CHAPTER - I

INTRODUCTION
1.1 GENERAL INTRODUCTION

Selection of regressors is an old and important problem in statistics as well as in any other field that uses regression analysis. The problem of variable selection in regression models has long been of interest to statisticians. In many regression situations, one is faced with a large number of independent variables which are potentially important for the model. Considerable attention has been given in the literature to the problem of selecting a subset of variables for the case of the general linear model. The problem of reducing the number of independent variables in the prediction equation of multiple regression analysis has received and shall continue to receive considerable attention in the statistical analysis.

Selection of regressors should be based on economic-theoretic considerations as well as on statistical evidence. Economic theory can often indicate what regressors should be included in any equation, and sometimes, but less often, a class of functional forms that should be considered. In addition, it can often tell the econometrician the likely sign of a coefficient and sometimes even a shorter range within which the coefficient is likely to lie.

The problem of selecting the best subset or subsets of explanatory variables in a multiple linear regression analysis is two-fold. The first important problem is the development of criterion for choosing between two contending subsets. Applying these criteria to all possible subsets, if the number of independent variables is large, may not be economically feasible and so the second problem is concerned with decreasing the computational effort.
Recently, some attempts have been made to generalize the concept to more
general model selection problems. An interesting account of model selection
problems can be found in Linhart and Zucchini (1986). Traditionally the focus of
research has been on the application of selection criteria to data analysis.

For a long time, computer-based stepwise procedures based on testing
hypotheses have been the dominant approach in the literature. Subset regression
procedures are widely available in statistical computer packages. These
procedures can be useful in model fitting when a subset of given regressors is
assumed to be adequate for describing a dependent variable. Evaluating the
goodness of fit of subset regression procedures, however, is problematic. In
ordinary least squares, the null distribution of the $R^2$ is readily transformed to an $F$
distribution under the usual assumptions. In subset regression, on the other hand,
the null distribution of the $R^2$ between the dependent variable and the selected
subset of regressors can not be transformed to an $F$-distribution.

A popular method for selecting regressors is forward selection with a
stopping rule. Most automated programs denote this stopping rule as the F-to-
enter. This rule is based on the usual test statistic for the hypothesis that the partial
correlation between the dependent variable and predictor to be added is zero,
given the previously selected predictors. Wilkinson and Dallal (1981) have
examined the distribution of the sample $R^2$ when forward selection of regressors is
governed by the F-to-enter stopping rule. When compared with other selection procedures, Forward selection does not necessarily find the minimum residual sum of squares for a given subset size. Several other procedures such as Backward elimination, Stepwise regression, and all possible subsets regression can produce different subsets on the same sample. Since these other procedures are widely used, knowledge of the distribution of the sample $R^2$ would be helpful in each case.


In the present study, an attempt has been made to develop some new criteria for selecting regressors in regression analysis.

1.2 STATEMENT OF THE PRESENT STUDY

The problem of selecting the 'best' subset of variables in a linear regression context has long been of special interest to theoretical and applied statisticians. There is considerable literature on subject of the selection of the best subset of independent variables in a multiple regression or general linear univariate model
setting. Draper and Smith (1996) discussed several criteria and procedures; Gorman and Toman (1966) discussed Mallows' Cp Criterion for variable selection; Allen (1971) discussed Mean Squared Error of prediction as a criterion for selecting variables; Lindley (1968) emphasized the Bayesian procedure for selection of variables; Helms (1974) has introduced the Average Estimated Variance (AEV) as a criterion for the selection of variables, when the model is intended for use as a substitute for the real, but unknown, function over a particular region of interest (set of values of the independent variables). Hocking (1976) reviewed the various concepts and a number of computational methods associated with variable selection in linear regression models.

In applications of regression analysis for prediction purposes, large number of independent variables is often available. There may be uncertainty as to which of these independent variables should be included in the final analysis as adequate prediction may be possible using only a subset of those available. Many methods of variable selection have been proposed in the literature. In deciding on a technique, it is necessary to evaluate the criterion of goodness of prediction on which it is based, and to some extent, the computational effort involved.

Thompson (1978) reviewed some important methods and evaluated them critically and selected the most appropriate methods for the selection criteria. This criteria considered for the two types of linear models:
(i) The fixed linear model, relates to the function

\[ f(y; X_1, X_2, \ldots, X_K; \beta_1, \beta_2, \ldots, \beta_K) \]

Where \( y \) is an observed random variable depending on \( K \) known regressors \( X_1, X_2, \ldots, X_K \); such that

\[ E(y) = \sum_{i=1}^{K} \beta_i X_i \quad \ldots \quad (1.2.1) \]

Where \( \beta_1, \beta_2, \ldots, \beta_K \) are unknown parameters and \( \text{Var}(y) = \sigma^2 \).

Here \( \sigma^2 \) is unknown but independent of \( X_1, X_2, \ldots, X_K; \beta_1, \beta_2, \ldots, \beta_K \).

(ii) The random linear model, Where it is assumed that \( y, x_1, x_2, \ldots, x_K \) have jointly a \((K+1)\)-dimensional normal distribution with unknown mean vector and covariance matrix. Then the conditional expectation of \( y \), given \( x_1 = X_1, \ldots, x_K = X_K \), is given by

\[ E[y/x_1 = X_1, \ldots, x_K = X_K] = \sum_{i=1}^{K} \beta_i X_i \quad \ldots \quad (1.2.2) \]

and the conditional variance is given by

\[ \text{Var}[y/x_1 = X_1, x_2 = X_2, \ldots, x_K = X_K] = \sigma^2 \quad \ldots \quad (1.2.3) \]

There is a fundamental difference in the effect on the fixed and random models of the elimination of some of the variables. In the case of fixed linear regression model, for instance, the variable \( X_K \) is omitted in a situation in which \( \beta_K \neq 0 \), the result is no longer a model of the form:
1.3 OBJECTIVES OF THE PRESENT STUDY

The main objective of the present study is to develop some new selecting regressors in linear regression models.

The specific aims of the study are:

i. to review the various existing methods in the literature for the selection of regressors in multiple linear regression models;

ii. to explain some variable subset selection techniques in multiple regression analysis;

iii. to propose some new criteria for selection of regressors in econometrics;

and iv. to give stepwise procedures of reducing the number of regressors in the prediction equation of multiple regression analysis.

1.4 ORGANIZATION OF THE PRESENT RESEARCH STUDY

The organization of the present research work itself reveals how the objectives of the study have been achieved within the described framework.

Chapter-I is an introductory one. It contains general introduction about the problem of selection of regressors and statement of the research problem besides the objectives of the present study. It also brings out the organization of the present research study and chapter scheme.
Chapter-II describes the various criteria for selection of regressors in the multiple regression analysis existing in the literature. A number of basic and advanced criteria function such as $R^2$ and $\bar{R}^2$, $C_p$ criterion, Amemiya's unconditional MSE criterion, $S_p$ criterion, Akaike's Information criterion, Sawa's BIC criterion, Stein-Rule Bivar criterion etc have been explained in this chapter.

Chapter-III deals with the basic stepwise regression procedures for variable selection in multiple regression analysis. The mean square error of prediction criterion has been discussed along with a similar average estimated variance criterion for the selection of variables in the general linear model. It brings out some influence measures along with variable selection in discriminant analysis.

Chapter-IV presents the various methods for choosing variable subsets in multiple linear regression analysis under these methods, the mean squared prediction error has been considered as basis of the criteria. The regression diagnostics has been discussed by giving PRESS and other diagnostic measures such as $Q^2$-statistics based on predicted residuals.

Chapter-V proposes some new criteria for selection of regressors in econometrics based on different types of residuals such as Ordinary Least Squares,
Studentized and Predicted residuals. A modified stepwise regression technique for selection of regressors has been proposed by giving certain stopping rules by using coefficients of multiple determination based on Internally Studentized residuals. For model building, a distance measure has been suggested to find Influence on the predicted variables in the regression analysis. Using principal component of analysis, a stepwise ridge regression method has been explained for variable selection in regression analysis. PRESS statistic has been computed for the restricted and unrestricted linear models to test the hypothesis of subset selection of variables in the regression analysis. It also presents Wald, Likelihood Ratio and Lagrange Multiplier test statistic with some modifications for the selection of subset of regressors.

Chapter-VI depicts the main conclusions of the present research study. It also narrates the plan for future research as an extension in the lines of present research work.

Several relevant references regarding the present research study have been documented under a separate title "BIBLIOGRAPHY".
The contents of the present research study have been presented under the following heads:

CHAPTER I : INTRODUCTION

CHAPTER II : SELECTION OF REGRESSORS IN MULTIPLE REGRESSION

CHAPTER III : VARIABLE SELECTION IN MULTIPLE LINEAR REGRESSION

CHAPTER IV : CHOICE OF VARIABLE SUBSETS IN MULTIPLE LINEAR REGRESSION

CHAPTER V : SOME NEW CRITERIA FOR SELECTION OF REGRESSORS IN ECONOMETRICS

CHAPTER VI : CONCLUSIONS

BIBLIOGRAPHY