Chapter 1

Subset Selection Problem in Regression

1.1 Introduction

Regression analysis is a statistical data analysis tool that is useful in all areas like engineering, sciences, medical sciences, agricultural sciences, environmental sciences, business and marketing research etc. It is widely used statistical technique in data analysis. Linear regression uses data to construct a mathematical expression capable of estimating the actual, but unknown relationship between a set of independent or predictor variables and their corresponding response values.

Suppose that the yield of a crop (Y) depends on a number of factors such as soil fertility(X_1), weather conditions(X_2), water availability(X_3), amount of fertilizers used (X_4) , cultivation practices (X_5) . Assuming the linear relationship between the predictor variables and the response variable, we can write a multiple linear regression model as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_5 X_{i5} + \varepsilon_i \qquad (1.1.1)$$

This is multiple linear regression model with five predictor variables. The term linear is used because response variable (Y) is a linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_4$ and β_5. The parameter β_0 is the intercept of the model. The parameter $\beta_j, j = 1, 2, \ldots, 5$ indicate the expected change in response (Y) per unit change in X_j when all other regressors are held constant. ε_i denotes the error term.

In matrix form, the multiple linear regression model for n observations and $k - 1$ regressor variables can be described as

$$Y = X\beta + \varepsilon \qquad (1.1.2)$$

where Y is an $n \times 1$ vector of observed response values, X is an observed $n \times k$ matrix of regressor variables augmented with a column of ones, β is $k \times 1$ vector of unknown regression coefficients, ε is an $n \times 1$ vector of

random error variable representing the unexplained random variation in response.

The goal of regression analysis is to find a good estimate of the unknown regression coefficients β from the observed sample or data. The assumptions that are made in study of regression analysis are:

1. The errors are statistically independent from one another with expected value zero and constant variance, that is, $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$.
2. The errors are normally distributed. In notation, $\varepsilon \sim N_n(0,\ \sigma^2 I)$.

In most of the problems, the values of the parameters (regression coefficients) and the error variance σ^2 will not be known and they must be estimated from sample data. The method of ordinary least squares is widely used for parameter estimation in the regression. In the ordinary least squares method, the unknown parameters are estimated by minimizing the sum of the squared deviations between the actual Y values and the values of Y predicted by the equation. The ordinary least squares (OLS) estimator of regression parameter β given in Eq. (1.1.2) is defined as

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y. \tag{1.1.3}$$

Some properties of OLS estimator are listed as follows:

1. $\hat{\beta}_{OLS}$ is an unbiased estimator of regression parameter β.
2. The covariance matrix of OLS estimator is given by
   $$Cov(\hat{\beta}_{OLS}) = \sigma^2(X'X)^{-1}.$$
3. $\hat{\beta}_{OLS}$ is maximum likelihood estimator of β (Under normality assumption).

Standard text books like Montgomery, Peck and Vining (2006) give detailed description about the properties of the OLS estimator.

Another aspect of interest of many researchers is testing the significance of regression coefficients. Suppose, one might be interested in testing the hypothesis H_0: $\beta_2 = 0$ where β is partitioned as $\beta = [\beta_1' \mid \beta_2']$ and β_1 is of order $(k - r) \times 1$ and β_2 is of order $r \times 1$.

The model may be written as

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon \qquad (1.1.4)$$

where X_1 is of order $n \times (k-r)$ and the order of X_2 matrix is $n \times r$.

For the full model,

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

Regression sum of squares is

$$SS_R(\beta) = \hat{\beta}'_{OLS}X'Y$$

with k degrees of freedom.

Assuming the null hypothesis H_0 is true, model given in Eq. (1.1.4) becomes

$$Y = X_1\beta_1 + \varepsilon.$$

This model is called as reduced model and the OLS estimator of β_1 for the reduced model is

$$\hat{\beta}_{1OLS} = (X_1'X_1)^{-1}X_1'Y.$$

The regression sum of squares is $SS_R(\beta_1) = \hat{\beta}'_{1OLS}X_1'Y$ with $k-r$ degrees of freedom.

The regression sum of squares due to β_2 given that β_1 is already in the model is $SS_R(\beta_2/\beta_1) = SS_R(\beta) - SS_R(\beta_1)$ with r degrees of freedom. This sum is called the extra sum of squares due to β_2.

The test statistic to test H_0: $\beta_2 = 0$ is

$$F_0 = \frac{SS_R(\beta_2/\beta_1)}{SS_{Res}/(n-k)}$$

follows F distribution with r and $(n-k)$ degrees of freedom where $SS_{Res} = Y'Y - \hat{\beta}'_{OLS}X'Y$ (with $n-k$ degrees of freedom). Hence, we reject H_0 at $\alpha\%$ level of significance if $F_0 > F_{\alpha,r,n-k}$. In the next section, we discuss the subset selection problem in regression.

1.2 Subset Selection Problem

Many analysts work with the massive data sets believe that the more information analyzed the better the results will be. That means the prediction

will be better if a large number of regressor variables are considered for model building. However, in modeling and analyzing the data, we find unimportant regressor variables that can decrease precision and complicate interpretations of the study. The following problems occur when one includes large number of variables in a regression model (Miller, 1990):

1. The variance of the predicted values becomes large and the prediction accuracy decreases.
2. The interpretation becomes extremely difficult; the cost of the data collection goes up and also it takes more time for computations.

In the above viewpoint, to improve the prediction accuracy, one approach is to retain only subset of relevant regressors and eliminate the rest from the model. The problem of selecting a subset of variables that contains the most of the relevant information and which explains the response variable is called 'subset (or variable or model) selection'. Many studies have shown that the subset selection procedures makes model more economical and the prediction accuracy may be improved through exclusion of irrelevant regressors. Miller (1984) provides some reasons of selection of subset of the regressors from a large number of regressors:

1. to estimate or predict at a lower cost by reducing the number of variables on which data are to be collected.
2. to predict more accurately by eliminating uninformative variables.
3. to describe multivariate data set parsimoniously and
4. to estimate regression coefficients with smaller standard error.

Some standard texts like Chatterjee and Hadi (2006) summarize the advantages of deletion of regressor variables as follows:

1. The variance of predicted value based on subset model is less than or equal to the variance of predicted values based on full model.
2. The variance of the estimates of regression coefficients of regressor variables in the subset model is not larger than the variances of the corresponding estimates for the full model.

A quite natural but important question arises after taking the decision of reduction of the number of regressors, which regressor variables should be included in the subset model. In the literature, many subset selection methods are available; most of them are categorized in to the following two main categories.

1. Exhaustive Search methods: These methods are based on examining all possible subsets of regressor variables with respect to some criterion like R^2, $adj\ R^2$, Mallows' C_p criterion etc.
2. Sequential (Systematic) selection algorithms: This category contains methods such as the Forward Selection, Backward Elimination, Stepwise Method etc.

1.3 Brief Review of Subset Selection Techniques in Regression

Research in variable selection started in the early 1960s. Many existing subset selection techniques which are based on residuals are available in the literature. Efroymson (1960) proposed an algorithm for selection of variables. Mantel (1970) showed that backward elimination procedure is performing better than forward selection method. Miller (1990) provides a good description of subset selection methodologies and pointed out that backward elimination is usually not feasible when there are more variables than observations and it requires far more computations than forward selection. Pope and Webster (1972) give detailed description about the method of addition and deletion of variables. Some well known references in this context are given in Hocking (1976), Berk (1978), Thomson (1978a, b).

In 1983, Hawking published a review article on historical development in linear regression methodology during the period 1959 – 1982. Many standard texts like Miller (2002), Draper and Smith (2003), Montgomery, Peck and Vining (2006), discuss the subset selection problem in classical regression setup.

Many criteria are evolved for selection of appropriate subset model. Mallows' C_p statistic is one of the most popular criteria in subset selection procedure (see Mallows, 1973). The purpose of C_p statistic is to guide the researchers in the process of subset selection (Mallows, 1995; George, 2000). The bounds on C_p statistic are obtained by Hocking in 1976. Weisberg (1981) proposed case-by-case version of Mallows' C_p statistic. A modified C_p has also been proposed by Gilmour (1996). A computationally efficient branch-and-bound algorithm for computing the best subset regression models is devised by Gatu and Kontoghiorghes (2006), Gatu, Yanev and Kontoghiorghes (2007). To date, almost in all statistical packages like Minitab, SYSTAT etc Mallows' C_p is incorporated.

Haitovsky (1969) and Valentine (1969) proposed subset selection method based on $Adj\, R^2$ and F-statistic respectively. Subset selection using the prediction sum of squares (PRESS) statistic has been proposed in 1974 by Allen.

One of the popular criteria often used in practice is Akaike information criterion (AIC, Akaike, 1970, 1974). Hurvich and Tsai (1989) have shown that AIC is biased and they proposed corrected version of AIC called AICc. Schwarz information criterion is derived using Bayesian arguments, this criterion is also known as Bayesian information criterion (BIC, Schwarz, 1978). Apart from the above mentioned subset selection methods, a few names are Mills and Prasad (1992), Rahman and King (1997, 1999) and McQuarrie (1999).

Various shrinkage methods became popular due to its selection consistent property when sample size is much larger than the number of predictors. Some of these methods are: The LASSO (Tibshirani, 1996), bridge regression (Fu, 1998; Huang, Horowitz, and Ma, 2008), the SCAD (Fan and Li, 2001; Fan and Peng, 2004), the elastic net (Zou and Hastie, 2005; Zou and Zhang, 2009), the Adaptive LASSO (Zou, 2006; Zhang and Lu, 2007), the penalized one-step estimator (Zou and Li, 2008), and others. All those methods are very useful

and can be formulated as penalized optimization problems (Fan and Peng, 2004; Huang, Horowitz, and Ma, 2008; Zou and Zhang, 2009).

In the 1984 article, "Selection of Subsets of Regression Variables", Miller claimed that only by an exhaustive search of all $2^k - 1$ possible subsets could one be assured of finding the best-fitting model. Also, he suggested that all subset procedure is practically possible only when the number of variables is less than twenty. But, now a days with the help of computers, it might be feasible to test every subset for more than twenty variables. However, there is limitation to obtaining the best subset by examining all possible subsets when the number of regressors is too large. Hocking and Leslie (1962) and LaMotte and Hocking (1970) deal with such a computational problem and proposed a branch and bound procedure for obtaining the subset without considering all possible regression equations. Many interesting and attractive methods and criteria dealing the subset selection problem in linear regression are proposed in the literature (see Bremian and Freedman, 1983; Hjorth, 1982; Roecker, 1991; Chatterjee and Samant, 1994; Brieman, 1992; Drezner, Marcoulides and Salhi, 1999).

It is interesting to see that almost all above mentioned techniques are based on ordinary least squares estimation of regression parameters in the linear regression set up, which in turn leads to the following problems.

1.4 Problems with Existing Subset Selection Methods

Majority of the subset selection methods reviewed above assume that the form of the model is linear and the method of ordinary least squares is most appropriate method for estimation of regression coefficients. The assumption that makes ordinary least squares estimator so attractive in terms of general hypothesis testing and parameter significance testing is that the distribution of error is normal. Violation of this assumption may results in less reliable estimates of the model parameters. Consequently, subset selection methods

based on ordinary least squares estimator may select wrong subset. Here, we consider the following situations.

1. Outliers
2. Multicollinearity
3. Non-normality

Below, we present the problems of existing methods for the above situations.

1.4.1 Problem of Subset Selection in the Presence of Outlier

An outlier is an observation that is inconsistent with the remainder of the data. Though outliers are considered as an error or noise, they may carry important information. Outliers have a strong impact on the slope of the regression. Even a single outlying observation may drastically change the values of ordinary least squares estimates (Huber, 1998). It pulls the regression line towards itself. Various outlier detection methods are available in the literature (see Rousseeuw and Leroy, 1987; Moen, Griffin and Kalantar, 1993; Hodge and Austin, 2004).

In the presence of outlier, subset selection method based on ordinary least squares estimator gives misleading results. Many researchers have proposed subset selection methods to deal with such situation. Ronchetti (1985) proposed robust version of AIC called RAIC, Ronchetti and Staudte (1994) proposed robust C_p known as RC_p, Kim and Hwang (2000) proposed modified C_p, called $C_{p(k)}$ statistic, by deleting k outlier observations from the data. A new selection criteria S_p, has been proposed by Kashid and Kulkarni (2002) based on M-estimator. Sommer and Huggins (1996) proposed RT_p criterion for subset selection based on Wald test statistic.

1.4.2 Problem of Subset Selection in the Presence of Multicollinearity

In this subsection, first we discuss the problem of multicollinearity and its detection methods. The near-linear dependency among the regressors termed

as multicollinearity. Several methods have been proposed for detection of multicollinearity in the data. Some of them are discussed briefly as follows.

1.4.2.1 Examining the Correlation Structure

This is a very simple and primitive measure of detection of multicollinearity. If absolute value of the correlation coefficient between the variables X_i and X_j is close to unity, it implies that the corresponding regressor variables are nearly linearly dependent.

1.4.2.2 Variance Inflation Factor (VIF)

The VIF for each term in the model measures the combined effect of the dependency among the regressors on the variance of the term (see Montgomery, Peck and Vining, 2006). It is defined as

$$VIF_j = \left(1 - R_j^2\right)^{-1}, \qquad j = 1, 2, \ldots, k-1.$$

where R_j^2 is coefficient of determination obtained when X_j is regressed on the remaining regressors. If X_j is nearly linearly dependent on some of the remaining regressors, corresponding R_j^2 is close to unity and VIF_j is large. Practical experience indicates that if any of the VIFs exceeds 5 or 10, then the problem of multicollinearity exists.

1.4.2.3 Condition Indices

The condition indices of $X'X$ matrix are defined as

$$k_j = \frac{\lambda_{max}}{\lambda_j} \qquad j = 1, 2, \ldots, k-1.$$

where λ_j is the jth characteristic root or eigen value of $X'X$ matrix. The number of condition indices that are large (say ≥ 1000) measures the number of near-linear dependency in $X'X$ matrix. The largest k_j is known as condition number (k) defined as $k = \frac{\lambda_{max}}{\lambda_{min}}$. The problem of multicollinearity is not serious when k is less than 100. Severity of multicollinearity increases as value of condition

number increases from 100 to 1000. k greater than 1000 indicates severe multicollinearity.

Multicollinearity tends to inflate the variance and absolute value of the ordinary least squares coefficients. Sometimes, near-linear dependency among the regressors may results in parameter estimates that differ from the true values by orders of magnitude or have the incorrect sign. Thus, subset selection methods based on OLS estimator leads to selection of inadequate subset in the presence of multicollinearity. A very few methods to deal with such situation are available in the literature (see Wu, 2004; Chong and Jun, 2005). Very recently, Dorugade and Kashid (2010a) proposed R_p criterion for subset selection in the presence of multicollinearity.

1.4.3 Problem of Subset Selection when Error follows Non-normal Distribution

The performance of OLS estimator is not reasonably well when the error distribution is non-normal. The OLS fit may be spoiled by small but reasonable deviation from normal error structure (see Huber, 1972; Andrew, 1974). A rank based parameter estimation procedure is intended to perform well for almost any possible distribution of the errors (Birkes and Dodge, 1993). The rank based estimator [Jaeckel, 1972; Hettmansperger and McKean, 1977; Hettmansperger, 1984) has been proposed to provide more robust estimates of regression coefficients. Few subset selection methods using rank based estimator are available in literature like N_p criterion. This criterion is proposed by Kashid and Kulkarni (2003).

In the next section, we demonstrate the performance of subset selection method based on OLS estimator for the above situations.

1.5 Examples:

We demonstrate the performance of Mallows' C_p in above three situations through simulated data. Mallows' C_p statistic is defined as

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

$$C_p = \frac{RSS_p}{\sigma^2} - (n - 2p) \qquad (1.5.1)$$

where RSS_p is the residual sum of squares based on subset model, p is the number of parameters in subset model and σ^2 is error variance and is replaced by its suitable estimate based on full model.

Note that, C_p statistic is based on OLS estimator which is sensitive to presence of outlier or multicollinearity or violation of normality assumption of error. We demonstrate this fact, through the examples 1.5.1 to 1.5.3. Example 1.5.1 illustrates how the outlying observation affects the values of C_p statistic for subset models. Example 1.5.2 shows that the C_p statistic selects wrong subset in the presence of multicollinearity in the data. In Example 1.5.3, performance of C_p statistic is evaluated when error term follows non-normal distribution.

Example 1.5.1:

To illustrate the performance of C_p statistic in presence of outlier, we used an artificial data from Abt, K. (1967). It contains three regressors with 15 observations. The Dataset is given in Table 1.5.1.

Table 1.5.1
The Dataset

Sr. No	Y	X_1	X_2	X_3	Sr. No	Y	X_1	X_2	X_3
1.	13.69	04	08	26	9.	29.46	12	16	09
2.	18.22	04	12	17	10.	17.07	12	04	19
3.	17.68	06	10	22	11.	28.61	13	13	11
4.	22.43	07	13	11	12.	22.08	13	07	17
5.	26.70	08	16	8	13.	25.45	15	10	15
6.	14.08	08	04	21	14.	33.44	15	15	05
7.	23.04	10	10	14	15.	24.81	16	08	10
8.	19.62	10	07	15					

The values of C_p statistic for all possible subset models (7 subset models) for actual data and introducing an outlier observation $Y_{14} = 100$ are computed. This data is analyzed by Kashid and Kulkarni (2002) and these results are reported in the following Table.

Table 1.5.2
Values of C_p Statistic for all subset models

Regressors in The Model	C_p	
	For Actual Data	For Outlier Data
X_1	460.00	4.70
X_2	362.00	4.50
X_3	124.00	1.10
X_1 X_2	3.80	2.10
X_1 X_3	103.00	2.80
X_2X_3	108.00	3.00
$X_1X_2X_3$	4.00	4.00

It clearly indicates that value of C_p Statistic is drastically changed when single outlying observation is present in the data. Hence, in the presence of outlier, C_p Statistic gives misleading results.

Example 1.5.2:

This example demonstrates the performance of C_p Statistic when the data suffers from the problem of multicollinearity. We consider the example from Dorugade and Kashid (2010b). In this example, random sample of size 20 was generated from $N_4(0,\Sigma)$ on X_1, X_2, X_3, and X_4 and 20 observations on error variable were generated from $N(0,8)$, where

$$\Sigma = \begin{bmatrix} 1 & 0.2290 & -0.8240 & -0.2450 \\ 0.2290 & 1 & -0.139 & -0.973 \\ -0.8240 & -0.139 & 1 & -0.030 \\ -0.2450 & -0.973 & -0.030 & 1 \end{bmatrix}$$

Subset model considered is

$$Y = 10 + X_2 + X_3 + \varepsilon$$

Since, the VIF values are 50.9, 391.3, 68.6 and 427.8; it clearly indicates the presence of multicollinearity. The results are reported in the following Table 1.5.3 (see Dorugade and Kashid (2010b))

Table 1.5.3
Values of C_p Statistic for all subset models in presence of Multicollinearity

Regressors in the model	C_p- Statistic
X_1	26.54
X_2	19.18
X_3	19.77
X_4	16.09
X_1X_2	04.65
X_1X_3	18.45
X_1X_4	02.43
X_2X_3	01.04
X_2X_4	13.88
X_3X_4	01.34
$X_1X_2X_3$	03.03
$X_1X_2X_4$	03.55
$X_1X_3X_4$	03.16
$X_2X_3X_4$	03.04
$X_1X_2X_3X_4$	5

From Table 1.5.3, we observe that the C_p statistic select the subset other than $\{X_2, X_3\}$. It implies C_p statistic fails to select proper subset when the multicollinearity present in the data.

Example 1.5.3:

In this example, violation of normality assumption is considered and the performance of C_p statistic is studied through the simulated data. We generated random sample of size 30 from U(0, 1) on X_1, X_2, X_3, and X_4. Response was generated by using the following relation

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

where $\varepsilon \sim f(.)$. For illustration purpose, we consider the distribution of error is student's t distribution with 2 degrees of freedom. Using the simulated data, the values of C_p statistic for all possible subset models are computed and reported in Table 1.5.4.

Table 1.5.4
Values of C_p Statistic for all subset models when error distribution follows Student's t with 2 degrees of freedom

Regressors in the model	C_p- Statistic
X_1	16.10
X_2	01.90
X_3	15.50
X_4	16.20
X_1X_2	02.10
X_1X_3	17.40
X_1X_4	18.00
X_2X_3	02.90
X_2X_4	03.10
X_3X_4	17.40
$X_1X_2X_3$	03.50
$X_1X_2X_4$	03.40
$X_1X_3X_4$	19.40
$X_2X_3X_4$	04.30
$X_1X_2X_3X_4$	5

From Table 1.5.4, it is clear that the C_p statistic selects subset with single variable $\{X_2\}$. We conclude that, C_p fails to selects a proper subset $\{X_1, X_2\}$.

1.6 Discussion

Examples given in Section 1.5 indicate that, if the estimation of regression coefficient is poor or unreliable, then the subset selection method based on these estimators may leads to select wrong subset. Hence, before proceeding to develop new subset selection method in presence of outlier and multicollinearity, in the next chapter, we propose new estimator of regression parameters when outliers and multicollinearity occurs simultaneously in the data.