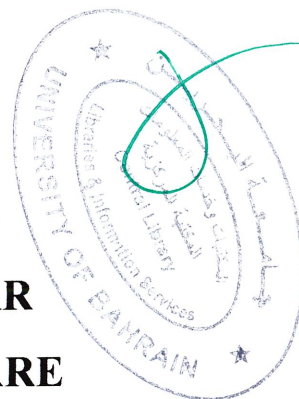


UNIVERSITY OF BAHRAIN



**ESTIMATION OF PARAMETERS IN LINEAR
REGRESSION MODELS WHEN THE DATA ARE
INCOMPLETE**

BY

SADIQ S.MOHSEN ALMOSAWI

**A Thesis Submitted to the Department of Mathematics as
a Partial Fulfillment of the Requirements for the Degree**

of

Master of Science

**Under the Supervision of
Dr. Moh'd Akkas Ali**

December 1996



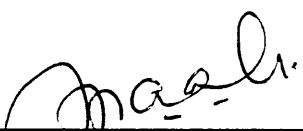
2 June 1997

The Thesis Defence Committee considers the thesis entitled :

"Estimation of Parameters regression Models when Data are incomplete"

submitted by *Sadiq S. Mohsen Almosawi* to the Department of Mathematics
satisfactory and acceptable for the Master of Science Degree in Mathematics.

Thesis Defence Committee :


Dr. M. Akkas Ali
Supervisor

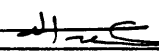
Department of Mathematics
University of Bahrain


Dr. Aziz Lahji
Member

Department of Mathematics
University of Bahrain


Prof. A.C. Atkinson

Department of Statistical and
Mathematical Science
London School of Economics and
Political Science
University of London


Dr. A.S.A. Al-Hammadi
Chairman

Department of Mathematics
University of Bahrain

Acknowledgment

By the grace of almighty Allah, I have succeeded in completing the thesis for partial fulfillment of requirements of the degree of Master of Science. I express my profound thanks to my supervisor Dr. Moh'd. Akkas Ali. I am grateful to my wife and relatives for their continuous encouragement.

Contents

Acknowledgment	iii
Preface	vi
Chapter 1 : Key terms and terminologies	
1.1 Introduction	1
1.2 The variate	1
1.3 Measurement scales	2
1.4 Measurement error	4
1.5 Missing data	5
1.6 Examining the patterns of missing data	6
1.7 Approaches for dealing with missing data	8
1.8 Replacement of missing data	9
1.9 Key terms	11
1.10 Regression	13
1.11 Selection of variables	19
1.12 Conclusion	19
Chapter 2 : Biased estimation in regression	
2.1 Introduction	20
2.2 Ridge regression estimators	21
2.3 Properties of ridge regression estimator	22
2.4 Shrinkage estimators	29
2.5 Principal component regression	32
Chapter 3 : Estimation of parameters of a linear regression model when some values of the independent variables are missing	
3.1 Introduction	36
3.2 Methods of estimation	45
3.3 Behavior of b_L , b_m and b_I	50

Chapter 4 : Use of mixed regression approach when some	53
values of the regressors are missing	55
4.1 Introduction	60
4.2 Comparison	62
4.3 Mixed estimation	
4.4 Properties of mixed regression estimator	
 Chapter 5 : A class of jackknifed estimators for the vector	
of unknown regression coefficients when some	69
observations are missing	71
5.1 Introduction	75
5.2 Jackknifing	78
5.3 Properties of $\tilde{\beta}$	
5.4 Relative bounds	
	83

References

Preface

Science is today becoming increasingly quantitative. Scientists are being confronted more and more with large amounts of numerical data gathered from their laboratories, field experiments and surveys. But the mere collecting and recording of data achieves nothing. Data that have been collected, must be analyzed and interpreted. Over the last several decades the linear regression model has surely become among the most widely employed quantitative tools of applied social sciences and many of the physical and chemical sciences. In the vast majority of regression analyses the least squares approach is used to fit a linear model to the data. The coefficients of regression play an important role in fitting a linear model to the data. The accuracy of the model depends on the accuracy of the estimation of the regression coefficients. The estimation of the parameters can be affected by several factors, such as influential observations and errors of measurements.

In this thesis the problem of estimation of parameters of a linear regression model when some values of the independent variables are missing is considered. This thesis consists of five chapters. Below a brief description of the contents of the thesis is given.

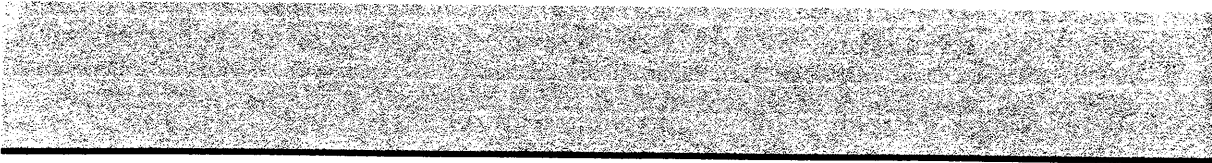
In chapter 1 , we review the key terms to develop an understanding of the concepts and terminologies to be used. Also, we present a simplified overview of data analysis and describe the methods currently available to examine our data.

Chapter 2 is devoted to biased estimation in regression. Here we discuss three methods of estimation, namely : ridge regression, shrinkage and principal component regression. Some properties of ridge regression estimators are also studied.

In chapter 3 , we discuss the estimation of parameters of a simple linear regression model when some values of the independent variables are missing. Here a special case of regression is considered, namely, a dependent variable y regresses on two independent variables x_1 and x_2 , where there are missing values on x_2 only. The behavior of available estimators is also discussed briefly.

The use of mixed regression approach when some values of the regressors are missing is discussed in chapter 4 . We show that the mixed regression estimator for the problem in hand is superior, in terms of squared bias and variance, to the least squares estimator.

Chapter 5 deals with a class of Jackknifed estimators for the vector of unknown regression coefficients when some observations are missing. Upper bounds for the relative squared bias and the relative variance of the vector of predicted responses are presented.



Chapter 1

Key Terms And Terminologies

1.1 INTRODUCTION

In this chapter we review the key terms to develop an understanding of the concepts and terminologies to be used . We present a simplified overview of data analysis and describe the methods currently available to examine our data. Examination of data is a time consuming but necessary step that is sometimes overlooked by analysts. Careful analysis of data leads to better results.

The assumptions underlying regression analysis are also reviewed. Researchers wishing to apply regression analysis should be particularly interested in assessing the assumptions of normality, independence of error, and linearity. The applicability of these issues varies with each data analysis scheme. However, they represent a process that should be performed to some extent for each application of regression.

1.2 THE VARIATE

The building block of multivariate analysis is the variate, a linear combination of variables with empirically determined weights. The variables are specified by the researcher, while the weights are determined by the specific

objective of the multivariate technique. A variate of m weighted variables (X_1 to X_m) can be stated mathematically as :

$$\text{Variate value} = W_1X_1 + W_2X_2 + \dots + W_mX_m ,$$

where X_j is the observed variable and W_j is the weight determined by the multivariate technique.

The result is a single value representing a combination of the entire set of variables that best achieves the objective of the specific multivariate analysis . The variate is a focal point of the analysis in many respects. We must understand not only its collective impact in meeting the technique's objective but also the contribution of each separate variable to the overall variate effect.

1.3 MEASUREMENT SCALES

Data analysis involves the partitioning, identification, and measurement of variation in a set of variables, either among themselves or between a dependent variable and one or more independent variables. The key word here is measurement because the researcher cannot partition or identify variation unless it can be measured. Measurement is important in accurately representing the concept of interest and is instrumental in the selection of the appropriate multivariate method of analysis.

There are two basic kinds of data : nonmetric (qualitative) and metric (quantitative). Nonmetric data describe differences in type or kind by indicating the presence or absence of characteristic or property. Metric data are interval and ratio data. Metric data measurements are made so that subjects may be identified as differing in amount or degree.

Nonmetric Measurement Scales:

Nonmetric measurements can be made with either a nominal or an ordinal scale. Measurement with a nominal scale assigns numbers used to label or identify subjects or objects. Nominal scales, also known as categorical scales, provide the number of occurrences in each class or category of the variable being studied. Therefore, the numbers or symbols assigned to the objects have no quantitative meaning beyond indicating the presence or absence of the attribute or characteristic under investigation.

Ordinal scales are the next higher level of measurement precision. Variables can be ordered or ranked with ordinal scales in relation to the amount of the attribute possessed. Every subclass can be compared with another in terms of a "greater than" or "less than relation".

Metric Measurement Scales:

Interval scale and ratio scale (both metric) provide the highest level of measurement precision, permitting nearly all mathematical operations to be performed. These two scales have constant units of measurement. The only real difference between interval and ratio scales is that interval scales have an arbitrary zero point, while ratio scales have an absolute zero point. The most familiar interval scales are the Fahrenheit and Celsius temperature scales. Both have a different arbitrary zero point, and neither indicate a zero amount or lack of temperature, because we can register temperatures below the zero point on each scale. Therefore, it is not possible to say that any value on an interval scale is a multiple of some other point on the scale.

Ratio scales represent the highest form of measurement precision, because they possess the advantage of all lower scale plus an absolute zero point. All

mathematical operations are permissible with ratio scale measurements. For example, 100 pounds is twice as heavy as 50 pounds.

Understanding the different types of measurement scale is important for two reasons. First, the researcher must identify the measurement scale of each variable used, so that nonmetric data are not incorrectly used as metric data and vice versa. Second, the measurement scale is critical in determining which multivariate techniques are the most applicable to the data.

1.4 MEASUREMENT ERROR

Measurement error is the degree to which the observed values are not representative of the "true" values. Measurement errors have many sources, ranging from data entry error to the imprecision of the measurement. All variables used in multivariate techniques are assumed to have some degree of measurement error.

The researcher's goal of reducing measurement error can follow several paths. In assessing the degree of measurement error present in any measure, the analyst must address both the validity and reliability of the measure. Validity is the degree to which a measure accurately represents what it is supposed to. If validity is assured, the researcher must still consider the reliability of the measurements. Reliability is the degree to which the observed variable measures the "true" value and is "error free"; thus it is the opposite of measurement error.

The impact of measurement error and poor reliability cannot be directly seen, because they are embedded in the observed variable. Thus researchers must always work to increase reliability and validity, which in turn will result in a "truer" portrayal of the variable of interest. Poor results are not always

due to measurement error , but the presence of measurement error is guaranteed to distort the observed relationships and make multivariate techniques less powerful . Reducing measurement error , while taking effort , time , and additional resources , may improve weak or marginal results and strengthen proven results as well .

1.5 MISSING DATA

Missing data are a fact of life in multivariate analysis ; in fact it is rare for the analyst to avoid facing some form of missing data problem in almost every situation .

A missing data process is any

- (a) Systematic event external to the respondent , such as data entry errors or data collection problems, or
- (b) action on the part of the respondent , such as refusal to answer , that leads to missing values.

The effects of some missing data processes are known and directly accommodated in the research plan . But others , particularly those based on actions by the respondent are rarely known . When the missing data processes are unknown , the analyst attempts to identify any patterns in the missing data that would characterize the missing data process . The concern for understanding the missing data processes is similar to the need to understand the cause of nonresponse in the data collection process . For example , are those individuals who did not respond different from those who did ? If so , do these differences have any impact on the analysis , the results or their interpretation ? Concerns similar to these also arise from missing response to individual variables.

In many multivariate analyses , particularly survey research applications , missing data eliminate so many observations that what was an adequate sample is reduced to an inadequate sample . In such situations , the researcher must either gather additional observations or find a remedy for the missing data in the original sample . Although finding a remedy for missing data is the most practical solution , few guidelines exist pertaining to the diagnosis and remedy of missing data . For this reason , the following sections discuss the different types of missing data processes , methods to identify the nature of the missing data process and available remedies for accommodating missing data into the multivariate analysis .

Ignorable Missing Data :

One example of an ignorable missing data process is the missing data of these observations in a population that are not included in the sample. While the analyst has a priori knowledge of certain characteristics about the entire population (i.e. the variable that define the sampling frame) , the purpose of the multivariate technique is to generalize from the sample observations to the entire population.

Another instance of ignorable missing data occurs when the data are censored. Censored data are observations not complete because of their stage in the missing data process. A typical example is an analysis of the cause of death. Respondents who are still living cannot provide complete information (i.e cause of death) and are thus censored.

1.6 EXAMINING THE PATTERNS OF MISSING DATA

The next step is to examine the patterns of missing data and to determine whether the missing data are distributed randomly across the cases and the

variables. The first test for assessing randomness is to compare the observations with and without missing data for each variable on the other variables. For example, the observations with missing data on X_1 are placed in one group and those observations with valid response for X_1 are placed in another group. Then, these two groups are compared to identify any difference on the remaining variables. Once comparisons have been made on the other variables, new groups are formed based on the missing data for the next variable and the comparisons are performed again on the remaining variables. This process continues until each variable has been examined for any difference. The objective is to identify any systematic missing data process that would be reflected in patterns of significant difference.

A second test for examining the patterns of missing data involves the use of correlations between dichotomous variables. The dichotomous variables are formed by replacing valid values with a value of one and missing data with a value of zero. The resulting correlations between the dichotomous variables indicate the extent to which missing data are related in pairs of variables. Low correlations indicate low association between the missing data process for those two variables. Absence of correlations indicates that no single missing data process is significantly affecting a substantial number of variables. If randomness is indicated for all pairs of variables, then the analyst can assume that the missing data can be classified as "missing completely at random". If significant correlations exist between some pairs of variables, then the analyst may have to assume that the data are only missing at random and these relationships must be accommodated in any remedies that are applied.

1.7 APPROACHES FOR DEALING WITH MISSING DATA

The simplest and most direct approach for dealing with missing data is to include only those observations with complete data. This method is available in all statistical programs. However, this approach should be used only if the missing data are "missing completely at random", because missing data that are not missing completely at random have non random elements that bias the result. Thus, even though only valid observations are used, the results are not generalizable to the population.

Another simple remedy for missing data is to delete the offending case and / or variable. In this approach, the analyst determines the extent of missing data on each case and variable and then deletes the case or variables exceeding a specified level. In many cases where a non random pattern of missing data is present, this may be the most efficient solution. The analyst may find that the missing data are concentrated in a small subset of cases and / or variables, with their exclusion substantially reducing the extent of the missing data. No firm guidelines exist on the necessary level exclusion, but any decision should be based on both empirical and theoretical considerations.

Imputation Methods:

A third approach to handling missing data is through one of the many imputation methods. Imputation is the process of estimating the missing value based on valid values of other variables and / or cases in the sample. The objective is to employ known relationships that can be identified in the valid value of the sample to assist in estimating the missing values. However, the analyst should carefully consider the use of imputation in each instance because of its potential impacts on the analysis.

1.8 REPLACEMENT OF MISSING DATA

The issue in imputation is the actual method for replacement of the missing value with some estimated value based on other information available in the sample. This can take many forms, varying from a direct substitution of values to estimation processes based on relationships among the variables. The following discussion focuses on the more widely used methods, although many other forms of imputation are available, see Little and Rubin (1987).

(i) Substitution:

In this method, observations with missing data are replaced by an estimate. A common example is to replace a sampled household that cannot be contacted or that has extensive missing data with another household not in the sample, preferably very similar to the original observation. This method is most widely used to replace missing data.

(ii) Mean Substitution:

One of the more widely used methods, this replaces the missing value for a variable with mean value of that variable based on all valid response. In this manner, the valid sample responses are used to calculate the replacement value. The rationale of this approach is that the mean is the best single replacement value. This approach, while used extensively, has three disadvantages. First, it make the variance estimate derived from the standard variance formula invalid by understanding the true variance in the data. Second, the actual distribution of values is distorted by substituting the mean for the missing value. Third, this method depresses the observed correlation because all missing data will have a single constant value.

(iii) Cold Deck Imputation:

In this method, the analyst substitutes a constant value derived from external sources or previous research for the missing value. This is similar in nature to the mean substitution method, differing only in the source of the substitution value. Cold deck imputation has the same disadvantage as the mean substitution method, and the analyst must be sure that the replacement value from an external source is more valid than the internally generated value, the mean.

(iv) Regression Imputation:

In this method, regression analysis is used to predict the missing value of a variable based on its relationship to other variables in the data set. While having the appeal of using relationships already existing in the sample as the basis of prediction, several disadvantages are also associated with this method. First, it reinforces the relationships already in the data. Second, this method assumes that the variable with missing data has substantial correlations with the other variables. If these correlations are not sufficient to produce a meaningful estimate, then other methods, such as mean substitution, are preferable. Finally, the regression procedure is not constrained in the estimate it makes. Thus, the predicted value may not fall in the valid range for the variable, thereby requiring some form of additional adjustment.

(v) Multiple Imputation:

The final imputation method is actually a combination of several methods. In this approach, two or more methods of imputation are used to derive a composite estimate, usually the mean of the various estimates, for the missing value. The rationale of this approach is that the use of the multiple approach minimizes the specific concerns with any single method and the composite will be the best possible estimate. The choice of this approach is primarily based on

the trade-off between the analyst's perception of the potential benefits weighted against the substantially higher effort required in making and combining the multiple estimate.

1.9 KEY TERMS

Linearity:

Linearity is assumed to express the concept that the model possesses the properties of additivity and homogeneity. In a simple sense, linear models predict values that fall in a straight line by having a constant unit change of the dependent variable (slope) for a constant unit change of the independent variable.

Normality:

The most fundamental assumption in regression analysis is the normality of the data. The assumption is needed for testing the hypotheses about the unknown regression coefficients and confidence intervals for them. If the variation from the normal distribution is sufficiently large, all resulting statistical tests are invalid, as normality is required to use the F and t statistics.

The simplest diagnostic for normality is a visual check of the histogram that compares the observed data with a distribution approximating the normal distribution. Although appealing because of its simplicity, this method is problematic for smaller samples, where the construction of the histogram can distort the visual portrayal to such an extent that the analysis is useless. A more reliable approach is the normal probability plot, which compares the cumulative distribution of actual data values with the cumulative distribution of the normal distribution.

In addition to examining the normal probability plot, one can also use statistical tests to assess normality. The simplest test is a rule of thumb based on the skewness value. The statistic Z is calculated as

$$Z = \text{skewness} / \sqrt{\sigma/n},$$

where n is the sample size. If the calculated Z value exceeds a critical value, then the distribution is nonnormal.

Homoscedasticity :

When the variance of the error term of a model appears constant over a range of predictor variables, the data are said to be homoscedastic. The assumption of equal variance of the population error is critical to the proper application of linear regression. When the error terms have increasing variance, the data are said to be heteroscedastic.

Residual:

Portion of a dependent variable not explained by a regression model. The residual represents the unexplained portion of the dependent variable. Residuals can be used in diagnostic procedures that examine the unexplained portion to identify problems in the estimation technique or unidentified relationships.

Multicollinearity:

Extent to which a variable can be explained by the other variables in the analysis. As multicollinearity increases, it complicates the interpretation of the variate as it is more difficult to ascertain the effect of any single variable, owing to their interrelationships.

1.10 REGRESSION

The objective of regression analysis is to predict a single dependent variable from the knowledge of one or more independent variables. When the problem involves a single independent variable, the statistical technique is called simple regression. When the problem involves two or more independent variables, it is called multiple regression. The following discussion is divided into three parts. The parts are

- (i) Prediction without an independent variable.
- (ii) Prediction using a single independent variable (simple regression).
- (iii) Prediction using several independent variables (multiple regression).

Prediction Without an Independent Variable :

Let us start with a simple measure. Suppose that we surveyed 8 families and asked how many cars were held by the entire family . The data are shown in columns 1 and 2 of Table 1.1 . If we were asked to predict how many cars a family holds using only these data, we could use any of several descriptive measures - the arithmetic average , the median or even the mode . For each of these statistics , we could make a prediction as to the number of cars held by the families . Because we are basing this prediction on only the information about cars usage , we must make a single prediction and apply it to all families . For example , our prediction with the arithmetic average (the mean) would be seven cars . Our prediction would be stated as " The average number of cars held by a family is seven " .

TABLE 1.1
Survey Results For Cars .

Family Number	Actual Number of cars	Average Number of cars	Error
1	4	7	+3
2	6	7	+1
3	6	7	+1
4	7	7	0
5	8	7	-1
6	7	7	0
7	8	7	-1
8	10	7	-3

But the analyst must still answer one question . How accurate is the prediction for each measure ? The customary way to assess the adequacy of a predictor variable is to examine the errors in predicting the criterion variable when it is used for prediction . For example , if we predict that family 1 has 7 cars , we overestimate by 3 . Thus the error is +3 . If this procedure were followed for each family , some estimates would be too high, others would be too low, and still others would give the correct number of cars held. We square each error and then add the results together. The total, referred to as the sum of squared errors, provides a measure of the prediction accuracy. The objective is to obtain the smallest possible sum of squared errors, because this would mean our predictions would be most accurate.

For a single set of observations, the arithmetic average (mean) will produce a smaller sum of squared errors than any other measure of central tendency. Therefore, for our set of 8 observations, the average is the best single predictor of the number of cars held by families. The sum of squared errors for our example when we use the mean as our single predictor is 22. In our discussion of simple and multiple regression, we will use this prediction by the mean as baseline for comparison because it represents the best prediction without using any independent variables.

Prediction Using A Single Independent Variable:

The average is the best predictor if we do not use any independent variable. But in our survey (Table 1.1) we also collected information on other measures that could act as independent variable. Let us determine whether knowledge of another measure- the number of people in each family- will help our predictions. This procedure involves a single independent variable and is referred to as simple regression.

Simple regression is another procedure for predicting data and it is based on the principle of minimizing the sum of squared errors of prediction. We know that without using family size we can best describe the number of cars held as the mean value, seven. Another way to write the prediction is as follows:

$$(\text{Predicted number of cars}) = \text{Average number of cars held}$$

or

$$\hat{Y} = \bar{y} .$$

Using our information on family size , we could try to improve our predictions by reducing our prediction errors . To do so , the prediction errors in the number of cars held must be correlated with family size . The concept of correlation

represented by the correlation coefficient(s) is fundamental to regression analysis and describes the relationship between two variables . Two variables are said to be correlated if changes in one variable are associated with changes in the other variable. In this way, as one variable changes, we would know how the other variable is changing. If the family size is correlated with car usage, we would then describe the relationship as follows:

$$\begin{bmatrix} \text{Predicated} \\ \text{number} \\ \text{of cars} \end{bmatrix} = \begin{bmatrix} \text{change in number of cars} \\ \text{held associated with unit} \\ \text{change in family size} \end{bmatrix} * \begin{bmatrix} \text{family size} \end{bmatrix}$$

or

$$\hat{Y} = b_1 X_1 .$$

For example, if we find that for each additional member in a family, the number of cars increases (on the average) by two, we would predict that families of four would have 8 cars and families of five would have 10 cars . Thus the number of cars = 2 * (family size) . However , we often find that the prediction is improved by adding a constant value. For example, the following pattern may be found:

$$(\text{Predicted number of cars}) = 2 + 2 * (\text{family size}) .$$

We take this approach with our sample of 8 families (see Table 1.2 below) and observe how well the description fits our data . The procedure is as follows :

$$\begin{bmatrix} \text{Predicted number} \\ \text{of cars held} \end{bmatrix} = \text{constant} + \begin{bmatrix} \text{change in number} \\ \text{of cars with differing} \\ \text{family size} \end{bmatrix} * \begin{bmatrix} \text{family size} \end{bmatrix}$$

or

$$\hat{Y} = b_0 + b_1 X_1 .$$

The value of b_0 and b_1 that minimize the sum of the squared errors of prediction are given by $b_0 = 2.87$ and $b_1 = 0.97$.

Therefore

$$\hat{Y} = 2.87 + 0.97 * [\text{family size}] .$$

We can determine whether our knowledge of family size has helped us better predict car holdings when compared with our prediction using only the arithmetic average . The sum of squared errors using the average is 22 , and using our new approach with a single independent variable , we find that the sum of squared errors is 5.50 (see Table 1.2 below) .

TABLE 1.2
Survey Results Relating Number of Cars to Family Size

Family Size (X_i)	Number of cars (Y_i)	Prediction $\hat{Y}_i = 2.87 + 0.97X_i$	Error ($\hat{Y}_i - Y_i$)
2	4	4.81	+ 0.81
2	6	4.81	- 1.19
4	6	6.75	+ 0.75
4	7	6.75	- 0.25
5	8	7.72	- 0.28
5	7	7.72	+ 0.72
6	8	8.69	+ 0.69
6	10	8.69	- 1.31

Using the least squares procedure and a single independent variable, we see that our new approach, simple regression, is better than using just the average.

Multiple regression:

We previously demonstrated how simple regression helped improve our prediction of car holdings. By using data on family size, we predicted the number of cars a family would own more accurately than we could by simply using the arithmetic average. This result raises the question of whether we could improve our prediction even further by using additional data obtained from the families. Would our prediction of the number of cars be improved if we used data on family size in addition to data on another variable, family income?

The ability of an additional independent variable to improve the prediction of the dependent variable is related not only to the correlation of the additional independent variable to the dependent variable but also to the correlations of the additional independent variables to the independent variable already in the regression equation. Collinearity is the association measured as the correlation between two independent variables. Multicollinearity refers to the correlation among three or more independent variables (evidenced when one is regressed against the others). The impact of multicollinearity is to reduce any individual independent variables predictive power by the extent to which it is associated with the other independent variables.

Purpose of Multiple Regression:

One fundamental purpose of multiple regression is the prediction of the dependent variable with a set of independent variables. In doing so, multiple regression fulfills one of two objectives. The first objective is to maximize the

overall predictive power of the independent variable as represented in the variate. Multiple regression can also meet the second objective of comparing two or more sets of independent variables to ascertain the predictive power of each variate.

1.11 SELECTION OF VARIABLES

The ultimate "success" of any multiple regression analysis starts with the selection of the variables to be used in the analysis. Because multiple regression is a dependence technique, the analyst must specify which variable is the independent variable and which variables are used as predictor variables. The selection of both types of variables should be based principally on conceptual or theoretical grounds. The fundamental task of variable selection must be performed by the analyst, even though the analyst has many options.

1.12 CONCLUSION

The procedures available for handling missing data are varied in form, complexity, and intent. The analyst must always be prepared to assess and deal with missing data, as it is frequently encountered in regression analysis. The decision to use only observations with complete data may always seem to be one of a conservative and safe nature but there are inherent limitations in this approach. The analyst has no single method best suited in every situation but instead must make a reasoned judgment of the situation, considering the objective under investigation.

Chapter 2

Biased Estimation in Regression

2.1 INTRODUCTION

We consider the general linear regression model

$$Y = X\beta + \mathbf{e} \quad , \quad \dots(2.1.1)$$

$$E(\mathbf{e}) = 0 \quad , \quad E(\mathbf{e}\mathbf{e}') = \sigma^2 I_n \quad ,$$

where $\mathbf{Y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ matrix of rank $p < n$, β is a $p \times 1$ vector of unknown regression coefficients and $\mathbf{e}$ is an $n \times 1$ vector of random errors. The least squares estimator of β is given by

$$\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\mathbf{e} \quad . \quad \dots(2.1.2)$$

The estimator $\hat{\beta}$ is unbiased for β and its variance is given by

$$\begin{aligned}\text{Var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] \\ &= \sigma^2 \text{Tr}(X'X)^{-1},\end{aligned}\quad \dots(2.1.3)$$

where $\text{Tr}(F)$ denotes the trace of a square matrix F . Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be the eigenvalues of $(X'X)$. Then the variance of $\hat{\beta}$ can be rewritten as

$$\text{Var}(\hat{\beta}) = \sigma^2 \sum_{i=1}^p \left(\frac{1}{\lambda_i} \right) \quad \dots(2.1.4)$$

If one or more eigenvalues of $(X'X)$ are very small, then the variance of $\hat{\beta}$ will be very large. The general principle is that it might be preferable to trade off a small amount of bias in order to substantially reduce the variance of the estimate of β . There are several methods of estimation which are based on trading off bias for variance. This chapter describes three of these : ridge regression, shrinkage and principal component regression.

2.2 RIDGE REGRESSION ESTIMATORS

Let $\tilde{\beta}$ be any estimate of β . Then the residual sum of squares can be written as:

$$\begin{aligned}\tilde{e}'\tilde{e} &= (Y - X\tilde{\beta})'(Y - X\tilde{\beta}) \\ &= [(Y - X\hat{\beta}) + X(\hat{\beta} - \tilde{\beta})]'[(Y - X\hat{\beta}) + X(\hat{\beta} - \tilde{\beta})] \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}) \\ &\quad + (Y - X\hat{\beta})'X(\hat{\beta} - \tilde{\beta}) + (\hat{\beta} - \tilde{\beta})'X'(Y - X\hat{\beta})\end{aligned}$$

Since $X'(Y - X\hat{\beta}) = 0$, therefore ,

$$\begin{aligned}\tilde{e}'\tilde{e} &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}) \\ &= \hat{e}'\hat{e} + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}),\end{aligned}$$

where $\hat{e} = (Y - X\hat{\beta})$.

Clearly, $(X'X)$ is positive definite, therefore, $(\hat{\beta} - \tilde{\beta})'(X'X)(\hat{\beta} - \tilde{\beta}) \geq 0$, and hence

$$\tilde{e}'\tilde{e} \geq \hat{e}'\hat{e}.$$

Now let us minimize $\tilde{\beta}'\tilde{\beta}$ subject to the condition $(\tilde{\beta} - \hat{\beta})'X'X(\tilde{\beta} - \hat{\beta}) = Q_0$, where Q_0 is a given constant. Let

$$F = \tilde{\beta}'\tilde{\beta} + \left(\frac{1}{k}\right)[(\tilde{\beta} - \hat{\beta})'X'X(\tilde{\beta} - \hat{\beta}) - Q_0].$$

Hence $(1/k)$ is the multiplier. Now $(dF / d\tilde{\beta}) = 0$ gives

$$\tilde{\beta} + \left(\frac{1}{k}\right)X'X(\tilde{\beta} - \hat{\beta}) = 0$$

or

$$\tilde{\beta} = (X'X + kI)^{-1}X'Y, \quad k > 0, \quad \dots(2.2.1)$$

which is the ridge regression estimator of β suggested by Hoerl and Kennard (1970). In the following section, we shall study some properties of $\tilde{\beta}$.

2.3 PROPERTIES OF RIDGE REGRESSION ESTIMATOR

The relationship of a ridge regression estimator to an ordinary least squares estimator is given by the alternative form

$$\begin{aligned}\tilde{\beta} &= (X'X + kI)^{-1}X'X\hat{\beta} \\ &= [I + k(X'X)^{-1}]^{-1}\hat{\beta} = Z\hat{\beta}, \quad \dots(2.3.1)\end{aligned}$$

where $Z=[I+k(X'X)^{-1}]^{-1}$. Clearly,

$$\tilde{\beta}'\tilde{\beta} = \hat{\beta}'Z'Z\hat{\beta} \leq \delta_1^2 \hat{\beta}'\hat{\beta} .$$

Here δ_1^2 is the largest eigenvalue of $(Z'Z)$. Since Z is a symmetric positive definite matrix,

$$\begin{aligned} \delta_1^2 &= \text{Largest eigenvalue of } Z'Z \\ &= \text{Largest eigenvalue of } Z^2 \\ &= [\text{Largest eigenvalue of } Z]^2 \\ &= [\lambda_1/(\lambda_1+k)]^2 \leq 1 , \end{aligned}$$

where λ_1 is the largest eigenvalue of $(X'X)$. Therefore,

$$\tilde{\beta}'\tilde{\beta} \leq \hat{\beta}'\hat{\beta} .$$

That is , $\tilde{\beta}$ for $k \neq 0$ is shorter than $\hat{\beta}$.

We observe that if $k = 0$, $Z = I$ and if $k \rightarrow \infty$, $Z \rightarrow 0$. Moreover,

$$\begin{aligned} Z &= (X'X+kI)^{-1}X'X = (X'X+kI)^{-1}[X'X+kI-kI] \\ &= I-k(X'X+kI)^{-1} . \end{aligned}$$

Hence

$$I-Z = k(X'X+kI)^{-1} .$$

For an estimate $\tilde{\beta}$ the residual sum of squares is

$$\begin{aligned} \phi(k) &= (Y-X\tilde{\beta})'(Y-X\tilde{\beta}) \\ &= Y'Y - \tilde{\beta}'X'Y - (Y'X\tilde{\beta} - \tilde{\beta}'X'X\tilde{\beta}) \\ &= Y'Y - \tilde{\beta}'X'Y - (\hat{\beta}'X'X\tilde{\beta} - \tilde{\beta}'X'X\tilde{\beta}) \end{aligned}$$

$$\begin{aligned}
&= Y'Y - \tilde{\beta}' X'Y - (\hat{\beta} - \tilde{\beta})' X'X \tilde{\beta} \\
&= Y'Y - \tilde{\beta}' X'Y - \hat{\beta}' (I-Z) X'X \tilde{\beta} \\
&= Y'Y - \tilde{\beta}' X'Y - k \hat{\beta}' (X'X + kI)^{-1} X'X \tilde{\beta} \\
&= Y'Y - \tilde{\beta}' X'Y - k \hat{\beta}' Z \tilde{\beta} \\
&= Y'Y - \tilde{\beta}' X'Y - k \tilde{\beta}' \tilde{\beta} .
\end{aligned}$$

The above expression shows that $\phi(k)$ is the total sum of squares less the regression sum of squares for $\tilde{\beta}$ with a modification depending upon the squared length of $\tilde{\beta}$.

The expected value of $\tilde{\beta}$ is given by

$$E(\tilde{\beta}) = \beta - k(X'X + kI)^{-1} \beta.$$

Clearly, the bias and squared bias of $\tilde{\beta}$ are

$$\text{Bias}(\tilde{\beta}) = E(\tilde{\beta}) - \beta = -k(X'X + kI)^{-1} \beta ,$$

$$\text{Bias}^2(\tilde{\beta}) = \{E(\tilde{\beta}) - \beta\}' \{E(\tilde{\beta}) - \beta\}$$

$$= k^2 \beta' (X'X + kI)^{-2} \beta$$

respectively. We define a function of k by

$$V_1(k) = \text{Bias}^2(\tilde{\beta}) = k^2 \beta' (X'X + kI)^{-2} \beta . \quad \dots(2.3.2)$$

The variance of $\tilde{\beta}$ is given by

$$\text{Var}(\tilde{\beta}) = E[\{\tilde{\beta} - E(\tilde{\beta})\} \cdot \{\tilde{\beta} - E(\tilde{\beta})\}]$$

$$= E[e'X(X'X+kI)^{-2}Xe]$$

$$= \sigma^2 \text{Tr}[(X'X+kI)^{-2}X'X] .$$

Hence, the mean square error of $\tilde{\beta}$ is given by

$$\text{MSE}(\tilde{\beta}) = \sigma^2 \text{Tr}[(X'X+kI)^{-2}X'X] + k^2 \beta' (X'X+kI)^{-2} \beta . \quad \dots(2.3.3)$$

Let

$$V_2(k) = \sigma^2 \text{Tr}[(X'X+kI)^{-2}X'X] .$$

In the following theorem, the behavior of $V_2(k)$ is examined.

Theorem (2.1) :

The total variance $V_2(k)$ is a continuous monotonically decreasing function of k .

Proof : Let $Q = (X'X+kI)^{-1}$. Therefore,

$$\begin{aligned} Q^2 X'X &= (X'X+kI)^{-2} (X'X+kI-kI) \\ &= Q - k Q^2 . \end{aligned}$$

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be eigenvalues of $X'X$. Hence

$$\left(\frac{1}{\lambda_1 + k} \right) \leq \left(\frac{1}{\lambda_2 + k} \right) \leq \dots \leq \left(\frac{1}{\lambda_p + k} \right)$$

are the eigenvalues of Q , and

$$\left(\frac{1}{\lambda_1 + k} \right)^2 \leq \left(\frac{1}{\lambda_2 + k} \right)^2 \leq \dots \leq \left(\frac{1}{\lambda_p + k} \right)^2$$

are the eigenvalues of Q^2 . Hence

$$\begin{aligned} \text{Tr}(Q^2 X^* X) &= \text{Tr}(Q) - k \text{Tr}(Q^2) \\ &= \sum_{i=1}^p \left(\frac{1}{\lambda_i + k} \right) - k \sum_{i=1}^p \left(\frac{1}{\lambda_i + k} \right)^2 \\ &= \sum_{i=1}^p \lambda_i / (\lambda_i + k)^2 . \end{aligned}$$

Therefore,

$$V_2(k) = \sigma^2 \sum_{i=1}^p \lambda_i / (\lambda_i + k)^2 . \quad \dots(2.3.4)$$

Clearly, each term in $V_2(k)$ decreases as k increases. This completes the proof.

Next,

$$V'_2(k) = \frac{d}{dk} V_2(k) = -2\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} .$$

Note that $V'_2(k) \rightarrow -\infty$, as $k \rightarrow 0^+$ and $\lambda_p \rightarrow 0$.

Theorem (2.2) :

The squared bias $V_1(k)$ is a continuous monotonically increasing function of k

Proof : Since $X'X$ is a positive definite matrix, there exists an orthogonal matrix T such that

$$T' X'X T = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p) .$$

Hence, $T'(X'X + kI)T = D + kI$.

Therefore,

$$\begin{aligned} V_1(k) &= k^2 \beta' T T' (X'X + kI)^{-2} T T' \beta \\ &= k^2 \beta' T T' (X'X + kI)^{-1} T T' (X'X + kI)^{-1} T T' \beta \\ &= k^2 \beta' T (D + kI)^{-1} (D + kI)^{-1} T' \beta \\ &= k^2 \alpha' (D + kI)^{-2} \alpha \\ &= k^2 \sum_{i=1}^p \alpha_i^2 / (\lambda_i + k)^2 , \end{aligned} \quad \dots(2.3.5)$$

where $\alpha' = \beta' T = (\alpha_1, \alpha_2, \dots, \alpha_p)$ an $1 \times p$ vector. Since $\lambda_i > 0$, there is no singularity in (2.3.5). Note, also that $V_1(0) = 0$. Therefore, $V_1(k)$ is a continuous function for $k \geq 0$. For $k > 0$, (2.3.5) can be written as

$$V_1(k) = \sum_{i=1}^p \alpha_i^2 / (1 + \lambda_i / k)^2 . \quad \dots(2.3.6)$$

The functions $(1 + \lambda_i / k)$, $i=1, 2, \dots, p$, are monotone decreasing for increasing k . Each term of $V_1(k)$ is monotone increasing and so $V_1(k)$ is monotone increasing. This completes the proof.

Differentiating (2.3.5) with respect to k , we obtain

$$\begin{aligned}
V_1'(k) &= 2k \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + k)^2} + k^2 \sum_{i=1}^p \alpha_i^2 (-2)(\lambda_i + k)^{-3} \\
&= 2k \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i + k)^2} \left\{ 1 - \frac{k}{\lambda_i + k} \right\} \\
&= 2k \sum_{i=1}^p \frac{\alpha_i^2 \lambda_i}{(\lambda_i + k)^3} .
\end{aligned}$$

Clearly, $V_1'(k) \rightarrow 0$ as $k \rightarrow 0$.

Theorem (2.3) :

There are always exists a $k > 0$ such that

$$\frac{d}{dk} \text{MSE}(\tilde{\beta}) < 0 \quad \dots(2.3.7)$$

Proof : From (2.3.3), (2.3.2) and (2.3.4) we have

$$\begin{aligned}
\text{MSE}(\tilde{\beta}) &= V_1(k) + V_2(k) \\
&= k^2 \sum_{i=1}^p \alpha_i^2 / (\lambda_i + k)^2 + \sigma^2 \sum_{i=1}^p \lambda_i / (\lambda_i + k)^2 .
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{d \text{MSE}(\tilde{\beta})}{dk} &= 2k \sum_{i=1}^p \frac{\alpha_i^2 \lambda_i}{(\lambda_i + k)^3} - 2\sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} \\
&= 2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^3} \{ k\alpha_i^2 - \sigma^2 \} .
\end{aligned}$$

So if $k < (\sigma^2 / \alpha_{\max}^2)$, then (2.3.7) is hold. This completes the proof.

One particular choice for k is

$$\begin{aligned} k &= \sigma^2 / \alpha' \alpha = \sigma^2 / (\beta' T T' \beta) \\ &= (\sigma^2 / \beta' \beta) \end{aligned}$$

which is not operational. To make it operational Hoerl and Kennard (1970) suggested to use

$$\hat{k} = \hat{\sigma}^2 / (\hat{\beta}' \hat{\beta}).$$

2.4 SHRINKAGE ESTIMATORS

Another class of biased estimators introduced by James and Stein (1961). This class is known as shrinkage estimators. While the original thrust of ridge regression was to alleviate effects of multicollinearity that of the shrinkage estimator was aimed solely at reducing the mean square error. For the model (2.1.1) a class of shrinkage estimators of the parameter vector β is

$$\hat{\beta}_s = \left[1 - \frac{k(n-p)S^2}{\hat{\beta}' X' X \hat{\beta}} \right] \hat{\beta}, \quad \dots(2.4.1)$$

where $\hat{\beta}$ is the least squares estimator of β , $0 < k < 1$, and

$$S^2 = (n-p)^{-1}[(Y - X\hat{\beta})'(Y - X\hat{\beta})].$$

The estimator $\hat{\beta}_s$ is biased. Thus, in order to compare it with the least squares estimator $\hat{\beta}$, we need to consider the mean squared error matrix,

$$M(\hat{\beta}_s) = E [(\hat{\beta}_s - \beta)(\hat{\beta}_s - \beta)']$$

and compare it with $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$, the covariance matrix of $\hat{\beta}$. It is well known that

$$E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] = \sigma^2 (X'X)^{-1}.$$

However, it is more convenient to compare the scalar quantities

$\text{Tr}[(X'X) M(\hat{\beta}_s)]$ and $\text{Tr}[(X'X) M(\hat{\beta})]$. Note that

$$\text{Tr}[(X'X) M(\hat{\beta}_s)] = E[(\hat{\beta}_s - \beta)'X'X(\hat{\beta}_s - \beta)]$$

and

$$\begin{aligned} \text{Tr}[(X'X) M(\hat{\beta})] &= E[(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)] \\ &= \sigma^2 \text{Tr}(X'X)^{-1}(X'X) \\ &= p\sigma^2 \end{aligned}$$

It has been shown by Stein that if $0 < k < 2(p-2)(n+p+2)^{-1}$ and $p \geq 3$, the former quantity- i.e $E[(\hat{\beta}_s - \beta)'X'X(\hat{\beta}_s - \beta)]$ is always less than $p\sigma^2$ and attains its minimum value at $k = (p-2)/(n-p+2)$. It may mentioned that this value of k does not depend on any parameter.

Ali(1989) studied a class of shrinkage estimators of the vector of regression coefficients. He defined a linear estimator

$$b = L'Y,$$

where L' is an arbitrary $p \times n$ matrix, of β . The squared bias and the trace of the dispersion matrix of b are give by

$$\begin{aligned} B &= [\{E(b) - \beta\}' \{E(b) - \beta\}] \\ &= \beta' (L'X - I)' (L'X - I) \beta \end{aligned}$$

and

$$\begin{aligned} V &= E[\{b - E(b)\}' \{b - E(b)\}] \\ &= \sigma^2 \text{Tr}(L'L) \end{aligned}$$

respectively. He obtained a generalized Stein estimator of β by minimizing

$$F = \beta' (L'X - I)' (L'X - I) \beta + w\sigma^2 \text{Tr}(L'L) \quad \dots(2.4.2)$$

where w is an arbitrary non-negative scalar constant. We observe that (2.4.2) consists of two parts, one representing the bias and the other variance. Here the choice of w in F represents the weight we attach to variance. w may be suitably chosen depending on the importance of variance in the estimator. Note that if $w=0$, then $F=B$ and if $w=1$, then $F=B+V$. The case $w=1$ was considered by Hosmane (1988). Ali (1989) showed that $w=1$ is not the best choice of w .

The matrix L that minimizes F is given by

$$L^* = [X\beta\beta'X' + (w\sigma^2)I]^{-1}X\beta\beta'.$$

If $w > 0$, then we can rewrite L^* as

$$L^* = (w\sigma^2)^{-1} \left[I - \frac{X\beta\beta'X'}{w\sigma^2 + \beta'X'X\beta} \right] X\beta\beta'. \quad \dots(2.4.3)$$

See, Rao (1973, p.33). Substituting L^* given by (2.4.3) for L in $b = L'Y$ we obtain

$$b_* = [w\sigma^2 + \beta'X'X\beta]^{-1} (\beta'X'X\hat{\beta})\beta, \quad \dots(2.4.4)$$

where $\hat{\beta}$ is the least squares estimator of β . The estimator b_* is not operational. To make it operational Ali(1989) replaced β and σ^2 in(2.4.4) by their unbiased estimators $\hat{\beta}$ and

$$\hat{\sigma}^2 = (Y - X\hat{\beta})'(Y - X\hat{\beta}) / (n-p)$$

respectively. This gives

$$b_+ = [1 - w\hat{\sigma}^2 (w\hat{\sigma}^2 + \hat{\beta}'X'X\hat{\beta})^{-1}] \hat{\beta}.$$

Ali (1989) examined the properties of b_+ . He found that b_+ , in the sense of mean squared error, dominates the least squares estimator as well as the generalized Stein estimator developed by Hosmane (1988).

2.5 PRINCIPAL COMPONENT REGRESSION

Principal component regression is an approach to the problem of multicollinearity, which, like other approaches has advantages and

disadvantages. We reconsider the model (2.1.1). Let T be a $p \times p$ orthogonal matrix with the property

$$T'X'X T = D = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$$

where the λ_i , $i=1, 2, \dots, p$, are the characteristic roots of $(X'X)$ arranged so that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. Then we reparameterize the model (2.1.1),

$$\begin{aligned} Y &= X T T' \beta + e \\ &= Z \alpha + e, \end{aligned} \quad \dots(2.5.1)$$

where $Z = XT$ and $\alpha = T' \beta$. The least square estimator of α , $\hat{\alpha}$ is given by

$$\hat{\alpha} = (Z'Z)^{-1} Z'Y = D^{-1} Z'Y$$

and

$$\text{Cov}(\hat{\alpha}) = \sigma^2 D^{-1}.$$

From the estimate $\hat{\alpha}$, we can return to $\hat{\beta}$ by the relationship

$$\hat{\beta} = T \hat{\alpha}.$$

Let $T = [T_1 \ T_2 \ \dots \ T_p]$ and $Z_i = X T_i$, where $T_1, T_2, \dots, T_p$ are the columns of T . The Z_i are called the principal components of the matrix $X'X$, and the $\hat{\alpha}_i = (1/\lambda_i) Z_i' Y$ are the principal component regression coefficients. From $\hat{\beta} = T \hat{\alpha}$ we have

$$\hat{\beta} = [T_1 \ T_2 \ \dots \ T_p] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{bmatrix} = \sum_{i=1}^p \hat{\alpha}_i T_i,$$

where $\hat{\alpha}_i$ is the i th component of $\hat{\alpha}$. Clearly, $\hat{\beta}$ may be regarded as a weighted average of the T_i with the $\hat{\alpha}_i$'s serving as weights.

To determine the effect of small characteristic roots on the variance of $\hat{\beta}$, we recall

$$\text{Cov}(\hat{\beta}) = \sigma^2 T D^{-1} T' = \sigma^2 \sum_{i=1}^p (T_i T_i') / \lambda_i.$$

In this form, it may be seen that a small value of λ_i , which leads to a large value of $(1/\lambda_i)$, results in large value of the diagonal elements of $\text{Cov}(\hat{\beta})$; these, of course, are the $\text{var}(\hat{\beta}_i)$, the variance of the i th component of $\hat{\beta}$. To avoid such large variances, it has been proposed to estimate β from the first $r < p$ components—those associated with the r largest eigenvalues of $X'X$. Specifically, let

$$T_{(r)} = [T_1 \ T_2 \ \dots \ T_r], \quad \bar{T}_{(r)} = [T_{r+1} \ T_{r+2} \ \dots \ T_p]$$

and

$$\hat{\alpha}'_{(r)} = (\hat{\alpha}_1 \ \hat{\alpha}_2 \ \dots \ \hat{\alpha}_r).$$

The estimate of β based on the first r principal components, $\hat{\beta}_+$ is

$$\hat{\beta}_+ = T_{(r)} \hat{\alpha}_{(r)} = T_{(r)} T'_{(r)} \hat{\beta}$$

$$= \mathbf{T}_{(r)}' \mathbf{T}_{(r)} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} \hat{\boldsymbol{\beta}} .$$

$$= \mathbf{T}_{(r)}' \mathbf{D}_{(r)}^{-1} \mathbf{T}_{(r)} \mathbf{X}'\mathbf{X} \hat{\boldsymbol{\beta}} .$$

The bias and the variance of $\hat{\boldsymbol{\beta}}_+$ are given by

$$\text{Bias}(\hat{\boldsymbol{\beta}}_+) = E(\hat{\boldsymbol{\beta}}_+) - \boldsymbol{\beta} = (\mathbf{T}_{(r)}' \mathbf{T}_{(r)} - \mathbf{I}) \boldsymbol{\beta}$$

and

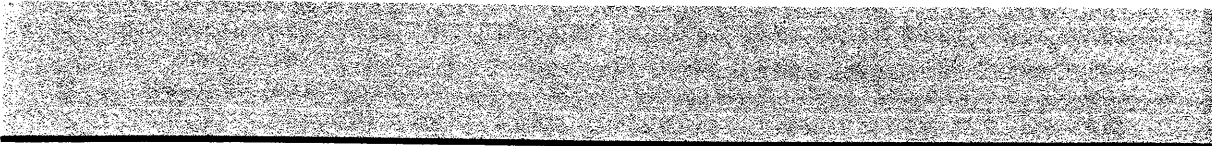
$$\text{Var}(\hat{\boldsymbol{\beta}}_+) = \text{Tr Cov}(\hat{\boldsymbol{\beta}}_+)$$

$$= \sigma^2 \sum_{i=1}^r (1 / \lambda_i) ,$$

respectively.

It is clear that

$$\text{Var}(\hat{\boldsymbol{\beta}}_+) \leq \text{Var}(\hat{\boldsymbol{\beta}}) .$$



Chapter 3

Estimation of Parameters of a Linear Regression Model When Some Values of the Independent Variables are Missing

3.1 INTRODUCTION

Even in a well planned experiment sometimes it so happens that the yeilds on some plots are not available. They may be missing due to various reasons such as **(a)** natural causes like pests, flood, fire, etc. **(b)** man-made causes such as investigators failure to note the yeild, or theft, etc.

In an industrial experiment some results can be missing because of mechanical break downs unrelated to the experimental process. In an opinion survey some individuals may be unable to express a preference for one candidate over another. Further, respondents in a household survey may refuse to report income. It is natural to treat the values that are not observed as missing, in the sense that there are true underlying values that would have been observed if the technique involved had been better or equipment had been better maintained. The lack of a response is essentially an additional point in the sample space of

the variable being measured, which identifies a " don't know " stratum of the population.

Woolson and Clarke (1984) analyze data from the Muscative Coronary Risk Factor study, where responses are categorized by the pattern of missing data. Here we reproduce their data. Table 3.1 below summarizes their data matrix. Five variables (gender, age and obesity for three rounds of the survey) are recorded for 4856 cases - gender and age are completely recorded, but the three obesity variables are sometimes missing. Since age is recorded in five categories and the obesity variables are binary, the data can be displayed as counts in a contingency table. Tables 3.2 and 3.3 below display the data in this form, with missingness of obesity treated as a third category of the variable, where 0 = obese, N = not obese, and M = missing. Thus the pattern "MON" denotes missing at the first round, obese at the second round, and not obese at the third round, and other patterns are defined similarly.

TABLE 3.1

Data Matrix for Children in a Survey Summarized by the Pattern of Missing

Data: Here 1= Observed, 0=Missing.

Pattern	Variables					Number of children with Pattern
	Age	Gender	Weight 1	Weight 2	Weight 3	
A	1	1	1	1	1	1770
B	1	1	1	1	0	631
C	1	1	1	0	1	184
D	1	1	0	1	1	645
E	1	1	1	0	0	756
F	1	1	0	1	0	370
G	1	1	0	0	1	500

TABLE 3.2

Number of Males Classified by Population and Relative Weight Category in Three Rounds of Survey.

Response Category	Age Group of Male Children				
	5-7	7-9	9-11	11-13	13-15
NNN	90	150	152	119	101
NNO	9	15	11	7	4
NON	3	8	8	8	2
NOO	7	8	10	3	7
ONN	0	8	7	13	8
ONO	1	9	7	4	0
OON	1	7	9	11	6
OOO	8	20	25	16	15
NNM	16	38	48	42	82
NOM	5	3	6	4	9
ONM	0	1	2	4	8
OOM	0	11	14	13	12
NMN	9	16	13	14	6
NMO	3	6	5	2	1
OMN	0	1	0	1	0
OMO	0	3	3	4	1
MNN	129	42	36	18	13
MNO	18	2	5	3	1
MON	6	3	4	3	2
MOO	13	13	3	1	2
NMM	32	45	59	82	95
OMM	5	7	17	24	23
MNM	33	33	31	23	34
MOM	11	4	9	6	12
MMN	70	55	40	37	15
MMO	24	14	9	14	3

TABLE 3.3

Number of Females Classified by Population and Relative Weight Category in Three Rounds of a Survey.

Response Category	Age Group of Female Children				
	5-7	7-9	9-11	11-13	13-15
NNN	75	154	148	129	91
NNO	8	14	6	8	9
NON	2	13	10	7	5
NOO	4	19	8	9	3
ONN	2	2	12	6	6
ONO	2	6	0	2	0
OON	1	6	8	7	6
OOO	8	21	27	14	15
NNM	20	25	36	36	83
NOM	0	3	0	9	15
ONM	0	1	7	4	6
OOM	4	11	17	13	23
NMN	7	16	8	31	5
NMO	2	3	1	4	0
OMN	0	0	1	2	0
OMO	1	4	4	6	1
MNN	109	47	39	19	11
MNO	22	4	6	1	1
MON	7	1	7	2	2
MOO	24	8	13	2	3
NMM	23	47	53	58	89
OMM	5	7	16	37	32
MNM	27	23	25	21	43
MOM	5	5	9	1	15
MMN	65	39	23	23	14
MMO	19	13	8	10	5

Source: Woolson and Clarke (1984)

*NNN indicates not obese in 1977, 1979 and 1981; O indicates obese, N indicates not obese and M indicates missing in a given year.

Woolson and Clarke analyzed these data by fitting multinomial distributions over $3^3 - 1 = 26$ response categories for each column of Tables 3.1 and 3.2. That is, missingness was regarded as defining strata of the population. Most statistical software packages allow the identification of non-response strata by creating one or more special codes for those entries of the data matrix that are not observed. More than one code might be used to identify a particular type of nonresponse such as "don't know" or refusal to answer or out of legitimate range. Statistical packages typically exclude units that have missing value codes for any of the variables involved in an analysis. This strategy is generally inappropriate, since the investigator is usually interested in making inferences about the entire target population, rather than the proportion of the target population that would provide responses to all relevant variables in the analysis.

To analyze data with a missing observation in experimental designs, the frequently used method is missing plot technique, in which an estimate of the missing observation is obtained by using the available data. Then the usual analysis of variance is carried out. The standard procedure for estimating the missing values is to minimize the residual sum of squares. Allan and Wishart (1930) were the first to introduce a formula for estimating a missing value in randomised block designs and Latin square designs. Later, Yates (1933) introduced an iterative procedure for estimating several missing observations in experimental designs. The literature shows that a number of attempts have been made to obtain a simplified procedure for estimating more than one missing

value. When the number of missing observations is at most two, explicit expressions are available to obtain the estimate for missing value. However, if more than two observations are missing Yates' iterative procedure is recommended (Kshirsagar, 1983). For a detailed discussion of missing value problems, see Rubin (1972), Seber (1977) and Dodge (1985).

In this chapter the problem of estimation of parameters of a simple linear regression model when some values of the independent variables are missing is considered.

Statement of the Problem

A problem frequently occurring in statistical practice is that of dealing with regression when some of the values of the independent variables are missing. We discuss in this chapter the special case of regression, a dependent variable y regress on two independent variables x_1 and x_2 , where there are missing values on x_2 only. Such a situation may arise in chemical experiments, for example, where destruction or removal of part of an experimental units prevents the investigator from making complete measurements on x_2 . The question then is how to deal properly with the incomplete sample.

We consider the linear model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \dots, N, \quad \dots(3.1.1)$$

where the e_i are identical independent distributed random variables with mean zero and variance σ^2 , and $(N-n)$ values on x_2 are unobserved.

The simplest approach is to apply the usual least squares analysis to the subset of complete case , an approach equivalent to applying the method of least squares to all available observations while regarding the missing values as additional parameters to be estimated. However, many alternatives to the complete case method have been proposed.

Dear (1959) studied the substitution of the sample means for missing observations, followed by the usual application of least squares . Walsh (1959) and Buck (1960) investigated the substitution of predicted values from simple linear regression among the observed values. Beale and Little (1975) applied the method of maximum likelihood to deal with the problem arising from missing data.

Properties of these procedures have been compared extensively, mainly through simulation studies by Beale and Little (1975), Donner and Rosner (1982), Haitovsky (1968), Heiberger (1977) and Timm (1970). Beale and Little (1975) have found that for multinormal distributions the method of maximum likelihood is preferable to the complete case method. Donner and Rosner (1982) showed that the method of maximum likelihood is superior to other methods in estimating β_1 and β_2 of the model (3.1.1).

Analytic comparisons have generally not been performed, largely because of the complexity of the problem that arises when missing values are scattered over several explanatory variables. However, Donner (1982) has derived expressions for the bias and variance associated with procedures frequently used to deal with the problem. He has found that in terms of mean squared error

(MSE), the mean substitution method is more effective for estimating β_2 , the coefficient of the incompletely observed variable when the correlations involving these variables are small and the proportion of missing cases is fairly high. In estimating β_1 , the coefficient of the completely observed variable, this method provides estimates which may have substantial bias unless the correlations involving these variables are small. Further, he showed that the method of linear prediction is advantageous over the complete case method for estimating β_1 while the two methods are equivalent for estimating β_2 .

In this chapter analytic comparisons of all available approaches are considered.

Description of Procedures Investigated:

The procedures considered are mentioned below with the resulting estimate of $\beta = (\beta_1, \beta_2)'$ denoted in parentheses:

(i) ***The complete case method*** in which the usual least squares analysis is applied to the cases for which both x_1 and x_2 are completely observed, $b_c = (b_{1c}, b_{2c})'$.

(ii) ***The method of linear prediction*** in which a missing x_2 is replaced by its predicted value, obtained from the simple linear regression of x_2 on x_1 over the complete cases followed by the usual analysis, $b_L = (b_{1L}, b_{2L})'$.

(iii) *The mean substitution method* in which a missing x_2 is replaced by the mean of x_2 over the complete case followed by the usual analysis, $b_m = (b_{1m}, b_{2m})'$.

Notation Used :

In presenting the results the following notations is used :

$f = n/N$: proportion of cases observed on x_2 , $0 < f < 1$.

r : simple correlation coefficient between x_1 and x_2 over the n complete cases.

r^* : simple correlation coefficient between x_1 and x_2 over all N cases, including those for which x_2 is not actually observed.

$\text{Tr}(Z)$: trace of a square matrix Z .

$|Z|$: Determinant of a square matrix Z .

$\|\beta\|^2 = \beta'\beta$: the square of the norm of any vector β .

$\text{Bias}(b) = E(b) - \beta$: the bias of any estimator b of β .

$\text{Bias}^2(b) = \|E(b) - \beta\|^2$: the square of the bias for any estimator b of β .

$D(b) = E\{[b - E(b)][b - E(b)]'\}$: the dispersion matrix of any vector of estimators b of β .

$\text{MSE}(b) = \text{Bias}^2(b) + \text{Tr}(D(b))$: the mean square error for an estimator b of β .

$$D = (r^* - r) .$$

$$H = 1 - f(1 - r^2) .$$

$$H_1 = 1 - fr^2 .$$

3.2 METHODS OF ESTIMATION

Assumptions :

We are primarily concerned with estimating β_1 and β_2 under the assumptions that the $(N-n)$ values of x_2 are missing. We assume that the values (x_{1i}, x_{2i}) are fixed with zero means and unit variances over all N cases, and over the n complete cases. We also assume that the y_i are centered about zero over all N cases and over the n complete cases. Writing the observations with missing x_{2i} last, we separate the complete case from the incomplete case in the linear model :

$$Y = X\beta + e, \quad \dots(3.2.1)$$

where Y is an $N \times 1$ vector of observations, X is an $N \times 2$ input matrix, $\beta = (\beta_1, \beta_2)'$ is a 2-dimensional vector parameters, and e is an $N \times 1$ vector of random errors.

We partition Y as follows :

$$Y = \begin{bmatrix} Y_c \\ \dots \\ Y_I \end{bmatrix}, \quad \dots(3.2.2)$$

where Y_c is the vector of observations on n complete cases, Y_I is the vector of observations for incomplete cases.

The input matrix X and the vector of random errors, e , are partitioned accordingly i.e

$$X = \begin{bmatrix} X_c \\ \dots \\ X_I \end{bmatrix} = \begin{bmatrix} X_{1c} & X_{2c} \\ \dots & \dots \\ X_{1I} & X_{2I} \end{bmatrix},$$

with $X_c = (X_{1c} \ X_{2c})$, $X_I = (X_{1I} \ X_{2I})$, and $e = \begin{bmatrix} e_c \\ \dots \\ e_I \end{bmatrix}$.

Our assumptions can be summarized as follows:

$$J'Y = J'Y_c = J'Y_I = 0 ,$$

$$J'X_{1c} = J'X_{2c} = J'X_{1I} = J'X_{2I} = 0 ,$$

$$n = X'_{1c} X_{1c} = X'_{2c} X_{2c} ,$$

$$N-n = X'_{1I} X_{1I} = X'_{2I} X_{2I} ,$$

$$r \ n = X'_{1c} X_{2c}$$

and $r^*N = X'_{1c} X_{2c} + X'_{1I} X_{2I} , \quad \dots(3.2.3)$

where J is a column vector of ones.

Methods of Estimation

1. Complete Case Method :

In the complete case method the usual least squares analysis is applied to the model

$$Y_c = X_c \beta + e_c ,$$

$$E(\mathbf{e}_c) = 0, E(\mathbf{e}_c \mathbf{e}_c') = \sigma^2 \mathbf{I},$$

where $\mathbf{X}_c = (\mathbf{X}_{1c} \mathbf{X}_{2c})$ and $\mathbf{I}$ is the identity matrix. For the complete case, the estimator of β is given by

$$\mathbf{b}_c = (\mathbf{X}_c' \mathbf{X}_c)^{-1} \mathbf{X}_c' \mathbf{Y}_c, \quad \dots(3.2.4)$$

where

$$(\mathbf{X}_c' \mathbf{X}_c) = \begin{bmatrix} n & r n \\ r n & n \end{bmatrix}$$

and

$$(\mathbf{X}_c' \mathbf{X}_c)^{-1} = \frac{1}{n(1-r^2)} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}.$$

The estimator $\mathbf{b}_c$ is unbiased for β and its dispersion matrix is given by

$$D(\mathbf{b}_c) = \sigma^2 (\mathbf{X}_c' \mathbf{X}_c)^{-1}.$$

The MSE of $\mathbf{b}_c$ is given by

$$\text{MSE}(\mathbf{b}_c) = \sigma^2 \text{Tr} (\mathbf{X}_c' \mathbf{X}_c)^{-1} = 2\sigma^2 / \{n(1-r^2)\}. \quad \dots(3.2.5)$$

Alternative Methods :

2. *Linear Prediction Method :*

In the linear prediction method we replace X_{2I} by its predicted value obtained by the relation

$$X_{2I} = r X_{1I}. \quad \dots(3.2.6)$$

3. The Mean Substitution Method :

In the mean substitution method we substitute X_{2I} by the mean of n available observation on x_2 . Since x_{2i} are observed with zero mean, thus for this method we substitute

$$X_{2I} = 0 . \quad \dots(3.2.7)$$

For simplicity we consider, in general, a prediction model of the form

$$X_{2I} = kr X_{1I} , k \geq 0 . \quad \dots(3.2.8)$$

Note that, the relations (3.2.6) and (3.2.7) can easily be obtained from the relation (3.2.8) by setting $k=1$ and 0 respectively.

Let

$$X_k = \begin{bmatrix} X_{1c} & X_{2c} \\ X_{1I} & krX_{1I} \end{bmatrix}$$

and

$$S = (X'_k X_k) = \begin{bmatrix} N & r n + kr(N - n) \\ r n + kr(N - n) & n + k^2 r^2 (N - n) \end{bmatrix} . \quad \dots(3.2.9)$$

We define b_k , an estimator of β by

$$b_k = (b_{1k} , b_{2k})' = S^{-1} X'_k Y . \quad \dots(3.2.10)$$

So

$$\text{Bias}(b_k) = S^{-1} X'_k (X - X_k) \beta ,$$

where

$$S^{-1} = \frac{N}{|S|} \begin{bmatrix} f + k^2 r^2 (1 - f) & -rf - kr(1 - f) \\ -rf - kr(1 - f) & 1 \end{bmatrix}, \quad \dots(3.2.11)$$

$$X'_k (X - X_k) = G \begin{bmatrix} 0 & 1 \\ 0 & kr \end{bmatrix}, \quad \dots(3.2.12)$$

$$|S| = Nn [(1-r^2)+r^2(1-k)^2(1-f)], \quad \dots(3.2.13)$$

and

$$G = N[D+r(1-k)(1-f)]. \quad \dots(3.2.14)$$

Straight forward algebraic simplification gives us

$$\text{Bias}(b_k) = \frac{n\beta_2 G}{|S|} \begin{bmatrix} 1 - kr^2 \\ r(k - 1) \end{bmatrix}, \quad \dots(3.2.15)$$

and

$$\text{Bias}^2(b_k) = \frac{n^2 \beta_2^2 G^2}{|S|^2} [(1 - kr^2)^2 + r^2 (k - 1)^2], \quad \dots(3.2.16)$$

where $|S|$ and G are given in (3.2.13) and (3.2.14) respectively.

From the relation (3.2.10) we can write

$$b_k - E(b_k) = S^{-1} X'_k e.$$

Therefore, the dispersion matrix of b_k is given by

$$D(b_k) = E\{(b_k - E(b_k)) (b_k - E(b_k))'\} = \sigma^2 S^{-1}, \quad \dots(3.2.17)$$

where S^{-1} is given by (3.2.11). The MSE of b_k is given by

$$\text{MSE}(b_k) = \text{Bias}^2(b_k) + \text{Tr } D(b_k)$$

$$= \frac{n^2 \beta_2^2 G^2}{|S|^2} \{(1 - kr^2)^2 + r^2(k-1)^2\} + \frac{\sigma^2 N}{|S|} \{f + k^2 r^2(1-f) + 1\} \dots (3.2.18)$$

For $k = 1/r^2$, we denote b_k by $b_I = (b_{1L}, b_{2I})'$. Again, note that the estimators b_L and b_m of β can be obtained from the relation (3.2.10) by setting $k=1$ and 0 respectively.

We discuss the behaviors of b_I , b_L and b_m in the next section.

3.3 BEHAVIOUR OF b_L , b_m AND b_I

The expressions for biases and dispersion matrices of b_L , b_m and b_I can be obtained from the relations (3.2.15) and (3.2.17) by setting $k=1, 0$ and $1/r^2$ respectively. We summarize these results in table 3.4

Donner (1982) has shown that for estimating β_1 , the coefficient of the completely observed variable, b_{1L} offers advantage over b_{1c} and b_{1m} . For estimating β_2 , the coefficient of the incompletely observed variable, the method of linear prediction and the complete case method are equivalent, i.e $b_{2L} \equiv b_{2c}$. Since $b_{2L} \equiv b_{2c}$, the only attractive alternative to b_{2c} in this case is the estimator b_{2m} , which has lower MSE if r is small and the proportion of missing cases is moderately high.

Replacing k in the relation (3.2.10) by $1/r^2$ and using the assumptions (3.2.3) we have

$$\begin{aligned}
b_I &= S^{-1} X'_k Y|_{k=I/r^2} \\
&= \frac{1}{n(1-r^2)} \begin{bmatrix} 1 & -r \\ -r & r^2 / H \end{bmatrix} \begin{bmatrix} X'_{1c} Y_c + X'_{1I} Y_I \\ X'_{2c} Y_c + X'_{1I} Y_I / r \end{bmatrix}.
\end{aligned}$$

Therefore,

$$b_{II} = \frac{1}{n(1-r^2)} \{X'_{1c} Y_c - rX'_{2c} Y_c\}.$$

For the complete case estimator

$$\begin{aligned}
b_c &= (X'_c X_c)^{-1} X'_c Y_c \\
&= \frac{1}{n(1-r^2)} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix} \begin{bmatrix} X'_{1c} Y_c \\ X'_{2c} Y_c \end{bmatrix}.
\end{aligned}$$

Therefore,

$$b_{1c} = \frac{1}{n(1-r^2)} \{X'_{1c} Y_c - rX'_{2c} Y_c\},$$

hence

$$b_{II} \equiv b_{1c}. \quad \dots(3.3.1)$$

TABLE 3.4

Bias and Dispersion Matrix of Estimators of β .

Estimator	Bias	Dispersion Matrix
$b_c = \begin{bmatrix} b_{1c} \\ b_{2c} \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\frac{\sigma^2}{n(1-r^2)} \begin{bmatrix} 1 & -r \\ -r & 1 \end{bmatrix}$
$b_L = \begin{bmatrix} b_{1L} \\ b_{2L} \end{bmatrix}$	$\begin{bmatrix} \beta_2 D \\ 0 \end{bmatrix}$	$\frac{\sigma^2}{n(1-r^2)} \begin{bmatrix} f + r^2(1-f) & -r \\ -r & 1 \end{bmatrix}$
$b_m = \begin{bmatrix} b_{1m} \\ b_{2m} \end{bmatrix}$	$\frac{\beta_2(D + r(1-f))}{H_1} \begin{bmatrix} 1 \\ -r \end{bmatrix}$	$\frac{\sigma^2}{n H_1} \begin{bmatrix} f & -rf \\ -rf & 1 \end{bmatrix}$
$b_I = \begin{bmatrix} b_{1I} \\ b_{2I} \end{bmatrix}$	$\frac{\beta_2(rr^* - H)}{H} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	$\frac{\sigma^2}{n(1-r^2)} \begin{bmatrix} 1 & -r \\ -r & r^2 / H \end{bmatrix}$

Note from table 3.4 that

$$\text{Bias}(b_{2c}) = 0 ,$$

$$\text{Bias}(b_{2I}) = \beta_2(r^*-H)/H ,$$

$$\text{Bias}(b_{2m}) = -\beta_2 r \{D+r(1-f)\}/H_1 ,$$

$$\text{Var}(b_{2c}) = \sigma^2 / \{n(1-r^2)\} ,$$

$$\text{Var}(b_{2I}) = \sigma^2 r^2 / \{n(1-r^2)H\}$$

and

$$\text{Var}(b_{2m}) = \sigma^2 / \{nH_1\} .$$

It can be show that if $r^2 \leq 1/2$ then

$$\text{Var}(b_{2I}) \leq \text{Var}(b_{2m}) , \quad \dots(3.3.2)$$

and if

$$\{1/r-(D+r(1-f))/H_1\} H \leq r^* \leq \{1/r+(D+r(1-f))/H_1\} H , \quad \dots(3.3.3)$$

then

$$\text{Bias}^2(b_{2I}) \leq \text{Bias}^2(b_{2m}) . \quad \dots(3.3.4)$$

Thus b_{2I} may be considered an effective alternative to b_{2m} . From the relations (3.3.2) and (3.3.4) it follows that b_{2I} offers advantages in estimating β_2 , while in estimating β_1 , $b_{1I} = b_{1c}$.

Chapter 4

Use of Mixed Regression Approach When Some Values of the Regressors are Missing

4.1 INTRODUCTION

We consider the model

$$Y_1 = X_1 \beta + e_1, \quad \dots(4.1.1)$$

$$E(e_1) = 0, \quad E(e_1 e_1') = \sigma^2 I_n,$$

where Y_1 is an $n \times 1$ vector of observations, X_1 is an $n \times p$ design matrix with rank $(X_1) = p < n$, β is an $p \times 1$ vector of unknown parameters, e_1 is an $n \times 1$ vector of random errors, I_k is the $k \times k$ identity matrix, 0 is the n -dimensional null vector and σ^2 is an unknown positive constant.

Further we assume that m additional observations are available on the dependent variable y , but some of the regressor values are missing. That is, additional observations are of the form

$$Y_2 = X^* \beta + e_2,$$

$$E(e_2) = 0, \quad E(e_2 e_2') = \sigma^2 I_m,$$

and $E (\mathbf{e}_1 \mathbf{e}_2') = E (\mathbf{e}_2 \mathbf{e}_1') = 0 ,$

where $\mathbf{Y}_2$ is an $m \times 1$ vector of additional observations, $\mathbf{X}_*$ is an $m \times p$ incomplete design matrix, that is some values of $\mathbf{X}_*$ are missing, and $\mathbf{e}_2$ is an $m \times 1$ vector of random errors.

Suppose that an $m \times p$ non-stochastic matrix $\mathbf{X}_2$ is substituted for the unknown design matrix $\mathbf{X}_*$ so that

$$\begin{aligned} \mathbf{Y}_2 &= \mathbf{X}_* \boldsymbol{\beta} + \mathbf{X}_2 \boldsymbol{\beta} - \mathbf{X}_2 \boldsymbol{\beta} + \mathbf{e}_2 \\ &= \mathbf{X}_2 \boldsymbol{\beta} + \boldsymbol{\delta}_2 + \mathbf{e}_2 , \end{aligned} \quad \dots(4.1.2)$$

where $\boldsymbol{\delta}_2 = (\mathbf{X}_* - \mathbf{X}_2) \boldsymbol{\beta}$. Combining (4.1.1) and (4.1.2) we get

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\delta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} . \quad \dots(4.1.3)$$

It may be mentioned that to replace $\mathbf{X}_*$ by $\mathbf{X}_2$ the experimenter may employ the method of mean substitution or the method of linear prediction.

The least squares estimator of $\boldsymbol{\beta}$ based on the combined model (4.1.3) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_1' \mathbf{X}_1 + \mathbf{X}_2' \mathbf{X}_2)^{-1} (\mathbf{X}_1' \mathbf{Y}_1 + \mathbf{X}_2' \mathbf{Y}_2) . \quad \dots(4.1.4)$$

The expected value, the bias and the variance - covariance matrix of $\hat{\boldsymbol{\beta}}$ are given by

$$E (\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbf{W}^{-1} \mathbf{X}_2' \boldsymbol{\delta}_2 ,$$

$$\text{Bias}(\hat{\beta}) = W^{-1}X_2'\delta_2,$$

and

$$\begin{aligned} D(\hat{\beta}) &= E[\{\hat{\beta} - E(\hat{\beta})\}\{\hat{\beta} - E(\hat{\beta})\}'] \\ &= \sigma^2 W^{-1}, \end{aligned} \quad \dots(4.1.5)$$

where $W = (X_1'X_1 + X_2'X_2)$, and δ_2 is given in (4.1.2). The mean square error matrix of $\hat{\beta}$ is given by

$$\begin{aligned} M(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[\{\hat{\beta} - E(\hat{\beta})\}\{\hat{\beta} - E(\hat{\beta})\}' \\ &\quad + \{E(\hat{\beta}) - \beta\}\{E(\hat{\beta}) - \beta\}'] \\ &= \sigma^2 W^{-1} + W^{-1}X_2'\delta_2\delta_2'X_2W^{-1}. \end{aligned} \quad \dots(4.1.6)$$

The least squares estimator of β based on the complete model (4.1.1) is

$$b = (X_1'X_1)^{-1}X_1'Y_1 = \beta + (X_1'X_1)^{-1}X_1'e_1. \quad \dots(4.1.7)$$

This estimator is unbiased for β . The mean square error matrix of b is given by

$$M(b) = \sigma^2 (X_1'X_1)^{-1}. \quad \dots(4.1.8)$$

4.2 COMPARISON

In this section we shall compare the mean squared error matrices of b and $\hat{\beta}$. Clearly,

$$M(b) - M(\hat{\beta}) = \sigma^2 (X_1'X_1)^{-1} - \sigma^2 W^{-1} - W^{-1}X_2'\delta_2\delta_2'X_2W^{-1}. \quad \dots(4.2.1)$$

$$\begin{aligned}
\text{Now, } W^{-1} &= (X_1' X_1 + X_2' X_2)^{-1} \\
&= (X_1' X_1)^{-1} - (X_1' X_1)^{-1} X_2' [I_m + X_2 (X_1' X_1)^{-1} X_2']^{-1} X_2 (X_1' X_1)^{-1} \\
&= (X_1' X_1)^{-1} - F' G^{-1} F, \quad \dots (4.2.2)
\end{aligned}$$

where $F = X_2 (X_1' X_1)^{-1}$, and $G = I_m + X_2 (X_1' X_1)^{-1} X_2'$. Using (4.2.2) in (4.2.1) we obtain

$$\begin{aligned}
M(b) - M(\hat{\beta}) &= \sigma^2 F' G^{-1} F - W^{-1} X_2' \delta_2 \delta_2' X_2 W^{-1} \\
&= \sigma^2 W^{-1} [W F' G^{-1} F W - X_2' \delta_2 \delta_2' X_2 / \sigma^2] W^{-1}. \quad \dots (4.2.3)
\end{aligned}$$

We observe that

$$\begin{aligned}
W F' &= (X_1' X_1 + X_2' X_2) (X_1' X_1)^{-1} X_2' \\
&= X_2' + X_2' X_2 (X_1' X_1)^{-1} X_2' \\
&= X_2' [I_m + X_2 (X_1' X_1)^{-1} X_2'] \\
&= X_2' G. \quad \dots (4.2.4)
\end{aligned}$$

It follows from (4.2.3) and (4.2.4) that

$$\begin{aligned}
M(b) - M(\hat{\beta}) &= \sigma^2 W^{-1} X_2' [G - \delta_2 \delta_2' / \sigma^2] X_2 W^{-1} \\
&= \sigma^2 W^{-1} X_2' G^{-1} [I_m - G^{-1} \delta_2 \delta_2' G^{-1} / \sigma^2] G^{-1} X_2 W^{-1}.
\end{aligned}$$

Obviously, $M(b) - M(\hat{\beta})$ will be positive semi-definite iff the matrix

$$I_m - G^{-1} \delta_2 \delta_2' G^{-1} / \sigma^2$$

is positive semi-definite.

Since $\{I_m - G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2\}$ is a symmetric matrix, it will be positive semi-definite iff the smallest eigenvalue of this matrix is non-negative.

The eigenvalues of this matrix are the roots $\lambda_1, \lambda_2, \dots, \lambda_m$ of the equation

$$| I_m - G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2 - \lambda I | = 0 . \quad \dots (4.2.5)$$

where $|E|$ denotes the determinant of the square matrix E . The equation (4.2.5) can be rewritten as

$$| (G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2) - (1 - \lambda) I_m | = 0 .$$

Writing $(1 - \lambda) = \lambda^*$ we obtain

$$| (G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2) - \lambda^* I_m | = 0 .$$

Since $(G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2)$ is a positive semi-definite matrix of rank one, it has only one non-zero eigenvalue. The remaining eigenvalues of $(G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2)$ are zero.

The non-zero eigenvalue of $(G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2)$ is given by

$$\begin{aligned} \lambda^* &= \text{Tr} [G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2] \\ &= (1/\sigma^2) \delta_2' G^{-1} \delta_2 , \end{aligned} \quad \dots (4.2.6)$$

where $\text{Tr} (F)$ denotes the trace of a square matrix F . Since $\lambda^* = 1 - \lambda$, therefore, $\lambda = 1 - \lambda^*$. Hence the smallest eigenvalue of

$$(I_m - G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2) \text{ is } (1 - \delta_2' G^{-1} \delta_2 / \sigma^2) .$$

The remaining eigenvalues of $(I_m - G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2)$ are 1 with multiplicity $(m-1)$.

Therefore , a necessary condition for $(I_m - G^{-*} \delta_2 \delta_2' G^{-*} / \sigma^2)$ to be positive semi-definite is that its smallest eigenvalue is non-negative. That is

$$\delta_2' G^{-1} \delta_2 / \sigma^2 \leq 1 . \quad \dots(4.2.7)$$

If the inequality in (4.2.7) holds , then accommodation of Y_2 , the vector of additional observations, will increase the precision of the least squares estimator of β . Otherwise, accommodation of Y_2 in the analysis will result in poorer least squares estimation.

The inequality in (4.2.7) is a function of unknown parameters δ_2 and σ^2 , therefore, the experimenter will never know when the condition (4.2.7) is satisfied. Here a formal test is needed to test the null hypothesis

$$H_0 : \delta_2' G^{-1} \delta_2 / (2 \sigma^2) \leq 1/2 \quad \dots(4.2.8)$$

against the alternative hypothesis

$$H_1 : \delta_2' G^{-1} \delta_2 / (2 \sigma^2) > 1/2 .$$

Yancy, Judge and Bock (1974) and Ali (1994) developed similar tests in different contexts.

Assumption

To test the hypothesis (4.2.8) we assume that the vector of random errors, $e = (e_1' \mid e_2')'$, of the model (4.1.3) follows an $(n + m)$ dimensional multivariate normal distribution with $E (e) = 0$ and $E (e e') = \sigma^2 I_{m+n}$.

It can be shown easily that if the above assumption is true, then the statistic $(Y_2 - X_2 b)$, where b is given by (4.1.7), has an m dimensional multivariate normal distribution with $E [Y_2 - X_2 b] = \delta_2$

and $E [\{ (Y_2 - X_2 b) - \delta_2 \} \{ (Y_2 - X_2 b) - \delta_2 \}'] = \sigma^2 G$.

Let $M_1 = I_n - X_1(X_1' X_1)^{-1} X_1'$,

$$\hat{e}_1 = Y_1 - X_1 b = Y_1 - X_1 (X_1' X_1)^{-1} X_1' Y_1 = M_1 Y_1 = M_1 e_1, \quad \dots(4.2.9)$$

where G is given in (4.2.2). Next, from (4.1.7) and (4.2.9) we obtain

$$\begin{aligned} E[(b - \beta) \hat{e}_1'] &= E[(X_1' X_1)^{-1} X_1' e_1 \cdot e_1' M_1] \\ &= \sigma^2 (X_1' X_1)^{-1} X_1' M_1 = 0. \end{aligned} \quad \dots(4.2.10)$$

Since $\hat{e}_1$ and b have multivariate normal distributions and $E[(b - \hat{\beta}) \hat{e}_1'] = 0$,

therefore, $\hat{e}_1$ and b are independently distributed. Further, let

$$\hat{\sigma}^2 = e_1' M_1 e_1 / (n-p) = \hat{e}_1' \hat{e}_1 / (n-p).$$

Since $\hat{e}_1$ and b are independently distributed, therefore, $\hat{\sigma}^2$ and b are also independently distributed. It is well known that

$$(n-p) \hat{\sigma}^2 / \sigma^2 \sim \chi_{n-p}^2.$$

Further, the statistic $(Y_2 - X_2 b)' G^{-1} (Y_2 - X_2 b) / \sigma^2$ has a non-central chi-square distribution with m degrees of freedom and non-centrality parameter

$$\lambda = \delta_2' G^{-1} \delta_2 / (2 \sigma^2), \quad \dots(4.2.11)$$

(See, Graybill (1961), corollary 4.7.1, page 83).

The two chi-square variables are independent. Hence the random variable

$$W = (Y_2 - X_2 b)' G^{-1} (Y_2 - X_2 b) / (m \hat{\sigma}^2)$$

has an $F(m, n-p, \lambda)$, where λ is given by (4.2.11), for all values of δ_2 not identically zero. Thus W provides a test statistic for testing the hypothesis

$$H_0 : \delta_2' G^{-1} \delta_2 / (2 \sigma^2) = \lambda \leq 1/2$$

against

$$H_1 : \delta_2' G^{-1} \delta_2 / (2 \sigma^2) = \lambda > 1/2 .$$

To use the test, the investigator would calculate W and reject the hypothesis if W exceeds the critical value provided for m and $(n-p)$ in the extensive non-central F table developed by Wallace and Toro-Vizcarrondo (1969).

4.3 MIXED ESTIMATION

Theil and Goldberger (1961) and Theil (1963) founded the mixed regression approach. Their mixed regression estimator is essentially a large class of estimators that includes ridge, and shrinkage estimators. Properties of these estimators when values of the explanatory variables are missing have not been examined. The present investigation shows that the mixed regression estimator for the situations in hand, is superior, in terms of squared bias and the variance, to the least squares estimator.

Suppose that in addition to having observations in (4.1.3) we have auxiliary information on the vector of unknown regression coefficients. We assume that the auxiliary information can be expressed as

$$r = R \beta + \phi , \quad \dots(4.3.1)$$

$$E(\phi) = 0 , E(\phi \phi') = \sigma^2 I_k / c ,$$

Here $\mathbf{r}$ and $\mathbf{R}$, known from the prior information are formulated as $k \times 1$ vector and a $k \times p$ matrix respectively, ϕ is a $k \times 1$ vector of unobservable random variables, independent of $\mathbf{e} = (\mathbf{e}'_1 \mid \mathbf{e}'_2)'$ of the model (4.1.3). The scalar ' c ' can be interpreted as a precision parameter of the prior information (4.3.1).

Under these specifications, the mixed regression estimator that combines the sample information (4.1.3) and the prior information (4.3.1) is

$$\hat{\beta}(c) = (\mathbf{X}'\mathbf{X} + c\mathbf{R}'\mathbf{R})^{-1} (\mathbf{X}'\mathbf{Y} + c\mathbf{R}'\mathbf{r}), \quad \dots(4.3.2)$$

where $\mathbf{Y} = (\mathbf{Y}'_1 \mid \mathbf{Y}'_2)'$ and $\mathbf{X} = [\mathbf{X}'_1 \mid \mathbf{X}'_2]'$.

Toutenburg (1982 , p.61) obtained an estimator similar to the one given by (4.3.2). The class of estimators defined by (4.3.2) contains a variety of estimators, for instance :

(a) If $c \rightarrow 0$, then $\hat{\beta}(c) \rightarrow \hat{\beta}(0) = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$.

$\hat{\beta}(0)$ is the **least squares estimator of β** based on the sample information (4.1.3).

(b) If $\mathbf{r} = 0$, $k = p$ and $\mathbf{R} = \mathbf{I}_p$, then

$$\hat{\beta}(c) = (\mathbf{X}'\mathbf{X} + c\mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{Y},$$

which is the **ordinary ridge regression estimator** originally developed by Hoerl and Kennard (1970). Toutenburg (1982, p.64) compared the mixed and the ordinary ridge regression estimators. When $\mathbf{r} = 0$, the auxiliary information in (4.3.1) may be called fictitious. Askin and Montgomery (1980) advocated the augmentation of sample data with fictitious observations.

(C) If $r = 0$, then we have

$$\hat{\beta}(c) = (X'X + cR'R)^{-1} X'Y$$

which is the **generalized ridge estimator** proposed by Rao (1975) . Further, if $r = 0$, then $\hat{\beta}(c)$ in (4.3.2) can be written as

$$\begin{aligned}\hat{\beta}(c) &= (X'X + cR'R)^{-1} X'X \hat{\beta}(0) \\ &= C_* \hat{\beta}(0),\end{aligned}\quad \dots(4.3.3)$$

where $\hat{\beta}(0) = (X'X)^{-1}X'Y$ and $C_* = (X'X + cR'R)^{-1}(X'X)$.

The class of estimators given by (4.3.3) are called **shrinkage estimators**.

4.4 PROPERTIES OF MIXED REGRESSION ESTIMATOR.

In this section we shall study the principal properties of both the mixed regression and the least squares estimators of β and $X\beta$. The mixed regression and the least squares estimators of $X\beta$ are given by

$$\hat{Y}(c) = X \hat{\beta}(c) = X (X'X + cR'R)^{-1} (X'Y + cR'r) \quad \dots(4.4.1)$$

and

$$\hat{Y}(0) = X \hat{\beta}(0) = HY, \quad \dots(4.4.2)$$

respectively where $\hat{\beta}(c)$ is given by (4.3.2), $\hat{\beta}(0) = (X'X)^{-1} X'Y$ and $H = X(X'X)^{-1}X'$. Clearly,

$$E [\hat{\beta}(c) - \beta] = (X'X + cR'R)^{-1} X' \delta, \quad \dots(4.4.3)$$

where $\delta = (0' \mid \delta_2')'$. Therefore,

$$\{E[\hat{\beta}(c) - \beta]\}' \{E[\hat{\beta}(c) - \beta]\} = \delta' X A^{-2} X' \delta, \quad \dots(4.4.4)$$

where $A = (X'X + cR'R)$.

The variance - covariance matrix of $\hat{\beta}(c)$ is

$$\begin{aligned} D(\hat{\beta}(c)) &= E[\{\hat{\beta}(c) - E[\hat{\beta}(c)]\} \{\hat{\beta}(c) - E[\hat{\beta}(c)]\}'] \\ &= \sigma^2 A^{-1}. \end{aligned} \quad \dots(4.4.5)$$

Setting $c = 0$ in (4.4.3), (4.4.4) and (4.4.5) we obtain

$$\begin{aligned} E[\hat{\beta}(0) - \beta] &= (X'X)^{-1} X' \delta, \\ \{E(\hat{\beta}(0)) - \beta\}' \{E(\hat{\beta}(0)) - \beta\} &= \delta' X (X'X)^{-2} X' \delta \end{aligned} \quad \dots(4.4.6)$$

and

$$D(\hat{\beta}(0)) = \sigma^2 (X'X)^{-1}. \quad \dots(4.4.7)$$

Theorem (4.1)

For any full rank design matrix X , non-null vector δ and $c > 0$,

$$(i) \quad \frac{\{E(\hat{\beta}(0)) - \beta\}' \{E(\hat{\beta}(0)) - \beta\}}{\{E(\hat{\beta}(c)) - \beta\}' \{E(\hat{\beta}(c)) - \beta\}} \geq 1$$

(ii) $[D(\hat{\beta}(0)) - D(\hat{\beta}(c))]$ is non-negative definite.

(iii) $MSE(\hat{\beta}(0)) - MSE(\hat{\beta}(c)) \geq 0$.

where $MSE(\hat{\beta}(c)) = \text{Tr}\{D(\hat{\beta}(c))\} + \{E(\hat{\beta}(c)) - \beta\}' \{E(\hat{\beta}(c)) - \beta\}$

and $MSE(\hat{\beta}(0)) = \text{Tr}\{D(\hat{\beta}(0))\} + \{E(\hat{\beta}(0)) - \beta\}' \{E(\hat{\beta}(0)) - \beta\}$.

Proof: Part (i) :

From (4.4.4) and (4.4.6) we have

$$\begin{aligned} \frac{\{E(\hat{\beta}(0)) - \beta\}' \{E(\hat{\beta}(0)) - \beta\}}{\{E(\hat{\beta}(c)) - \beta\}' \{E(\hat{\beta}(c)) - \beta\}} &= \frac{\delta'X (X'X)^{-2} X'\delta}{\delta'X A^{-2} X'\delta} \\ &= \frac{V'A (X'X)^{-2} AV}{V'V}, \end{aligned} \quad \dots(4.4.8)$$

where $V = A^{-1}X'\delta$ and $A = (X'X + cR'R)$.

It follows from (4.4.8) that

$$\frac{\{E(\hat{\beta}(0)) - \beta\}' \{E(\hat{\beta}(0)) - \beta\}}{\{E(\hat{\beta}(c)) - \beta\}' \{E(\hat{\beta}(c)) - \beta\}} \geq \text{Chmin} (A(X'X)^{-2} A), \dots(4.4.9)$$

where $\text{Chmin} (G)$ denotes the smallest eigenvalue of the square matrix G . Since the eigenvalues of $A(X'X)^{-1} = [I + cR'R(X'X)^{-1}]$ are the same as the eigenvalues of $(X'X)^{-1} A = [I + c(X'X)^{-1}R'R]$, therefore, the eigenvalues of $A(X'X)^{-2}A$, are the square of the eigenvalues of $A(X'X)^{-1}$. This statement can be proved as follows : let λ_* be an eigenvalue of $(X'X)^{-1}A$ and U_* be the corresponding eigenvector. Hence

$$(X'X)^{-1}AU_* = \lambda_*U_*.$$

Pre-multiplying both sides by $A(X'X)^{-1}$, we get

$$\begin{aligned} A(X'X)^{-2}AU_* &= \lambda_*(A(X'X)^{-1}U_*) \\ &= \lambda_*^2U_* \end{aligned}$$

showing that λ_*^2 is an eigenvalue of $A(X'X)^{-2}A$. Hence

$$\begin{aligned} \text{Chmin} (A(X'X)^{-2}A) &= [\text{Chmin} A(X'X)^{-1}]^2 \\ &= \{\text{Chmin} [I + cR'R(X'X)^{-1}]\}^2 \end{aligned}$$

$$= \{ \text{Chmin} [I + c(X'X)^{-1} R'R(X'X)^{-1}] \}^2 .$$

Since $(X'X)^{-1} R'R(X'X)^{-1}$ is a positive semi-definite matrix, therefore, all eigenvalues of this matrix are non-negative. Hence

$$\text{Chmin } A(X'X)^{-2}A \geq 1 . \quad \dots(4.4.10)$$

It follows from (4.4.9) and (4.4.10) that

$$\frac{\{ E(\hat{\beta}(0)) - \beta \}' \{ E(\hat{\beta}(0)) - \beta \}}{\{ E(\hat{\beta}(c)) - \beta \}' \{ E(\hat{\beta}(c)) - \beta \}} \geq 1$$

This complete the proof of Part(i).

Proof : of part (ii) :

From (4.4.5) and (4.4.7) we have

$$[D(\hat{\beta}(0)) - D(\hat{\beta}(c))] / \sigma^2 = (X'X)^{-1} - A^{-1} . \quad \dots(4.4.11)$$

$$\begin{aligned} \text{Now, } A^{-1} &= (X'X + cR'R)^{-1} \\ &= (X'X)^{-1} - (X'X)^{-1}R'[R(X'X)^{-1}R' + (cI)^{-1}]^{-1}R(X'X)^{-1} \\ &= (X'X)^{-1} - c(X'X)^{-1}R'[I + cR(X'X)^{-1}R']^{-1}R(X'X)^{-1}. \quad \dots(4.4.12) \end{aligned}$$

(See Rao (1973, P.33)).

Using (4.4.12) in (4.4.11) we obtain

$$[D(\hat{\beta}(0)) - D(\hat{\beta}(c))] / \sigma^2 = c(X'X)^{-1}R'[I + cR(X'X)^{-1}R']^{-1}R(X'X)^{-1}.$$

Hence for all $c > 0$, $[D(\hat{\beta}(0)) - D(\hat{\beta}(c))] / \sigma^2$ is a non-negative definite matrix. This completes the proof of part (ii). The proof of Part (iii) of Theorem (4.1) follows from parts (i) and (ii).

It follows from (4.4.1), (4.4.3), (4.4.4) and (4.4.5) that

$$\begin{aligned} E[\hat{Y}(c) - X\beta] &= XE[\hat{\beta}(c) - \beta] \\ &= XA^{-1}X'\delta = M\delta, \\ \{E(\hat{Y}(c)) - X\beta\}' \{E(\hat{Y}(c)) - X\beta\} &= \delta'M^2\delta, \end{aligned} \quad \dots(4.4.13)$$

and

$$D(\hat{Y}(c)) = XD(\hat{\beta}(c))X' = \sigma^2XA^{-1}X' = \sigma^2M, \quad \dots(4.4.14)$$

where $M = XA^{-1}X'$ and $A = (X'X + cR'R)$.

Setting $c = 0$ in (4.4.13) and (4.4.14) we get

$$\{E(\hat{Y}(0)) - X\beta\}' \{E(\hat{Y}(0)) - X\beta\} = \delta'H\delta, \quad \dots(4.4.15)$$

and

$$D(\hat{Y}(0)) = \sigma^2X(X'X)^{-1}X' = \sigma^2H, \quad \dots(4.4.16)$$

where $H = X(X'X)^{-1}X'$ is an idempotent matrix.

Theorem (4.2) :

For any full rank design matrix X , non-null vector δ and $c > 0$,

$$\begin{aligned} (i) \quad \{E(\hat{Y}(0)) - X\beta\}' \{E(\hat{Y}(0)) - X\beta\} \\ \geq \{E(\hat{Y}(c)) - X\beta\}' \{E(\hat{Y}(c)) - X\beta\} \end{aligned}$$

$$(ii) \quad [D(\hat{Y}(0)) - D(\hat{Y}(c))] \text{ is non-negative definite,}$$

$$(iii) \{MSE(\hat{Y}(0)) - MSE(\hat{Y}(c))\} \geq 0,$$

where

$$MSE(\hat{Y}(c)) = \text{Tr } D(\hat{Y}(c)) + \{E(\hat{Y}(c)) - X\beta\}' \{E(\hat{Y}(c)) - X\beta\}$$

and $MSE(\hat{Y}(0))$ can be obtained by setting $c = 0$ in $MSE(\hat{Y}(c))$.

Proof: Part (i) :

From (4.4.13) and (4.4.15) we have

$$\begin{aligned} & \{E(\hat{Y}(0)) - X\beta\}' \{E(\hat{Y}(0)) - X\beta\} - \{E(\hat{Y}(c)) - X\beta\}' \{E(\hat{Y}(c)) - X\beta\} \\ &= \delta'(H - M^2) \delta = \delta'[X(X'X)^{-1}X' - XA^{-1}X'XA^{-1}X']\delta \\ &= \delta'X[(X'X)^{-1} - A^{-1}(X'X)A^{-1}]X'\delta \\ &= \delta'XA^{-1}[A(X'X)^{-1}A - X'X]A^{-1}X'\delta \\ &= \delta'XA^{-1}E^{-1}[(EAE)^2 - I_p]E^{-1}A^{-1}X'\delta, \end{aligned} \quad \dots(4.4.17)$$

where $E = (X'X)^{-1}$. Now $EAE = I_p + cER'RE$. Since $ER'RE$ is a non-negative definite matrix, then all eigenvalues of $(ER'RE)$ are non-negative. Hence

$$\text{Chmin}(EAE) = 1 + c \text{Chmin}(ER'RE) \geq 1, \quad \dots(4.4.18)$$

where $\text{Chmin}(F)$ denotes the smallest eigenvalue of F . Since $\text{Chmin}(EAE) \geq 1$, therefore $\text{Chmin}(EAE)^2 \geq 1$ and

$$\text{Chmin}(EAE)^2 - 1 \geq 0. \quad \dots(4.4.19)$$

see Searle (1982, P.276).

It follows from (4.4.17) and (4.4.19) that

$$\text{Chmin } \{ (EAE)^2 - I_p \} = \text{Chmin } (EAE)^2 - 1 \geq 0.$$

Thus, all eigenvalues of $[(EAE)^2 - I_p]$ are non-negative. Further, $[(EAE)^2 - I_p]$ is symmetric. Hence $[(EAE)^2 - I_p]$ is a non-negative definite matrix. The result is obvious.

Proof: Part (ii) :

From (4.4.14) and (4.4.16) we have

$$\begin{aligned} [D(\hat{Y}(0)) - D(\hat{Y}(c))] / \sigma^2 &= H - M \\ &= X [(X'X)^{-1} - A^{-1}] X'. \end{aligned}$$

Since

$$\begin{aligned} A^{-1} &= (X'X + c R'R)^{-1} \\ &= (X'X)^{-1} - c (X'X)^{-1} R' [I + c R (X'X)^{-1} R']^{-1} R (X'X)^{-1}, \end{aligned}$$

Here $(H - M)$ is a non-negative definite matrix. Hence

$$[D(\hat{Y}(0)) - D(\hat{Y}(c))] / \sigma^2$$

is a non-negative definite matrix. This completes the proof of part (ii). Part (iii) follows from part (i) and part (ii).

Chapter 5

A class of Jackknifed Estimators for the Vector of Unknown Regression Coefficients When Some Observations are Missing

5.1 INTRODUCTION

The method of least squares has dominated the literature of regression analysis . The least squares estimators of the unknown regression coefficients enjoy the property of minimum variance unbiased estimators if the observations satisfy certain desirable properties . If some values of the predictor variables are missing and these missing values are replaced by some auxiliary values computed by an imputation method , then the least squares estimator of the vector of unknown regression coefficients no longer enjoys the property of unbiased estimator . The bias of the estimators are unknown and may be very large . Of course , the bias term depends on the efficiency of the imputation method . If the imputation method provides very good estimates of the missing values , then the bias associated with the least squares estimators may not be very large . Otherwise , the bias can be very large . It is well known that a highly biased estimator may render the point estimate dubious . Reduction of bias in

estimation , therefore , often calls for the attention of investigators . In developing the design methodology for response surfaces Box and Draper (1959) suggested to minimize the squared bias alone . Here an attempt is made to reduce the bias of the least squares estimator of the vector of unknown regression coefficients when missing regressor values are replaced by some suitable values.

Quenouille (1956) introduced the jackknife technique for reducing bias in estimation. An extensive literature dealing with this technique has been developed since then. An excellent review of the literature was presented by Miller (1974 a) . In a separate article Miller (1974 b) gave a detailed account of jackknifing linear model estimators. He showed that the jackknife procedures yield consistent results in large samples. Hinkley (1977) investigated the standard jackknife and a weighted jackknife in the general linear model. He found that for unbalanced data the weighted jackknife is superior to the standard jackknife.

In this chapter, we use the jackknife approach to reduce the influence of missing observations on the least squares estimator of the vector of unknown regression coefficients. A class of weighted jackknife estimators of the vector of unknown regression coefficients when the data are incomplete is proposed. Properties of these estimators are examined. Our approach reduces the squared bias at the cost of increase in variance. Upper bounds for the relative squared bias and the relative increase in variance of the vector of predicted responses are obtained. The investigator can easily obtain a weighted jackknifed estimator for the vector of unknown regression coefficients by fixing one of these bounds at any desired level. Our approach is easy to handle and no iteration is needed to employ it.

5.2 JACKKNIFING

To make this chapter self-sufficient we reconsider the general linear-model

$$Y_1 = X_1\beta + e_1 \quad \dots(5.2.1)$$

$$E(e_1) = 0, \quad E(e_1 e_1') = \sigma^2 I_n,$$

where Y_1 is an $n \times 1$ vector of observations, X_1 is an $n \times p$ design matrix with $\text{rank}(X_1) = p < n$, β is a $p \times 1$ vector of unknown regression coefficients, e_1 is an $n \times 1$ vector of random errors, I_k denotes the $k \times k$ identity matrix, σ^2 is an unknown positive constant, and 0 stand for the null vector of appropriate dimension.

Further, we assume that m additional observations are available on the dependent variable, but some of the regressor values are missing. That is, the additional observations are of the form :

$$Y_2 = X_*\beta + e_2, \quad \dots(5.2.2)$$

$$E(e_2) = 0, \quad E(e_2 e_2') = \sigma^2 I_m$$

and
$$E(e_1 e_2') = E(e_2 e_1') = 0,$$

where Y_2 is an $m \times 1$ vector of additional observations, X_* is an $m \times p$ incomplete design matrix. Some elements of X_* are missing, e_2 is an $m \times 1$ vector of random errors, and 0 is the $n \times m$ null matrix.

Suppose that an $m \times p$ non-stochastic matrix X_2 is substituted for the unknown design matrix X_* , so that

$$\begin{aligned}
Y_2 &= X_1\beta + X_2\beta - X_2\beta + e_2 \\
&= X_2\beta + \delta_2 + e_2, \quad \dots(5.2.3)
\end{aligned}$$

where $\delta_2 = (X_1 - X_2)\beta$. Combining (5.2.1) and (5.2.3) we get

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \begin{bmatrix} 0 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \quad \dots(5.2.4)$$

Let $X = [X_1' | X_2']'$ and $\text{rank}(X) = p < n$. The least squares estimator of β based on all $(n+m)$ observations is given by

$$\hat{\beta} = W^{-1}(X_1'Y_1 + X_2'Y_2), \quad \dots(5.2.5)$$

where $W = X'X = X_1'X_1 + X_2'X_2$. Now, we follow Hinkley (1977) to propose a class of weighted jackknifed estimators of β . Let Y_{2-i} and X_{2-i}' denote respectively the vector Y_2 with its i th co-ordinate deleted and the matrix X_2' with its i th column deleted and let

$$\begin{aligned}
\hat{\beta}_{-i} &= (X_1'X_1 + X_{2-i}'X_{2-i})^{-1}(X_1'Y_1 + X_{2-i}'Y_{2-i}) \\
&= (W - x_{2i}x_{2i}')^{-1}(X_1'Y_1 + X_2'Y_2 - x_{2i}y_{2i}), \quad \dots(5.2.6)
\end{aligned}$$

where W is defined in (5.2.5), x_{2i} is the i th column of X_2' and y_{2i} is the i th coordinate of Y_2 .

Next,

$$(W - x_{2i}x_{2i}')^{-1} = W^{-1} + \frac{W^{-1}x_{2i}x_{2i}'W^{-1}}{1 - x_{2i}'W^{-1}x_{2i}}$$

see , Rao (1973, p.33) ; and

$$(X_1'Y_1 + X_2'Y_2) = W\hat{\beta}.$$

The relation (5.2.6) can be written as

$$\begin{aligned}
 \hat{\beta}_{-i} &= \left[W^{-1} + \frac{W^{-1} x_{2i} x'_{2i} W^{-1}}{(1 - x'_{2i} W^{-1} x_{2i})} \right] [W\hat{\beta} - x_{2i} y_{2i}] \\
 &= \hat{\beta} - W^{-1} x_{2i} y_{2i} + \frac{W^{-1} x_{2i} (x'_{2i} \hat{\beta} - x'_{2i} W^{-1} x_{2i} y_{2i})}{(1 - x'_{2i} W^{-1} x_{2i})} \\
 &= \hat{\beta} - \frac{W^{-1} x_{2i} (y_{2i} - x'_{2i} \hat{\beta})}{(1 - x'_{2i} W^{-1} x_{2i})} \\
 &= \hat{\beta} - W^{-1} x_{2i} V_i, \quad i = 1, 2, \dots, m, \quad \dots(5.2.7)
 \end{aligned}$$

where $V_i = (y_{2i} - x'_{2i} \hat{\beta}) / (1 - x'_{2i} W^{-1} x_{2i})$. Following Miller (1974 b) we define pseudo values as

$$\begin{aligned}
 U_i &= m\hat{\beta} - (m-1) \hat{\beta}_{-i}, \\
 &= \hat{\beta} + (m-1) W^{-1} x_{2i} V_i, \quad \dots(5.2.8)
 \end{aligned}$$

$i=1, 2, \dots, m$.

From (5.2.8) a Jackknife estimator of β can be defined as

$$\bar{U} = \sum_{i=1}^m U_i / m = \hat{\beta} + [(m-1)/m] W^{-1} \sum_{i=1}^m x_{2i} V_i.$$

This is in line with the jackknife method originally developed by Quenouille (1956). Now, we compute the variance of V_i defined in (5.2.7).

Let

$$J = (0, 0, \dots, 0, 1, 0, \dots, 0)'$$

be an $(n+m) \times 1$ vector with unity in the $(n+i)$ th position and all other elements of J are zero. We observe that $J' X = x'_{2i}$ and $J' X W^{-1} x_{2i} = x'_{2i} W^{-1} x_{2i}$. Therefore, letting $Y = (Y'_1 | Y'_2)'$, we have

$$\begin{aligned}
V_i &= (y_{2i} - x'_{2i}\hat{\beta}) / (1 - x'_{2i}W^{-1}x_{2i}) \\
&= (J'Y - x'_{2i}W^{-1}X'Y) / (1 - x'_{2i}W^{-1}x_{2i}) \\
&= (J' - x'_{2i}W^{-1}X')Y / (1 - x'_{2i}W^{-1}x_{2i}).
\end{aligned}$$

Hence, the variance of V_i is given by

$$\begin{aligned}
\text{Var}(V_i) &= \frac{\sigma^2 (J' - x'_{2i}W^{-1}X')(J' - x'_{2i}W^{-1}X')'}{(1 - x'_{2i}W^{-1}x_{2i})^2} \\
&= \sigma^2 / (1 - x'_{2i}W^{-1}x_{2i}).
\end{aligned}$$

Therefore, the variance of V_i , $i=1,2,\dots,m$ is a decreasing function of $(1 - x'_{2i}W^{-1}x_{2i})$. In defining pseudo values we want to take the variance of V_i into consideration. We define the pseudo values as

$$\begin{aligned}
t_i &= \hat{\beta} - cW^{-1}x_{2i}(1 - x'_{2i}W^{-1}x_{2i})V_i \\
&= \hat{\beta} - cW^{-1}x_{2i}(y_{2i} - x'_{2i}\hat{\beta}), \\
i &= 1, 2, \dots, m,
\end{aligned}$$

where c is an arbitrary non-negative scalar constant. The non-negative constant c is introduced to de-emphasize the impact of missing regressor values. Now, we define a class of weighted jackknifed estimators of β by

$$\begin{aligned}
\tilde{\beta} &= \sum_{i=1}^m t_i/m = \hat{\beta} - (c/m)W^{-1} \sum_{i=1}^m x_{2i}(y_{2i} - x'_{2i}\hat{\beta}) \\
&= \hat{\beta} - (c/m)W^{-1}X'_2(Y_2 - X_2\hat{\beta}). \quad \dots(5.2.9)
\end{aligned}$$

we observe that if $c = 0$, then $\tilde{\beta} = \hat{\beta}$, the least squares estimator of β .

In the following section we shall study some properties of $\tilde{\beta}$.

5.3 PROPERTIES OF $\tilde{\beta}$

It follows from (5.2.4) and (5.2.5) that

$$\begin{aligned}\hat{\beta} &= W^{-1}(X_1'Y_1 + X_2'Y_2) \\ &= W^{-1}[X_1'(X_1\beta + e_1) + X_2'(X_2\beta + \delta_2 + e_2)] \\ &= \beta + W^{-1}X_1'e_1 + W^{-1}X_2'(\delta_2 + e_2),\end{aligned}$$

and

$$\begin{aligned}Y_2 - X_2\hat{\beta} &= X_2\beta + (\delta_2 + e_2) - X_2\beta - X_2W^{-1}X_1'e_1 - X_2W^{-1}X_2'(\delta_2 + e_2) \\ &= (I - X_2W^{-1}X_2')(\delta_2 + e_2) - X_2W^{-1}X_1'e_1\end{aligned}$$

Therefore, from (5.2.9) we have

$$\begin{aligned}\tilde{\beta} &= \beta + W^{-1}X_1'e_1 + W^{-1}X_2'(\delta_2 + e_2) - (c/m)W^{-1}X_2'[(I - X_2W^{-1}X_2')(\delta_2 + e_2)] \\ &\quad + (c/m)W^{-1}X_2'X_2W^{-1}X_1'e_1 \\ &= \beta + [I + (c/m)W^{-1}X_2'X_2]W^{-1}X_1'e_1 \\ &\quad + [I - (c/m)I + (c/m)W^{-1}X_2'X_2]W^{-1}X_2'(\delta_2 + e_2)\end{aligned}$$

Since $X_2'X_2 = W - X_1'X_1$, hence,

$$\begin{aligned}\tilde{\beta} &= \beta + [I + (c/m)W^{-1}(W - X_1'X_1)]W^{-1}X_1'e_1 \\ &\quad + [\{1 - (c/m)\}I + (c/m)W^{-1}(W - X_1'X_1)]W^{-1}X_2'(\delta_2 + e_2)\end{aligned}$$

$$\begin{aligned}
&= \beta + W^{-1/2}[(c/m)I + Q]W^{-1/2}X'_1e_1 \\
&\quad + W^{-1/2}QW^{-1/2}X'_2(\delta_2 + e_2), \quad \dots(5.3.1)
\end{aligned}$$

where $Q = [I - (c/m)F]$, and $F = W^{-1/2}X'_1X_1W^{-1/2}$. It follows from (5.3.1) that

$$E(\tilde{\beta} - \beta) = W^{-1/2}QW^{-1/2}X'_2\delta_2. \quad \dots(5.3.2)$$

The dispersion matrix of $\tilde{\beta}$ is given by

$$\begin{aligned}
D(\tilde{\beta}) &= \sigma^2 W^{-1/2}[(c/m)I + Q]W^{-1/2}X'_1X_1W^{-1/2}[(c/m)I + Q]W^{-1/2} \\
&\quad + \sigma^2 W^{-1/2}QW^{-1/2}X'_2X_2W^{-1/2}QW^{-1/2} \\
&= \sigma^2 \{ (c/m)^2 W^{-1}X'_1X_1W^{-1} + (c/m)W^{-1/2}QW^{-1/2}X'_1X_1W^{-1} \\
&\quad + (c/m)W^{-1}X'_1X_1W^{-1/2}QW^{-1/2} + W^{-1/2}QW^{-1/2}X'_1X_1W^{-1/2}QW^{-1/2} \} \\
&\quad + \sigma^2 W^{-1/2}QW^{-1/2}X'_2X_2W^{-1/2}QW^{-1/2}.
\end{aligned}$$

We observe that

$$\begin{aligned}
W^{-1/2}(X'_1X_1 + X'_2X_2)W^{-1/2} &= W^{-1/2}WW^{-1/2} = I, \\
W^{-1}X'_1X_1W^{-1/2}QW^{-1/2} &= W^{-1}X'_1X_1W^{-1/2}[I - c/m W^{-1/2}X'_1X_1W^{-1/2}]W^{-1/2} \\
&= W^{-1}X'_1X_1W^{-1} - (c/m)W^{-1}X'_1X_1W^{-1}X'_1X_1W^{-1}
\end{aligned}$$

which is symmetric,

and

$$\begin{aligned}
QQ &= [I - (c/m)W^{-1/2}X'_1X_1W^{-1/2}][I - (c/m)W^{-1/2}X'_1X_1W^{-1/2}] \\
&= I - 2(c/m)W^{-1/2}X'_1X_1W^{-1/2} + (c/m)^2 W^{-1/2}X'_1X_1W^{-1}X'_1X_1W^{-1/2}.
\end{aligned}$$

Hence

$$\begin{aligned}
D(\tilde{\beta}) &= \sigma^2 \{W^{-1/2} Q W^{-1/2} (X_1' X_1 + X_2' X_2) W^{-1/2} Q W^{-1/2}\} + \sigma^2 (c/m)^2 W^{-1} X_1' X_1 W^{-1} \\
&\quad + 2\sigma^2 (c/m) \{W^{-1} X_1' X_1 W^{-1} - (c/m) W^{-1} X_1' X_1 W^{-1} X_1' X_1 W^{-1}\} \\
&= \sigma^2 \{W^{-1} - 2(c/m) W^{-1} X_1' X_1 W^{-1} + (c/m)^2 W^{-1} X_1' X_1 W^{-1} X_1' X_1 W^{-1}\} \\
&\quad + \sigma^2 (c/m)^2 W^{-1} X_1' X_1 W^{-1} \\
&\quad + 2\sigma^2 (c/m) W^{-1} X_1' X_1 W^{-1} - 2\sigma^2 (c/m)^2 W^{-1} X_1' X_1 W^{-1} X_1' X_1 W^{-1} \\
&= \sigma^2 W^{-1} + \sigma^2 (c/m)^2 \{W^{-1} X_1' X_1 W^{-1} - W^{-1} X_1' X_1 W^{-1} X_1' X_1 W^{-1}\} \\
&= \sigma^2 W^{-1} + \sigma^2 (c/m)^2 W^{-1} X_1' [I - X_1 W^{-1} X_1'] X_1 W^{-1}. \quad \dots(5.3.3)
\end{aligned}$$

The bias and the dispersion matrix of $\hat{\beta}$ can be obtained by setting $c = 0$ in (5.3.2) and (5.3.3) respectively.

The bias of the vector of predicted responses, $X\tilde{\beta}$, is given by

$$E[X(\tilde{\beta} - \beta)] = XW^{-1/2}QV \quad \dots(5.3.4)$$

where $V = W^{-1} X_2' \delta_2$. Hence the squared bias of $X\tilde{\beta}$ is

$$\begin{aligned}
\{E[X(\tilde{\beta} - \beta)]\}' \{E[X(\tilde{\beta} - \beta)]\} &= V' Q W^{-1} X' X W^{-1} Q V \\
&= V' Q^2 V. \quad \dots(5.3.5)
\end{aligned}$$

The dispersion matrix of the predicted responses, $X\tilde{\beta}$, is given by

$$\begin{aligned}
D(X\tilde{\beta}) &= X D(\tilde{\beta}) X' \\
&= \sigma^2 X W^{-1} X' + \sigma^2 (c/m)^2 X W^{-1} X_1' [I - X_1 W^{-1} X_1'] X_1 W^{-1} X'. \quad \dots(5.3.6)
\end{aligned}$$

5.4 RELATIVE BOUNDS

In this section we shall obtain upper bounds for the relative squared bias and the relative variance of the vector of predicted responses. Setting $c = 0$ in (5.3.5) we obtain

$$\{E[X(\hat{\beta} - \beta)]\}' \{E[X(\hat{\beta} - \beta)]\} = V'V.$$

Therefore,

$$\begin{aligned} \sup_V \frac{\{E[X(\tilde{\beta} - \beta)]\}' \{E[X(\tilde{\beta} - \beta)]\}}{\{E[X(\hat{\beta} - \beta)]\}' \{E[X(\hat{\beta} - \beta)]\}} &= \sup_V \frac{V' Q^2 V}{V'V} \\ &\leq \lambda_{\max}(Q^2), \end{aligned} \quad \dots(5.4.1)$$

where $\lambda_{\max}(G)$ denotes the largest eigenvalue of the square matrix G . We now study the behaviour of the eigenvalues of Q and Q^2 .

Lemma (5.1)

If $X = [X'_1 | X'_2]'$ is a matrix of full rank, then all eigenvalues of $F = W^{-1}X'_1X_1W^{-1}$ lie in $(0,1)$.

Proof: We observe that $W^{-1}X'_1X_1W^{-1}$ and $W^{-1}X'_2X_2W^{-1}$ are positive semi-definite matrices. Therefore, all eigenvalues of these two matrices are non-negative. Further,

$$\begin{aligned} W^{-1}X'_1X_1W^{-1} + W^{-1}X'_2X_2W^{-1} &= W^{-1}(X'_1X_1 + X'_2X_2)W^{-1} \\ &= W^{-1}WW^{-1} \\ &= I_p. \end{aligned}$$

Hence,

$$W^{-1}X_2'X_2W^{-1} = I_p - W^{-1}X_1'X_1W^{-1}. \quad \dots(5.4.2)$$

Let λ_1 be the largest eigenvalue of $W^{-1}X_1'X_1W^{-1}$. Then it follows from (5.4.2) that $(1 - \lambda_1)$ is the smallest eigenvalue of $W^{-1}X_2'X_2W^{-1}$. Since $(1 - \lambda_1) \geq 0$, hence $\lambda_1 \leq 1$. This complete the proof.

Lemma (5.2)

If $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ are the eigenvalues of $F = W^{-1}X_1'X_1W^{-1}$, then

$$(1 - (c/m)\lambda_1)^2 \leq (1 - (c/m)\lambda_2)^2 \leq \dots \leq (1 - (c/m)\lambda_p)^2$$

are the eigenvalues of $Q^2 = [I - (c/m)F]^2$.

Proof: Let α be an eigenvalue of Q . Then α must satisfy the characteristic equation

$$|Q - \alpha I| = 0$$

$$\text{or, } |I - (c/m)F - \alpha I| = 0$$

$$\text{or, } |-(c/m)F - (\alpha - 1)I| = 0$$

$$\text{or, } |F + (m/c)(\alpha - 1)I| = 0, \text{ provided } c \neq 0,$$

$$\text{or, } |F - \lambda I| = 0,$$

where $\lambda = -(m/c)(\alpha - 1)$ is an eigenvalue of F . This gives $\alpha = 1 - (c/m)\lambda$.

Let V be the eigenvector of Q corresponding to the eigenvalue $\alpha = 1 - (c/m)\lambda$.

Therefore,

$$QV = \alpha V$$

$$\text{or } Q \cdot (QV) = \alpha(QV) = \alpha(\alpha V) = \alpha^2 V.$$

This shows that $\alpha^2 = \{1 - (c/m)\lambda\}^2$ is an eigenvalue of Q^2 . This completes the proof.

Theorem (5.3):

Let λ_p be the smallest eigenvalue of $F = W^{-1} X' X W^{-1}$.

If $0 \leq (c/m)\lambda_p \leq 2$, then

$$\sup_V \frac{[E\{X(\tilde{\beta} - \beta)\}]' [E\{X(\tilde{\beta} - \beta)\}]}{[E\{X(\hat{\beta} - \beta)\}]' [E\{X(\hat{\beta} - \beta)\}]} \leq 1$$

where $V = W^{-1} X' \delta_2$.

Proof: It follows from Lemma (5.2) and the relation (5.4.1) that

$$\sup_V \frac{[E\{X(\tilde{\beta} - \beta)\}]' [E\{X(\tilde{\beta} - \beta)\}]}{[E\{X(\hat{\beta} - \beta)\}]' [E\{X(\hat{\beta} - \beta)\}]} \leq [1 - (c/m)\lambda_p]^2.$$

If $0 \leq (c/m)\lambda_p \leq 2$, then $[1 - (c/m)\lambda_p]^2 \leq 1$. This completes the proof.

Theorem (5.4)

The matrix $\{D(\tilde{\beta}) - D(\hat{\beta})\}$ is semi-positive definite, where $D(\tilde{\beta})$ and $D(\hat{\beta})$ are the dispersion matrices of $\tilde{\beta}$ and $\hat{\beta}$ respectively.

Proof: set $c=0$ in the relation (5.3.3) to get $D(\hat{\beta}) = \sigma^2 W^{-1}$. Hence, from the relation (5.3.3) we get

$$D(\tilde{\beta}) - D(\hat{\beta}) = \sigma^2 (c/m)^2 W^{-1} X' [I - X W^{-1} X'] X W^{-1}. \quad \dots(5.4.3)$$

Obviously, the matrix $\{D(\tilde{\beta}) - D(\hat{\beta})\}$ will be semi-positive definite iff the smallest eigenvalue of the symmetric matrix $[I - X W^{-1} X']$ is non-negative.

We make use of the following result: Any non-zero eigenvalues of the matrix $X'X$ are the same as the non-zero eigenvalues of XX' . See(Rao; p.68).

Now,

$$X_1 W^{-1} X_1' = (X_1 W^{-1})(W^{-1} X_1') = (X_1 W^{-1})(X_1 W^{-1})'.$$

So the non-zero eigenvalues of $X_1 W^{-1} X_1'$ are the same as non-zero eigenvalues of $W^{-1} X_1' X_1 W^{-1}$. But, by Lemma (5.1), all eigenvalues of $W^{-1} X_1' X_1 W^{-1}$ lie in $(0,1)$, hence, the smallest eigenvalues of $[I - X_1 W^{-1} X_1']$ is non-negative. This complete the proof.

The following theorem gives an upper bound for the total variance of $X\tilde{\beta}$.

Theorem (5.5)

For any full rank design matrix X

$$\{\text{Tr}[D(X\tilde{\beta})] / \text{Tr}[D(X\hat{\beta})]\} \leq 1 + 1/4(c/m)^2.$$

Proof: It follows from (5.3.6) that

$$\begin{aligned} \text{Tr}[D(X\tilde{\beta})] &= \sigma^2 \text{Tr}(W^{-1} X' X) \\ &\quad + \sigma^2 (c/m)^2 \text{Tr}\{W^{-1} X_1' [I - X_1 W^{-1} X_1'] X_1 W^{-1} W^{-1} X' X W^{-1}\} \\ &= \sigma^2 \text{Tr}(W^{-1} X' X W^{-1}) + \sigma^2 (c/m)^2 \text{Tr}[W^{-1} X_1' X_1 W^{-1} \\ &\quad - (W^{-1} X_1' X_1 W^{-1})(W^{-1} X_1' X_1 W^{-1})] \\ &= \sigma^2 \text{Tr}(I_p) + \sigma^2 (c/m)^2 [\text{Tr}(F) - \text{Tr}(F^2)] \\ &= \sigma^2 P + \sigma^2 (c/m)^2 \left[\sum_{i=1}^p \lambda_i - \sum_{i=1}^p \lambda_i^2 \right] \\ &= P\sigma^2 + \sigma^2 (c/m)^2 \sum_{i=1}^p \lambda_i (1 - \lambda_i), \end{aligned} \quad \dots(5.4.4)$$

where λ_i are the eigenvalues of F . Since $\lambda_i (1 - \lambda_i) \leq 1/4$, therefore

$$\begin{aligned} \text{Tr}[D(X\tilde{\beta})] &\leq P\sigma^2 + \sigma^2 (c/m)^2 P/4 \\ &= P\sigma^2 [1 + (c/m)^2/4]. \end{aligned} \quad \dots(5.4.5)$$

Setting $c = 0$ in (5.4.4) we obtain

$$\text{Tr}[D(X\hat{\beta})] = P \sigma^2. \quad \dots(5.4.6)$$

From (5.4.5) and (5.4.6) we have

$$\frac{\text{Tr}[D(X\tilde{\beta})]}{\text{Tr}[D(X\hat{\beta})]} \leq [1 + (c/m)^2 / 4] .$$

This completes the proof.

Again from (5.4.5) and (5.4.6) we get

$$\frac{\text{Tr}[D(X\tilde{\beta})] - \text{Tr}[D(X\hat{\beta})]}{\text{Tr}[D(X\hat{\beta})]} \leq (1/4)(c/m)^2 . \quad \dots(5.4.7)$$

The inequality in (5.4.7) gives an upper bound for the relative increase in total variance of $X\tilde{\beta}$. We observe that the relative increase in total variance of $X\tilde{\beta}$ is small if $0 < (c/m) < 1$. Further, since all the eigenvalues of $F = W^{-1}X'X_1W^{-1}$ lie in $(0,1)$, and if $0 < (c/m) < 1$, then $(c/m)\lambda_p < 1$. This satisfies the condition stated in Theorem (5.1). For such choice of (c/m) a reduction in bias of $X\tilde{\beta}$ will be guaranteed. Moreover, the experimenter may fix the upper bound given in (5.4.7) to a desired level and obtain a non-negative value of (c/m) . This value of (c/m) can be used in (5.2.9) to define a weighted Jackknifed estimator of β . If this value of (c/m) lies in the interval $(0,1)$, then a reduction in bias of $X\tilde{\beta}$ will be obvious. In such a case increase in the total variance of $X\tilde{\beta}$ will also be insignificant.

References

- Allan, F. E and Wishart, J. (1930): "A method of estimating yield of missing plot in field experimental work." *Journal of Agricultural Science*, 20, 399-406.
- Ali, M. A. (1989): "On a class of shrinkage estimators of the vector of regression coefficients." *Commun. Statist. Theor. Meth.* 18(12), 4491-4500.
- Ali, M. A. (1994): "Identifying the location shift outliers that matter." *Sankhya, Series A*, vol. 56, Part 3, 500-511.
- Askin, R.G. and Montgomery, D.C. (1980): "Augmented robust estimators." *Technometrics*, 22, 333-341.
- Beale, E. M. L, and Little, R. J. A. (1975): "Missing values in multivariate analysis." *Journal of the Royal Statistical Society, Series B*, 37, 129-145.
- Box, G. E. P. and Draper, N. R. (1959): "A basis for the selection of a response surface design." *J. Amer. Statist. Assoc.* 54, 622-654.
- Buck, S. F. (1960): "A method of estimation of missing values in multivariate data, suitable for use with an electronic computer." *Journal of the Royal Statistical Society, Series B*, 22, 302-307.
- Dear, R. E. (1959): "A principal component missing data method for multiple regression models." SP-86, Santa Monica, California; System Development Corporation.
- Dodge, Y. (1985): "Analysis of experiments with missing data." John Wiley, New York.
- Donner, A. (1982): "The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values." *The American Statistician* 36, 378-381.

- Donner, A. and Rosner, B. (1982): "Missing value problems in multiple linear regression with two independent variables." *Communications in Statistics A, Theory and Methods* 11, 127-140.
- Graybill, F. A. (1961): "An introduction to linear statistical models." McGraw-Hill Book Company, New York.
- Haitovsky, Y. (1968): "Missing data in regression analysis." *Journal of the Royal Statistical Society, Series B*, 30, 67-82.
- Heiberger, R. M. (1977): "Regression with the pairwise present covariance matrix; a dangerous practice." Technical Report No. 19, Department of Statistics, University of Pennsylvania.
- Hinkley, D. V. (1977): "Jackknifing in unbalanced situations." *Technometrics*, 19, 285-292.
- Hoerl, A. E. and Kennard, R. W. (1970): "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics*, 12, 55-67.
- Hosmane, B. S. (1988): "On a generalized Stein estimator of regression coefficients." *Communications in Statistics, Theory and Methods*, 17(6), 1735-1740.
- James, W. and Stein, C. (1961): "Estimation with quadratic loss." *Proc. Fourth Berkely Symp. Math. Statist. Prob.* 1, 361-379.
- Kshirsagar, A. M. (1983): "A course in linear models." Marcel Dekker, New York.
- Little, R. J. A. and Rubin, D. B. (1987): "Statistical analysis with missing data." John Wiley & Sons.
- Miller, R. G. (1974a): "The jackknife: a review." *Biometrika* 61, 1-15.
- Miller, R. G. (1974b): "An unbiased jackknife." *Ann. Statist.* 2, 880-891.

- Quenouille, M. H. (1956): "Notes on bias in estimation." *Biometrika* 43, 353-360.
- Rao, C. R. (1973): "Linear statistical inference and its applications." Second Edition. John Wiley & Sons, New York.
- Rubin, D. B. (1972): "A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design." *Applied Statistics* 21, 136-141.
- Searle, S. R. (1982): "Matrix algebra useful for statistics." John Wiley & Sons, New York.
- Seber, G. A. F. (1977): "Linear regression analysis." John Wiley, New York.
- Theil, H. and Goldberger, A. S. (1961): "On pure and mixed estimation in economics." *Inter. Economic Review*, 2, 65-78.
- Theil, H. (1963): "On the use of incomplete prior information in regression analysis." *J. Amer. Statist. Assoc.* 58, 401-411.
- Timm, N. H. (1970): "The estimation of variance covariance matrices from incomplete data." *Psychometrika* 35, 417-437.
- Yancy, T. A. Judge, G. G. and Bock, M. E. (1974): "A mean square error test when stochastic restrictions are used in regression." *Communications in Statistics*, 3, 755-768.
- Yates, F. (1933): "The analysis of replicated experiments when the field results are incomplete." *The Empire Journal of Experimental Agriculture*, 1, 129-142.
- Wallace, T. D. and Toro-Vizcarrondo, C. E. (1969): "Tables for the mean square error test for exact linear restrictions in regression." *J. Amer. Statistics Assoc.* 64, 1649-1663.

Walsh, J. F. (1959): "Computer feasible general method for fitting and using regression function when data is incomplete." SP-71, Santa Monica, California; System Development Corporation.

Woolson, R. F. and Clarke, W. R. (1984): "Analysis of categorical incomplete longitudinal data." J. Royal Statist. Soc. A 147, 87-99.