T_h = \frac{(r_i - r_j)}{(w_i - w_j)}$ and weights $W_h^* = \frac{|w_i - w_j|}{\Sigma|w_i - w_j|},\ 1 \le i < j \le n.$

Step VII. Arrange the $T_h,\ h = 1, 2, \ldots, n(n+1)/2$ in an increasing order with corresponding weights W_h^* and calculate the cumulative sum that first exceeds 0.5. Then T_h is the weighted median and let it be denoted by t*.

Step VIII. The current vector of estimates is $\tilde{\beta}^*_{current} = \tilde{\beta}^* + t^* d.$

Step IX. Repeat Step II to Step VIII until

$$\max_j \left|\tilde{\beta}^*_{jcurrent} - \tilde{\beta}^*_j\right| \le \delta$$

where δ is a small specified number. At convergence, $\tilde{\beta}^*_{current}$ is the solution of the $\tilde{\beta}^*$.

Step X. Obtain residual $r_i = \left(Y_i - \hat{Y}_i\right)$ and hence obtain the median of residuals, which is an estimate of β_0.

In the next section, we develop new subset selection method using rank based estimator of regression parameters.

4.3 Proposed Method for Subset Selection

First, we define the residual and sum of rank weighted residual.

Let $\tilde{\beta}$ be the rank based estimator of β for the model (4.2.1). The i[th] residual for this model is given by,

$$r_{ik} = Y_i - X_i'\tilde{\beta}, \qquad i = 1, 2, \ldots, n. \tag{4.3.1}$$

Definition: Sum of Rank Weighted Residual:

Sum of rank weighted residuals ($SRWR$) for the model (4.2.1) is defined as,

$$SRWR = \sum_{i=1}^{n} a(r_{ik})\, r_{ik}$$

Using Wilcoxon score, the $SRWR$ can be written as

$$SRWR = \frac{\sqrt{12}}{(n+1)} \sum_{i=1}^{n} \left[rank\,(r_{ik}) - \frac{(n+1)}{2} \right] r_{ik} \qquad (4.3.2)$$

The sum of rank weighted residual is obtained using full model. Therefore, it is denoted by $SRWR_k$.

Consider the subset model 'A' based on $(p-1)$ regressor variables $(p < k)$ is,

$$Y = X_1\beta_1 + \varepsilon, \qquad (4.3.3)$$

where X_1 is an $n \times p$ matrix of regressor variables with 1's in first column. $\beta_1 = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ is $p \times 1$ vector of β. Suppose, $\tilde{\beta}_1$ be rank based estimator of β_1 based on subset model (4.3.3). The ith residual for subset model is given by

$$r_{ip} = Y_i - X_{1i}'\tilde{\beta}_1, \qquad i = 1, 2, \ldots, n. \qquad (4.3.4)$$

The sum of rank weighted residuals $(SRWR)$ for subset model based on $(p-1)$ regressor variables with Wilcoxon score function is given by,

$$SRWR_p = \frac{\sqrt{12}}{(n+1)} \sum_{i=1}^{n} \left[rank\,(r_{ip}) - \frac{(n+1)}{2} \right] r_{ip} \qquad (4.3.5)$$

It is well known that the sum of rank weighted residuals is used to measure goodness of fit of the model to data. To measure the discrepancy, we use the scaled difference between the sum of rank weighted residuals of subset model and full model, which is given by $\frac{SRWR_p - SRWR_k}{c\tau}$, τ is known or it is replaced by suitable estimate.

If the quantity $\frac{SRWR_p - SRWR_k}{c\tau}$ is small, it indicates that subset model with p parameters well fits to the data and so it will be enough to consider $(p-1)$ regressors for future prediction. Conversely, if $SRWR_p$ is far away from $SRWR_k$, it indicates the subset model is not an appropriate fit to the data. Based on this reasoning, we propose the following criterion for subset selection in linear regression with non-normal error terms.

4.3.1 Definition of NR_p Statistic

The NR_p statistic is defined as,

$$NR_p = \frac{SRWR_p - SRWR_k}{c\tau} - (k - 2p) \qquad (4.3.6)$$

where $c = 1/2$, τ is a scale parameter and is replaced by its suitable estimate. In the following subsection, we consider a simulation experiment to study the distributional behavior of NR_p statistic.

4.4 Distribution of NR_p Statistic

In this section, we study the distributional behavior of proposed statistic. If the error distribution is normal and $\rho(\varepsilon_i) = \varepsilon_i$ then NR_p statistic excluding constant term $(k - 2p)$ is a classical dispersion measure and it follows chi-square distribution with $(k - p)$ degrees of freedom for proper subset. Except by analogy with the classical methods and without much justification, McKean and Hettmansperger (1976) state the distribution of $\frac{SRWR_p - SRWR_k}{c\tau}$ for proper subset model as chi-square with $(k - p)$ degrees of freedom. To support this, we have carried out simulation study for different non-normal symmetric error distributions. The simulation design is given below.

The four regressor variables X_1 , X_2 , X_3 and X_4 were generated independently from $U(0,1)$. The subset model specification and error distribution considered in this experiment are listed in the Table 4.4.1. Note that the choice of subset model given in Table 4.3.1 is arbitrary and for the sake of illustration, it is used here.

Table 4.4.1
Subset model specification

Model No.	Subset model specification	Error distribution $f_\varepsilon(.)$
4.4.1	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Laplace (0, 1)
4.4.2	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Laplace (0, 2)
4.4.3	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Students t (2)
4.4.4	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Students t (4)
4.4.5	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Slash
4.4.6	$Y = 15 + 5X_1 + 8X_2 + \varepsilon$	Cauchy (0, 1)

We have used sample sizes n = 30, 50, 75 and 100. For each model, 200 samples were generated. The NR_p statistic is computed only for the proper subset $\{X_1, X_2\}$ excluding the constant term $(k - 2p)$. Q-Q plot (see D'Agostino and Stephens, 1986) and chi-square goodness of fit test are used to analyze the goodness of fit of chi-square distribution to NR_p statistic excluding constant term $(k - 2p)$.

4.4.1 Q-Q plots

A simple graphical tool Q-Q plot is used for understanding the nature of distribution of NR_p statistic excluding constant term $(k - 2p)$. For each sample size, Q-Q plots were obtained for the models given in Table 4.4.1. For the sake of brevity, we present Q-Q plots for sample size n = 30 and 50 in Figures 4.4.1- 4.4.12. From these plots, it can be clearly seen that NR_p statistic excluding constant term $(k - 2p)$ has approximately chi-square distribution with $(k - p)$ degrees of freedom for proper subset model.

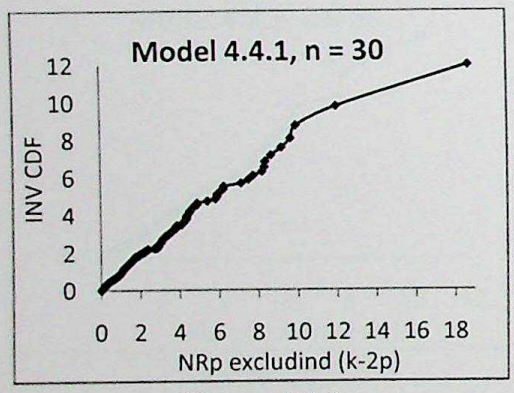


Figure 4.4.1

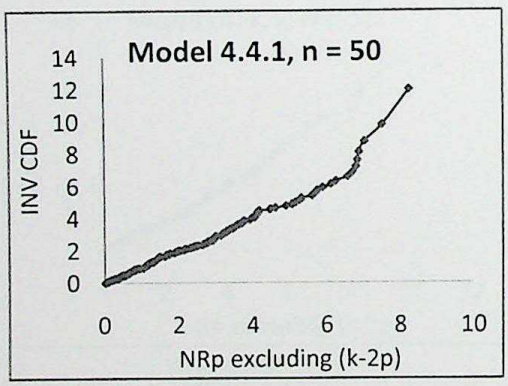


Figure 4.4.2

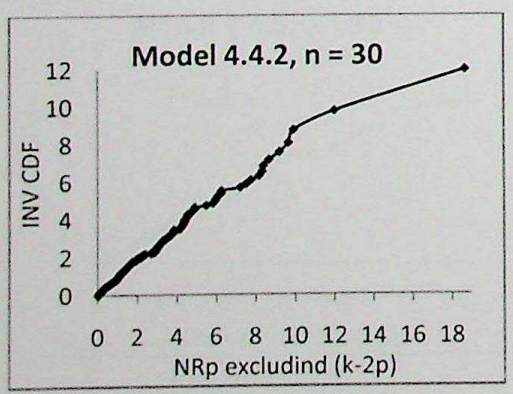


Figure 4.4.3

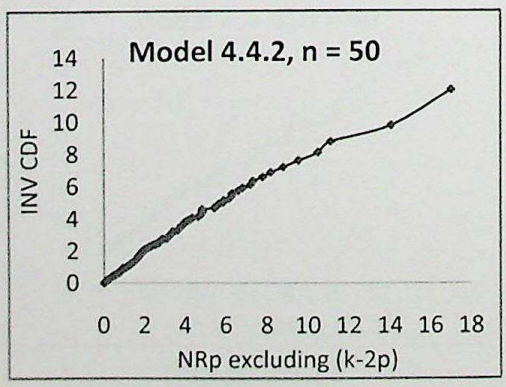


Figure 4.4.4

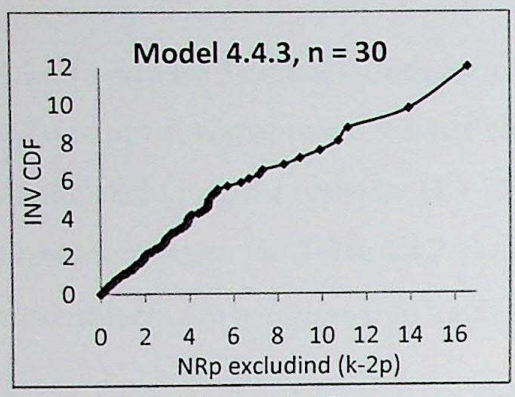


Figure 4.4.5

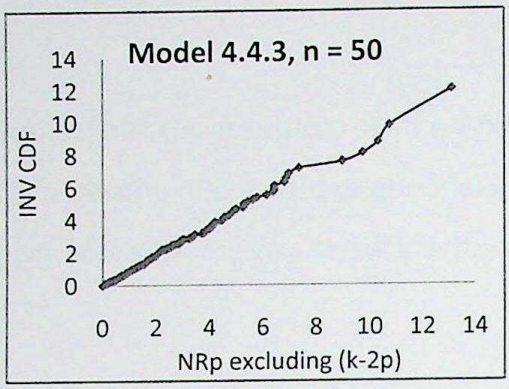


Figure 4.4.6

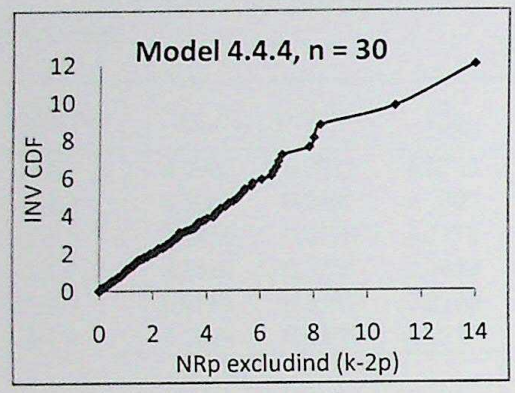


Figure 4.4.7

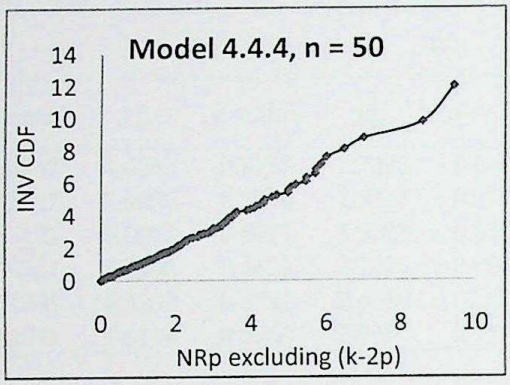


Figure 4.4.8

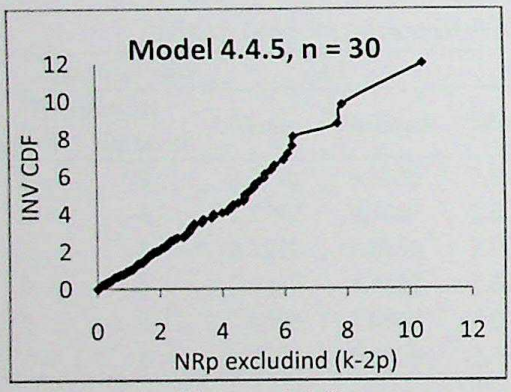


Figure 4.4.9

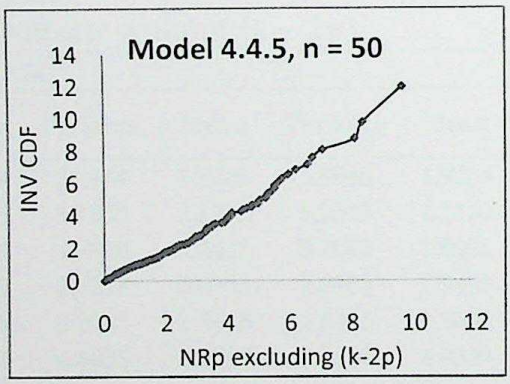


Figure 4.4.10

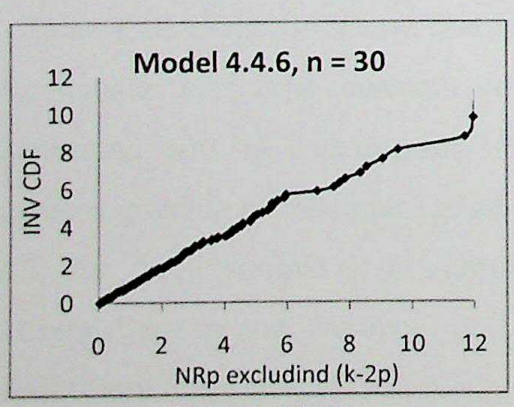


Figure 4.4.11

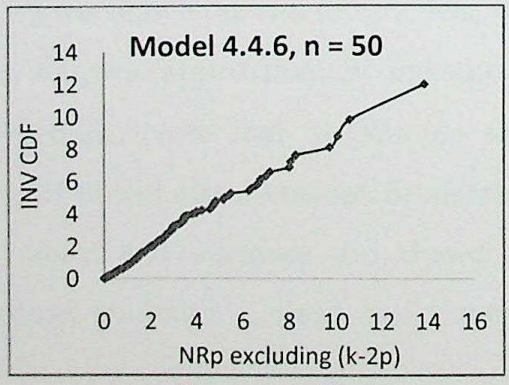


Figure 4.4.12

4.4.2 Chi-square Test for Goodness of Fit

To support above result we use chi-square test for goodness of fit for NR_p statistic excluding the constant $(k-2p)$. The results of chi-square goodness of fit test are reported in Table 4.4.2. Summary statistics of NR_p statistic without constant term are mentioned in Table 4.4.3.

Table 4.4.2
Chi-square goodness of fit test for NR_p statistic without $(k-2p)$.

Sample Size (n)	30		50		75		100	
Model No.	χ^2_{cal}	p-value	χ^2_{cal}	p-value	χ^2_{cal}	p-value	χ^2_{cal}	p-value
4.4.1	4.2967	0.2312	6.5724	0.4747	4.0063	0.6758	2.7422	0.8404
4.4.2	4.2064	0.2400	2.1309	0.7117	2.0039	0.7350	2.5354	0.7712
4.4.3	3.5336	0.4728	4.2274	0.5172	2.0061	0.5711	3.3022	0.6535
4.4.4	4.1686	0.5254	4.1819	0.6521	3.2406	0.6629	2.2424	0.8961
4.4.5	6.8084	0.3389	5.5140	0.4798	4.3209	0.5042	2.4539	0.7834
4.4.6	4.7629	0.4455	4.4317	0.4891	1.6149	0.6560	2.0557	0.7255

Table 4.4.3
Mean and variance of NR_p statistic without $(k-2p)$.

Sample Size			30		50		75		100	
Model No.	Theoretical Mean	Theoretical Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
4.4.1	2	4	2.2770	6.3335	1.9939	3.2444	1.8705	3.9840	1.9234	3.6502
4.4.2	2	4	2.2792	6.3407	2.2215	6.1538	2.1783	5.3303	2.1160	4.9459
4.4.3	2	4	2.1272	5.8198	2.2068	4.7320	2.2415	5.7065	1.9026	4.2803
4.4.4	2	4	2.0267	4.4122	1.8435	2.8218	2.0393	4.0784	1.9433	3.5821
4.4.5	2	4	1.8890	3.0948	2.0416	3.3848	1.9648	4.8355	2.0658	4.8974
4.4.6	2	4	2.2093	5.1970	2.1233	4.6035	2.0097	5.2669	1.9329	4.2411

Table 4.4.2 clearly indicates that for all the above considered models, the NR_p statistic excluding constant term follows approximately chi-square distribution with $(k-p)$ degrees of freedom. Note that, as sample size increases, p-value of chi-square goodness of fit test also increases. From table 4.4.3, for large sample size, sample mean and variance are closed to theoretical mean and variance for Laplace, student's t, slash and Cauchy distribution.

4.5 Some Results

In this section, we present the result which gives the relationship between NR_p statistic and Mallows' C_p.

Result 4.5.1: If $a(\varepsilon_i) = \varepsilon_i$, $c = \tau$ and error follows normal distribution then the proposed statistic NR_p is equivalent to Mallows' C_p.

Proof. Consider the identity score function $a(\varepsilon_i) = \varepsilon_i$, then equation defined in (4.2.2) becomes,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - X_i'\beta^*\right)^2 \tag{4.5.1}$$

Minimizing (4.5.1) with respect to β_0 and β^* leads to ordinary least squares estimator and is given by

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y.$$

From Eq. (4.3.2),

$$SRWR_k = \sum_{i=1}^{n} a(r_i)\, r_i$$
$$= \sum_{i=1}^{n} r_i^2.$$

This is residual sum of squares based on LS estimator and it is denoted by RSS.

Hence,

$$SRWR_k = RSS_k$$

Similarly, for subset model

$$SRWR_p = RSS_p.$$

Hence the NR_p statistic defined in Equation (4.3.6) becomes,

$$NR_p = \frac{RSS_p - RSS_k}{\tau^2} - (k - 2p) \tag{4.5.2}$$

where τ is a scale parameter. Under normality of error, τ^2 is replaced by its unbiased estimator $\hat{\tau}^2 = \frac{RSS_k}{(n-k)}$. Then,

$$NR_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \tag{4.5.3}$$
$$= C_p$$

Thus, NR_p equivalent to Mallows' C_p- statistic when the error structure is normal and score function is identity.

Result 4.5.2: If the subset model is adequate then

$$E[NR_p] \cong p \tag{4.5.4}$$

Proof. From Section 4.4, it is clear that the distribution of $\frac{SRWR_p - SRWR_k}{c\tau}$ is approximately chi-square with $(k-p)$ degrees of freedom. Hence,

$$\begin{aligned} E(NR_p) &= E\left[\frac{SRWR_p - SRWR_k}{c\tau} - (k-2p)\right] \\ &\cong (\mathrm{k}-\mathrm{p}) - (k-2p) \\ &= p. \end{aligned}$$

It indicates that a subset model for which the value of NR_p is close to p is the best subset model and it is adequate for prediction. Thus, using this theorem we develop the following algorithm for subset selection based on NR_p statistic.

Algorithm for Subset Selection Based on NR_p Statistic

Step 1.

Compute the NR_p statistic corresponding to all $2^{(k-1)} - 1$ possible subsets, where $(k-1)$ denotes the number of regressors in the full model.

Step 2.

Select a subset of minimum size for which

$$NR_p \leq p.$$

To implement NR_p criterion, we need an estimate of τ. Below we discuss the procedure of estimation of τ.

Estimator of τ:

Let $\hat{Y}_{ik}, i = 1, 2, \ldots, n$ be the predicted values of Y obtained from the full model. The i^{th} residual is given by

$$r_i = (Y_i - \hat{Y}_{ik}), \qquad i = 1, 2, \ldots, n$$

Let A_{ij} denote the pair-wise averages of the residuals that is,

$$A_{ij} = \frac{(r_i + r_j)}{2} \quad \text{for } i \leq j \leq n.$$

We observe that, there are $N = n(n+1)/2$ pairs of residual averages. Arrange these pairs in an ascending order as $A(1),\ A(2), \ldots, A(N)$.

We denote,

$$a = \frac{n(n+1)}{2}$$

$$b = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

c_1 = the closest integer to ½ + a - 1.645b

c_2 = the closest integer to ½ + a + 1.645b

$$f_c = \left[\frac{n}{(n-p)}\right]^{(1/2)}$$

Then the estimator of τ is given by

$$\hat{\tau} = f_c \sqrt{n}\left[\frac{A_{c_2} - A_{c_1}}{2}\right] \times 1.645 \qquad (4.5.5)$$

The detail description about parameter estimation procedure and scale parameter estimation procedure is given by Birkes and Dodge (1993).

4.6 Performance Evaluation of NR_p statistic

In this section, we have considered simulated examples to illustrate the use of NR_p statistic in subset selection. In both examples, we consider a simulated data with different distributions for error term. The model selection ability of NR_p is also evaluated through simulation study.

4.6.1. Numerical Examples

In Example 4.6.1, we compare performance of NR_p statistic and Mallows' C_p statistic for normal error term. In Example 4.6.2, we have considered non-normal distribution of error terms.

Example 4.6.1: In this example, the regressor variables X_1, X_2, X_3 and X_4 are generated from $U(0,1)$ and ε is generated from $N(0,1)$. Sample of size 30 is used and response Y is generated using the following models.

$$\text{Model I}: \quad Y = 5 + 2X_1 + 3X_2 + 4X_3 + 0X_4 + \varepsilon$$

$$\text{Model II}: \quad Y = 1 + 4X_1 + 2X_2 + 0X_3 + 0X_4 + \varepsilon$$

The values of NR_p and C_p statistics for all possible subsets are computed and presented in Table 4.6.1.

Table 4.6.1
Values of NR_p and C_p

Subset Model	Model I		Model II	
	NR_p	C_p	NR_p	C_p
X_1	44.1425	69.3170	19.9484	20.0854
X_2	33.8721	50.0449	52.4946	69.6637
X_3	29.0386	40.2889	57.6377	81.0226
X_4	45.7195	76.4886	57.6888	81.0342
X_1X_2	30.7913	41.4550	**2.8235**	**2.9043**
X_1X_3	23.5183	28.9632	21.9364	22.0569
X_1X_4	41.1343	64.3803	20.2022	20.7533
X_2X_3	14.8306	17.2789	54.1328	71.6262
X_2X_4	34.4314	51.8108	54.3763	71.6518
X_3X_4	30.1535	41.4258	59.5668	83.0226
$X_1X_2X_3$	**3.3164**	**3.2132**	4.7731	4.8983
$X_1X_2X_4$	27.0976	38.4272	3.0209	3.0461
$X_1X_3X_4$	25.3346	30.3029	22.1247	22.6710
$X_2X_3X_4$	15.7757	17.3480	55.9782	73.6179
$X_1X_2X_3X_4$	5	5	5	5

From Table 4.6.1, it is clear that the NR_p statistic and C_p statistic select proper subset for both models when the error distribution is normal. That is, performance of NR_p statistic and C_p statistic equally good for normal error distribution.

Example 4.6.2: In this example, we consider the random sample of size 30 and generated from $U(0,1)$ on X_1, X_2, X_3, X_4 and X_5. The response variable is generated using the following model as

$$Y = 1 + 3X_1 + 4X_2 + 5X_3 + \varepsilon,$$

where $\varepsilon \sim f_\varepsilon(.)$. Using the simulated data, the values of NR_p and C_p statistic for all possible subset models are computed for different error distributions and reported in Table 4.6.2.

Table 4.6.2
Values of NR_p and C_p

Regressors in the model	Distribution of error variable $(f_\varepsilon(.))$							
	L(0,2)		student (2)		slash		Cauchy (0,1)	
	NR_p	C_p	NR_p	C_p	NR_p	C_p	NR_p	C_p
X_1	4.3700	2.7525	18.2623	7.5054	21.1210	-0.0366	11.3718	0.6077
X_2	6.3025	5.7193	21.2441	9.0061	20.3778	2.1019	4.6185	0.0552
X_3	6.5748	3.5468	13.3355	3.1155	21.8905	1.6655	9.3454	0.7076
X_4	5.2612	2.5346	21.9252	9.5309	28.1813	1.8143	11.6784	0.1089
X_5	7.9224	5.6407	20.7217	8.6447	27.6863	1.7140	11.1322	-0.7080
X_1X_2	5.1941	4.7460	16.1185	7.4790	12.3566	1.6753	4.7727	1.9550
X_1X_3	4.2099	2.1414	6.7166	1.5926	9.5623	0.8123	10.3435	2.5644
X_1X_4	4.2191	2.3121	19.8321	9.1840	23.1027	1.7267	13.3739	2.0526
X_1X_5	6.1680	4.5847	18.0505	8.2421	23.1088	1.7713	12.6603	0.8900
X_2X_3	5.2778	4.7115	12.5671	4.1202	18.9685	3.6473	5.2075	2.0467
X_2X_4	6.0971	4.4765	21.8023	10.2271	22.3415	3.7885	5.0642	1.8180
X_2X_5	8.2743	7.5663	21.0935	9.7236	21.7203	3.6782	6.3160	0.7145
X_3X_4	6.1090	2.9311	13.5033	4.4920	23.8203	3.2511	10.9890	2.0465
X_3X_5	7.6585	4.7246	11.4820	3.4408	22.1895	2.9914	11.3368	1.0752
X_4X_5	7.2599	4.5331	20.5260	9.6783	29.6716	3.4300	12.7146	1.0355
$X_1X_2X_3$	**4.0540**	3.7655	**3.3089**	2.4159	**3.2828**	2.6035	**4.3047**	3.9313
$X_1X_2X_4$	5.0591	4.3121	17.9857	9.4098	14.1367	3.4931	5.3520	3.7549
$X_1X_2X_5$	7.1310	6.5840	16.9545	8.9429	14.3364	3.5174	6.6679	2.3577
$X_1X_3X_4$	4.6847	2.3951	8.3971	3.5584	11.4661	2.4866	12.1028	3.9667
$X_1X_3X_5$	5.1961	3.4067	5.9899	2.7193	9.8981	2.4046	12.3286	2.4685
$X_1X_4X_5$	6.1000	4.3107	19.3320	9.9636	25.0928	3.5599	14.4634	2.7439
$X_2X_3X_4$	5.6861	4.4406	13.5773	5.7333	20.9742	5.2472	4.9415	3.7851
$X_2X_3X_5$	6.9776	6.1595	11.4288	4.9993	18.5326	4.9855	7.2110	2.5469
$X_2X_4X_5$	8.0511	6.4760	21.5432	11.0323	23.6979	5.4110	5.8148	2.6723
$X_3X_4X_5$	7.6036	4.7271	12.0885	4.9846	24.1869	4.6317	12.9783	2.7798
$X_1X_2X_3X_4$	4.4716	4.1743	5.3074	4.4145	5.2249	4.3348	3.9957	5.7039
$X_1X_2X_3X_5$	5.4838	5.1985	4.0830	4.0002	4.0305	4.2454	6.2172	4.0153
$X_1X_2X_4X_5$	7.0052	6.3105	18.7195	10.8607	16.1345	5.3527	6.5603	4.3508
$X_1X_3X_4X_5$	6.1142	4.1541	7.6858	4.6936	11.8249	4.1137	14.0821	4.3102
$X_2X_3X_4X_5$	7.6016	6.3238	12.6330	6.6667	20.5194	6.6315	6.7926	4.4813
$X_1X_2X_3X_4X_5$	6	6	6	6	6	6	6	6

From Table 4.6.2, it seems that the NR_p selects proper subset $\{X_1, X_2, X_3\}$ for all error distributions considered whereas C_p fails to select the proper subset for Laplace (0, 2), Students t with 2 degrees of freedom, slash and Cauchy (0, 1) distribution. The C_p statistic selects more than one under fitted subset models for all considered error distributions. In the following, we present the model selection ability of NR_p.

4.6.2 Model Selection Ability

The following simulation study illustrates the performance of the proposed statistic in model selection problem. Four independent regressor variables, X_1, X_2, X_3 and X_4 are generated from $U(0,1)$ and error term generated from different distributions. Response variable Y is obtained using the regression model

$$Y_i = 15 + 5X_{i1} + 8X_{i2} + 0X_{i3} + 0X_{i4} + \varepsilon_i, \qquad i = 1, 2, \ldots, n.$$

For sample size, $n = 30, 50, 75$ and 100, the experiment is repeated 500 times and the number of times the correct subset model selected using N_p statistic and NR_p statistic is reported in Table 4.6.3.

Table 4.6.3
Model selection ability (in %)

$f_\varepsilon(.)$	$n = 30$		$n = 50$		$n = 75$		$n = 100$	
	N_p	NR_p	N_p	NR_p	N_p	NR_p	N_p	NR_p
Normal (0, 1)	73	76	72	77	72	76	72	75
Laplace (0, 1)	73	76	68	75	74	78	75	77
Laplace (0, 2)	60	65	72	75	74	78	77	80
Students t (2)	73	77	73	77	77	82	70	76
Students t (4)	69	76	73	78	72	77	74	78
Slash	55	60	69	72	75	78	74	78
Cauchy (0, 1)	68	71	72	76	79	82	73	76

It seems that, for large sample size ($n \geq 50$), NR_p statistic selects the correct subset model $\{X_1, X_2\}$ more than 72% times for all considered distributions for error terms. The correct model selection ability of NR_p criterion is uniformly greater than N_p. For small sample size with Slash and Cauchy error distribution, the performance of NR_p is moderate but better than N_p statistic.

In the subset selection procedure, we have to examine all possible subsets and then select a proper subset. But in some applications, to examine all possible subsets is laborious work. For example, if we have more than 30 regressors, we have more than billion subsets of regressors. Hence, it is not practically feasible to examine all possible subsets. So, it is reasonable to use stepwise procedure for variable selection. In the following section, we propose the stepwise procedure based on NR_p statistic.

4.7 Stepwise Procedure for Subset Selection Using NR_p Criterion

We propose stepwise procedure based on three criterions called (A) *Addition Criterion*, (B) *Deletion Criterion*, and (C) *Stopping Criterion*. The relation between NR_p and F- statistic is used for addition and deletion of variables. After each variable is added to the set of selected variables, a test is made to see whether the selected variable reduces $SRWR$ appreciably. A test is also made to see if any of the previously selected variables can be deleted without appreciably increasing sum of rank weighted residual. This procedure incorporates criterion for the addition and deletion of variables as follows.

(A) Addition Criterion

Let $SRWR_p$ denote the sum of rank weighted residuals based on $(p-1)$ variables and constant term in the model. Suppose $SRWR_{p+1}$ be the sum of rank weighted residuals corresponding to minimum NR_p value obtained by adding another variable to the present subset. The ratio

$$R_1 = \frac{SRWR_p - SRWR_{p+1}}{c\hat{\tau}_{p+1}}$$

is calculated and compared with $F_{1,\ n-p-1,\ \alpha}$, the upper α^{th} quantile of F distribution with 1 and $n-p-1$ degrees of freedom. If $R_1 \geq F_{1,n-p-1,\alpha}$ then the corresponding variable is added to the selected subset of variables.

(B) Deletion Criterion

With $p-1$ variables and a constant in the selected subset, let $SRWR_{p-1}$ be the sum of rank weighted residuals corresponding to minimum NR_p value obtained after deleting any variable from the previously selected variable. The ratio

$$R_2 = \frac{SRWR_{p-1} - SRWR_p}{c\hat{\tau}_p}$$

is calculated and compared with an $F_{1,\ n-p,\ \alpha}$. If $R_2 \leq F_{1,n-p,\alpha}$ then the corresponding variable is deleted from the selected subset model.

(C) Stopping Criterion

We use two stopping criteria as follows:

1. If $NR_p \leq P$ then stop the procedure. OR
2. Stop the procedure when no further addition of variables is possible.

Algorithm

Let S denote the set of regressor variables included in the subset model. Initially, let $S = \{X_0\}$. The stepwise procedure is as follows.

Step I.

Regress Y on each variable and compute NR_p statistic for each submodel. Select a variable corresponding to minimum NR_p.

Step II.

Add the selected variable in S and check the conditions. If minimum $NR_p \leq p$ then terminate the procedure otherwise go to next step.

Step III.

Regress response variable Y on previously selected variables and a variable from a set of unselected variables. Obtain the NR_p statistic. Repeat the same for all unselected variables. Select that subset which corresponds to minimum $\mathrm{NR_p}$. If minimum $NR_p \leq p$ then terminate the procedure otherwise go to next step.

Step IV.

Obtain R_1 corresponding to subset model selected in step III. If $R_1 \leq F_{1,n-p-1,\alpha}$ terminate the procedure otherwise to go next step.

Step V.

Regress response variable Y on currently selected variable and eliminating single variable from set of selected variables and obtain $\mathrm{NR_p}$. Repeat this for all previously selected variables. Compute

R_2 for the subset corresponding to minimum NR_p. If $R_2 \leq F_{1,n-p,\alpha}$ then remove that variable from S and or otherwise go to next step.

Step VI.

Repeat Step III to V until $NR_p \leq p$ or all the variables are included in S.

Illustration

To illustrate the above stepwise procedure, we considered a simulated data. Eight regressor variables are generated from $U(0,1)$ with sample of size 50. The response variable Y is generated using the following model

$$Y_i = 8 + 5X_{i1} + 6X_{i2} + 3X_{i3} + 4X_{i4} + 2X_{i5} + 0X_{i6} + 0X_{i7} + 0X_{i8} + \varepsilon_i, \quad i = 1, 2, \ldots, 50.$$

where $\varepsilon_i \sim C(0,1)$. The results are presented in Table 4.7.1. The total number of subset models is 255 and the number of subsets required to arrive at the decision is 33. This implies that, this stepwise method reduces approximately 87 percent computations for this example.

Table No. 4.7.1.
Stepwise procedure for subset selection

Step	Minimum NR_p	Minimum value of $NR_p \leq p$?	R_1	R_2	$F_{1,n-p-1,\alpha}$	$F_{1,n-p,\alpha}$	S	Decision
I	{1}* 57.5235	No ($p = 2$)	-	-	-	-	{1}	continue
II	{1,4} 37.8812	No ($p = 3$)	12.7791	25.9490	4.0471	4.0471	{1,4}	continue
III	{1,2,4} 17.6815	No ($p = 4$)	16.9330	18.2465	4.0517	4.0517	{1,2,4}	continue
IV	{1,2,3,4} 7.6399	No ($p = 5$)	11.7017	26.7865	4.0566	4.0566	{1,2,3,4}	continue
V	{1,2,3,4,5} 4.7179	Yes ($p = 6$)	-	-	-	-	{1,2,3,4,5}	Terminate

* { i, j, k, … } are the subscripts of X_i, X_j, X_k, ….

4.8 Conclusion

In this chapter, a subset selection method using rank based estimators of unknown regression coefficients is proposed. Proposed method performs well for all considered non-normal distributions of error terms and also for normal error distribution. Therefore, it doesn't need to modify the statistic when error follows normal distribution. The stepwise procedure based on NR_p reduces cost and time when model contains large number of variables. R-code for evaluation of performance of proposed method is given in the Appendix – III.

In the Chapter 3 and 4, we have proposed subset selection in multiple linear regression model under some different situations. These methods are appropriate when response variable is continuous. One more problem in regression, how select the subset model when response variable is count or categorical. This problem is addressed in the next chapter.