IMPUTATION TECHNIQUES IN MULTIVARIATE ANALYSIS

BY

YASIR YAGOUB ALI

THIS THESIS IS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN MATHEMATICAL STATISTICS IN THE DEPARTMENT OF MATHEMATICS,

UNIVERSITY OF NAIROBI,

FEBRUARY, 1999.

DECLARATION

This thesis is my original work and has not been presented for a degree in any other University.

Signature: Y. Y. Ali.

Yasir Yagoub Ali

This thesis has been submitted for examination with my approval as University supervisor.

Signature:

Prof. J.A.M. Ottieno

Date: 10/12/99

DEPARTMENT OF MATHEMATICS,
UNIVERSITY OF NAIROBI,
P.O. BOX 30197,
NAIROBI,
KENYA.

LIST OF CONTENTS

LIST OF TABLES

ABSTRACT

This research deals with the problem of missing data in multivariate analysis, in the sense that not all variables of interest are measured on every unit or element of the sample. The emphasis of the thesis is on imputation techniques as a method of handling missing data problem in multivariate analysis. Special attention is paid to the method of Buck (1960) as a pioneering imputation method for estimating the covariance matrix of any k-variate population in the presence of missing values.

We have extended Buck's method to the case of units with more than one missing value and obtained the properties of the resulting estimators. A simplified procedure for the estimation of the bias of the variances of the observed and imputed data has also been developed. On the basis of the simplified procedure, a functional relationship between the relative bias and the coefficient of determination has been established. It has also been shown that for some patterns of missingness, Buck's method makes maximum use of the available information.

The problems caused by imputation via Buck's method in regression analysis are studied. It has been shown that the presence of the imputed values create serious biases in the obtained estimates.

For the case of the model-based strategy it has been shown that the factorization method of Anderson (1957) is equivalent to the special case of Buck's method where units have one missing value subject to one variable. We have also shown that this equivalence of the two methods does not hold for the case of units with more than one missing value. It is also shown that, under normality assumptions, the EM algorithm is equivalent to an iterated version of Buck's method.

Finally we have made an attempt to lay a foundation for extending Buck's method to handle non-randomly missing data. With that in mind, the work of Nordheim (1978, 1984) has been extended by considering the case of non-random misclassification.

Throughout the thesis numerical illustrations and validation of the obtained theoretical results are given using real data. The data are analyzed using *SPSS* and *STATGRAPHICS* statistical computer softwares.

ACKNOWLEDGEMENTS

Various institutions and individuals have contributed to the successful completion of this work. I would therefore, like to acknowledge the University of Nairobi through the Board of Postgraduate Studies (BPS) and the Department of Mathematics for offering me an admission that enabled me to pursue this study.

I would also like to thank the entire staff of the Department of Mathematics, University of Nairobi for their encouragement and enthusiasm. In particular, I wish to express my profound gratitude to Professor J.A.M. Ottieno for his guidance and moral support that made this study possible. I also wish to extend my sincere appreciation to Dr. J.O. Owino, Mr. Khogali A. Khogali, Mr. C. Achola, Professors G.P. Pokharyal and J.W. Odhiambo, Dr. C. Abungu and Ms. Irene Karimi for their help and encouragement.

A special tribute goes to Abdelwahab S.M. Sinnary who provided unlimited support, insightful comments resulting from long hours of discussion and was instrumental in helping me keep my sanity throughout the whole ordeal.

In addition, I wish to thank the German Academic Exchange Service (DAAD) for sponsoring the study together with a six-month study visit to Germany. The kind assistance and encouragement of DAAD's Nairobi Office staff cannot go without appreciation.

Thanks must also go to Professor Walter Kraemer for hosting me in the Department of Statistics at Dortmund University (Germany), providing me with the necessary facilities to finish my study and commenting on an earlier draft of the work.

Many colleagues have influenced and enriched the content of this work through conversations and other means. Of those, I would particularly like to express appreciation to Ali A. Rabah, Dr. Abdul L. Bello and Kepher H. Makambi.

Finally, I wish to thank my friends Mohammed D. Eisa, Ibrahim A. Elrayah, Hanan O. Ali and Hamid H. Ezairig for their wonderful companionship throughout my study years in Nairobi.

CHAPTER I

INTRODUCTION

1.1 AN OVERVIEW OF MISSING-DATA PROBLEM

Standard statistical methods have been developed to analyze data sets arranged in rectangular forms, often called *data matrices*. The rows of the data matrix represent individuals, cases, units or observations, depending on the context, and the columns represent variables measured for each unit. The entries in the data matrix are real numbers representing continuous or categorical measurements.

Missing values are phenomena in data that occur when measurements on some variables for some individuals (or units) are not available for whatever reason. Such data are referred to as incomplete data, fragmentary data, spoilt data, omitted data, missing plots, partial data, scarce data or missing data, depending on the statistical context. Statisticians have long appreciated that the existence of missing data can change an ordinary simple statistical analysis into a complex one.

Although the problem of missing data was discovered by Fisher and Yates in the 1920s, serious research in the area has flourished later in the early 1970s. This is mainly due to the developments in the computer technology that facilitated the previously laborious numerical computations of the subject. Since then the area has witnessed rapid advances in three areas of statistics, namely Survey Sampling, Experimental Design and Multivariate Analysis. Although our emphasis in this thesis is on the last area, for completeness, we shall give a brief review of the problems created by the existence of missing values in the three areas.

1.1.1 MISSING VALUES IN DESIGNED EXPERIMENTS

In many designed experiments, it may happen by chance or by some other reason that some of the observations are missing because they cannot be collected or they are simply not obtainable. For example, crops destroyed in some plots, a patient withdraws from the treatment, one or more animals die in the course of the experiment before measuring the treatment effect on them, etc., are common phenomena.

In such situations the designed experiment is no longer balanced and loses the orthogonality that it possesses when all observations are present. As a result the proper complete-data least squares analysis becomes complicated. The least squares normal equations that are constructed for the analysis of the data, will not have the terms corresponding to the missing observations. Therefore some of the required parameters may not be estimable. The question of interest is how to analyze the data in such situations.

One possible solution to the problem is to repeat the experiment under similar conditions and obtain the values of the missing observations. However, such a solution, though ideal, may not be feasible economically and physically. The statistical literature provides two options for dealing with this problem.

The first option starts by assuming that the reasons for the occurrence of missing values in the response variable (Y) does not depend on any Y value. Then the complete-data least squares method is applied to the complete rows of the matrix of the factors ($\mathbf{X}$) by simply ignoring the rows of $\mathbf{X}$ corresponding to missing y_i. The second option starts by imputing (filling-in) the missing values to restore the balance of the design and then proceed with the standard analysis. The first attempt to obtain imputed values for missing data was made by Allan and Wishart (1930). Three years later,

Yates (1933) devised a least squares approach that enhances the analysis of replicated experiments when field results are incomplete. The method usually requires adjusting the degrees of freedom used in obtaining residual mean square error. Specifically, the imputed values are chosen in such a way that the standard residual sum of squares is minimized with respect to missing values. To illustrate the method, suppose that m of the n intended observations are missing. Without loss of generality, we take these to be the last m components of the observation vector $\underset{\sim}{y}$, which now become unknown, $u_1, u_2, \ldots, u_m$. Thus we may write

$$\underset{\sim}{y}_t = \begin{bmatrix} \underset{\sim}{z} \\ \underset{\sim}{u} \end{bmatrix}, \tag{1.1}$$

where $\underset{\sim}{z}((n-m)\text{x}1)$ contains the actually observed values of $\underset{\sim}{y}$ and $\underset{\sim}{u}(m\text{x}1)$ the unknown observations. In effect, we are presented with a fresh set of unknowns to estimate, in addition to the parameters of the model. To estimate the value of $\underset{\sim}{u}$ we apply the LS method to minimize the sum of squared residuals given by

$$S = (\underset{\sim}{y} - \mathbf{X}\Theta)'(\underset{\sim}{y} - \mathbf{X}\Theta). \tag{1.2}$$

S must now be minimized not only for variations in Θ but also for variations of $\underset{\sim}{u}$. Partitioning $\mathbf{X}$ into $(x_z, x_u)'$ conformably with the partition $\underset{\sim}{y} = (\underset{\sim}{z}, \underset{\sim}{u})'$, we have

$$S = (\underset{\sim}{z} - \underset{\sim}{x}_z\Theta)'(\underset{\sim}{z} - \underset{\sim}{x}_z\Theta) + (\underset{\sim}{u} - \underset{\sim}{x}_u\Theta)'(\underset{\sim}{u} - \underset{\sim}{x}_u\Theta). \tag{1.3}$$

Since only the second of the two non-negative terms of (1.3) depends on $\underset{\sim}{u}$, we reduce it to zero by putting

$$\underline{u} = \underline{x}_u \Theta; \tag{1.4}$$

thus S at (1.3) is reduced to its first term, which may then be minimized w.r.t. Θ.

From (1.2) an estimate $\hat{\Theta}(u)$ of Θ is obtained as

$$\hat{\Theta}(\underline{u}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\underline{y}) = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'_z\underline{z} + \mathbf{X}'_u\underline{u}), \tag{1.5}$$

and using this in conjunction with (1.4), we have

$$\underline{u} = \underline{x}_u \hat{\Theta}(\underline{u}). \tag{1.6}$$

Expression (1.6) states that each missing observation is to be equated to its estimated expectation in the original LS analysis. However, the degrees of freedom for the residual sum of squares must obviously be reduced, since we now have only $(n - m)$ observations.

The literature provides ample discussion of the various strategies for dealing with the problem in designed experiments, e.g., Bartlett (1937), Tocher (1952), Dodge (1985) and Alvo and Cabilio, (1991, 1995).

1.1.2 MISSING VALUES IN SAMPLE SURVEYS

Nonresponse in Survey samples is a phenomena that occurs when a respondent, for some reasons, refuses to answer some or all the questions of a survey. We can distinguish three categories of nonresponse. These include noncoverage, complete nonresponse, and item nonresponse. Noncoverage refers to the failure of the experimenter to include some units of a survey population in the sampling frame. Complete nonresponse is the case where a unit in the frame refuses to participate in the survey. Item nonresponse occurs when a respondent does not give answers to some of the survey questions. Details of the types of nonresponse and their remedies are discussed

by Kish (1965) and Cochran (1977). Of these three categories of nonresponse only item nonresponse will provide usable data for analysis.

To illustrate the problem created by the presence of nonresponse in sample surveys, we start by considering the complete-data randomization inference where

i- Units be selected by probability sampling where the sampling distribution is determined before the actual sample selection, and

ii- Each unit has a probability (strictly greater than zero) of being selected in the sample.

Let $\mathbf{X} = (x_{ij})$, $i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, k$ where k variables are measured on the i-th unit. Suppose inferences are required for the population of N units. For the i-th unit, define the sample indicator function

$$I_i = \begin{cases} 1, & \text{if unit 'i' is included in the sample;} \\ 0, & \text{otherwise,} \end{cases} \tag{1.7}$$

and

$$\underline{I} = (I_1, \ldots, I_N)', \tag{1.8}$$

then the sampling distribution for a SRS of size n can be defined, before any x value is selected, by

$$f(I) = \begin{cases} \binom{N}{n}^{-1}, & \text{if } \sum_{i=1}^{N} I_i = n; \\ 0, & \text{otherwise.} \end{cases} \tag{1.9}$$

where $\binom{N}{n}$ is the number of ways n units can be chosen from the population, and,

$$\pi_i = \Pr(I_i = 1) > 0, \text{ for all i.} \tag{1.10}$$

Then the objective of randomization inference is to estimate population parameters, such as the population mean $\bar{X}$, by sample functions such as the sample mean $\bar{x}$. This is based on the distribution of the sample quantities in repeated sampling from the distribution of I, f(I).

The fundamental property of the randomization approach, a known probability distribution governing which values are observed and which are not, is lost when some of the selected units refuse to respond. Suppose that n units are selected by SRS and let

$$R_i = \begin{cases} 1, & \text{if } x_i \text{ is selected in the sample and responds;} \\ 0, & \text{otherwise,} \end{cases} \tag{1.11}$$

and

$$\underline{R} = (R_1, \ldots, R_N)'. \tag{1.12}$$

Then the values of X are observed iff $R_i = I_i = 1$. Thus, in the presence of nonresponse, it turns out to be impossible to define a statistic that is a function of the recorded values and is unbiased estimate of the population quantity with respect to the distribution of I. It follows that the complete-data randomization approach, outlined above is no longer valid in the presence of nonresponse.

The literature provides three main methods of solutions. The first is the quasi-randomization approach where a distribution for R is assumed. This is a direct extension of the randomization inference discussed above. The second method of solution is the model-based ML approach to survey nonresponse. Here, some modeling assumptions are made about the nonresponding portion of the population (e.g., the means of X in the responding and nonresponding units are equal). Details of this approach are given in

Little and Rubin (1987). For a thorough discussion of the survey nonresponse from both randomization and modeling perspectives, see Madow *et al* (1983), volume II. The third strategy is the multiple imputation technique which will be discussed in chapter II. A comprehensive discussion of the latter strategy is given by Rubin (1987). Other references include Efron (1994), Fay (1996) Shao and Sitter (1996) and Cook (1997).

1.1.3 MISSING VALUES IN MULTIVARIATE ANALYSIS

Many multivariate analysis techniques assume that one starts with an array of numbers x_{ij} representing the values of the j-th variable in the i-th observation. This will be for $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, N$ if we have N observations and k variables. From these raw data one then forms a square matrix (a_{ij}) of sum of squares and products defined by

$$a_{ij} = \sum_r x_{ir} x_{rj} \tag{1.13}$$

or more usually, by

$$a_{ij} = \sum_r (x_{ir} - \bar{x}_r)(x_{rj} - \bar{x}_j), \tag{1.14}$$

where

$$\bar{x}_j = \sum_i x_{ij}/N. \tag{1.15}$$

One then can proceed to a multiple regression analysis or any of the more specialized analyses such as principal component analysis, discriminant analysis, factor analysis, or interdependence analysis.

Problems arise when there are missing entries in the original data matrix; that is, if individual variables are missing in some observations. In particular,

the existence of missing values in multivariate contexts, where more than one variable have missing values, creates more difficulties. The degree of difficulty depends on many factors that include the proportion of missing values, their pattern (distribution) among variables and the independence of the missing values of one variable from those of other variables including the variable itself.

Since most of the multivariate statistical analyses are based on an initial reduction of the data to the sample mean vector and sample covariance matrix of the variables, the question of how to estimate these quantities from incomplete multivariate data is therefore an important one. The available literature suggests various strategies for dealing with this problem. One such strategy is to drop from analysis all units with missing values and to base the analysis on the completely recorded units. Obviously, this strategy is only valid under the assumption that the missing values are missing at random. An alternative approach is to impute (fill-in) the missing values and then proceed with the analysis using both the observed and imputed data. A third approach is to obtain the estimates of the population parameters from the incomplete data by maximizing the likelihood function under certain modeling assumptions. In chapter II we shall discuss, in some detail, the various strategies for handling missing values in multivariate data analysis.

We conclude this overview by contrasting the problem of missing data in multivariate analysis and its counterparts in sample surveys and designed experiments. The following are some of the major differences:

1- Since in designed experiments the levels of the factors in an experiment are fixed by the experimenter, the missing values, if they occur, are more likely to be in the response variable Y, than in the factors, X. Thus they are mostly univariate in nature. Moreover, the estimation of

missing values is only needed for the restoration of the balance of the design. No extra information is contributed by the imputed values.

2- In sample surveys, though the missing data is multivariate in nature (occurring in more than one variable), yet the method of data collection plays an important role in guiding the analysis. For example in the randomization inference the population values are treated as fixed quantities and inferences are based on the distribution that determines the sample selection.

3- In multivariate analysis the MLE of the parameters from incomplete data (discussed in chapter II) are obtained by specifying a probability distribution for the incomplete data and the missingness mechanism. The method of sample selection enters the analysis only indirectly through its influence on the choice of the distribution.

4- In sample surveys the population of interest is explicitly finite, thus the estimates are often finite population quantities, e.g., population means or totals. In multivariate analysis it is the estimation of population parameters which is of primary interest. Often inferences about parameters differ from inferences about finite-population quantities by finite-population correction factors.

As mentioned earlier, in this thesis we shall be concerned with the problem of missing values in multivariate data analysis. Therefore, in the next section we give a brief literature review of the subject.

1.2 BRIEF LITERATURE REVIEW

While the widespread occurrence of incomplete data may be regarded as a "necessary evil" realistically associated with data collection, statisticians have responded to this challenge by developing methods which are suitable for the statistical analysis with missing data. In a broad sense, the studies

mentioned in this literature review are grouped according to their underlying methods of analysis; that is as to whether they are using deletion, imputation or maximum likelihood techniques. Further classifications within each group are based on the nature of the missingness mechanism.

Afifi and Elashoff (1966), Kim and Curry (1977) and Bello (1992) are some of the studies that give surveys of the literature on multivariate statistical analysis with missing data.

Afifi and Elashoff (1966) provide a survey of methods dealing with missing observations in a regression context. The focus is on the method of least squares and the method of maximum likelihood. The underlying model is

$$y = v + \mathbf{X}\beta + e,$$

where y, v, and e are n-dimensional column vectors, $\mathbf{X}$ is an nxp design matrix, and β is a p-dimensional column vector.

The vector v has identical elements each equal to $\mu_y - \mu_1\beta_1 - \ldots - \mu_p\beta_p$. The random variables $e_1, \ldots, e_n$ have means 0, common variances σ^2, and are mutually independent of X. Further, it is assumed that all the parameters are estimable.

A variant of least squares in which imputed values for the missing values are used is then discussed.

In a series of subsequent papers Afifi and Elashoff (1967, 1969a, 1969b) compared the efficiency and sampling properties of various simple estimators in simple linear regression with missing values. Toutenburg *et al* (1995) dealt with the problem of missing values in regression analysis with nonstochastic regressor matrix. The mixed regression framework was the central method of the paper. Recently, Rao and Toutenburg (1995, chapter 8) gave a good

coverage of methods that deal with the problem of missing data in regression contexts.

The use of the estimating equations method in regression analysis with missing values has recently been considered by many scholars, e.g., Robins and Rotnitzky (1995), Robins *et al* (1995) and Zhao *et al* (1996).

Liu (1996) considered the Bayesian estimation of multivariate linear regression with missing data using the multivariate t-distribution. A monotone data augmentation algorithm for posterior simulation of the parameters and missing data imputation was presented. He considered the case of fully observed predictor variables and possibly missing values from outcome variables. Other studies in this area include Garrett (1996) and Jones (1996).

Kim and Curry (1977) reviewed the literature up to 1977. They also introduced a method for testing whether the missing observations are missing at random or not.

Buck (1960) devised a method for imputing missing values in multivariate data. The method is based on regression techniques. An interesting property of the method is that it is a combination of deletion and imputation strategies.

Dear (1959) developed a method for imputing multivariate missing values using principal component analysis.

Haitovsky (1968) compared the efficiency of complete-case analysis and available-case analysis. Using Monte Carlo simulations, various patterns of missing values were artificially created. The author then compared the biases of the two procedures in estimating regression coefficients. His conclusion is that the complete-case analysis is relatively better than available-case analysis.

Jackson (1968) and Chan and Dunn (1972) dealt with the problem of missing values in discriminant analysis. Jackson (1968) seems to be the first author to signal out the inappropriateness of the missing at random (MAR) assumption in her studies, and hence the need for developing more realistic assumptions for the missingness mechanism. Chan and Dunn (1972), in a simulation study, compared the performance of various imputation techniques and deletion strategies in linear discriminant analysis. The objective was to determine which method gives minimum probabilities of misclassification.

Bello (1992), in a simulation study, compared the performance of five single (deterministic) imputation techniques in regression and discriminant analyses. One of his conclusions is that none of the considered methods is the best overall in all circumstances.

Rubin (1978) developed the multiple (stochastic) imputation technique which is a powerful method specially in sample surveys. Application of multiple imputation techniques to incomplete multivariate normal data has been considered by Rubin and Schafer (1990). Brownstone (1991) considered the application of multiple imputations in linear regression models. A comprehensive study of multiple imputation techniques is given by Rubin (1987).

A common property of imputation techniques is that they impute the missing values first and then estimate the population parameters from the completed data. They also assume that the missing observations are missing at random (MAR). However, Greenlees *et al* (1982) use the stochastic censoring approach to impute non-randomly missing values. Troxel *et al* (1997), considered a method of weighted estimating equations for dealing with non-random missingness in regression analysis. The weights were taken to be equal to the inverse probability of being observed. Comparisons between

the weighted method and the unweighted (complete-case) analysis were performed. The weighted method was found to give asymptotically unbiased estimates of the regression coefficients when the missingness probabilities depend only on the covariates.

Estimation of population parameters from incomplete data under the multivariate normality assumptions has been considered by many authors. Wilks (1932), and Rao (1952, pp. 161–165) have considered the estimation of the parameters of the bivariate normal distribution. Matthai (1951) considered the general multivariate normal population but gave results for the trivariate case. The obtained solutions in the above studies were not explicit. Edgett (1956) considered the trivariate normal population where only one variable is subject to missingness. However, he arrived at an explicit form of solution for the maximum likelihood equations. Anderson (1957) considered the same problem and showed that the methods of Lord (1955) and Edgett (1956) can be obtained using a factorization approach. Other studies in this area are given by Hocking and Smith (1968), Hartley and Hocking (1971), Orchard and Woodbury (1972) and Beale and Little (1975).

Dempster, Laird and Rubin (1977) formalized the ideas of Orchard and Woodbury (1972) and Beale and Little (1975) in what is now known as the EM algorithm.

In all the above mentioned ML-based studies, the assumption about the process causing missing data seems to be that each value in the data set is equally likely to be missing. However, Nordheim (1978, 1984) dealt with the problem of estimating the population proportion from categorical data with non-random missingness.

The statistical literature also discusses missing data that arise intentionally (not by accident, but by design), e.g., Trawinski and Bargmann (1964)

and Hocking and Oxspring (1971).

Unlike the imputation methods, we note that the maximum likelihood approach is concerned with the estimation of the parameters and not the missing observation itself. It is worth noting that Little and Rubin (1983) have suggested a method for the joint estimation of the parameters and missing observations by maximizing the complete-data likelihood. However, the method has no or very little practical significance due to so many limitations. Details of the method and these limitations are also given in Little and Rubin (1987, chapter 5).

1.3 MECHANISMS THAT LEAD TO MISSING DATA

A common practice in data analysis involves making structural and distributional assumptions about the data at hand at the outset of statistical analysis. When the data matrix is incomplete, further assumptions will be required: one for the missing values mechanism and the other for the distribution of incomplete data.

Basically, there are two possible mechanisms that can lead to missing values, namely, missing at random (MAR) or missing completely at random (MCAR).

Let $\mathbf{Y} = (\mathrm{Y_{obs}}, \mathrm{Y_{mis}})$ where $\mathrm{Y_{obs}}$ denotes the observed values and $\mathrm{Y_{mis}}$ denotes the missing values. If $\mathrm{f}(\mathbf{Y} \mid \Theta) = \mathrm{f}(\mathrm{Y_{obs}}, \mathrm{Y_{mis}} \mid \Theta)$ denotes the joint probability density function of Y_{obs} and Y_{mis}, then the marginal density of $\mathrm{Y_{obs}}$ is given by

$$\mathrm{f}(\mathrm{Y_{obs}} \mid \Theta) = \int_{\mathrm{Y_{mis}}} \mathrm{f}(\mathrm{Y_{obs}}, \mathrm{Y_{mis}}) \mathrm{dY_{mis}}, \tag{1.16}$$

and the likelihood function of Θ based on Y_{obs} is

$$L(Y_{obs}, \Theta) \propto f(Y_{obs}, \Theta). \tag{1.17}$$

More generally, we can include in the model the distribution of a variable indicating whether each component of Y is observed or missing. For example, suppose $\mathbf{Y} = (y_{ij})$, an nxk matrix of n observations measured for k variables, define the indicator $\mathbf{R} = (R_{ij})$ such that

$$R_{ij} = \begin{cases} 1, & \text{if } y_{ij} \text{ is observed;} \\ 0, & \text{if } y_{ij} \text{ is missing.} \end{cases} \tag{1.18}$$

The model treats $\mathbf{R}$ as a random variable and specifies the joint distribution of $\mathbf{R}$ and $\mathbf{Y}$ as:

$$f(\mathbf{Y}, \mathbf{R} \mid \Theta, \Psi) = f(\mathbf{Y} \mid \Theta).f(\mathbf{R} \mid \mathbf{Y}, \Psi), \tag{1.19}$$

where, $f(\mathbf{R} \mid Y, \Psi)$ denotes the distribution for the missing-data mechanism. The distribution of the actual data $(Y_{obs}, \mathbf{R})$ is obtained as follows:

$$f(Y_{obs}, \mathbf{R}, \Theta, \Psi) = \int_{Y_{mis}} f(Y_{obs}, Y_{mis}, \Theta).f(\mathbf{R} \mid Y_{obs}, Y_{mis}, \Psi) dY_{mis}, \tag{1.20}$$

and the likelihood of Θ and Ψ is

$$L(Y_{obs}, \mathbf{R}, \Theta, \Psi) \propto f(Y_{obs}, \mathbf{R}, \Theta, \Psi). \tag{1.21}$$

Then Rubin (1976) defines the missing data to be missing at random (MAR) if,

$$f(\mathbf{R} \mid Y_{obs}, Y_{mis}, \Psi) = f(\mathbf{R} \mid Y_{obs}, \Psi), \tag{1.22}$$

and missing completely at random (MCAR) if,

$$f(\mathbf{R} \mid Y_{obs}, Y_{mis}, \Psi) = f(\mathbf{R}, \Psi). \tag{1.23}$$

A direct consequence of the MAR assumption is that the inference for Θ can be based on (1.17) rather than (1.21), i.e., the missing data mechanism is ignorable in that the resulting likelihoods given by (1.17) and (1.21) are proportional.

In statistical analysis with missing data, the definition of the missingness mechanism by the simple dichotomy (random versus non-random) is not sufficient. Knowledge of the nature of the random missingness (MAR or MCAR) is a key element in choosing an appropriate analysis (Rubin, 1976; Little and Rubin, 1987; and Bello, 1992). For example the deletion-pairwise strategy requires the *strong* MCAR assumption while the model-based ML approach is valid under the *weak* MAR assumption. This point will be discussed further in chapter II.

Unfortunately, in practice, the question of missing data mechanism is rarely answered carefully, in such cases an assumption is being made that the mechanism is ignorable, i.e., MAR or MCAR. Such an assumption, if proved incorrect, may have serious implications for the type of method adopted for analysis and subsequently on the final conclusions.

We should note that the methods of analyzing incomplete data are heavily dependent on the assumptions of MAR and MCAR for the missingness mechanism. Questions that arise are:

i- What models of missingness and methods of analysis are to be adopted for the analysis of non-randomly missing data?

ii- A more complicated case may arise if the missingness mechanism is not

the same for all incomplete data, i.e., a mixture of MAR and/or MCAR, and non-random missingness. The literature seems to be rare in such aspects.

1.4 TESTS FOR MISSING VALUES MECHANISMS

There are many approaches for testing the missing values assumptions MAR and MCAR. Commonly used methods are:

1- **Examination of missingness pattern:**

By the pattern of missing data we mean the frequency distribution of different categories of missingness such as missing only on X_i, missing on both X_i and X_j, and so on. For example, in an income survey, it could happen that high income earners exhibit a similar pattern of missing variables and low income earners another pattern of missing variables. The systematic difference among income earners, as may be revealed by the pattern of missing values, may be a pointer to a non-missing-at-random mechanism.

2- **Kim and Curry's (1977) approach:**

This method is suggested for testing for MAR assumption. The idea is to consider the (k + 2) patterns of missing data where k is the number of variables with a substantial number of missing values. The patterns to consider are:

i- β_i^o, $(i = 1, 2, \ldots, \mathrm{k})$ which denotes the number of cases for which variable i is missing,

ii- β_{k+1}^o which denotes the number of cases with missing values on two or more variables, and

iii- β_{k+2}^o which denotes the number of cases with complete variables.

Then the expected frequencies of the (k+2) patterns of missing values, under the assumption of MAR, are defined as follows

$$\beta_i^e = n(q_i \prod_{i \neq j=1}^{k} p_j), \tag{1.24}$$

$$\beta_{k+1}^e = n(1 - \prod_{j=1}^{k} p_j) - \sum_{j=1}^{k} \beta_j^e, \tag{1.25}$$

$$\beta_{k+2}^e = n(\prod_{j=1}^{k} p_j), \tag{1.26}$$

where q_i and p_i denote the proportion of missing and non-missing cases on variable i respectively, and n is the sample size of the data matrix. Significant difference between the observed β_i^o and the expected β_i^e, $(i = 1, 2, \ldots, k+2)$ frequencies, evaluated by the ordinary $\chi^2(k+1)$, will indicate non-randomness.

3- **Frane's (1978) approach:**

This is the first approach suggested for testing MCAR assumption. The method consists of the following steps:

i- Pick the variable that is closest to being missing for half of the cases.

ii- Divide all cases into two groups on the basis of whether the chosen variable is observed or missing for the cases.

iii- Then perform t-tests on the two groups formed in (ii) above.

4- **Ratio of determinants method:**

By MCAR assumption, there is no difference between cases with complete variables and cases with partially observed variables. It therefore follows that the quantity

$$\lambda = \frac{|\mathbf{S}_c|}{|\mathbf{S}_p|}, \tag{1.27}$$

must be equal to or near unity, for MAR or MCAR assumption to hold. Here, $\mathbf{S}_p$ denotes the sample covariance matrix of the incomplete data computed by all-available data method, and, $\mathbf{S}_c$ denotes the same matrix computed by the case-wise-deletion method. If λ is sufficiently greater than unity, MCAR assumption is not appropriate.

5- **The approach of Little (1988):**

Let

$J \equiv$ the number of unique patterns of missing-values,

$m_j \equiv$ the number of cases with missing-values pattern j, $(j = 1, 2, \ldots, J)$,

and,

$X_{obs,i} \equiv$ the vector of values of observed variables in case i.

It follows that,

$$\overline{X}_{obs,j} = m_j^{-1} \sum_{i=1}^{m_j} X_{obs,i}, \tag{1.28}$$

is the mean vector of observed variables for pattern j, then Little (1988) proposes the statistic

$$d^2 = \sum_{j=1}^{J} m_j (\overline{X}_{obs,j} - \hat{\mu}_{obs,j}) \tilde{\Sigma}_{obs,j}^{-1} (\overline{X}_{obs,j} - \hat{\mu}_{obs,j})', \tag{1.29}$$

where

$$\tilde{\Sigma}_{obs,j} = \left(\frac{m_j}{m_j - 1} \right) \hat{\Sigma}_{obs,j}, \tag{1.30}$$

and $\hat{\mu}_{obs,j}$ and $\hat{\Sigma}_{obs,j}$ are respectively the mean vector and covariance matrix of the observed variables in pattern j.

Asymptotically (as the sample size becomes very large), d^2 follows χ^2 distribution with $\sum_j^J (p_j - p)$ degrees of freedom with p_j being the number of observed variables for cases in the j-th missing-values pattern and p is the total number of variables. MCAR assumption is rejected for large values of d^2. Evidently, d^2 tests for systematic differences among all patterns of missing values.

Unfortunately, no single approach is completely adequate on its own. Perhaps a combination of two or more approaches may prove useful in enhancing the credibility of a final decision about the missingness mechanism. Rubin (1978) noted that the success of any of the above methods in revealing the randomness nature of missing data depends on the experience with, and historical knowledge of, the data. After all, the reason for missing data within a defined experiment may vary from one individual to another, as is often found in surveys where respondents are presumed to have answered questionnaires independently. Since no one can possibly ferret out the reason why a respondent has refused to supply information to certain variables, the tests for missing-data mechanisms discussed above should be seen as a complementary means of getting an overall picture of randomness and not as an absolute criterion.

1.5 OBJECTIVES OF THE STUDY

This thesis is concerned with the study of imputation techniques in multivariate analysis to achieve the following objectives.

1- To extend the method of Buck (1960) to the case of units with more than one missing value.

2- To prove that the conditioning of Buck's method calculations upon the complete vector observations is valid under the missing completely at

random (MCAR) assumption for the missing observations.

3- To set connections between imputation techniques and maximum likelihood methods of estimation from incomplete data.

4- To develop some models that take account of the non-random nature of the missingness mechanism.

1.6 SUMMARY OF WORK DONE IN THIS THESIS

The focus of this thesis is to re-examine the method of Buck (1960) for estimating the covariance matrix of any k-variate population in the presence of missing values. A fundamental property of the method is that it combines two different strategies for handling missing data, that is, deletion strategy and imputation techniques. It is therefore a good representative of the two strategies.

To achieve the above mentioned objectives of the study, our approach is based on both theoretical investigations and numerical validation. The theoretical investigations concentrate on the study of the statistical properties of the post-imputation covariance matrix. These include the biasedness and the statistical consistency. Multivariate regression and multivariate analysis of variance (MANOVA) are the theoretical tools used throughout the thesis. As for the numerical validations we have considered some data collected by Bumpus (1898). Considerations that led to the adoption of Bumpus data for numerical illustrations, rather than other recent data sets, include the fact that it has a reasonable number of variables and cases. The more recent data that exist tend to have relatively small number of variables and cases. For this then Bumpus data was deemed to be more suitable for considering the various patterns of missingness required for numerical illustrations. The data consists of 49 samples of birds on which eight morphological measurements on each bird were taken. The data we have used are for five variables as shown in Manly (1986).

Throughout the thesis numerical illustrations and validation of the obtained theoretical results are given as subsections in their respective chapters. These are obtained by considering seven patterns of missingness from the

data of Bumpus (1898). The specific computations required for imputation of the missing values of each pattern, Bumpus (1898) data, and the data of the seven patterns of missingness are given in the appendix. The data are analyzed using *SPSS* and *STATGRAPHICS* statistical computer softwares.

We have started our study by giving an overview of the problem of missing data in three important areas of statistics. These are: Sample Surveys, Designs of Experiments and Multivariate Analysis. Contrasts and comparisons between the problems created by the presence of missing data in these three areas have been considered.

We have then discussed various strategies for handling the problem of missing data in multivariate analysis. These strategies are; the deletion-pairwise strategy, imputation strategy and model-based strategy.

In the imputation strategy we have re-examined the method of Buck (1960) in some detail to allow for alternatives, modifications and extensions. The results are as follows:

For some patterns of missingness, Buck's method makes maximum use of the available information.

A simplified procedure for the estimation of the bias of the variances of the observed and imputed data has been developed. Unlike Buck's procedure, the simplified procedure does not require the computation of the inverse of the covariance matrix. Apart from its relative ease of computation, the developed procedure has the advantage of giving a functional relationship between the relative bias and the coefficient of determination. The statistical consistency of the estimate is also highlighted.

We have also investigated the biasedness in the covariance matrix of the observed and imputed data in the case of units with more than one missing value. The conclusion arrived at is that both the variances and covariances are biased. The bias of the covariances is found to be a function of the sample covariance of the residuals of the multivariate regression. The bias of the variances is shown to be inversely proportional to the coefficient of determination. These results are shown to be a general case of which Buck's

(1960) results for the case of units with one missing value are a special case. The statistical consistency of the estimates is also discussed.

The problems caused by imputation via Buck's method in estimating the coefficient of determination, the t-statistic for testing the significance of the regression coefficients and the standard errors of the regression coefficients are studied. The conclusion arrived at is that the imputed values create serious biases in these estimates. It is therefore recommended that the presence of imputed values as well as the method of imputation must be clearly documented.

For the case of the model-based strategy we have studied, in detail, the generalizations of Anderson's (1957) factorization method to the multivariate normal distribution with one variable subject to missingness. We have also studied the generalization of Anderson's method to the case of units with more than one missing value. These generalizations are later used to obtain some equivalence relations between imputation techniques and ML methods of estimation from incomplete data.

In order to set connections between the various strategies for handling the problem of missing data, we have shown that the factorization method of Anderson (1957) is equivalent to the special case of Buck's method where units have one missing value subject to one variable. A necessary condition for this equivalence is the normality assumption. We have also shown that this equivalence of the two methods does not hold for the case of units with more than one missing value. It is also shown that, under normality assumptions, the EM algorithm is equivalent to an iterated version of Buck's method.

Finally we have made an attempt to lay a foundation for extending Buck's method to handle non-randomly missing data. With that in mind, we have reviewed Nordheim's work of 1978 and 1984 on the estimation of the population proportions from categorical data with non-random missingness. We have then extended this work by considering the case of non-random misclassification.

CHAPTER II

STRATEGIES FOR HANDLING MISSING VALUES IN MULTIVARIATE DATA ANALYSIS

2.1 INTRODUCTION

Unfortunately, despite great efforts by statisticians to solve the problem of missing data, there is yet no universal solution. This is mainly due to the fact that the choice of appropriate missing values strategy and its workability depends to a great extent on the nature of the data missing and the statistical analysis required. The most crucial question is: when and under which conditions can one safely consider a missing data problem to be trivial?. Obviously, the smaller the proportion of missing values; the larger the sample size; and the more random the missing values, the less complicated the missing data problems.

The choice of one missing-value strategy rather than another depends on many factors that include: the number of variables under consideration, the number of cases (units) with missing values, the interdependence between variables, the nature of the missingness mechanism, and whether the required statistical analysis involves recourse to the data after extracting summary statistics from them or not.

In this chapter we shall review more technically the procedures for handling missing values in multivariate analysis. For each strategy we highlight the conditions and assumptions that validate its use. More importantly, we shall emphasize the statistical properties of the resulting estimates. This is particularly important since the wide spread of statistical computer packages have led to the blind adoption of missing-values strategies without knowing their theoretical background.

We present the strategies for handling missing data problem in multivariate analysis under three main headings: deletion-pairwise strategy, imputation strategy, and model-based ML approach.

2.2 DELETION-PAIRWISE STRATEGY

The basic idea of this approach is to estimate the population parameters (not the missing values) from the available information ignoring the missing values. Rubin (1976) showed that the application of this approach is only valid under the restrictive *strong* MCAR assumption for the missingness mechanism. Generally, if the sample size is large, as is often the case in sample surveys and the proportion of missing values is relatively small, probably the first options to consider for handling missing data is the deletion-pairwise strategy (also referred to as historical strategy). These are: case-wise-deletion method, variable-wise-deletion method, and all-available-data method.

2.2.1 CASE-WISE-DELETION METHOD

Consider the data matrix $\mathbf{X}=(x_{ij})$, where x_{ij} is the value of the j-th variable for observation i, $i = 1,2,\ldots,n$; $j = 1,2,\ldots,k$. Rearrange $\mathbf{X}$ in such a way that the first m rows of $\mathbf{X}$ represent the cases for which all variables are recorded (complete cases) and the next $(n-m)$ rows represent the cases with missing values, i.e.,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \ldots & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & \ldots & x_{mk} \\ \ldots & ? & \ldots & ? & \ldots \\ \ldots & \ldots & ? & \ldots & \ldots \\ x_{n1} & ? & ? & \ldots & x_{nk} \end{bmatrix} \tag{2.1}$$

where, ? denotes a missing value. Then, analyses using the case-wise-deletion

method (also called list-wise or complete-case analysis) confine the calculation of summary statistics to cases where all variables are present, ignoring all cases with one or more missing variables.

Clearly this method allows for the use of the complete data statistical analyses without modification. Also it allows for comparability of univariate statistics, since these are obtained on the same sample base. The disadvantage of this method is that it sacrifices information , since each unit with at least one missing value is discarded from the analysis. To illustrate the point, assume that k=10 variables and that each variable in each case is observed or missing independently according to a Bernoulli process with probability of missingness equals to 0.1. Then the probability that case i is complete is $.9^{10} = 0.35$. That is, 65% of the cases will be deleted and only 35% of the observed data will be retained for the analysis. Similarly, for k=20, only about 12% of the data will be retained. Thus the consequent reduction in sample size may be serious, particularly if k is large. Moreover, we note that this method may break down altogether, when, for example, every case has at least one missing variable in it.

A crucial concern is whether or not the selection of complete cases leads to biases in sample estimates. Under the MCAR assumption, the complete cases are effectively a random sub-sample of the original cases, and thus discarding data in the incomplete cases does not bias estimates. However, in many real life situations the incomplete portion of the data differs systematically from the complete portion. For instance, in medical follow-up studies those individuals who missed the follow-up are often different from those who attended. In this case complete-case analysis is seriously biased and invalid since the MCAR assumption is violated. Generally, the nature of bias depends on the missingness mechanism that leads to the specific selection of

the complete-cases and the type of the required analysis.

To illustrate this point, let, k=2 denote the number of variables where Y_1=age and Y_2=income are two variables measured in a survey. Suppose that either Y_1 or Y_2 may be missing such that

$$\Pr(y_{i1} \text{ missing}, y_{i2} \text{ present} \mid y_{i1}, y_{i2}) = \phi_{01}(y_{i2}) \tag{2.2}$$

and

$$\Pr(y_{i1} \text{ present}, y_{i2} \text{ missing} \mid y_{i1}, y_{i2}) = \phi_{10}(y_{i2}) \tag{2.3}$$

hence,

$$\Pr(y_{i1} \text{ present}, y_{i2} \text{ present} \mid y_{i1}, y_{i2}) = 1 - \phi_{01}(y_{i2}) - \phi_{10}(y_{i2}), \tag{2.4}$$

where, ϕ_{01} and ϕ_{10} are functions of y_{i2} but not y_{i1}. Hence, according to (1.22), Y_1 is missing at random (MAR) but Y_2 is not.

Further, assume that the forms of ϕ_{01} and ϕ_{10} are such that high– and low–income cases are more likely to be incomplete than those for middle–income individuals. Then marginal distributions of income and age based on a casewise-deletion method are severely biased towards the middle-income earners. Also this knowledge of missingness process indicates that the estimation of the linear regression of Y_2 on Y_1 based on the complete cases is subject to bias since Y_2 is non-randomly missing. However, the linear regression of Y_1 on Y_2, is not subject to selection bias, since the selection is a function of the independent variable Y_2 and not the dependent variable Y_1. Hence, analyses based on the deletion-pairwise strategy are quite sensitive to the MCAR assumption for the missingness mechanism. A strategy for adjusting for the bias in the selection of the complete cases is to assign them

case weights for use in subsequent analyses. Details of this is given by Little and Rubin (1987, chapter 4).

2.2.2 VARIABLE-WISE-DELETION METHOD

Here, all variables with missing values in (2.1) above are discarded from the data. This method may be reasonable to adopt if missing values are confined to a few number of variables. The application of this method seems to be very rare. Its most popular application is found in Brothwell and Krzanowski (1974). This method of analysis can be very useful for some specific statistical analyses. For example, in regression analysis if a fully observed dependent variable is highly correlated with a partially observed variable, this method can help in eliminating the collinearity problem. However, in analyses that require a fixed number of variables measured on all cases in the population, the use of this method is not possible. For example, in discriminant analysis, the construction of Fisher's linear discriminant function using variable-wise-deletion method is impossible. This is because the method might retain q_i variables in the i-th population and q_j in the j-th population, hence a discriminant function will be impossible.

2.2.3 ALL-AVAILABLE-DATA METHOD

The essential idea of this method is as follows: The covariance (or correlation) between any two variables in (2.1) above is computed from as many cases as have values for both variables. This implies that cases with missing values on either (or both) variables are excluded from the computations. In particular, the pairwise covariances between the j-th and k-th variables are obtained as:

$$s_{jk}^{jk} = \frac{1}{n^{(jk)} - 1} \sum^{(jk)} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad (2.5)$$

where, $n^{(jk)}$ is the number of cases with both x_j and x_k observed and the means $\bar{x}_j$, $\bar{x}_k$ and the summation in (2.5) are calculated over those $n^{(jk)}$ cases.

Let s^j_{jj} and s^k_{kk} denote sample variances of x_j and x_k from available cases. Combining these with s^{jk}_{jk} yields the following estimate of correlations:

$$r^*_{jk} = \frac{s^{jk}_{jk}}{\sqrt{s^j_{jj} s^k_{kk}}}. \tag{2.6}$$

The problem with (2.6) is that the estimated correlations can lie outside the range (-1,1). This difficulty is avoided by computing pairwise correlations – where variances are estimated from the same sample base as the covariances– as follows:

$$r^{jk}_{jk} = \frac{s^{jk}_{jk}}{\sqrt{s^{jk}_{jj} s^{jk}_{kk}}}. \tag{2.7}$$

The most serious problem of this approach is that it can yield covariance and correlation matrices which are non-positive definite (has negative eigenvalues). To illustrate the point, consider the following hypothetical case on three variables each with 8 recorded observations and 4 missing values denoted by '?' as follows:

X_1	1	2	3	4	1	2	3	4	?	?	?	?
X_2	1	2	3	4	?	?	?	?	1	2	3	4
X_3	?	?	?	?	1	2	3	4	4	3	2	1.

Expression (2.7) yields $r^{12}_{12} = 1$, $r^{13}_{13} = 1$, $r^{23}_{23} = -1$. These estimates are clearly unsatisfactory, since

$$\text{Corr}(x_1, x_2) = \text{Corr}(x_1, x_3) \Rightarrow \text{Corr}(x_2, x_3) = 1, \text{ not } -1. \tag{2.8}$$

In the same way, covariance matrices based on (2.5) are not necessarily positive definite.

The problem of non-positive definite (NPD) matrices in all-available-case analysis is a major impediment to the use of the method since many statistical analyses, including multiple regression require positive definite matrices. Details of the problem of non-positive definite matrices as well as some suggested remedies are given by Dong (1985); and Knol and Ten Berge (1989).

Bello (1992) suggested the following method for handling an NPD matrix (correlation or covariance):

Let $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_p$ be the eigenvalues of any symmetric *pxp* matrix and **S** is a correlation or covariance matrix. Typically, if **S** is an NPD matrix, some λ_i will be negative. Denote the non-positive eigenvalues by λ_i^- and the positive eigenvalues by λ_i^+. Then,

1- Choose a minimum acceptable eigenvalue as $\delta = min_i \mid \lambda_i^- \mid$, $(i = 1, \ldots, q), q < p$.

2- Construct a modified eigenvalue λ_i^* as follows:

$$\lambda_i^* = \begin{cases} \lambda_i - \frac{(1+\epsilon+\eta)\delta\lambda_i}{\sum \lambda_j^+}, & \mid \lambda_i \mid > \delta \\ \lambda_i + \delta, & 0 \leq \mid \lambda_i \mid < \delta \\ \delta, & \lambda_i < 0, \end{cases} \tag{2.9}$$

where, η is the number of eigenvalues that lies in $(0, \delta)$ range and ϵ is the number of negative eigenvalues.

3- And the modified matrix is

$$\mathbf{S}^* = \sum_{i=1}^{p} \lambda_i^* \gamma_i \gamma_i', \tag{2.10}$$

which is always positive definite and non-singular. Here, γ_i's are the eigenvectors associated with λ_i of the NPD matrix **S**.

All-available-data estimates such as those given by (2.5)–(2.7) try to improve on the complete-case analysis by incorporating, in the analysis, some of the partially observed data. Thus one might expect the resulting estimates to be better than those obtained via casewise and variable-wise-deletion methods. Kim and Curry (1977) supported this conclusion in a simulation study under MCAR assumption and modest correlations. However, Haitovsky (1968), also in a simulation study had indicated the superiority of the complete-case analysis when the correlations are large.

In conclusion, we note that the problem with the use of casewise-deletion method is the considerable loss of data. On the other hand the use of all-available data method yields inconsistent covariance and correlation matrices in a multivariate context. More importantly, both methods do not address the problem of missing data itself. Thus it does not allow the data to be fully used for explanatory purposes. To this end, a more powerful strategy which overcomes the pitfalls of the previous strategy is needed. This involves estimating the missing values themselves via the imputation strategy.

2.3 IMPUTATION STRATEGY

This is an approach which tries to overcome the limitations of the deletion strategy. The basic idea of this approach is to estimate in the first step the missing values (not parameters) and then proceed to the estimation of the parameters. To have complete data sets is sometimes impossible due to cost or time constraints. Thus it would be economical and time saving to impute (fill-in) the missing observations. The advantages of this strategy include:

1- The parameter estimation is more efficient since a greater amount of the data is restored for the analysis.

2- It allows the use of the data for explanatory purposes.

The main drawback of this approach is that it requires iterative numerical solutions, that makes it out-of-reach to most data analyst. This is because most of the imputation methods are not programmed in the existing statistical computer packages, e.g., *SAS*, *SPSS*, *BMDP*, ...etc.

The area of experimental designs had witnessed the discovery of the problem of missing data. Allan and Wishart (1930) seem to be the first to have used the idea of analyzing the experiment by estimating the missing values. They used a procedure of fitting constants to estimate the missing values. Yates (1933) is the first statistician who used the least squares method to estimate the missing values. The basic idea of Yates's method is that the residual sum of squares is to be minimized with respect to both the regression coefficients and the missing values. This approach was extended to the problem of missing data in multivariate analysis by treating the missing values as parameters. Let

$$L(Y_{obs} \mid Y_{mis}, \theta) \propto f(Y_{obs}, Y_{mis} \mid \theta), \tag{2.11}$$

be a function of (θ, Y_{mis}) for fixed Y_{obs}, under MAR assumption. An estimate of θ can be obtained by maximizing (2.11) over both θ and Y_{mis}. The problem of this approach is that the number of parameters increases with the number of missing observations. Details of this approach and its limitations are given by Little and Rubin (1983).

To date, there is a wide variety of imputation techniques which could possibly be categorized as *STOCHASTIC* or *DETERMINISTIC*.

2.3.1 SINGLE IMPUTATION TECHNIQUES

Deterministic imputation techniques are usually referred to as single imputation techniques. This is because it imputes a single value for each missing observation. The most popular deterministic techniques include: Imputing unconditional means, Imputing conditional means (Buck's method), Dear's Principal Component Method (DPC), General Iterative Principal Component Method (GIPC) and the Singular Value Decomposition Method (SVD).

2.3.1.1 UNCONDITIONAL IMPUTATION

This is the first imputation method to appear in the statistical literature. It is also referred to as the mean substitution method (MSM). The originator of this method is not known, but it is often attributed to Wilks. The basic idea involves replacing the missing values in a particular variable by the mean of the available data on the variable. Although many authors have adopted this method in their works, apart from Wilks (1932), there seems to be no serious investigation into the statistical properties of the method.

In this section we shall try to study the biasedness in the variance-covariance matrix obtained from the data completed via the MSM method.

Let $\bar{x}_j$ and s_{jj} be the mean and variance of the j-th variable after imputing the missing observations using the MSM method. Similarly, let $\bar{x}_j^*$ and s_{jj}^* be the All-available-data estimate of the mean and variance of the j-th variable. Then we have the following results:

Lemma 2.1

Using MSM, the estimated post-imputation mean of the j-th variable is equal to the same estimate obtained via all-available-data method, i.e.,

$$\bar{x}_j = \bar{x}_j^*$$

Proof

If the j-th variable is recorded for $n^{(j)}$ out of n observations, then the average of the observed and imputed values is

$$\bar{x}_j = \frac{1}{n}\left\{\sum_{n^{(j)}}(x_{ij}) + (n - n^{(j)})\bar{x}_j^*\right\} \tag{2.12}$$

where

$$\bar{x}_j^* = \frac{1}{n^{(j)}}\sum_{n^{(j)}} x_{ij}, \tag{2.13}$$

is the All-available-data estimate of the mean. Substituting (2.13) in (2.12), we have

$$\begin{aligned}\bar{x}_j &= \frac{1}{n}\left\{n^{(j)}\bar{x}_j^* + (n - n^{(j)})\bar{x}_j^*\right\} \\ &= \bar{x}_j^*.\end{aligned} \tag{2.14}$$

Note that the above lemma says: The estimated mean vector via the unconditional imputation method is the same as the one obtained via All-available-data method. However, for the covariance matrix this is not the case as is shown by the following theorem:

Theorem 2.1

Under MCAR assumption for the missing values, the variances and covariances obtained via the MSM method are biased.

Proof

The All-available-data estimate of the variance of the j-th variable is

$$s_{jj}^* = \frac{1}{n^{(j)} - 1}\left\{\sum_{n^{(j)}}(x_{ij} - \bar{x}_j^*)^2\right\}. \tag{2.15}$$

And the variance of the j-th variable after imputing the missing values using the MSM method is

$$s_{jj} = \frac{1}{n-1}\left\{\sum_{n}(x_{ij} - \bar{x}_j)^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{n}(x_{ij} - \bar{x}_j^*)^2\right\}$$

(since from (2.14), $\bar{x}_j = \bar{x}_j^*$)

$$\therefore\ s_{jj} = \frac{1}{n-1}\left\{\sum_{n^{(j)}}(x_{ij} - \bar{x}_j^*)^2 + \sum_{n-n^{(j)}}(x_{ij} - \bar{x}_j^*)^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{n^{(j)}}(x_{ij} - \bar{x}_j^*)^2\right\}$$

(since $\sum_{n-n^{(j)}}(x_{ij} - \bar{x}_j^*)^2 = 0$ due to imputation).

Hence

$$s_{jj} = \frac{(n^{(j)} - 1)}{(n-1)} s_{jj}^* \tag{2.16}$$

which implies that the sample variance from the filled-in data underestimates the variance by a factor $(n^{(j)} - 1)/(n-1)$.

Similarly, for the covariances we have,

$$s_{jk}^* = \frac{1}{n^{(jk)} - 1}\left\{\sum_{n^{(jk)}}(x_{ij} - \bar{x}_j^*)(x_{ik} - \bar{x}_k^*)\right\} \tag{2.17}$$

which is the All-available-data estimate of the covariance of the j-th and k-th variables, where $\bar{x}_j^*$ and $\bar{x}_k^*$ are the all-available-data estimates of the means.

$n^{(jk)}$ is the number of cases with both x_{ij} and x_{ik} observed. And the same estimate from the data imputed via the MSM method is

$$
\begin{aligned}
s_{jk} &= \frac{1}{n-1}\left\{\sum_{n}(x_{ij}-\overline{x}_j)(x_{ik}-\overline{x}_k)\right\} \\
&= \frac{1}{n-1}\left\{\sum_{n}(x_{ij}-\overline{x}_j^*)(x_{ik}-\overline{x}_k^*)\right\} \\
&= \frac{1}{n-1}\left\{\sum_{n^{(jk)}}(x_{ij}-\overline{x}_j^*)(x_{ik}-\overline{x}_k^*)+\sum_{n-n^{(jk)}}(x_{ij}-\overline{x}_j^*)(x_{ik}-\overline{x}_k^*)\right\} \\
&= \frac{1}{n-1}\left\{\sum_{n^{(jk)}}(x_{ij}-\overline{x}_j^*)(x_{ik}-\overline{x}_k^*)\right\}
\end{aligned}
$$

(since $\sum_{n-n^{(jk)}}(x_{ij}-\overline{x}_j^*)(x_{ik}-\overline{x}_k^*)=0$ due to imputation)

$$
\therefore\ s_{jk} = \frac{n^{(jk)}-1}{n-1}s_{jk}^* \tag{2.18}
$$

which implies that the sample covariance from the filled-in data underestimates the covariance by a factor $(n^{(jk)}-1)/(n-1)$.

Thus the method underestimates both the variances and covariances. The resulting estimate of the j-th variable's variance needs to be corrected for this bias by multiplying by the factor $(n-1)/(n^{(j)}-1)$. Similarly the estimated covariance of the j-th and k-th variables is to be multiplied by the factor $(n-1)/(n^{(jk)}-1)$.

The main criticism of this method is that it does not make use of the intercorrelation that often exists among multiple variables in obtaining the imputed values. Consideration of this has led statisticians to think of conditional imputation.

2.3.1.2 CONDITIONAL IMPUTATION: BUCK'S METHOD

Buck (1960) appears to be the first author to give a comprehensive regression method that estimates the missing values and adjusts for the resulting bias of the estimated parameters. Since then the method has been widely used by scholars dealing with the problem of missing data in multivariate context. An interesting property of the method is that it is a combination of deletion and imputation strategies. Buck starts by using the complete observations to estimate the means of all variables, and also the covariance matrix. These values can then be used to estimate any missing value x_{ij} as linear functions of the variables that are known for this observation. If we then substitute the estimates for the unknown variables, we can build up the means vector and covariance matrix for the completed data.

Buck's method is a useful improvement over the estimators found by unconditional imputation. However, the covariance matrix calculated by the method provides a biased estimate of the values that they would have taken if none of the data have been missing. Buck's method therefore estimates this bias, and adjusts for it.

We have in this work given special attention to the method of Buck as a pioneering method in the area that combines both deletion and imputation strategies. In the next two chapters we shall introduce the method in some detail and give some contributions in some of its aspects.

2.3.1.3 DEAR'S PRINCIPAL COMPONENT (DPC) METHOD

Dear (1959) developed a method for estimating multivariate missing values using principal component analysis. He obtained the first principal component from the n_c-complete cases of the sample (i.e, $\mathbf{X}_c$). The method can be described in the following steps

1- Calculate the sample covariance matrix, $S = (n_c - 1)^{-1}\mathbf{X}_c'\mathbf{X}_c$. Note that the principal components which are based on sample covariance and correlation matrices will produce different results for the same data matrix. To avoid this, Dear (1959) standardized the elements of $\mathbf{X}_c$ to $\mathbf{Z}_c$, where $z_{jk} = (x_{jk} - \bar{x}_k)/\sqrt{s_{kk}}$. Thus $\mathbf{S}$ is now the correlation matrix obtained from $\mathbf{Z}_c$.

2- Calculate the largest eigenvalue of $\mathbf{S}$, $\lambda_1 = \max_k(\lambda_k)$, and its associated eigenvector $\eta_{1k}, (k = 1, \ldots, p)$.

3- Let the first principal component for the i-th case be

$$\gamma_i = \sum_{k=1}^{p} \eta_{1k} z_{ik}, \qquad i = 1, 2, \ldots, n_c \tag{2.19}$$

so that the missing values in the i-th case are to be replaced by the nearest point on the first principal component as follows:

$$\hat{z}_{ik} = \begin{cases} z_{ik}, & \text{if } x_{ik} \text{ is observed,} \\ \eta_{1k}\gamma_i, & \text{if } x_{ik} \text{ is missing.} \end{cases} \tag{2.20}$$

4- De-standardize $\hat{\mathbf{Z}}_c$ to $\hat{\mathbf{X}}_c$.

2.3.1.4 GENERAL ITERATIVE PRINCIPAL COMPONENT (GIPC) METHOD

We can note that Dear's principal component method, discussed above may collapse altogether if all cases have missing values in it. On the other hand, the method will result in poor estimate of $\mathbf{S}$ if the number of complete cases is relatively small. To avoid these difficulties an iterative principal component method is suggested as follows:

1- Use the all-available-data method of section 2.2.3 to calculate $\mathbf{S}$. In case it is non-positive definite, modify it using the procedure given in the

same section. Alternatively, use the unconditional imputation method after adjusting for bias as discussed in section 2.3.1.1, and calculate S.

2- Construct the first principal component from S and estimate the missing values with (2.20) above.

3- Recalculate S from the imputed data matrix and repeat 2 above.

4- Iterate between 3 and 2 until successive imputed values do not change significantly according to a certain convergence criteria.

2.3.1.5 SINGULAR VALUE DECOMPOSITION (SVD) METHOD

The singular value decomposition method is a technique by which an arbitrary real (nxp) matrix **X**, of rank k, can be expressed as the sum of k matrices of rank one. Good (1969) discussed the use of the method in least squares and principal component analysis.

Krzanowski (1987, 1988) was the first author to suggest using the method to impute missing values in multivariate analysis. Suppose that p variables $x_1, \ldots, x_p$ are observed on each of n individuals ($n \geq p$), and the resultant values are displayed in an (nxp) data matrix **X**. Then the singular value decomposition of **X** is defined by

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}' \tag{2.21}$$

where $\mathbf{U}'\mathbf{U} = \mathbf{I}_p, \mathbf{V}'\mathbf{V} = \mathbf{I}_p$, and $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$ with $d_1 \geq \ldots \geq d_p \geq 0$.

If the (i,j)-th elements of the matrices **X** and **U** are denoted by x_{ij} and u_{ij}, respectively, decomposition (2.21) has its elementwise representation

$$x_{ij} = \sum_{t=1}^{p} u_{it} d_t v_{tj}. \tag{2.22}$$

Now assume that x_{ij} is missing. Denote by $\mathbf{X}^{(-i)}$ and $\mathbf{X}^{(-j)}$ the resulting matrices after deleting the i-th row and j-th column of $\mathbf{X}$ respectively. Then from (2.21), the singular value decompositions of the (n-1)xp matrix $\mathbf{X}^{(-i)}$ and the nx(p-1) matrix $\mathbf{X}^{(-j)}$ can be written as,

$$\mathbf{X}^{(-i)} = \overline{\mathbf{U}}\,\overline{\mathbf{D}}\,\overline{\mathbf{V}}' \text{ with } \overline{\mathbf{U}} = (\bar{u}_{st}), \overline{\mathbf{V}} = (\bar{v}_{st}), \text{ and } \overline{\mathbf{D}} = \text{diag}(\bar{d}_1, \ldots, \bar{d}_p), \tag{2.23}$$

and

$$\mathbf{X}^{(-j)} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}' \text{ with } \tilde{\mathbf{U}} = (\tilde{u}_{st}), \tilde{\mathbf{V}} = (\tilde{v}_{st}), \text{ and } \tilde{\mathbf{D}} = \text{diag}(\tilde{d}_1, \ldots, \tilde{d}_p). \tag{2.24}$$

An estimate of the missing value x_{ij} is obtained by combining (2.23) and (2.24), that is

$$\hat{x}_{ij} = \sum_{t=1}^{p-1} (\tilde{u}_{it}\sqrt{(\tilde{d}_t)})(\bar{v}_{tj}\sqrt{\bar{d}_t}). \tag{2.25}$$

For more than one missing value, the method starts with any initial imputed values and then uses (2.25) to update them. The process is then iterated until stability is achieved.

2.3.1.6 ON THE PERFORMANCE OF DETERMINISTIC IMPUTATION TECHNIQUES

Bello (1992), on the basis of a simulation study compared the performance of some of the above methods as applied to discriminant and regression analyses. The study concentrated on five deterministic imputation techniques: Mean Substitution Method (MSM), EM algorithm (EM), Dear's Principal Component Method (DPCM), General Iterative Principal Component Method (GIPCM) and Singular Value Decomposition Method (SVDM).

It is worthnoting that Bello's study considered the EM algorithm as an imputation technique and not ML method of estimation from incomplete data. This can be justified by the fact that Bello's objective was to estimate the missing observations and not the population parameters.

100 observations were generated from a five-variate normal distribution with a fixed mean vector of zeros and covariance matrix, $\Sigma = (\rho_{ij})$ $(i, j = 1, \ldots, p)$ using NAG subroutines G05EAF and G05EZF. k(%) of the (nxp) data matrix was randomly deleted, with MAR mechanism. Each of the above mentioned five imputation techniques was then applied to the resulting incomplete data. Some of the results obtained are:

1- Although no single imputation technique is the best overall in all circumstances, MSM and DPCM behave erratically when the intercorrelation among the variables is moderate or high. They performed worse than the iterative imputation techniques (SVD and GIPC) which, under this condition, are equally efficient.

2- For the estimation of mean vector, results indicated that there is little or no evidence of superiority of one imputation technique over another. They are all virtually equivalent.

3- For mixture of continuous and categorical variables, DPC, GIPC and SVD methods can still be achieved straightforwardly. The EM algorithm can still be used for this context, but unfortunately, this may involve quite a large number of parameters estimates and formidable computations.

4- The misclasification errors of the imputed data in linear, quadratic and kernel discriminant functions decrease with increasing sample size and decreasing proportion of missing values. This result was found to be true for the five imputation techniques considered in the study.

5- There is insufficient evidence to discredit the use of the EM algorithm when the data markedly deviate from normality assumptions.

2.3.2 MULTIPLE IMPUTATION

The theory underlying multiple (stochastic) imputation was first proposed in Rubin (1978), although the idea appears in Rubin (1977). The idea is to impute several times for each respondent, each imputation being a random draw (with replacement) from the set of respondents y_i using a specified probability sampling model. To illustrate the main idea, suppose we want to estimate $\overline{Y}$, the mean of a variable Y in a finite population of size N. Let $y_i, i = 1, \ldots, n$ be a SRS of size n from the population $Y_i, i = 1, \ldots, N$. Suppose now that because of nonresponse only m out of n values of y_i are observed. Then, using multiple imputation we impute I times for each of the $n - m$ missing values to form I complete-data sets. From these I complete-data sets we calculate I complete-data statistics. Denote the post-imputation sample means by $\overline{y}_{*1}, \ldots, \overline{y}_{*I}$ and the sample variances by $\hat{V}_{*1}, \ldots, \hat{V}_{*I}$.

With I imputations from the SRS model, the 95% interval for $\overline{Y}$ will be

$$\overline{y}_* \pm 2(W + \overline{V}_*/n)^{1/2}, \tag{2.26}$$

where $\overline{y}_*$, $\overline{V}_*$ and W are defined by employing the multiple imputation inference. Thus the center of the interval is the average of the I centers,

$$\overline{y}_* = \sum_1^I \overline{y}_{*i}/I, \tag{2.27}$$

and the variance defining the width of the 95% interval is the average variance within the imputations plus the variance across imputations of the I centers; the average variance within the imputations is

$$\overline{V}_*/n = \frac{1}{I}\sum_1^I \frac{\hat{V}_{*i}}{n}, \tag{2.28}$$

and the variance across imputations of the centers is

$$W = \sum_{1}^{I} (\bar{y}_{*i} - \bar{y}_{*})^2/(I-1). \tag{2.29}$$

Then Herzog and Rubin (1983) showed that using multiple imputation (outlined above) rather than single imputation reduces the estimated variance of $\overline{Y}$ (over repeated sampling and imputation procedures). This is because the mean of a multiple imputation interval has less variability than the mean of a single imputation interval. More importantly, they showed that the underestimation of the width of the 95% interval of $\overline{Y}$ when using single imputation is more serious than when using multiple imputation.

It is worth mentioning that multiple imputations are particularly useful in sample surveys and censuses where standard complete-data analyses are difficult to modify in the presence of nonresponse (Rubin, 1996). However, the principle of multiple imputation has been applied to multivariate analysis by Rubin and Schafer (1990) and Brownstone (1991).

2.4 MODEL-BASED ESTIMATION OF MISSING VALUES: MAXIMUM LIKELIHOOD APPROACH

The theory of ML estimation of the population parameters is clear in most of its aspects. Specifically, if $f(\underline{x}, \theta)$ is the joint p.d.f from which a sample of size n is generated, then the likelihood function is any function of θ proportional to the joint p.d.f with a factor which is independent of θ; i.e.,

$$L(x, \theta) = \prod_{i=1}^{n} f(x_i, \theta) = f(\underline{x}, \theta) \tag{2.30}$$

where

i- $L(x, \theta)$ is differentiable and bounded above

ii- The sample values $x_i \sim N(\mu, \sigma^2)$

iii- x_i's are *iid* random variables.

Then the ML estimate of θ is obtained by maximizing $L(x, \theta)$ or the loglikelihood function $\ell(x, \theta)$ with respect to the elements of θ. For the complete-data analysis, interval estimation is based upon the large sample property

$$(\theta - \hat{\theta}) \sim N(0, \mathbf{C}) \tag{2.31}$$

where $\mathbf{C}$ is the covariance matrix for $(\theta - \hat{\theta})$ such that

$$\mathbf{C} = [E\{I(\hat{\theta} \mid x)\}]^{-1}, \tag{2.32}$$

and

$$I(\hat{\theta} \mid x) = \frac{-\partial^2 \log L}{\partial \theta^2} \tag{2.33}$$

is the observed information.

The basic idea of the ML estimation for incomplete data is the same as that of the complete-data case. The likelihood for the parameters based on the incomplete data is derived and ML estimates are found by solving the likelihood equation. The main difficulties however, are:

1- Asymptotic standard errors obtained from the information matrix given by (2.33) above are somewhat questionable. This is because the observed data, in the presence of missing values, do not generally constitute an iid sample. Thus simple results based on (2.31) above that imply the large sample normality of the likelihood function do not immediately apply.

2- Complications arise from dealing with the missing data mechanism (MDM). Specifically, Rubin (1976) showed that ML estimation for incomplete data requires the *weak* (i.e, MAR) assumption discussed in

chapter I. The practical significance of this assumption is that it allows the ignorability of the distribution of the missingness mechanism (R). To illustrate this point consider a simple univariate incomplete exponential sample of size n where y_i, $i = 1,\ldots,m$ are observed and y_i, $i = m+1,\ldots,n$ are missing. Then the joint density of the n exponential random variables is given by

$$f(\underline{y} \mid \theta) = \theta^{-n} \exp\left[-(\sum_{i}^{n} y_i/\theta)\right], \quad i = 1,2,\ldots,n \tag{2.34}$$

from which

$$f(y_{obs} \mid \theta) = \theta^{-m} \exp\left[-(\sum_{i}^{m} y_i/\theta)\right] \tag{2.35}$$

and the likelihood function ignoring the missing data mechanism is given by

$$L(\theta \mid y_{obs}) \propto f(y_{obs} \mid \theta) = \theta^{-m} \exp\left[-(\sum_{i}^{m} y_i/\theta)\right]. \tag{2.36}$$

Now, to incorporate the missing data mechanism, let

$$R = R_i = \begin{cases} 1, & \text{for } i = 1,2,\ldots,m \\ 0, & \text{for } i = m+1,\ldots,n \end{cases} \tag{2.37}$$

Further assume each unit is observed with probability ψ, then the distribution of the MDM is given by

$$f(R \mid \underline{y}, \psi) = \psi^{m}(1-\psi)^{n-m}, \tag{2.38}$$

which is MAR since it is independent of y_{mis}.

Then the joint distribution of R and $\underline{y}$ can be written as

$$f(\mathrm{R}, \underset{\sim}{\mathrm{y}} \mid \theta, \psi) = f(\underset{\sim}{\mathrm{y}} \mid \theta)\mathrm{f}(\mathrm{R} \mid \underset{\sim}{\mathrm{y}}, \psi)$$
$$= \theta^{-n}\psi^{m}(1-\psi)^{n-m}\exp\left[-\sum_{i}^{n}(\mathrm{y_i}/\theta)\right] \tag{2.39}$$

Thus the distribution of the actual observed data consisting of the variables $(\mathrm{y_{obs}}, \mathrm{R})$ is obtained from (2.39) as

$$f(\mathrm{y_{obs}}, \mathrm{R} \mid \theta, \psi) = f(R \mid \mathrm{y_{obs}}, \psi)\sum_{\mathrm{y_{mis}}}\mathrm{f}(\mathrm{y_{obs}} \mid \theta)$$
$$= \psi^{m}(1-\psi)^{n-m}\theta^{-m}\exp\left[-\sum_{i}^{m}(\mathrm{y_i}/\theta)\right]. \tag{2.40}$$

Thus incorporating the distribution of the MDM, the likelihood function is given by

$$L(\theta, \psi \mid \mathrm{y_{obs}}, \mathrm{R}) \propto f(\mathrm{y_{obs}}, \mathrm{R} \mid \theta, \psi)$$
$$= \psi^{m}(1-\psi)^{n-m}\theta^{-m}\exp\left[-\sum_{i}^{m}(\mathrm{y_i}/\theta)\right]. \tag{2.41}$$

Now, the ML estimate $\hat{\theta}$ of θ obtained from (2.36), which ignores the distribution of the MDM is $\hat{\theta} = \sum_{m}\mathrm{y_i}/\mathrm{m}$, which is the same estimate obtained by maximizing (2.41) which incorporates the MDM. This is because the MAR assumption given by (2.38) has the effect of making (2.36) and (2.41) proportional (differing by a constant $\psi^{m}(1-\psi^{n-m})$) which is independent of θ. Thus the ignorability of the MDM is a direct consequence of the MAR assumption.

However, complications arise if the MAR assumption given by (2.38) is violated. To fix ideas let the incomplete data be created by censoring at

some known censoring point c, so that only values less than c are recorded. Then

$$f(\mathrm{R} \mid \underline{\mathrm{y}}, \psi) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{f}(\mathrm{R_i} \mid \mathrm{y_i}, \psi) \tag{2.42}$$

where

$$f(\mathrm{R_i} \mid \mathrm{y_i}, \psi) = \begin{cases} 1, & \text{if } (R_i = 1 \text{ and } y_i < c) \text{ or } (R_i = 0 \text{ and } y_i > c); \\ 0, & \text{otherwise.} \end{cases} \tag{2.43}$$

Hence the joint density of the actual observed data $\mathrm{f}(\mathrm{y_{obs}}, \mathrm{R})$ is given by

$$\begin{aligned} f(\mathrm{y_{obs}}, \mathrm{R} \mid \theta) &= \prod_{i=1}^{m} f(\mathrm{y_i}, \mathrm{R_i} \mid \theta) \\ &= \prod_{i=1}^{m} f(\mathrm{y_i}) \prod_{\mathrm{i}=\mathrm{m}+1}^{\mathrm{n}} \mathrm{f}(\mathrm{R_i} \mid \theta) \\ &= \prod_{i=1}^{m} f(\mathrm{y_i}) \mathrm{f}(\mathrm{R_i} \mid \mathrm{y_i}) \prod_{\mathrm{i}=\mathrm{m}+1}^{\mathrm{n}} \Pr(\mathrm{y_i} > \mathrm{c}) \\ &= \theta^{-m} \exp\left[-\sum_{i}^{m} (\mathrm{y_i}/\theta)\right] \exp[-(n-m)c/\theta] \end{aligned}$$

since for the exponential distribution $\Pr(\mathrm{y_i} > \mathrm{c}) = \exp^{-\mathrm{c}/\theta}$ and $\mathrm{f}(\mathrm{R_i} \mid \mathrm{y_i}) = 1$ from (2.43).

$$\therefore \quad f(\mathrm{y_{obs}}, \mathrm{R} \mid \theta) = \theta^{-\mathrm{m}} \exp\left[\left\{-\sum_{\mathrm{i}}^{\mathrm{m}} (\mathrm{y_i}) - (\mathrm{n}-\mathrm{m})\mathrm{c}\right\}/\theta\right] \tag{2.44}$$

from which a ML estimate of θ is obtained as

$$\hat{\theta} = \sum_{i}^{m} (\mathrm{y_i}/\mathrm{m}) + [(\mathrm{n}-\mathrm{m})\mathrm{c}]/\mathrm{m} \tag{2.45}$$

which is different from the one obtained from (2.36) by ignoring the distribution of the MDM. This is because the distribution of the MDM given by (2.43) is no longer an MAR, hence nonignorable. Note that the last term of (2.45) reflects the effect of the non-random missingness on the estimated parameter. Also note that in the above censoring illustration the MDM is nonignorable but known. A more serious complication arises when the MDM is nonignorable but ψ is unknown.

It is interesting to note that the MAR assumption in the ML estimation of missing data plays a similar role to that of the MCAR assumption in deletion-pairwise strategy. In the former, MAR justifies the ignorability of the MDM, whereas in the latter MCAR justifies the discarding of the incomplete cases.

2.4.1 FACTORIZATION METHOD

The factorization method in the model-based approach to missing data problem can best be described by reviewing three pioneering papers in the area. These are by Lord (1955), Edgett (1956) and Anderson (1957). Lord (1955) considered three variables u, v and w from the trivariate normal distribution. In the available data, variable w is completely recorded; variables u and v are never jointly recorded. He obtained ML estimates for eight of the nine parameters since there is no information for the estimation of the partial correlation between u and v. Edgett (1956) considered the same case where only u or v is subject to missingness. He found the ML solutions in an explicit form.

Anderson (1957) was the first author to consider the factorization method. He considered the bivariate normal distribution and made generalizations to the trivariate and multivariate normal cases. To describe the method suppose N observations on x and y have a bivariate normal distribution

with density $f(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho)$. Further assume that x is completely observed while x and y are jointly observed on n observations; that is, $N - n$ observations on y are missing. The data for this pattern of missingness are

$$x_1, \ldots, x_n, x_{n+1}, \ldots, x_N,$$

$$y_1, \ldots, y_n, \ ?, \ ?, \ldots, ?.$$

Then the key idea of the factorization method is that the joint density of x and y can be factorized into two terms as

$$f(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) = f(x \mid \mu_x, \sigma_x^2) f(y \mid v + \beta_{yx} x, \sigma^2) \tag{2.46}$$

where

$$v = \mu_y - \beta_{yx}\mu_x, \tag{2.47}$$

$$\beta_{yx} = \rho\sigma_y/\sigma_x, \tag{2.48}$$

$$\sigma^2 = \sigma_y^2(1 - \rho)^2. \tag{2.49}$$

Then the likelihood function can be written as:

$$\begin{aligned} L(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) &= \prod_{i}^{n} f(x_i, y_i \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) \\ &\quad \cdot \prod_{i=n+1}^{N} f(x_i \mid \mu_x, \sigma_x^2) \\ &= \prod_{i=1}^{N} f(x_i \mid \mu_x, \sigma_x^2) \prod_{i=1}^{n} f(y_i \mid v + \beta_{yx} x_i, \sigma^2). \end{aligned} \tag{2.50}$$

The ML estimates of $\mu_x, \sigma_x^2, v, \beta_{yx}$ and σ^2 are those values that maximize (2.50). A thorough discussion of the factorization method and its relation to imputation techniques (Buck's method) is given in chapter V.

Next, the generalization of the method to the trivariate normal case is considered. In Edgett's case there are three variables w, x, and y. The data are

$$\begin{aligned} &w_1, \ldots, w_n, w_{n+1}, \ldots, w_N, \\ &x_1, \ldots, x_n, x_{n+1}, \ldots, x_N, \\ &y_1, \ldots, y_n, ?, ?, \ldots, ?. \end{aligned}$$

The only difference between the previous case and this one is that x has been replaced by the pair (x, w). Since the same approach is used, we will only give the possible factorization of the likelihood function, which is,

$$\begin{aligned} L(w, x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho) = \prod_i^n f(w_i, x_i, y_i \mid \mu_w, \mu_x, \mu_y; \sigma_w^2, \sigma_x^2, \sigma_y^2; \rho_{wx}, \rho_{wy}, \\ \rho_{xy}) \cdot \prod_{i=n+1}^N f(w_i, x_i \mid \mu_w, \mu_x; \sigma_w^2, \sigma_x^2; \rho_{wx}) \\ = \prod_{i=1}^N f(w_i, x_i, \mu_w, \mu_x; \sigma_w^2, \sigma_x^2; \rho_{wx}) \\ \cdot \prod_{i=1}^n f(y_i \mid v + \beta_{yw.x} w_i + \beta_{yx.w} x_i, \sigma_{y.wx}^2) \quad (2.51) \end{aligned}$$

In Lord's case we have

$$\begin{aligned} &x_1, \ldots, x_n, x_{n+1}, \ldots, x_N, \\ &y_1, \ldots, y_n, \ ?, \ldots, \ ?, \\ &?, \ldots, \ ?, z_{n+1}, \ldots, z_N. \end{aligned}$$

The likelihood function can be factorized as follows:

$$L(x, y \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho_{xy}) = \prod_{i=1}^{n} f(x_i, y_i \mid \mu_x, \mu_y; \sigma_x^2, \sigma_y^2; \rho_{xy})$$
$$\cdot \prod_{i=n+1}^{N} f(x_i, z_i \mid \mu_x, \mu_z; \sigma_x^2, \sigma_z^2; \rho_{xz})$$
$$= \prod_{i=1}^{N} f(x_i \mid \mu_x, \sigma_x^2) \prod_{i=1}^{n} f(y_i \mid v_y + \beta_{yx} x_i, \sigma_{y.x}^2)$$
$$\cdot \prod_{i=n+1}^{N} f(z_i \mid v_z + \beta_{zx} x_i, \sigma_{z.x}^2). \qquad (2.52)$$

Generalizations to the multivariate normal case are possible. Unfortunately, the method is not general in the sense that it cannot handle all patterns of multivariate missing data. For instance the method is inapplicable to the following set of data

$$x_1, ?, ?, x_n, x_{n+1}, \ldots, x_N,$$

$$y_1, ?, ?, y_n, y_{n+1}, ?, \ldots, ?$$

since the bivariate density of x and y cannot be written as the product of the marginal density of x and the conditional density of y given x. Anderson does not elaborate on situations that cannot be handled by his method. He only gave an example of two patterns of missingness that can be handled by the method. Conditions that lead to the collapse of the method were later given by Rubin (1976) in a theorem that shall be given later in this section.

A common property of the methods of Lord, Edgett and the factorization method of Anderson is that they completely ignore the missingness mechanism. This property can only be explained by an implicit MAR assumption for the missing data. Otherwise the distribution of the MDM should be incorporated in the likelihood function as outlined in section 2.4. In fact, the

implicit MAR/MCAR assumptions in the field of statistical analysis with missing data is a common property shared by almost all studies conducted before 1976. It is Rubin (1976) who gave a precise mathematical treatment of the MDM that makes it an important component in the analysis. On the light of this, Rubin (1976) formalizes the theory of Anderson's factorization methods and its limitations as follows:

Theorem 2.2

Let $\ell(\theta \mid Y_{obs})$ be the loglikelihood function based on the incomplete data Y_{obs} under MAR assumption. Then for a variety of models and incomplete data problems, an alternative parameterization $\phi = \phi(\theta)$, where ϕ is a one-one monotone function of θ, can be found such that the loglikelihood decomposes into components

$$\ell(\phi \mid Y_{obs}) = \ell_1(\phi_1 \mid Y_{obs}) + \ell_2(\phi_2 \mid Y_{obs}) + \ldots + \ell_J(\phi_J \mid Y_{obs}) \qquad (2.53)$$

where

1- $\phi_1, \phi_2, \ldots, \phi_J$ are distinct parameters, in the sense that the joint parameter space of $\phi = (\phi_1, \phi_2, \ldots, \phi_J)$ is the product of the individual parameter spaces for ϕ_j, $j = 1, \ldots, J$.

2- The components $\ell_j(\phi_j \mid Y_{obs})$ correspond to loglikelihoods for complete data problems.

Then $\ell(\phi \mid Y_{obs})$ can be maximized by maximizing $\ell_j(\phi_j \mid Y_{obs})$ separately for each j.

Remark

By the invariance property of the ML estimates, it follows that if $\hat{\phi}$ is the resulting ML estimate of ϕ, hence the ML estimate of any function of ϕ is $\hat{\theta} = \theta(\hat{\phi})$. Differentiating (2.53) twice with respect to $\phi_1, \phi_2, \ldots, \phi_J$ yields a diagonal information matrix of the form

$$I(\phi \mid \mathrm{Y_{obs}}) = \begin{bmatrix} I(\phi_1 \mid \mathrm{Y_{obs}}) & & 0 \\ & I(\phi_2 \mid \mathrm{Y_{obs}}) & \\ 0 & & I(\phi_J \mid \mathrm{Y_{obs}}) \end{bmatrix}. \quad (2.54)$$

Hence the covariance matrix $\mathbf{C}$ is given by

$$\mathbf{C}(\hat{\theta} \mid \mathrm{Y_{obs}}) = \begin{bmatrix} I^{-1}(\hat{\theta}_1 \mid \mathrm{Y_{obs}}) & & 0 \\ & I^{-1}(\hat{\theta}_2 \mid \mathrm{Y_{obs}}) & \\ 0 & & I^{-1}(\hat{\theta}_J \mid \mathrm{Y_{obs}}) \end{bmatrix}. \quad (2.55)$$

The approximate covariance matrix of the ML estimate of a function $\theta = \theta(\phi)$ of ϕ can be found using the formula

$$\mathbf{C}(\hat{\theta} \mid Y_{obs}) = \underline{\mathrm{D}}(\hat{\theta})\underline{\mathrm{C}}(\hat{\theta} \mid \mathrm{Y_{obs}})\underline{\mathrm{D}}'(\hat{\theta}) \quad (2.56)$$

where $\mathbf{D}$ is the matrix of partial derivatives of θ with respect to ϕ. That is to say

$$\mathbf{D}(\theta) = \{d_{jk}(\theta)\} = \left\{\frac{\partial \theta_j}{\partial \theta_k}\right\}, \quad (2.57)$$

where θ is expressed as a column vector.

In practice the patterns of incomplete data often do not have the particular forms that allow the use of the factorization methods. Moreover, for some patterns of missingness a factorization may exist, but the parameters ϕ_j in the factorization may not be distinct. Consequently, maximizing the elements of (2.53) separately for each j does not maximize the likelihood. In such cases an alternative method of solution is the Expectation-Maximization (EM) algorithm.

2.4.2 ITERATIVE METHODS: THE EM ALGORITHM

Recalling our notation of section 1.3, let

$$L(\theta \mid Y_{obs}) = \int_{Y_{mis}} f(Y_{obs}, Y_{mis} \mid \theta) dY_{mis}, \tag{2.58}$$

be the marginal density of Y_{obs} under MAR assumption.

If (2.58) is differentiable and unimodal, ML estimates are obtained by solving the likelihood equation

$$S(\theta \mid Y_{obs}) \equiv \frac{\partial \ln L(\theta \mid Y_{obs})}{\partial \theta} = 0. \tag{2.59}$$

When a closed-form solution of (2.59) cannot be found, iterative methods can be applied. Let $\theta^{(0)}$ be an initial estimate of θ, based on the completely recorded units. Let $\theta^{(t)}$ be an estimate at the t-th iteration. The *Newton-Raphson* algorithm is defined by

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)} \mid Y_{obs}) S(\theta^{(t)} \mid Y_{obs}), \tag{2.60}$$

where $I(\theta \mid Y_{obs})$ is the observed information given by

$$I(\theta \mid Y_{obs}) = \frac{\partial^2 \ell(\theta \mid Y_{obs})}{\partial \theta \, \partial \theta}. \tag{2.61}$$

If the likelihood is concave and unimodal, then the sequence of iterates $\theta^{(t)}$ converges to the ML estimates $\hat{\theta}$ of θ. A variant of this procedure is the *Method of Scoring*, where the observed information in (2.60) is replaced by the expected information:

$$\theta^{(t+1)} = \theta^{(t)} + J^{-1}(\theta^{(t)}) S(\theta^{(t)} \mid Y_{obs}), \tag{2.62}$$

where

$$J(\theta) = E\{I(\theta \mid \mathrm{Y_{obs}}) \mid \theta\} = -\int \frac{\partial^2 \ell(\theta \mid \mathrm{Y_{obs}})}{\partial\theta\,\partial\theta} \mathrm{f}(\mathrm{Y_{obs}} \mid \theta)\mathrm{dY_{obs}}. \qquad (2.63)$$

Both these methods involve calculating the matrix of second derivatives of the likelihood function. For general patterns of missingness, the entries in this matrix tend to be complicated functions of θ especially when θ is of high dimensions. As a result, the above methods might be too cumbersome for practical purposes.

An alternative iterative approach, which does not require the calculation of the second derivatives, is the *Expectation-Maximization* (EM) algorithm which is introduced below.

We start our discussion of the EM algorithm by introducing the *Missing Information Principle* of Orchard and Woodbury (1972). This is because the theoretical basis of the EM algorithm were first set in this earlier work of Orchard and Woodbury as we shall see below. The formulation given here appears in Beale and Little (1975).

Consider the vectors $\mathrm{Y_{obs}}$ and $\mathrm{Y_{mis}}$ of random variables with a joint distribution depending on the vector θ of parameters, where

$\mathrm{Y_{obs}} \equiv$ The complete observations and the known variables in the incomplete observations,

$\mathrm{Y_{mis}} \equiv$ The missing values in the incomplete observations.

We wish to find $\hat{\theta}$, the estimate of θ which maximizes the likelihood function $\mathrm{L}(\mathrm{Y_{obs}}, \theta)$ or its log. However, $\hat{\theta}$ may not be obtained easily this way. Orchard and Woodbury (1972) came up with an alternative approach which is based on the *Missing Information Principle.* The central idea of this principle is *to find the value of* θ *which maximizes the expected value of* $\mathrm{L}(\mathrm{Y_{obs}}, \mathrm{Y_{mis}}, \theta)$ *by*

considering Y_{mis} *as a random variable with some known distribution.* The appropriate formulae can be derived by imagining that the sample is replicated an arbitrarily large number of times, with Y_{obs} taking the same value in all replications but with Y_{mis} having its known distribution. This idea is central to the *Missing Information Principle* which is now described.

Let $f(Y_{mis} \mid Y_{obs};\theta)$ denote the probability density function (PDF) for the conditional distribution of Y_{mis} given Y_{obs} and θ, and as usual let $\ell(Y_{mis} \mid Y_{obs};\theta)$ denote $\ln f(Y_{mis} \mid Y_{obs};\theta)$. Then

$$\ell(Y_{mis}, Y_{obs};\theta) = \ell(Y_{obs};\theta) + \ell(Y_{mis} \mid Y_{obs};\theta). \tag{2.64}$$

Now take any assumed value θ_A for θ. This, together with the observed value of Y_{obs}, defines a distribution for Y_{mis}, and we can now take the expectations of both sides of (2.64) by integrating out with respect to Y_{mis}.

If the distribution of Y_{mis} has a probability density element $f(Y_{mis} \mid Y_{obs};\theta_A)dY_{mis}$, then expectation of (2.64) becomes

$$\begin{aligned}\int \ell(Y_{mis}, Y_{obs};\theta) f(Y_{mis} \mid Y_{obs};\theta_A) dY_{mis} &= \int \ell(Y_{obs};\theta) f(Y_{mis} \mid Y_{obs};\theta_A) dY_{mis} \\ &\quad + \int \ell(Y_{mis} \mid Y_{obs};\theta) f(Y_{mis} \mid Y_{obs};\theta_A) dY_{mis} \\ &= \ell(Y_{obs};\theta) + E\{\ell(Y_{mis} \mid Y_{obs};\theta) \mid Y_{obs};\theta_A\}. \end{aligned} \tag{2.65}$$

We can now find the value θ_M of θ that maximizes the left-hand side of (2.65). This may depend on θ_A, so we can write

$$\theta_M = \Phi(\theta_A). \tag{2.66}$$

(2.66) represents a transformation from the vector θ_A to the vector θ_M. We now define the Missing Information Principle.

The *Missing Information Principle*:

Estimate θ by a fixed point of the transformation Φ, namely a value of θ such that

$$\theta = \Phi(\theta), \tag{2.67}$$

which is called the "fixed point equation".

We now apply the above theory to our problem of the multivariate normal missing data. Denote by **X** the (Nxn) matrix of the n_c-complete cases, by P_i the set of variables observed in the i-th unit, and by P_T the total set of variables observed, i.e., the n_c-complete cases plus the observed variables in the i-th unit (P_i). Then in the above notation

$$\theta = (\mu, \Sigma),\ \theta_A = (\mu_A, \Sigma_A),\ \theta_M = \Phi(\theta_A) = (\mu_M, \Sigma_M).$$

The loglikelihood for the multivariate normal distribution is

$$\ell(\mathrm{X}; \mu, \Sigma) = -\frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{N}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}(\mathrm{x_{ij}} - \mu_{\mathrm{j}})\sigma^{\mathrm{jk}}(\mathrm{x_{ik}} - \mu_{\mathrm{k}}) - \frac{1}{2}\mathrm{N}\log(\det\Sigma),$$

where σ^{jk} denotes the jk-th element of Σ^{-1}. Taking expectations with $\theta = \theta_A$ and the known variables fixed,

$$E\{\ell(\mathrm{X}; \mu, \Sigma \mid \mathrm{P_T}; \mu_\mathrm{A}, \Sigma_\mathrm{A})\} = \Bigg\{ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}\sum_{k=1}^{n}(\hat{\mathrm{x}}_{ijA} - \mu_j)(\hat{\mathrm{x}}_{ikA} - \mu_k) + \sigma_{jkA.P_i}\Bigg\}\sigma^{jk} - \frac{1}{2}N\log(\det\Sigma),$$

where

$$\hat{\mathrm{x}}_{ijA} = E\{\mathrm{x_{ij}} \mid \mathrm{P_i}; \mu_\mathrm{A}, \Sigma_\mathrm{A}\}$$

and

$$\sigma_{jkA.P_i} = \mathrm{cov}\left\{x_{ij}, x_{ik} \mid P_i; \mu_A, \Sigma_A\right\}.$$

Maximizing with respect to μ and Σ gives the analog of (2.66), i.e.,

$$\mu_{jM} = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_{ijA}$$

$$\sigma_{jkM} = \frac{1}{N}\sum_{i=1}^{N}\left\{(\hat{x}_{ijA} - \mu_{jM})(\hat{x}_{ikA} - \mu_{kM}) + \sigma_{jkA.P_i}\right\},$$

for $1 \leq j, k \leq n$. Now set $\mu_A = \mu_M = \mu$, $\Sigma_A = \Sigma_M = \Sigma$. The fixed point equations are

$$\hat{x}_{ij} = E\left\{x_{ij} \mid P_i; \mu, \Sigma\right\}, \tag{2.68}$$

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_{ij}, \tag{2.69}$$

$$\sigma_{jk} = \frac{1}{N}\sum_{i=1}^{N}\left\{(\hat{x}_{ij} - \mu_j)(\hat{x}_{ik} - \mu_k) + \sigma_{jk.P_i}\right\}, \tag{2.70}$$

$$\sigma_{jk.P_i} = \mathrm{cov}(x_{ij}, x_{ik} \mid P_i; \mu, \Sigma). \tag{2.71}$$

These are the formulae found by Orchard and Woodbury (1972). To find the maximum likelihood estimates we obtain initial estimates of μ and Σ and iterate between (2.68) and (2.71) until we find no significant changes in the estimates between two successive iterations. At each iteration the data are completed by (2.68), and the means and sum of squares and products matrix found for the variables. This matrix is adjusted by adding $\sigma_{jk.P_i}$ for every observation i to the jk-th element. This adjustment is zero unless both x_{ij} and x_{ik} are missing.

Noting that the above process involves an expectation step and a maximization step, Dempster, Laird and Rubin (1977) proved a couple of theorems that justify (2.68) and called it the expectation-maximization "EM" algorithm. They also discussed the application of the method to: grouped, censored or truncated data, finite mixture models, variance component estimation, hyper-parameter estimation, iteratively reweighted least squares and factor analysis.

Details of the theory and application of the method are given by Hartley (1958), Dempster, Laird and Rubin (1977), and McLachlan and Krishnan (1997).

An interesting property of the EM algorithm is that it combines two seemingly different strategies for handling missing data, that is, imputation strategy and ML approach. The relationship between the EM algorithm and imputation techniques (Buck's method) has been developed in chapter V.

2.5 CONCLUSIONS

Three main strategies for handling missing data in multivariate analysis have been discussed in this chapter; namely, deletion-pairwise strategy, imputation strategy, and maximum likelihood approach. Clearly, none of these strategies can be considered as a universal solution to the problem of missing data. The appropriate method of analysis depends on the specific situation under consideration. Buck's method and the EM algorithm can be considered as combinations of the three main strategies. The former is a combination of deletion and imputation strategies and the latter is a combination of imputation and ML strategies. These hybrid methods have been specifically designed to overcome some of the limitations of the individual strategies.

In the course of this thesis, special attention will be given to these hybrid methods. In the next two chapters a critical review of Buck's method as well as some contributions shall be considered. The EM algorithm and Anderson's factorization method and their relation to imputation techniques (Buck's method) will be dealt with in chapter V.

CHAPTER III

CRITICAL ANALYSIS OF BUCK'S METHOD

3.1 INTRODUCTION

Estimation of statistical parameters from multivariate data with missing observations via the deletion-case-wise strategy may result in wasted information. This is because units with incomplete data are rejected entirely as outlined in chapter II. This seems to be unsatisfactory especially if many variables are known for an incomplete unit. Also we have seen that the use of all-available-data method may result in inconsistencies in the covariance matrix. Moreover, this strategy does not address the problem of missing values itself in the sense that it offers no estimates for the missing values. Furthermore, the deletion strategies are only valid under the restrictive assumption that the missing observations are missing completely at random (MCAR). This assumption is rarely satisfied in reality.

Maximum likelihood approach, on the other hand, seeks to estimate the parameters, not the missing values, under certain distribution assumptions; in most cases the normality assumptions.
This seems to be satisfactory for types of analyses that do not make recourse to the original data after estimating the parameters from them. This is particularly true for many multivariate statistical analyses including multiple regression, factor analysis, canonical correlation and discriminant analysis, but not explanatory data methods, e.g., construction of histograms.
Other limitations of this approach are:

i) It gives maximum likelihood estimators for certain special patterns of missing observations; thus the method is not very general.

ii) As the number of variables increases the maximum likelihood solution

could become too involved for practical purposes and in most of the cases no explicit solutions can be found.

iii) The fundamental objection is that most multivariate data cannot be regarded as samples from multivariate normal distributions.

Taking into account the limitations given above, Buck (1960) developed a method for estimating missing values in multivariate data. An interesting property of the method is that it is a combination of deletion and imputation strategies. It therefore allows for the estimation of both the missing values and the parameters. The purpose of the method as stated by the author is "*to give a method of estimating the variance-covariance matrix of any k-variate population*" (Buck (1960), pp. 303, second paragraph). Thus the method does not make any assumptions about the distribution of the incomplete data nor about the missingness mechanism. The suggested method, as shown by the author, "*gives unbiased covariances but the variances need correction for bias*" (Buck (1960), pp. 302, first paragraph). Accordingly, he derived the bias and adjusted for it.

3.2 SPECIFIC ISSUES

The method of Buck, as a pioneering work in the area of missing data has, however, been a subject of debate among different scholars at various times. The following are some of the issues that have been raised in the literature about the method

1- Afifi and Elashoff (1966) stated that: "*Buck carries out his calculations conditional upon the complete vector observations. The particular reasons for this conditioning are not clear*" (see Afifi and Elashoff (1966), pp. 600, second paragraph).

2- Kim and Curry (1977) argued that: "*because his conclusion is based on the examination of a single data set (containing 72 cases and 4 variables*

from which he randomly deleted a few cases from each variable resulting in a total loss of 34 cases) and a single simulation, his conclusion should not be taken seriously" (see Kim and Curry (1977), pp. 222–223).

Little and Rubin (1987) stated that:

3- "*The filled-in data from Buck's method yield reasonable estimates of means, particularly if the multivariate normality assumptions are plausible*".

4- "*The sample covariance matrix from the filled-in data underestimates the sizes of the variances and covariances, ...*"

5- "*If we assume MCAR and ignore sampling variability of the estimates of μ and Σ based on the complete cases, then the conditional means imputed by Buck's method are the best point estimates of the missing values in the sense of minimizing the expected squared error*".

(for 3, 4, and 5 above see Little and Rubin (1987), pp. 44–47).

Most of the above mentioned statements given by the various scholars were not supplemented by either rigorous theoretical proofs or repeated simulation runs that make them globally accepted.

It is our objective in this work to enter the debate about Buck's method by trying to verify some of the above mentioned statements. Other results given by the author himself concerning the bias of the estimated parameters will also be examined.

To achieve this, in this chapter we introduce the method, illustrate its workability, and highlight the issues and reasons for disagreement. In the next chapter we start our investigation on some of the raised issues and give some contributions in some aspects of the method. Our approach will be based on both theoretical investigations and numerical validation.

For the theoretical investigations, multivariate regression and multivariate analysis of variance (MANOVA) are the main tools of analysis.

As for the numerical validation, we consider data collected by Bumpus (1898) who studied the effect of natural selection on sparrows by taking a sample of 21 survivors and 28 non-survivors. Eight morphological measurements on each bird were taken. In our study we shall use the data for five of the variables given in the appendix, table A(1), which is extracted from Manly (1986).

3.3 BASIC IDEA OF BUCK'S METHOD

Let (x_{ij}), $(i = 1,2,\ldots,n;\ j = 1,2,\ldots,k)$ represent the sample of n units, on each of which it is desired to have measurements on k variables. The observations x_{ij} can be represented in the form of an nxk matrix, $\mathbf{X}$, in which some of the elements are missing. Without loss of generality, assume that the last $n - n_c$ units have missing entries. Thus we write

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_c 1} & x_{n_c 2} & \cdots & x_{n_c k} \\ ? & x_{n_c 2+1} & \cdots & x_{n_c k+1} \\ \vdots & ? & \vdots & ? \\ x_{n1} & x_{n2} & ? & x_{nk} \end{bmatrix}, \tag{3.1}$$

where ? denotes a missing value, and

$n \equiv$ The total number of units (cases),

$n_c \equiv$ The number of complete cases (i.e., with all variables observed).

We can write (3.1) as

$$\mathbf{X} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k), \tag{3.2}$$

where

$\underline{x}_j$ is the j-th (column) vector, for $j = 1, 2, \ldots, k$.

Let $\mathbf{X}_{(j)}$ denote the matrix $\mathbf{X}$ without the j-th column vector, i.e.,

$$\mathbf{X}_{(j)} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{j-1}, \underline{x}_{j+1}, \ldots, \underline{x}_k).$$

We should note that $\mathbf{X}$ is nxk while $\mathbf{X}_{(j)}$ is $nx(k-1)$.
Similarly, $\mathbf{X}_{(jk)}$ denotes the matrix $\mathbf{X}$ without the j-th and k-th columns.

The matrix (3.1) can also be written, in a partitioned form, as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_c \\ \\ \mathbf{X}_{n-n_c} \end{bmatrix},$$

where $\mathbf{X}_c = (x_{ij})$, $i = 1, \ldots, n_c$; $j = 1, \ldots, k$ is obtained from $\mathbf{X}$ by considering its complete cases. And $\mathbf{X}_{n-n_c} = (x_{ij})$, $i = n_c + 1, \ldots, n$; $j = 1, \ldots, k$ gives the incomplete cases of $\mathbf{X}$. In most cases, we assume that the columns of both $\mathbf{X}_c$ and the completed (after imputation) data matrix $\mathbf{X}$ are centered so that $\bar{x}_j = 0$
Other notations are:

$n^{(j)} \equiv$ The number of the recorded (available) observations in the j-th variable

$n^{(jk)} \equiv$ The number of the recorded (available pairwise) observations in the j-th and k-th variables.

$m^{(j)} \equiv$ The number of missing observations in the j-th variable

$m^{(jk)} \equiv$ The number of missing (missing pairwise) observations in the j-th and k-th variables.

Then the problem is to devise a method for estimating the covariance matrix of the k-variate population on the basis of the incomplete data matrix $\mathbf{X}$.

The basic idea of Buck's method consists of estimating the missing values in the sample by regression techniques, based on the matrix of the complete cases ($\mathbf{X}_c$). The imputed values are then used to calculate a revised variance-covariance matrix. The method starts by calculating the expected values of x_{rj}, for $r = 1, 2, \ldots, n_c$, by forming, for each value of j in $\mathbf{X}_c$, the multiple regression of the j-th variable on the other $k - 1$ variables. Therefore, we obtain k regression equations which can be expressed in the form

$$E(x_{rj}) = f_j\{x_{r1}, x_{r2}, \ldots, x_{r(j-1)}, x_{r(j+1)}, x_{r(j+2)}, \ldots, x_{rk}\};\ j = 1, \ldots, k \quad (3.3)$$

where f_j's are taken to be linear functions.

Equations (3.3) are used to estimate the missing values as follows: If the i-th case has its j-th variable missing, we can estimate its value x_{ij}, by using one of the equations (3.3) substituting x_{ij} for x_{rj}, i.e.,

$$E(x_{ij}) = f_j\{x_{i1}, x_{i2}, \ldots, x_{i(j-1)}, x_{i(j+1)}, x_{i(j+2)}, \ldots, x_{ik}\};\ j = 1, \ldots, k \quad (3.4)$$

The method is easily extended to the case in which the units (cases) have more than one missing value, as follows: If v variates are missing, then from $\mathbf{X}_c$ we require to calculate the multiple regression equations for each missing variable on the remaining $k - v$ other variates. If any combination of v variates may be missing, then,

$$k\binom{k-1}{v-1} \quad (3.5)$$

possible regression equations have to be calculated, and a missing value is estimated by its expected value obtained from the correctly chosen regression

equation. Clearly, the method will collapse altogether if each case has a missing value.

Although the widespread use of personal computers and statistical packages have greatly facilitated the computation of the regression coefficients in Buck's method, yet the method of Woolf will be considered, in some detail, in the next section. This is because most of the theoretical derivations in the subsequent sections are based upon this method.

3.4 COMPUTATION OF THE REGRESSION COEFFICIENTS: WOOLF'S PROCEDURE

For the data matrix $\mathbf{X}$, let $\mathbf{A}$ be a kxk matrix that denotes the covariance matrix of $\mathbf{X}_c$, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kk} \end{bmatrix}$$

We can write $\mathbf{A}$ as

$$\mathbf{A} = (a_{sj}) = \begin{bmatrix} a_{11} & \underline{\alpha}_1' \\ \underline{\alpha}_1 & \mathbf{C} \end{bmatrix} \quad (s, j = 1, \dots, k) \tag{3.6}$$

where

$$a_{11} = V(\underline{x}_1)$$

$$\underline{\alpha}_1' = (a_{12}, a_{13}, \dots, a_{1k})$$

and

$$\mathbf{C} = \begin{bmatrix} a_{22} & \dots & a_{2k} \\ \vdots & \vdots & \vdots \\ a_{k2} & \dots & a_{kk} \end{bmatrix} = \mathbf{A}_{(1)}, \tag{3.7}$$

where $\mathbf{A}_{(1)}$ is obtained from $\mathbf{A}$ by deleting its first column. However, by symmetricity of $\mathbf{A}$, this implies deleting the first row of $\mathbf{A}$ also.
Then from $\mathbf{X}_c$, the regression coefficients of x_1 on $x_2, \ldots, x_k$ are given by

$$\underline{\hat{\beta}}_1' = \underline{\alpha}_1' \mathbf{C}^{-1}, \tag{3.8}$$

By writing $\mathbf{X}_{(1)} = (\underline{x}_2, \underline{x}_3, \ldots, \underline{x}_k)$, of order $nx(k-1)$, we can estimate the value of $\underline{x}_1$ for those cases which have $\underline{x}_1$ only missing by taking

$$\hat{\underline{x}}_1 = \mathbf{X}_{(1)} \hat{\underline{\beta}}_1, \tag{3.9}$$

Generally, we estimate the value of $\underline{x}_j$ for those cases which have $\underline{x}_j$ only missing by taking

$$\hat{\underline{x}}_j = \mathbf{X}_{(j)} \hat{\underline{\beta}}_j, \tag{3.10}$$

where $\mathbf{X}_{(j)} = (\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{j-1}, \underline{x}_{j+1}, \underline{x}_{j+2}, \ldots, \underline{x}_k)$.
And, the regression coefficients of x_j on the remaining $k-1$ other variables are obtained from $\mathbf{X}_c$ as

$$\underline{\hat{\beta}}_j' = \underline{\alpha}_j' \mathbf{C}^{-1} \tag{3.11}$$

with

$$\underline{\alpha}_j' = (a_{js}), \quad s(\neq j) = 1, 2, \ldots, k \tag{3.12}$$

and $\mathbf{C}$ is the covariance matrix of $\mathbf{X}_{(j)}$ based on the n_c complete-cases. We should note that we are using $\mathbf{C} = \mathbf{A}_{(j)}$ for any j. However, it will be clear from the context which row and column will be deleted. We should also note that the application of Woolf's method for the computation of the

regression coefficients of x_j on $\mathbf{X}_{(j)}$ requires the determination of $\underline{\alpha}_j$ and its corresponding $\mathbf{C}$. The vector $\underline{\alpha}_j$ is obtained from the j-th column of $\mathbf{A}$ by deleting its jj-th element. The corresponding $\mathbf{C}$ is obtained from $\mathbf{A}$ by deleting its j-th row and column.

3.5 DETERMINATION OF BIAS IN THE VARIANCE-COVARIANCE MATRIX

3.5.1 THE VARIANCES

From (3.8) and (3.9) above we have,

$$\begin{aligned} V(\hat{\underline{x}}_1) &= V(\mathbf{X}_{(1)}\hat{\underline{\beta}}_1) = V(\mathbf{X}_{(1)}\mathbf{C}^{-1}\underline{\alpha}_1) \\ &= \underline{\alpha}_1'\mathbf{C}^{-1}V(\mathbf{X}_{(1)})\mathbf{C}^{-1}\underline{\alpha}_1 \\ &= \underline{\alpha}_1'\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\underline{\alpha}_1 \\ &= \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1. \end{aligned} \tag{3.13}$$

If we write

$$\mathbf{A}^{-1} = \begin{bmatrix} c_{11} & \underline{e}_1' \\ \underline{e}_1 & \mathbf{F} \end{bmatrix}, \tag{3.14}$$

where c_{11} is the first element in $\mathbf{A}^{-1}$, then since $\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}$, it follows that

$$\begin{bmatrix} a_{11} & \underline{\alpha}_1' \\ \underline{\alpha}_1 & \mathbf{C} \end{bmatrix}\begin{bmatrix} c_{11} & \underline{e}_1' \\ \underline{e}_1 & \mathbf{F} \end{bmatrix} = \mathbf{I}_k. \tag{3.15}$$

Hence,

$$a_{11}c_{11} + \underline{\alpha}_1'\underline{e}_1 = 1, \tag{3.16}$$

and

$$\underline{\alpha}_1 c_{11} + \mathbf{C}\underline{e}_1 = \underline{0}. \tag{3.17}$$

From (3.17) $\underline{e}_1 = -\mathbf{C}^{-1}\underline{\alpha}_1 c_{11}$. Substituting this value in (3.16) we get

$$c_{11} = (a_{11} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1)^{-1}$$

or

$$c_{11}^{-1} = a_{11} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1, \tag{3.18}$$

where c_{11} is the first element in $\mathbf{A}^{-1}$.
Subsequently

$$V(\hat{\underline{x}}_1) = V(\underline{x}_1) - c_{11}^{-1}, \tag{3.19}$$

which implies that the variance of the estimated values of x_1 is less than the variance of the actual values of x_1 by the amount c_{11}^{-1}. This difference, in terms of expectations, gives the biasedness of the post-imputation variance of x_1. Therefore, if the value of $\underline{x}_1$ is missing for a proportion λ_1 of all units, and the predicted values are substituted and a new covariance matrix is calculated, then the obtained post-imputation variance a_{11}^* underestimates the actual variance, $V(\underline{x}_1)$, by the amount $\lambda_1 c_{11}^{-1}$. Thus the post-imputation variance of $\underline{x}_1$ is to be corrected for bias by adding to it the amount λ_1/c_{11}.

The general correction to this bias is as follows:
If $\underline{x}_j$ is missing for a proportion λ_j of all units, then the obtained post-imputation variance a_{jj}^* is adjusted to

$$a_{jj}^* + \lambda_j/c_{jj}, \tag{3.20}$$

where c_{jj} is a diagonal element in $\mathbf{A}^{-1}$ and,

$$\lambda_j = \frac{m^{(j)}}{n} = \frac{n - n^{(j)}}{n}$$

is the proportion of missing values in the j-th variable,

where

$$n^{(j)} = \text{The number of the recorded values in variable j}$$

$$m^{(j)} = \text{The number of missing values in variable j}$$

Remark

Note that for each x_j we have

$$0 \leq \lambda_j \leq 1,$$

thus the amount of bias $(\lambda_j c_{jj}^{-1})$ is such that

$$\text{Bias} = \begin{cases} 0, & \text{if } \lambda_j = 0, \\ c_{jj}^{-1}, & \text{if } \lambda_j = 1, \\ \lambda_j c_{jj}^{-1}, & \text{if } 0 < \lambda_j < 1. \end{cases}$$

Notice that $\lambda_j = 0$ corresponds to the case of no missing values in x_j, and $\lambda_j = 1$ corresponds to the case where all values of x_j are imputed. In other words, c_{jj}^{-1} can be viewed as the maximum amount of bias that occurs when all values of x_j are imputed.

3.5.2 THE COVARIANCES

To prove the unbiasedness of the covariances, Buck (1960) proceeded as follows: Recall that the covariance matrix of $\mathbf{X}_c$ is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

which can be written as

$$\mathbf{A} = (a_{sj}) = \begin{bmatrix} a_{11} & \underline{\alpha}_1' \\ \underline{\alpha}_1 & \mathbf{C} \end{bmatrix} \quad (s, j = 1, \ldots, k),$$

where

$$a_{11} = V(\underline{x}_1)$$

$$\underline{\alpha}_1' = (a_{12}, a_{13}, \ldots, a_{1k}) = \mathrm{Cov}(\underline{x}_1, \mathbf{X}_{(1)}), \tag{3.21}$$

thus $\underline{\alpha}_1'$ gives the covariance of x_1 and $\mathbf{X}_{(1)} = (\underline{x}_2, \underline{x}_3 \ldots, \underline{x}_k)$ computed from the n_c-complete cases. And

$$\mathbf{C} = \begin{bmatrix} a_{22} & \ldots & a_{2k} \\ \vdots & \vdots & \vdots \\ a_{k2} & \ldots & a_{kk} \end{bmatrix}$$

is the covariance matrix of the complete cases of $\mathbf{X}_{(1)} = (\underline{x}_2, \underline{x}_3 \ldots, \underline{x}_k)$, of order (k-1)x(k-1).

We estimate the value of x_1 for those cases which have $\underline{x}_1$ only missing, by taking

$$\hat{\underline{x}}_1 = \mathbf{X}_{(1)} \hat{\underline{\beta}}_1, \tag{3.22}$$

where the regression coefficients of $\underline{x}_1$ on $\mathbf{X}_{(1)} = (\underline{x}_2, \underline{x}_3 \ldots, \underline{x}_k)$ are obtained from $\mathbf{X}_c$ as

$$\hat{\underline{\beta}}_1' = \underline{\alpha}_1' \mathbf{C}^{-1} \tag{3.23}$$

From (3.22) and (3.23) we have

$$\hat{\underline{x}}_1 = \mathbf{X}_{(1)} \mathbf{C}^{-1} \underline{\alpha}_1. \tag{3.24}$$

from which the covariance of $\hat{\underline{x}}_1$ with $\mathbf{X}_{(1)} = (\underline{x}_2, \underline{x}_3 \ldots, \underline{x}_k)$ is obtained as

$$\begin{aligned} \text{Cov}(\hat{\underline{x}}_1, \mathbf{X}_{(1)}) &= \underline{\alpha}_1' \mathbf{C}^{-1} \text{Cov}(\mathbf{X}_{(1)}) \\ &= \underline{\alpha}_1' \mathbf{C}^{-1} \mathbf{C} \\ &= \underline{\alpha}_1'. \end{aligned} \tag{3.25}$$

Comparing (3.25) with (3.21) it follows that

$$\text{Cov}(\hat{\underline{x}}_1, \mathbf{X}_{(1)}) = \text{Cov}(\underline{x}_1, \mathbf{X}_{(1)}). \tag{3.26}$$

That is, the expected values of the covariance elements in **A** are the same for both actual and predicted values of $\mathbf{x}_1$.

It is from here that Buck (1960) concluded that "*...with this method the resultant covariances are unbiased, but the variances need correction for bias*". As a numerical illustration of his method, Buck considered a set of data consisting of 72 units on which 4 variables were measured. From the total 288(72x4) observations, 35 observations were picked at random and considered as missing. It is from here that Kim and Curry (1977) concluded that the results of Buck should not be taken seriously since it is based on a single set of data.

The method of Buck has been a subject of some modifications at various times. Kasap (1973), in an unpublished M.Sc. thesis suggested the following modification: "*after substituting the estimated values for the missing observations, repeat the process until successive iterations fail to change materially. The values obtained at the final iteration are then taken as estimates of the missing values*".

Another interesting imputation technique that looks very similar to the method of Buck was brought to our attention by Kasap (1973). According to Kasap, the method was studied in detail by Federspiel (1959) in an

unpublished Ph.D. thesis. Obviously, the method was not meant to be an extension of Buck's method as it was developed independently as indicated by the dates. However, for completeness, we present it here as a possible modification. The main idea of the method was summarized by Kasap (1973) as follows: "*A variable for which there are missing observations is treated as the dependent variable for a regression analysis, using all other variables as independent variables with mean values substituted for missing observations. A new estimate of each missing observation is calculated and substituted using this equation. This operation is done for each variable for which there are missing observations. The computation of regression equations and the substitution of new values is repeated until the coefficients for the regression equations in successive iterations converge. These procedures are known to have produced biased estimates of β's and its variances which cannot be regarded as sampling phenomena*".

We note that this method starts with unconditional imputation and then proceeds as a conditional imputation method until convergence occurs. On the other hand Buck's method starts with casewise-deletion method and proceeds as a conditional imputation method without convergence criteria.

It is clear that the modification of Kasap (1973) is an iterative Buck's method, that is, the original method plus a convergence criteria. Kasap does not elaborate on the theoretical rationale of his modification. However, he compared the predictive efficiency of the original method and his modified version and concluded that his modified version does not improve the efficiency of prediction.

3.6 UTILIZATION OF THE AVAILABLE DATA IN BUCK'S METHOD

In section 3.3 we have seen that the imputation process in Buck's method is based on the n_c completely observed units. The partially observed units play no role in the estimation of the missing values, they are completely discarded from the analysis. This can be seen as waste of the available (recorded) information. In some extreme cases this might lead to the collapse of the method altogether. For instance, for a pattern of missingness where each unit has a missing value ($n_c = 0$), the method is inapplicable. In this section we shall try to see how Buck's method makes use of the available (recorded) observations in each variable. This will be studied by considering the following patterns of missingness which are seen to be exhaustive.

i- One variable is subject to missingness

ii- More than one variable are subject to missingness, but units can have only one missing value. For simplicity, this case will be referred to as "the case of units with one missing value".

iii- Any combination of the k variables can be missing in each unit. This will be referred to as "the case of units with more than one missing value".

Consider a sample $\mathbf{X} = (x_{ij})$, $i = 1, \ldots, n$; $j = 1, \ldots, k$ of n units (cases), on each of which it is desired to have measurements on k variables. Assume that each unit can have more than one missing value, that is, case (iii) above. Recall our notation of section 3.3.

Then, for each $j = 1, \ldots, k$ we have

$$n^{(j)} = n - m^{(j)} \tag{3.27}$$

Now let $U^{(i)}$= Number of units with 'i' missing values, $i > 1$.

Then the number of units with the j-th variable observed and any combination of the remaining $k-1$ variables missing is given by

$$n - n_c - m^{(j)} = \sum_{\ell \neq j}^{k} m^{(\ell)} - \sum_{i=2}^{k-1} (i-1)U^{(i)}. \tag{3.28}$$

Thus

$$\begin{aligned} n_c &= n - \sum_{j=1}^{k} m^{(j)} + \sum_{i=2}^{k-1} (i-1)U^{(i)} \\ &= n - \{\sum_{\ell \neq j}^{k} m^{(\ell)} + m^{(j)}\} + \sum_{i=2}^{k-1} (i-1)U^{(i)} \\ &= n^{(j)} - \sum_{\ell \neq j}^{k} m^{(\ell)} + \sum_{i=2}^{k-1} (i-1)U^{(i)}, \end{aligned} \tag{3.29}$$

since $n^{(j)} = n - m^{(j)}$.

Then for the case of one variable (the j-th variable) subject to missingness we have

$$\sum_{\ell \neq j}^{k} m^{(\ell)} = \sum_{i=2}^{k-1} (i-1)U^{(i)} = 0. \tag{3.30}$$

Substituting (3.30) in (3.29) above we have

$$n_c = n^{(j)}.$$

Thus, for the case of one variable subject to missingness, Buck's method makes maximum utilization of the available (recorded) observations in that variable. In other words, all the available observations in the j-th variable participate in the imputation of the missing values.

For the case of units with one missing value, the number of complete cases is obtained from (3.29) by setting $\sum_{i=2}^{k-1}(i-1)U^{(i)} = 0$. That is,

$$n_c = n^{(j)} - \sum_{\ell \neq j}^{k} m^{(\ell)}. \tag{3.31}$$

Thus, in this case the number of complete cases is less than the available cases on the j-th variable by $(\sum_{\ell \neq j}^{k} m^{(\ell)})$ cases. In other words, the method of Buck discards from the analysis $\sum_{\ell \neq j}^{k} m^{(\ell)}$ available number of units on the j-th variable. These non-missing values of the j-th variable play no role in the estimation of the missing values. Thus Buck's method does not make full utilization of the available information. The same conclusion is also true for the case of units with more than one missing value as can be seen from (3.29) above.

The modification by Federspiel (1959), discussed in the previous section was actually an attempt to maximize the utilization of the available observations in Buck's method by incorporating the $(\sum_{\ell \neq j}^{k} m^{(\ell)} - \sum_{i=2}^{k-1}(i-1)U^{(i)})$ partially observed units in the imputation process. Clearly, this method is a good alternative for patterns of missingness where Buck's method collapses when all cases (units) are partially observed.

Here, we suggest the following modification which was later found to have been considered by Chan and Dunn (1972) in the context of discriminant analysis. The idea is also to improve the utilization of the available data by incorporating the partially observed units in the analysis. This can be summarized as follows:

1- On the basis of the n_c-complete units, impute the missing values for those units with one missing value $(U^{(1)})$.

2- Use $n_c + U^{(1)}$ to impute the missing values for those units with two

missing values, and so on until the missing values for all patterns are imputed.

Obviously, this modification can be made iteratively by repeating the whole process until the imputed values in two successive iterations are not materially different. A simulation study might then help to compare the performance of the original method and the modified version. It is worth mentioning that Chan and Dunn (1972) have conducted a simulation study to compare the performance of the above modification (without iterations) and the original method. Their conclusion is that there is no preference between the two methods in minimizing the probabilities of misclassification in discriminant analysis.

3.7 REAL DATA ILLUSTRATIONS

Adopting the notation of section 3.4, the use of Woolf's procedure in the method of Buck, and the determination of bias can be illustrated numerically as follows: From the data of table A(1) consider the following patterns of missingness (units with one missing value): Pattern (1), where 29 observations were picked at random and considered to be missing. Thus we were left with 20 complete samples, 6 cases with x_1 missing, 7 cases with x_2 missing, 5 cases with x_3 missing, 6 cases with x_4 missing and 5 cases with x_5 missing. Patterns (2), (3) and (4) were constructed respectively by randomly picking 4, 5 and 7 observations from each variable and consider them to be missing. Thus the number of completely recorded units for patterns (2), (3) and (4) were 33, 24 and 21 respectively.

The resulting incomplete data as well as the summary statistics (means and covariance matrix) for the complete cases of each of the above patterns are given in the appendix. Also displayed in the appendix are the results of multiple regression analysis, required for the estimation of missing values

in each pattern of missingness. The data are analyzed using *SPSS* and *STATGRAPHICS*.

For the missing data pattern (1) the regression coefficients of x_1 on $\mathbf{X}_{(1)} = (x_2, x_3, x_4, x_5)$ are obtained from the 20 complete cases by (3.8) as:

$$\underline{\hat{\beta}}'_{h1} = \alpha'_1 C^{-1}, \quad h = 2,3,4,5$$

where **C** is the covariance matrix of $\mathbf{X}_{(1)} = (x_2, x_3, x_4, x_5)$.
From table A(2.2) of the appendix we have

$$\mathbf{A} = \begin{bmatrix} 8.45000 & & & & \\ 4.6868 & 12.4500 & & & \\ .3982 & 1.0534 & .3510 & & \\ .2418 & .8813 & .1268 & .1571 & \\ .4887 & .8861 & .1856 & .0698 & .3373 \end{bmatrix}.$$

Therefore

$$\mathbf{C} = \begin{bmatrix} 12.4500 & & & \\ 1.0534 & .3510 & & \\ .8813 & .1268 & .1571 & \\ .8861 & .1856 & .0698 & .3373 \end{bmatrix}.$$

Thus, from **A** we have

$$\alpha'_1 = (4.6868 \quad .3982 \quad .2418 \quad .4887),$$

and

$$\mathbf{C}^{-1} = \begin{bmatrix} .15115 & -.103855 & -.675138 & -.200218 \\ -.103855 & 5.22325 & -2.72833 & -2.03668 \\ -.675138 & -2.92833 & 12.0035 & .790914 \\ -.200218 & -2.03668 & .790914 & 4.44772 \end{bmatrix}.$$

Thus for h=2,3,4,5

$$\underline{\hat{\beta}}'_{h1} = \underline{\alpha}'_1 \mathbf{C}^{-1} = (.406 \quad -.062 \quad -.961 \quad .615),$$

so that

$$\hat{\beta}_{01} = \overline{\mathrm{X}}_1 - \sum_{h=2}^{5} \hat{\beta}_{h1} \overline{\mathrm{X}}_h = 66.383,$$

which is the same result listed in table A(2.4) of the appendix obtained via the *SPSS*.

Similarly we obtain the regression coefficients of $\underline{x}_j$ on $\mathbf{X}_{(j)}$, for j = 2, 3, 4, 5, by choosing from **A** the appropriate $\underline{\alpha}$ and **C**.

To obtain the amount of bias in the estimation of $V(\underline{x}_j)$, using Buck's procedure, we have to obtain the covariance matrix of the n_c-complete cases and its inverse. Then the amount of bias is given by λ_j / c_{jj} where λ_j is the proportion of missing values in the j-th variable and c_{jj} is the first element in $\mathbf{A}^{-1}$, i.e.,

$$\mathbf{A} = \begin{bmatrix} 8.45000 & & & & \\ 4.6868 & 12.4500 & & & \\ .3982 & 1.0534 & .3510 & & \\ .2418 & .8813 & .1268 & .1571 & \\ .4887 & .8861 & .1856 & .0698 & .3373 \end{bmatrix}$$

and its inverse is

$$\mathbf{A}^{-1} = \begin{bmatrix} .153757 & & & & \\ -.062419 & .176489 & & & \\ .951513E-02 & -.107718 & 5.22384 & & \\ .147868 & -.735166 & -2.71918 & 12.1457 & \\ -.946304E-01 & -.161802 & -2.04254 & .699908 & 4.50596 \end{bmatrix}.$$

Thus the amount of bias in the estimation of $V(x_1)$ is given by

$$\lambda_1 c_{11}^{-1} = (6/49)/.153757 = .79638$$

where c_{11} is the first element in $\mathbf{A}^{-1}$.

Similarly, the biases of the variances of all variables for the incomplete data patterns (2), (3) and (4) are obtained and displayed in the following table:

Table 3(1): BIAS OF THE VARIANCES FOR ALL PATTERNS OF MISSINGNESS

	BIAS			
x_j	PATTERN (1)	PATTERN (2)	PATTERN (3)	PATTERN (4)
x_1	.7964	.3840	.4397	.3924
x_2	.8095	.7531	.6045	1.2358
x_3	.0195	.0211	.0226	.0256
x_4	.01009	.0070	.0060	.0106
x_5	.02265	.0537	.0508	.0532

3.8 CONCLUSIONS

In this chapter we have introduced the method of Buck and discussed the theoretical properties of the resulting estimates. The specific computations required for the application of the method have been illustrated using the procedure of Woolf (1951). Numerical illustrations of the method using Woolf's procedure are also given. We have also reviewed some studies that give either similar methods or modifications to Buck's method.

We have suggested a simple modification to Buck's method which is later found to have been considered by Chan and Dunn (1972). The utilization of the available observations by Buck's method is also studied. Moreover, we have enumerated and briefly described some of the major issues that have been raised in the literature about the method.

The verification of these issues of section 3.2 will be the subject matter of the next chapter.

CHAPTER IV

SOME CONTRIBUTIONS TO THE METHOD OF BUCK

4.1 INTRODUCTION

In this chapter we shall try to verify some of the unproved statements about the method of Buck given in chapter III. A simpler procedure for the determination of the biasedness in the variance-covariance matrix will also be derived. The developed procedure will enable us to determine the biasedness in the estimation of the variance of the j-th variable in the method of Buck with minimal computation. Apart from its relative ease of computation, the developed procedure has enabled us to create a functional relationship between the amount of bias, relative bias, and the coefficient of determination for each variable. Hence, the conditions under which the method of Buck gives better estimates (have minimum bias) can be highlighted on the basis of this functional relationship. Also some of the statistical properties of the obtained estimates will be discussed. For the covariances we shall re-examine the statement of Buck about its unbiasedness. The conditions under which the statement is true are given. A new formula for the correction of bias in the case of units with more than one missing value is outlined. The statistical consistency of the estimated covariances is also highlighted. The effect of the imputed data via Buck's method on estimating the correlation coefficient and the standard error of the regression coefficient are studied.

4.2 DETERMINATION OF BIAS IN THE VARIANCE-COVARIANCE MATRIX: UNITS WITH ONE MISSING VALUE

In this section we attempt to obtain and interpret the biasedness in the covariance matrix, obtained via Buck's method, from the standpoint of regression and analysis of variance techniques. The usefulness of this approach

is that it will enable us to examine the effect of imputed values on the variance by examining its effect on the components of the variance; that is, by examining the sample variance of the residuals and the variance of the estimated values. The case of units with more than one missing value will be studied separately as we shall approach it from the standpoint of multivariate regression and multivariate analysis of variance (MANOVA) techniques.

4.2.1 THE VARIANCES

The computation of the biasedness of the variances, as outlined in section 3.5.1, requires the computation of c_{jj}^{-1}, the inverse of the j-th diagonal element in $\mathbf{A}^{-1}$. Specifically, we have to compute the covariance matrix of the n_c-complete cases as well as its inverse. Then the amount of bias in the estimation of the variance of the j-th variable is taken as the product of the proportion of missing observations in the j-th variable and c_{jj}^{-1}. However, the same result can be obtained, with relative ease of computation, through a procedure given by the following lemma.

Lemma 4.1

The inverse of the j-th diagonal element in $\mathbf{A}^{-1}$, denoted by c_{jj}^{-1}, is the sample variance of the residuals of the regression of the j-th variable on the remaining k−1 other variables (Φ_{jj}); that is,

$$c_{jj}^{-1} = \Phi_{jj}$$

where

$$\Phi_{jj} = \frac{(\underline{x}_j'\underline{x}_j - \underline{\hat{\beta}}_j'\mathbf{X}_{(j)}'\underline{x}_j)}{n_c - 1}.$$

Proof

Consider the regression of $\underline{x}_1$ on $\mathbf{X}_{(1)} = (\underline{x}_2, \ldots, \underline{x}_k)$, i.e.,

$$\hat{\underline{x}}_1 = \mathbf{X}_{(1)}\hat{\underline{\beta}}_1,$$

which is based on the first n_c rows of $\mathbf{X}_{(1)}$. Then the total variability of the above model can be partitioned as

$$\underline{x}_1'\underline{x}_1 = \left[\hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1 + (\underline{x}_1'\underline{x}_1 - \hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1)\right], \tag{4.1}$$

where $\underline{x}_1'\underline{x}_1$ is the corrected total sum of squares, $\hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1$ is the component of variance explained by the regression of $\underline{x}_1$ on $\mathbf{X}_{(1)}$, and $(\underline{x}_1'\underline{x}_1 - \hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1)$ is the residual sum of squares.

Dividing (4.1) by $(n_c - 1)$ we get

$$\frac{\underline{x}_1'\underline{x}_1}{n_c - 1} = \frac{\hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1}{n_c - 1} + \frac{(\underline{x}_1'\underline{x}_1 - \hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1)}{n_c - 1}.$$

Therefore,

$$V(\underline{x}_1) = V(\hat{\underline{x}}_1) + \frac{(\underline{x}_1'\underline{x}_1 - \hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1)}{n_c - 1} \tag{4.2}$$

and from (3.18) we have

$$a_{11} = \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1 + c_{11}^{-1}$$

or

$$V(\underline{x}_1) = V(\hat{\underline{x}}_1) + c_{11}^{-1}. \tag{4.3}$$

Comparing (4.2) and (4.3) it follows that

$$c_{11}^{-1} = \frac{(\underline{x}_1'\underline{x}_1 - \hat{\underline{\beta}}_1'\mathbf{X}_{(1)}'\underline{x}_1)}{n_c - 1} = \Phi_{11}. \tag{4.4}$$

Hence, the amount of bias in the estimation of the variance of $\underset{\sim}{x}_1$ is actually a function of the sample variance of the residuals (based on $(n_c - 1)$ degrees of freedom) obtained from the regression of $\underset{\sim}{x}_1$ on $\mathbf{X}_{(1)}$. Thus if $\underset{\sim}{x}_1$ is missing for a proportion λ_1 of all units then the variance of the observed and imputed values of $\underset{\sim}{x}_1$ is to be corrected for bias by adding to it the amount $\lambda_1\Phi_{11}$. Here, note that Φ_{11} is readily obtained from the analysis of variance table of the regression of the first variable on the remaining $(k - 1)$ other variables by dividing the residual sum of squares by $(n_c - 1)$. Thus (4.4) will enable us to determine the bias without computing either the covariance matrix or its inverse.

Generally, the amount of bias in the estimation of the variance of $\underset{\sim}{x}_j$ can be written as

$$\lambda_j\Phi_{jj} = \lambda_j \frac{(\underset{\sim}{x}_j'\underset{\sim}{x}_j - \hat{\underset{\sim}{\beta}}_j'\mathbf{X}'_{(j)}\underset{\sim}{x}_j)}{n_c - 1}, \tag{4.5}$$

where $\mathbf{X}_{(j)} = (\underset{\sim}{x}_1, \ldots, x_{j-1}, x_{j+1}, x_{j+2}, \ldots, x_k)$ and λ_j is the proportion of missing values in the j-th variable.

Theorem 4.1

The amount of bias $(\lambda_j\Phi_{jj})$ of the variance of the j-th variable in the method of Buck is given by

$$(\Phi_{jj})^{(n_c)} - (\Phi_{jj})^{(n)} = \frac{n - n^{(j)}}{n - 1}\Phi_{jj}, \tag{4.6}$$

where $(\Phi_{jj})^{(n_c)}$ and $(\Phi_{jj})^{(n)}$ are the pre– and post-imputation sample variances of the residuals of the regression of the j-th variable on the remaining $k-1$ other variables respectively.

Proof

Let $\mathbf{X} = (\underline{x}_{il})$, $(i = 1, 2, \ldots, n; l = 1, 2, l, \ldots, k)$ represent the sample of n units, on each of which it is desired to have measurements on k variables. Further assume that the j-th variable is

$$x_{ij}, \ (i = 1, 2, \ldots, n_c, n_c + 1, \ldots, n),$$

where the first n_c units on the j-th variable are observed and the next $(n - n_c)$ units are missing. Then using Buck's method the $(n - n_c)$ missing observations on the j-th variable are to be estimated from the regression equation of the observed values, i.e.,

$$\hat{\underline{x}}_{ij} = \mathbf{X}_{(j)}\underline{\hat{\beta}}_j, \ \mathbf{X}_{(j)} = (\underline{x}_{i1}, \underline{x}_{i2}, \underline{x}_{i\,j-1}, \underline{x}_{i\,j+1}, \ldots, \underline{x}_{ik}), \ i = (1, 2, \ldots, n_c). \quad (4.7)$$

Using (4.7) we impute the $(n - n_c)$ missing values on the j-th variable. Now, consider the regression equation of both the observed and imputed values, i.e.,

$$\hat{\underline{x}}_{ij} = \mathbf{X}_{(j)}\underline{\hat{\beta}}_j, \ \mathbf{X}_{(j)} = (\underline{x}_{i1}, \underline{x}_{i2}, \underline{x}_{i\,j-1}, \underline{x}_{i\,j+1}, \ldots, \underline{x}_{ik}), \ i = (1, 2, \ldots, n). \quad (4.8)$$

From (4.7), the pre-imputation sample variance of the residuals is

$$(\Phi_{jj})^{(n_c)} = \frac{\sum_{i=1}^{(n_c)}(x_{ij} - \hat{x}_{ij})^2}{n_c - 1} = \Phi_{jj}, \quad (4.9)$$

and from (4.8) the post-imputation sample variance of the residuals is

$$\begin{aligned}(\Phi_{jj})^{(n)} &= \frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{n - 1} \\ &= \frac{\sum_{i=1}^{n_c}(x_{ij} - \hat{x}_{ij})^2}{n - 1} + \frac{\sum_{i=n_c+1}^{n}(x_{ij} - \hat{x}_{ij})^2}{n - 1} \\ &= \frac{\sum_{i=1}^{n_c}(x_{ij} - \hat{x}_{ij})^2}{n - 1} \end{aligned} \quad (4.10)$$

since $\sum_{n-n_c}(x_{ij} - \hat{x}_{ij})^2 = 0$ due to imputation. This is because the imputed values are projected exactly on the regression line given by (4.7).
Thus

$$
\begin{aligned}
(\Phi_{jj})^{(n_c)} - (\Phi_{jj})^{(n)} &= \frac{\sum_{i=1}^{n_c}(x_{ij} - \hat{x}_{ij})^2}{n_c - 1} - \frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2}{n - 1} \\
&= \sum_{i=1}^{n_c}(x_{ij} - \hat{x}_{ij})^2 \left(\frac{1}{n_c - 1} - \frac{1}{n - 1}\right) \\
&= \frac{n - n_c}{n - 1} \frac{\sum_{i=1}^{n_c}(x_{ij} - \hat{x}_{ij})^2}{n_c - 1} \\
&= \frac{n - n_c}{n - 1}(\Phi_{jj})^{(n_c)} \\
&= \frac{n - n_c}{n - 1}\Phi_{jj}. \qquad (4.11)
\end{aligned}
$$

It follows that if x_j is missing for $m^{(j)} = (n - n^{(j)})$ of all units then the amount of bias in the estimation of its variance from the completed data is given by

$$\frac{n - n^{(j)}}{n - 1}\Phi_{jj}. \qquad (4.12)$$

Note that (4.12) gives the difference between the pre– and post-imputation sample variances of the residuals. This difference, in terms of expectations, gives the bias of the post-imputation variance of the j-th variable.

Remark

Note that the only difference between this derivation and the one of Buck is that the adjusted sum of squares and products matrix is divided by $(n - 1)$ instead of n to derive the estimated covariance matrix. For large sample sizes with missing observations this difference is of no consequence. However, for small sample sizes ($n < 10$) the difference becomes more significant. Also

note that the amount of bias given by (4.12) above can be viewed as an estimate of the "missing" contribution of the imputed values to the sample variance of the residuals. It is worth noting that the contribution of the imputed values to the component of variance explained by regression is unaffected by imputation.

Theorem 4.2

$$\Phi_{jj} = a_{jj}(1 - R_j^2),$$

where a_{jj} is the variance of the j-th variable obtained from the n_c complete cases, and R_j^2 is the coefficient of determination of the regression equation $\hat{\underline{x}}_j = \mathbf{X}_{(j)}\hat{\underline{\beta}}_j$, where $\mathbf{X}_{(j)} = (\underline{x}_{i1}, \ldots, \underline{x}_{i(j-1)}, \underline{x}_{i(j+1)}, \underline{x}_{i(j+2)}, \ldots, \underline{x}_{ik})$, $i = 1, \ldots, n_c$.

Proof

From the regression equation

$$\underline{x}_j = \mathbf{X}_{(j)}\underline{\beta}_j + \underline{\epsilon}_j, \tag{4.13}$$

where $\mathbf{X}_{(j)} = (\underline{x}_{i1}, \ldots, \underline{x}_{i(j-1)}, \underline{x}_{i(j+1)}, \underline{x}_{i(j+2)}, \ldots, \underline{x}_{ik})$, $i = 1, \ldots, n_c$, we have

$$R_j^2 = 1 - \frac{(\underline{x}_j'\underline{x}_j - \hat{\underline{\beta}}_j'\mathbf{X}_{(j)}'\underline{x}_j)}{\underline{x}_j'\underline{x}_j}. \tag{4.14}$$

Substituting the value of $(\underline{x}_j'\underline{x}_j - \hat{\underline{\beta}}_j'\mathbf{X}_{(j)}'\underline{x}_j)$ from (4.5) we get

$$R_j^2 = 1 - \frac{(n_c - 1)\Phi_{jj}}{\underline{x}_j'\underline{x}_j},$$

which implies

$$\Phi_{jj} = \frac{\underline{x}_j'\underline{x}_j(1 - R_j^2)}{n_c - 1}. \tag{4.15}$$

We should note that $\underset{\sim}{x}_j'\underset{\sim}{x}_j$ is the corrected total sum of squares of the regression equation given by (4.13), hence

$$(\underset{\sim}{x}_j'\underset{\sim}{x}_j/n_c - 1) = a_{jj},$$

where a_{jj} is the variance of the j-th variable obtained from the n_c-complete cases. Therefore,

$$\Phi_{jj} = a_{jj}(1 - R_j^2). \tag{4.16}$$

Thus, from (4.16), the correction for bias is given by

$$\lambda_j\Phi_{jj} = \lambda_j a_{jj}(1 - R_j^2), \tag{4.17}$$

from which the relative bias is given by

$$\lambda_j \frac{\Phi_{jj}}{a_{jj}} = \lambda_j(1 - R_j^2). \tag{4.18}$$

Note that (4.16) gives the maximum amount of bias in the estimation of the variance of the j-th variable as a function of two forces, namely, R_j^2 and the variance of the j-th variable (a_{jj}) computed from the n_c-complete cases. This formula plays an important role in the investigation of the missingness mechanism in the method of Buck which will be discussed in section 4.6. Moreover, (4.18) gives the relative bias in the estimation of the variance of the j-th variable. Specifically, formula (4.18) says:

1- For fixed λ_j, the relative bias decreases as R_j^2 increases, and

2- For fixed R_j^2, the relative bias increases as λ_j increases.

Theorem 4.3

The estimated variances via Buck's method are inconsistent.

Proof

Since $\lambda_j = \frac{m^{(j)}}{n}$, the amount of bias in the estimation of the variance of the j-th variable can be written as

$$\lambda_j \Phi_{jj} = \frac{m^{(j)}}{n} \Phi_{jj}$$

$$= \left(\frac{n - n^{(j)}}{n} \right) \Phi_{jj}, \tag{4.19}$$

$$\therefore \lim_{n \to \infty} \lambda_j \Phi_{jj} = \lim_{n \to \infty} \left[\frac{n - n^{(j)}}{n} \Phi_{jj} \right] \neq 0, \tag{4.20}$$

unless $n^{(j)} = n$.

Thus the estimate of the variance given by Buck's method is generally inconsistent.

4.2.2 THE COVARIANCES

For this case of units with one missing value, the proof of the unbiasedness of the covariances was given by Buck (1960) as outlined in section 3.5.2. However, the biasedness of the covariances for the case of units with more than one missing value will be studied in section 4.4.1.

4.2.3 REAL DATA ILLUSTRATIONS

The equivalence of the results obtained via the alternative procedure and Buck's procedure, for the determination of bias, and the verification of the fact that the relative bias is a decreasing function of R^2 (for fixed λ)—can be illustrated numerically as follows:

To obtain the amount of bias in the estimation of $V(\bar{x}_j)$, using Buck's procedure, we have to obtain the covariance matrix of the n_c-complete cases and its inverse. Then the amount of bias is given by $\lambda_j c_{jj}^{-1}$ where λ_j is the proportion of missing values in the j-th variable and c_{jj}^{-1} is the inverse of the j-th element in $\mathbf{A}^{-1}$. These were obtained in section 3.7 and the results were listed in table 3(1). However, using theorem 4.2 above, the amount of bias can be directly obtained using (4.17). The Bias $= \lambda_j a_{jj}(1 - R_j^2)$ where the quantities, required for the estimation of the bias are readily available from the results of the regression and analysis of variance listed in the appendix.

The biases of the variances of all variables for the incomplete data patterns (1), (2), (3) and (4) are obtained and displayed in the following tables.

Table 4(1): COMPUTATION OF BIAS FOR THE MISSING DATA PATTERN (1)

x_j	a_{jj}	λ_j	R_j^2	Bias	%Bias
$\bar{x}_1$	8.4500	12.24%	.23032	.7964	9.4%
x_2	12.4500	14.29%	.54486	.8095	6.5%
x_3	.3510	10.20%	.45431	.0195	5.6%
x_4	.1571	12.24%	.47568	.01009	6.4%
x_5	.3373	10.20%	.34182	.02265	6.7%

Table 4(2): COMPUTATION OF BIAS FOR THE MISSING DATA PATTERN (2)

x_j	a_{jj}	λ_j	R_j^2	Bias	%Bias
$\bar{x}_1$	11.4803	8.18%	.62862	.3480	3%
x_2	25.7488	8.16%	.64170	..7531	2.9%
x_3	.4508	8.16%	.42763	.0211	4.7%
x_4	.2242	8.16%	.62023	.0070	3.1%
x_5	.9122	8.16%	.27861	.0537	5.9%

Table 4(3): COMPUTATION OF BIAS FOR THE MISSING DATA PATTERN (3)

x_j	a_{jj}	λ_j	R_j^2	Bias	%Bias
x_1	14.4620	10.2%	.70206	.4397	3%
x_2	33.3623	10.2%	.82242	.6045	1.8%
x_3	.6041	10.2%	.63316	.0226	3.7%
x_4	.3541	10.2%	.83673	.0060	1.7%
x_5	1.0100	10.2%	.50745	.0508	5%

Table 4(4): COMPUTATION OF BIAS FOR THE MISSING DATA PATTERN (4)

x_j	a_{jj}	λ_j	R_j^2	Bias	%Bias
x_1	5.8242	14.29%	.52838	.3924	6.7%
x_2	19.6703	14.29%	.56023	1.2358	6.3%
x_3	.3855	14.29%	.53545	.0256	6.6%
x_4	.2765	14.29%	.73103	.0106	3.8%
x_5	.7504	14.29%	.50366	.0532	7%

Comparing the figures displayed in the above tables for the bias with those of table 3(1), it follows that our procedure gives exactly the same result obtained via Buck's procedure.

From tables 4(1) through 4(4) it is clear that the % bias for fixed λ_j (given in the last column of each table) is a decreasing function of R_j^2 for all variables in all patterns of missingness. This result is particularly important for the users of the method. Since the bias is a decreasing function of R_j^2 then it is quite logical to impute the missing values of the j-th variable from the regression of that variable only on those variables which are highly correlated with it. This will ensure the minimization of bias as R_j^2

will be maximized. Also the volume of computation will be tremendously reduced since the imputation of the missing values of the j-th variable need not be based on the regression of the j-th variable on all the remaining $k-1$ variables, but only on a subset of it, say z where $z < k-1$. The subset of variables which is highly correlated with the j-th variable can be determined from the correlation matrix of the n_c-complete cases.

4.3 AN OVERVIEW OF MULTIVARIATE REGRESSION ANALYSIS

For the case of units with one missing value, we have seen that Buck's method uses multiple regression as a tool for imputing the missing observations. In the next section we shall study some of the statistical properties of the post-imputation variance-covariance matrix for the case of units with more than one missing value. To achieve this, we shall view the post-imputation covariance matrix from the standpoint of multivariate regression and multivariate analysis of variance (MANOVA). We shall therefore give a brief review of the theory of multivariate regression and MANOVA for the complete-data case.

Consider the model defined by

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U} \tag{4.21}$$

where

Y(nxp) is an observed matrix of p response variables on each of n individuals,

X(nxk) is a known matrix,

B(kxp) is a matrix of unknown regression parameters, and

U is a matrix of unobserved random disturbances whose rows for given **X** are uncorrelated, each with mean **0** and common covariance matrix $\mathbf{\Sigma}$.

$\mathbf{X}$ represents a matrix of p "independent" variables observed on each of the n individuals. Usually the first column of $\mathbf{X}$ equals 1, namely $\mathbf{X} = (1, \mathrm{X}_1)$ to allow for an overall mean effect.

The columns of $\mathbf{Y}$ represent "dependent" variables which are to be explained in terms of the "independent" variables given by the columns of $\mathbf{X}$. We should note that

$$\mathrm{E}(\mathrm{y}_{ij}) = \underline{\mathrm{x}}_i' \underline{\beta}_{(j)}, \tag{4.22}$$

so that the expected value of y_{ij} depends on the i-th row of $\mathbf{X}$ and the j-th column of the matrix of regression coefficients. The case of p=1, where there is only one dependent variable, is the familiar multiple regression model which we will write as

$$\underline{\mathrm{y}} = \mathbf{X}\underline{\beta} + \underline{\mathrm{u}} \tag{4.23}$$

In most applications $\mathbf{U}$ is assumed to be normally distributed, so that

$$\mathbf{U} \text{ is a data matrix from } \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{4.24}$$

where $\mathbf{U}$ is independent of $\mathbf{X}$. Under the assumption of normal errors, the loglikelihood for the data $\mathbf{Y}$ in terms of the parameters $\mathbf{B}$ and $\boldsymbol{\Sigma}$ is given by

$$\ell(\mathbf{B}, \boldsymbol{\Sigma}) = -\frac{1}{2}\mathrm{n} \log \mid 2\pi\boldsymbol{\Sigma} \mid -\frac{1}{2}\mathrm{tr}(\mathbf{Y} - \mathbf{XB})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{XB})' \tag{4.25}$$

For estimation of $\mathbf{B}$ to be unique we shall suppose that $\mathbf{X}$ has full rank p, so that the inverse $(\mathbf{X}'\mathbf{X})^{-1}$ exists. Mardia *et al* (1979, pp. 158–159) proved that, for the loglikelihood function (4.25) the maximum likelihood estimates of $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{4.26}$$

and

$$\hat{\mathbf{\Sigma}}^{-1} = n^{-1}\mathbf{Y}'\mathbf{P}\mathbf{Y} \tag{4.27}$$

where

$$\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{4.28}$$

The multivariate analysis of variance for the multivariate regression model given by (4.21) can be performed by noting that

$$\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} \tag{4.29}$$

therefore

$$\begin{aligned}\hat{\mathbf{U}}'\hat{\mathbf{U}} &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) \\ &= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}}\end{aligned}$$

which can be written as

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{U}}'\hat{\mathbf{U}} \tag{4.30}$$

In other words, the total sum of squares and products (SSP) matrix $\mathbf{Y}'\mathbf{Y}$ is partitioned into an SSP matrix $\hat{\mathbf{Y}}'\hat{\mathbf{Y}}$ due to the multivariate regression model and an SSP matrix $\hat{\mathbf{U}}'\hat{\mathbf{U}}$ about the regression model.

Let

$$\mathbf{Y}'\mathbf{Y} = \mathbf{Q}_t$$

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Q}_h$$

$$\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Q}_e$$

and assume that the values of $\mathbf{Y}$ are centered so that the columns of $\mathbf{Y}$ have zero means. Then the MANOVA table for the multivariate regression model given by (4.21) is given below.

Table 4(5): MANOVA TABLE FOR THE MULTIVARIATE REGRESSION OF $\mathbf{Y} = (x_1, x_2, \ldots, x_p)$ on $\mathbf{X} = (x_{p+1}, \ldots, x_k)$

S.V.	d.f.	SSP.
Reg.	$k - p$	$\mathbf{Q}_h$
Resid.	By Subt.	$\mathbf{Q}_e$
Total (corr.)	$n - 1$	$\mathbf{Q}_t$

Note that dividing the elements of $\mathbf{Q}_e$ by $(n-1)$, its diagonal and off-diagonal elements give the sample variances and covariances of the residuals of the elements of $\mathbf{Y}$ respectively. Similarly, $(n-1)^{-1}\mathbf{Q}_t$ gives the covariance matrix of the elements of $\mathbf{Y}$. Also note that $(n-1)^{-1}\mathbf{Q}_h$ is the covariance matrix of the estimated values of $\mathbf{Y}$.

Thus the multivariate linear regression and MANOVA will enable us to examine the expected effect of imputations on the elements of $\mathbf{Q}_t$ by examining its effect on the corresponding elements of $\mathbf{Q}_e$ and $\mathbf{Q}_h$.

In the univariate regression model defined by (4.23), that is when p=1, we know that the squared multiple correlation coefficient R^2 represents the proportion of variability in the dependent variable which is explained by the regression model, and from (4.23) and (4.30) we have

$$\underline{y}'\underline{y} = \hat{\underline{y}}'\hat{\underline{y}} + \hat{\underline{u}}'\hat{\underline{u}} \tag{4.31}$$

where $\underline{y}$ is centered so that $\bar{y} = 0$ and $\hat{\underline{y}}'\hat{\underline{y}}$ is the component of variance explained by the regression model given by (4.23), thus

$$R^2 = \hat{\underline{y}}'\hat{\underline{y}}(\underline{y}'\underline{y})^{-1}. \tag{4.32}$$

Substituting the value of $\hat{\underline{y}}'\hat{\underline{y}}$ from (4.31) in (4.32) we get

$$\begin{aligned} R^2 &= (\underline{y}'\underline{y} - \hat{\underline{u}}'\hat{\underline{u}})(\underline{y}'\underline{y})^{-1} \\ &= 1 - \hat{\underline{u}}'\hat{\underline{u}}(\underline{y}'\underline{y})^{-1} \end{aligned}$$

or

$$1 - R^2 = \hat{\underline{u}}'\hat{\underline{u}}/(\underline{y}'\underline{y}). \tag{4.33}$$

A similar measure for the multivariate correlation between matrices $\mathbf{X}$ and $\mathbf{Y}$ in the model $\mathbf{Y}=\mathbf{XB}+\mathbf{U}$ can be obtained as follows:

Note that

$$\begin{aligned} \hat{\mathbf{U}}'\hat{\mathbf{U}} &= (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\mathbf{B}} - \hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} + \hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}} \end{aligned}$$

and since $\hat{\mathbf{B}}'\mathbf{X}'\mathbf{Y} = \hat{\mathbf{B}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{B}}$, it follows that

$$\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\mathbf{B}}$$

$$= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$
$$= \mathbf{Y}'[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$$
$$= \mathbf{Y}'\mathbf{P}\mathbf{Y}. \tag{4.34}$$

Let

$$\mathbf{D} = (\mathbf{Y}'\mathbf{Y})^{-1}\hat{\mathbf{U}}'\hat{\mathbf{U}}. \tag{4.35}$$

The matrix $\mathbf{D}$ is a generalization of $1 - R^2$ in the univariate case. Mardia *et al* (1979, pp. 170–171) noted that $\hat{\mathbf{U}}'\hat{\mathbf{U}} = \mathbf{Y}'\mathbf{P}\mathbf{Y}$ ranges between zero, when all the variation in $\mathbf{Y}$ is explained by the model, and $\mathbf{Y}'\mathbf{Y}$ at the other extreme, when no part of the variation in $\mathbf{Y}$ is explained. Therefore

$$\mathbf{0} \leq \mathbf{I} - \mathbf{D} \leq \mathbf{I} \tag{4.36}$$

Remark

We should note that the relation given by (4.36) holds for the elements of the principal diagonal of $\mathbf{I} - \mathbf{D}$. To investigate the existence of a similar relation amongst the off-diagonal elements of $\mathbf{I} - \mathbf{D}$ we proceed as follows:

Recall that the total variation of the multivariate regression model given by (4.21) has been partitioned by (4.30) as follows:

$$\mathbf{Y}'\mathbf{Y} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\mathbf{U}}'\hat{\mathbf{U}}. \tag{4.37}$$

For simplicity, assume that $\mathbf{Y} = (\mathbf{x}_1, \mathbf{x}_2)$. Then the off-diagonal element of (4.37) is given by

$$\mathbf{x}_1'\mathbf{x}_2 = \hat{\mathbf{x}}_1'\hat{\mathbf{x}}_2 + \hat{\mathbf{u}}_1'\hat{\mathbf{u}}_2. \tag{4.38}$$

Formula (4.38) partitions the total covariability of the model given by (4.21) into two components. The first term of the RHS of (4.38) gives the covariability "explained" by the model while the second term gives the "unexplained" or residual covariability.

From (4.38) we can define the measure

$$R_{12}^2 = \frac{\hat{x}_1' \hat{x}_2}{x_1' x_2}$$

Substituting $\hat{x}_1' \hat{x}_2$ from (4.38) we get

$$R_{12}^2 = \frac{x_1' x_2 - \hat{u}_1' \hat{u}_2}{x_1' x_2},$$

or

$$R_{12}^2 = 1 - \frac{\hat{u}_1' \hat{u}_2}{x_1' x_2}. \tag{4.39}$$

Note that $\hat{u}_1' \hat{u}_2$ ranges between zero, when all the covariation in $\mathbf{Y} = (x_1, x_2)$ is explained by the model, and $x_1' x_2$ at the other extreme, when no part of the covariation in $\mathbf{Y}$ is explained. Therefore

$$0 \leq R_{12}^2 \leq 1. \tag{4.40}$$

Generally, for $\mathbf{Y} = (x_j, x_k)$, $j, k = 1, \ldots, p$ we have

$$0 \leq R_{jk}^2 \leq 1, \quad j < k = 1, \ldots p. \tag{4.41}$$

Given (4.41), it follows that (4.36) can now be written as

$$0 \leq \mathbf{I} - \mathbf{D} \leq \mathbf{1}\mathbf{1}', \tag{4.42}$$

where $\mathbf{1}\mathbf{1}'$ is a unit matrix, i.e., $\mathbf{1}' = (1 1 \ldots 1)$.

It is worth mentioning that we have sought to prove (4.42) by intuition. However, further investigation on (4.42) that may lead to a rigorous mathematical proof is perhaps worthwhile.

COMPUTATIONAL REMARKS

We should note that Buck's method, for the case of units with more than one missing value, applies the above mentioned complete-data multivariate regression technique to the complete-cases of the data matrix **X**. Missing values are then estimated by their expected values obtained from the correctly chosen regression equation.

It is worth mentioning that if any combination of v variables may be missing then the required number of regression equations in the method of Buck is given by

$$\text{Number of regression equations} = k\binom{k-1}{v-1}.$$

For the case of units with one missing value, v=1 or no combination of more than one variable are allowed to be missing in the same unit.

Note that the required number of regression equations increases rapidly as k increases, e.g., for k=5; v=3 this number is 30 whereas for k=6 and v=3 it increases to 60.

4.4 DETERMINATION OF BIAS IN THE VARIANCE-COVARIANCE MATRIX: UNITS WITH MORE THAN ONE MISSING VALUE

4.4.1 THE COVARIANCES

Buck (1960) proved the biasedness of the variances and the unbiasedness of the covariances for the case of units with one missing value. This was shown in chapter III. It is our objective in this section to study the biasedness in the

variance-covariance matrix for the case of units with more than one missing value.

Adopting our usual notation, let n be the total number of cases, n_c the number of complete cases and $m^{(jk)}$ the number of cases with both x_{ij} and x_{ik} missing. Further, let a^*_{jk} be the post-imputation covariance of the j-th and k-th variables and a_{jk} be the same estimate based on the n_c-complete cases.

The sample covariance of the j-th and k-th variables can be expressed, from the standpoint of the multivariate regression and MANOVA as follows

$$a^*_{jk} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(x_{ij}-\hat{x}_{ij})(x_{ik}-\hat{x}_{ik}) + \sum_{i=1}^{n}(\hat{x}_{ij}-\bar{x}_j)(\hat{x}_{ik}-\bar{x}_k)\right\}, \quad (4.43)$$

where the first term of the R.H.S. is the sample covariance of the residuals of the multivariate regression of $\underline{x}_j, \underline{x}_k$ on $\mathbf{X}_{(jk)} = (\underline{x}_1, \ldots, \underline{x}_{j-1}, \underline{x}_{j+1}, \ldots, \underline{x}_{k-1})$, and the second term is the component of variance explained by the regression of $\underline{x}_j, \underline{x}_k$ on $\mathbf{X}_{(jk)}$. Formula (4.43) can be written as

$$a^*_{jk} = \frac{1}{n-1}\left\{\sum_{n_c}(x_{ij}-\hat{x}_{ij})(x_{ik}-\hat{x}_{ik}) + \sum_{m^{(jk)}}(x_{ij}-\hat{x}_{ij})(x_{ik}-\hat{x}_{ik}) + \sum_{i=1}^{n}(\hat{x}_{ij}-\bar{x}_j)(\hat{x}_{ik}-\bar{x}_k)\right\}. \quad (4.44)$$

The contribution of the imputed values to the component of variance explained by regression is incorporated in the last term of (4.44).

But

$$\sum_{m^{(jk)}}(x_{ij}-\hat{x}_{ij})(x_{ik}-\hat{x}_{ik}) = 0 \quad (4.45)$$

since the imputed values of the j-th and k-th variables —$(\hat{\underline{x}}_{ij}, \hat{\underline{x}}_{ik})$— lie exactly on the regression line of $\underline{x}_j, \underline{x}_k$ on $\mathbf{X}_{(jk)}$. In other words, the contribution of the $m^{(jk)}$ imputed pairs of observations to the sample covariance of the residuals is effectively set to zero. Thus the residual covariance of the observed and imputed values of the j-th and k-th variables is less than the same estimate obtained from the complete cases. To estimate this reduction we proceed as follows:

Let

$$\mathbf{X} = (x_{il}) \ (i = 1, 2, \ldots, n; l = 1, 2, \ldots, k)$$

represent the sample of n units, on each of which it is desired to have measurements on k variables. Further assume that the j-th and k-th variables are

$$\underline{x}_{ij}, \underline{x}_{ik}, \ (i = 1, 2, \ldots, n_c, n_c + 1, \ldots, n),$$

where the j-th and k-th variables are observed on the first n_c units and missing on the next $(n - n_c)$ units. Then using Buck's method the $(n - n_c)$ missing observations on the j-th and k-th variables are to be estimated from the multivariate regression equation of the observed values, i.e.,

$$(\hat{\underline{x}}_{ij}, \ \hat{\underline{x}}_{ik}) = \mathbf{X}_{(jk)}\hat{\mathbf{B}}; \ \mathbf{X}_{(jk)} = (\underline{x}_{i1}, \underline{x}_{i2}, \ldots, \underline{x}_{i,j-1}, \underline{x}_{i,j+1}, \cdots, \underline{x}_{i,k-1}),$$
$$i = 1, 2, \ldots, n_c. \quad (4.46)$$

Using (4.46) we impute the $(n - n_c)$ missing values on the j-th and k-th variables to have a completed data matrix.

Now, consider the regression equation of both the observed and imputed values, i.e.,

$$(\hat{\underline{x}}_{ij}, \ \hat{\underline{x}}_{ik}) = \mathbf{X}_{(jk)}\hat{\mathbf{B}}; \ \mathbf{X}_{(jk)} = (\underline{x}_{i1}, \underline{x}_{i2}, \cdots, \underline{x}_{i,j-1}, \underline{x}_{i,j+1}, \cdots, \underline{x}_{i,k-1}),$$
$$i = 1, 2, \ldots, n. \quad (4.47)$$

From (4.46), the pre-imputation sample covariance of the residuals is

$$(\Phi_{jk})^{(n_c)} = \frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n_c - 1} \tag{4.48}$$

and from (4.47) the post-imputation sample covariance of the residuals is

$$\begin{aligned}(\Phi_{jk})^{(n)} &= \frac{\sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n-1} \\ &= \frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n-1} + \frac{\sum_{n_c+1}^{n}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n-1} \\ &= \frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n-1}\end{aligned} \tag{4.49}$$

since $\sum_{n_c+1}^{n}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik}) = 0$ due to imputation. This is because the imputed values are projected exactly on the regression line given by (4.46). Thus

$$\begin{aligned}(\Phi_{jk})^{(n_c)} - (\Phi_{jk})^{(n)} &= \frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n_c - 1} \\ &\quad - \frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n-1} \\ &= \sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})\left(\frac{1}{n_c - 1} - \frac{1}{n-1}\right) \\ &= \frac{n - n_c}{n-1}\frac{\sum_{n_c}(x_{ij} - \hat{x}_{ij})(x_{ik} - \hat{x}_{ik})}{n_c - 1} \\ &= \frac{n - n_c}{n-1}(\Phi_{jk})^{(n_c)}.\end{aligned} \tag{4.50}$$

It follows that if $\underset{\sim}{x}_j$ and $\underset{\sim}{x}_k$ are missing on $(n - n^{(jk)})$ of all units then the estimate of their residual covariance obtained from the completed data is less

than the same estimate obtained from the complete cases. The difference of the two estimates is given by

$$\frac{m^{(jk)}}{n-1}\left(\Phi_{jk}\right)^{(n_c)} = \frac{n-n^{(jk)}}{n-1}\left(\Phi_{jk}\right)^{(n_c)}. \tag{4.51}$$

This difference, in terms of expectations, gives the bias of the estimated covariances.

4.4.2 THE VARIANCES

Assume that the j-th and k-th variables are jointly missing on $m^{(jk)}$ units. Further, let $a_{jj}^{\star}$ and $a_{kk}^{\star}$ denote the estimates of the variances of the two variables obtained from the completed data. Then the sample variances of the j-th and k-th variables can be expressed, from the standpoint of the multivariate regression and MANOVA as follows:

$$a_{jj}^{\star} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(x_{ij}-\hat{x}_{ij})^2 + \sum_{i=1}^{n}(\hat{x}_{ij}-\bar{x}_j)^2\right\} \tag{4.52}$$

$$a_{kk}^{\star} = \frac{1}{n-1}\left\{\sum_{i=1}^{n}(x_{ik}-\hat{x}_{ik})^2 + \sum_{i=1}^{n}(\hat{x}_{ik}-\bar{x}_k)^2\right\}, \tag{4.53}$$

where the first terms on the R.H.S. of (4.52) and (4.53) are the sample variances of the residuals of the multivariate regression of $\underset{\sim}{x}_j$, $\underset{\sim}{x}_k$ on $\mathbf{X}_{(jk)} = (\underset{\sim}{x}_1 \cdots, \underset{\sim}{x}_{j-1}, \underset{\sim}{x}_{j+1} \cdots, \underset{\sim}{x}_{k-1})$, and the second terms are the components of variance explained by the regression. Then (4.52) can be written as:

$$a_{jj}^{\star} = \frac{1}{n-1}\left\{\sum_{n_c}(x_{ij}-\hat{x}_{ij})^2 + \sum_{n-n_c}(x_{ij}-\hat{x}_{ij})^2 + \sum_{i=1}^{n}(\hat{x}_{ij}-\bar{x}_j)^2\right\}, \tag{4.54}$$

where the contribution of the imputed values to the component of variance explained by regression is incorporated.

But

$$\sum_{n-n_c} (x_{ij} - \hat{x}_{ij})^2 = 0 \tag{4.55}$$

since the imputed values are projected on the regression line of $\underset{\sim}{x}_j$ on $\mathbf{X}_{(jk)}$. In other words, the contribution of the $n - n_c = m^{(jk)}$ imputed observations to the sample variance of the residuals is effectively set to zero. Thus the variance of the observed and imputed values of the j-th variable is less than the same estimate obtained from the complete cases.

Putting k=j in (4.48) through (4.51) the resultant reduction in the variance of the j-th variable is obtained as:

$$\frac{n - n^{(jk)}}{n-1} (\Phi_{jj})^{(n_c)}, \tag{4.56}$$

where $(\Phi_{jj})^{(n_c)}$ is the sample variance of the residuals of the regression of the j-th variable on the remaining $k-2$ variables. The same result applies to a^*_{kk}.

This reduction in terms of expectations, gives the amount of bias in the estimation of the variances of the j-th and k-th variables.

The above discussion of sections 4.4.1 and 4.4.2 can now be formalized in the following theorem:

Theorem 4.4

If the j-th and k-th variables are jointly missing on $m^{(jk)}$ units, then $\mathrm{Cov}(x_j, x_k)$ and $\mathrm{V}(x_j)$, obtained via Buck's method are biased. The bias of $\mathrm{Cov}(x_j, x_k)$ is given by

$$(\Phi_{jk})^{(n_c)} - (\Phi_{jk})^{(n)} = \frac{m^{(jk)}}{n-1} (\Phi_{jk})^{(n_c)},$$

and the bias of $\mathrm{V}(x_j)$ is given by

$$(\Phi_{jj})^{(n_c)} - (\Phi_{jj})^{(n)} = \frac{m^{(jk)}}{n-1}(\Phi_{jj})^{(n_c)},$$

where

$(\Phi_{jk})^{(n_c)}$ and $(\Phi_{jk})^{(n)}$ are the pre– and post-imputation sample covariances of the residuals of the multivariate regression of the j-th and k-th variables on the remaining k−2 variables, and

$(\Phi_{jj})^{(n_c)}$ and $(\Phi_{jj})^{(n)}$ are the pre– and post-imputation sample variances of the residuals of the multivariate regression of the j-th and k-th variables on the remaining k−2 variables.

The various issues discussed, separately, in sections 4.4.1 and 4.4.2 as well as the results of Buck (1960) for the case of units with one missing value can be dealt with jointly as follows:

Let **A** be the covariance matrix of the n_c-complete cases and $\mathbf{A}^{-1}$, its inverse. Further, partition these matrices as follows:

$$\mathbf{A} = \left[\begin{array}{cc|ccc} a_{11} & a_{12} & a_{13} & \cdots & a_{1k} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2k} \\ -- & -- & -- & -- & -- \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & a_{k3} & \cdots & a_{kk} \end{array}\right] = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{12} & \mathbf{D}_{22} \end{bmatrix} \qquad (4.57)$$

$$\mathbf{A}^{-1} = \left[\begin{array}{cc|ccc} c_{11} & c_{12} & c_{13} & \cdots & c_{1k} \\ c_{21} & c_{22} & c_{23} & \cdots & c_{2k} \\ -- & -- & -- & -- & -- \\ c_{31} & c_{32} & c_{33} & \cdots & c_{3k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{k1} & c_{k2} & c_{k3} & \cdots & c_{kk} \end{array}\right] = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{12} & \mathbf{B}_{22} \end{bmatrix} \qquad (4.58)$$

where $\mathbf{D}_{11}$ is the covariance matrix of those variables with jointly missing values. The corresponding elements of $\mathbf{D}_{11}$ in $\mathbf{A}^{-1}$ are given by $\mathbf{B}_{11}$. Then we can state the following:

Theorem 4.5

Under MCAR assumption for the missing values, the maximum biasedness of the post-imputation elements of $\mathbf{D}_{11}$ is

$$\mathbf{B}_{11}^{-1} = \begin{bmatrix} (a_{11} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1) & (a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1) \\ (a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1) & (a_{22} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2) \end{bmatrix}.$$

Proof

Recall that the variance-covariance matrix of the n_c-complete cases is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix}$$

We can write A as

$$\mathbf{A} = (a_{sj}) = \begin{bmatrix} a_{11} & a_{12} & \underline{\alpha}_1' \\ a_{12} & a_{22} & \underline{\alpha}_2' \\ \underline{\alpha}_1 & \underline{\alpha}_2 & \mathbf{C} \end{bmatrix} \quad (s, j = 1, \ldots, k), \tag{4.59}$$

where

$$a_{11} = V(\underline{x}_1)$$

$$a_{22} = V(\underline{x}_2)$$

$$a_{12} = \mathrm{Cov}(\underline{x}_1, \underline{x}_2)$$

$$\underline{\alpha}_1' = (a_{13}, a_{14}, \ldots, a_{1k})$$

$$\underline{\alpha}_2' = (a_{23}, a_{24}, \ldots, a_{2k}).$$

Thus the elements of $\underline{\alpha}'_j$, j = 1,2 give the covariances of $\underline{x}_j$ and $\mathbf{X}_{(12)} = (\underline{x}_3, \ldots, \underline{x}_k)$ computed from the n_c-complete cases, and

$$\mathbf{C} = \begin{bmatrix} a_{33} & \cdots & a_{3k} \\ \vdots & \cdots & \vdots \\ a_{k3} & \cdots & a_{kk} \end{bmatrix}.$$

Clearly, $\mathbf{C}$ is the covariance matrix of $\mathbf{X}_{(12)} = (\underline{x}_3, \underline{x}_4, \ldots, \underline{x}_k)$, of order (k−2)x(k−2).

Recalling our notation of section 3.4 and 4.3, we estimate the value of $\underline{x}_1$ and $\underline{x}_2$ for those cases which have $\underline{x}_1$ and $\underline{x}_2$ only missing, by estimating the multivariate regression equation

$$\hat{\mathbf{Y}} = \mathbf{X}_{(12)}\hat{\mathbf{B}}, \tag{4.60}$$

where $\hat{\mathbf{Y}} = (\hat{\underline{x}}_1, \hat{\underline{x}}_2)$.
The regression coefficients of x_j on $\mathbf{X}_{(12)} = (\underline{x}_3, \underline{x}_4, \ldots, \underline{x}_k)$ are obtained by Woolf's procedure as:

$$\hat{\underline{\beta}}'_j = \underline{\alpha}'_j \mathbf{C}^{-1}, \; j = 1, 2. \tag{4.61}$$

Thus we have

$$\hat{\underline{x}}_1 = \mathbf{X}_{(12)}\mathbf{C}^{-1}\underline{\alpha}_1, \tag{4.62}$$

and

$$\hat{\underline{x}}_2 = \mathbf{X}_{(12)}\mathbf{C}^{-1}\underline{\alpha}_2. \tag{4.63}$$

Therefore the variance of $\hat{\underline{x}}_j$, j=1,2 is given by

$$\begin{aligned} \mathrm{V}(\hat{\underline{x}}_j) &= \mathrm{V}(\mathbf{X}_{(12)}\hat{\underline{\beta}}_j) = \mathrm{V}(\mathbf{X}_{(12)}\mathbf{C}^{-1}\underline{\alpha}_j) \\ &= \underline{\alpha}'_j\mathbf{C}^{-1}\mathrm{V}(\mathbf{X}_{(12)})\mathbf{C}^{-1}\underline{\alpha}_j \\ &= \underline{\alpha}'_j\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\underline{\alpha}_j \\ &= \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_j, \end{aligned} \tag{4.64}$$

which can be written as

$$\text{Cov}(\hat{\underline{x}}_j, \hat{\underline{x}}_j) = \underline{\alpha}_j' \mathbf{C}^{-1} \underline{\alpha}_j. \tag{4.65}$$

It follows that for $k \neq j$ we have

$$\text{Cov}(\hat{\underline{x}}_j, \hat{\underline{x}}_k) = \underline{\alpha}_j' \mathbf{C}^{-1} \underline{\alpha}_k, \tag{4.66}$$

where $(\underline{x}_j, \underline{x}_k)$ are those variables with jointly missing values.
Thus,

$$\text{Cov}(\hat{\underline{x}}_1, \hat{\underline{x}}_2) = \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_2. \tag{4.67}$$

If we write

$$\mathbf{A}^{-1} = \begin{bmatrix} c_{11} & c_{12} & \underline{e}_1' \\ c_{12} & c_{22} & \underline{e}_2' \\ \underline{e}_1 & \underline{e}_2 & \mathbf{F} \end{bmatrix}, \tag{4.68}$$

then, since $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_k$, we have from (4.59) and (4.68),

$$a_{11}c_{11} + a_{12}c_{12} + \underline{\alpha}_1' \underline{e}_1 = 1 \tag{4.69}$$

$$a_{12}c_{11} + a_{22}c_{12} + \underline{\alpha}_2' \underline{e}_1 = 0 \tag{4.70}$$

$$\underline{\alpha}_1 c_{11} + \underline{\alpha}_2 c_{12} + \mathbf{C}\underline{e}_1 = \underline{0} \tag{4.71}$$

and

$$a_{11}c_{12} + a_{12}c_{22} + \underline{\alpha}_1' \underline{e}_2 = 0 \tag{4.72}$$

$$a_{12}c_{12} + a_{22}c_{22} + \underline{\alpha}_2' \underline{e}_2 = 1 \tag{4.73}$$

$$\underline{\alpha}_1 c_{12} + \underline{\alpha}_2 c_{22} + \mathbf{C}\underline{e}_2 = \underline{0}. \tag{4.74}$$

From (4.71) we have

$$\mathbf{C}\underline{e}_1 = -\underline{\alpha}_1 c_{11} - \underline{\alpha}_2 c_{12}$$

or

$$\underline{e}_1 = -\mathbf{C}^{-1}\underline{\alpha}_1 c_{11} - \mathbf{C}^{-1}\underline{\alpha}_2 c_{12}. \tag{4.75}$$

Substituting (4.75) into (4.70), we have

$$a_{12}c_{11} + a_{22}c_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1 c_{11} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2 c_{12} = 0.$$

That is,

$$a_{12}c_{11} + a_{22}c_{12} - c_{11}(\underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1) - c_{12}(\underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2) = 0$$

or

$$c_{11}(a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1) + c_{12}(a_{22} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2) = 0$$

from which we get

$$c_{12} = \frac{-c_{11}(a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1)}{(a_{22} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2)}. \tag{4.76}$$

Substituting (4.75) and (4.76) in (4.69), we get

$$a_{11}c_{11} - \frac{a_{12}c_{11}(a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1)}{(a_{22} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2)} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1 c_{11}$$
$$+ \frac{\underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_2 c_{11}(a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1)}{(a_{22} - \underline{\alpha}_2\prime\mathbf{C}^{-1}\underline{\alpha}_2)} = 1;$$

that is

$$c_{11}(a_{11} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_1) - \frac{c_{11}(a_{12} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_1)[a_{12} - \underline{\alpha}_1'\mathbf{C}^{-1}\underline{\alpha}_2]}{(a_{22} - \underline{\alpha}_2'\mathbf{C}^{-1}\underline{\alpha}_2)} = 1$$

from which we get

$$c_{11} = \frac{(a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2)}{(a_{11} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1)(a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2) - (a_{12} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_2)^2}, \quad (4.77)$$

since

$$\alpha_2' \mathbf{C}^{-1} \alpha_1 = \alpha_1' \mathbf{C}^{-1} \alpha_2.$$

Substituting the value of c_{11} in (4.76), we get

$$c_{12} = \frac{-(a_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1)}{(a_{11} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1)(a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2) - (a_{12} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_2)^2}. \quad (4.78)$$

To obtain the value of c_{22}, we have from (4.74)

$$\mathbf{C} \underline{e}_2 = -\underline{\alpha}_1 c_{12} - \underline{\alpha}_2 c_{22}$$

or

$$\underline{e}_2 = -\mathbf{C}^{-1} \underline{\alpha}_1 c_{12} - \mathbf{C}^{-1} \underline{\alpha}_2 c_{22} \quad (4.79)$$

substituting the value of $\underline{e}_2$ into (4.73), we have

$$a_{12} c_{12} + a_{22} c_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1 c_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2 c_{22} = 1,$$

that is

$$c_{12}(a_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1) + c_{22}(a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2) = 1$$

or

$$c_{22} = \frac{1 - c_{12}(a_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1)}{(a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2)}.$$

Substituting the value of c_{12} from (4.78), we get

$$c_{22} = \frac{(a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1)}{(a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1)(a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2) - (a_{12} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_2)^2}. \tag{4.80}$$

Let

$$k = (a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1)(a_{22} - \underline{\alpha}_2 C^{-1} \underline{\alpha}_2') - (a_{12} - \underline{\alpha}_1 C^{-1} \underline{\alpha}_2)^2$$

then

$$c_{11} = \frac{a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2}{k} \tag{4.81}$$

$$c_{12} = \frac{-(a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1)}{k} \tag{4.82}$$

$$c_{22} = \frac{a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1}{k}. \tag{4.83}$$

Therefore

$$B_{11} = \begin{bmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{bmatrix} = \begin{bmatrix} (a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2)/k & -(a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1)/k \\ -(a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1)/k & (a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1)/k \end{bmatrix} \tag{4.84}$$

Inverting B_{11}, we have

$$\begin{aligned} B_{11}^{-1} &= \frac{1}{c_{11}c_{22} - c_{12}^2} \begin{bmatrix} c_{22} & -c_{12} \\ -c_{12} & c_{11} \end{bmatrix} \\ &= \begin{bmatrix} (a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2)/k & -(a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1)/k \\ -(a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1)/k & (a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1)/k \end{bmatrix}^{-1} \\ &= \frac{k^2}{k} \frac{1}{k} \begin{bmatrix} (a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1) & (a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1) \\ (a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1) & (a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2) \end{bmatrix} \\ &= \begin{bmatrix} (a_{11} - \underline{\alpha}_1' C^{-1} \underline{\alpha}_1) & (a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1) \\ (a_{12} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_1) & (a_{22} - \underline{\alpha}_2' C^{-1} \underline{\alpha}_2) \end{bmatrix}. \end{aligned} \tag{4.85}$$

This ends the proof.

Remark

Theorem 4.5 can be generalized to the case of units with more than two missing values. Note that this corresponds to the case of more than two variables jointly missing on the i-th unit. For the case of units with ℓ missing values, $2 < \ell < k$, theorem 4.5 can be generalized as follows:

Re-arrange the elements of **A** given by (4.57) in such a way that the ℓ variables with jointly missing values belong to $\mathbf{D}_{11}$. Recall that $\mathbf{D}_{11}$ is the covariance matrix of the ℓ variables with jointly missing values. Then the corresponding elements of $\mathbf{D}_{11}$ in $\mathbf{A}^{-1}$ give $\mathbf{B}_{11}$. Hence theorem 4.5 can be generalized by the $\ell x \ell$ matrix

$$\mathbf{B}_{11}^{-1} = \begin{bmatrix} (a_{11} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1) & & & \\ (a_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1) & (a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2) & & \\ \vdots & \vdots & \ddots & \\ (a_{1\ell} - \underline{\alpha}_\ell' \mathbf{C}^{-1} \underline{\alpha}_1) & (a_{2\ell} - \underline{\alpha}_\ell' \mathbf{C}^{-1} \underline{\alpha}_2) & \cdots & (a_{\ell\ell} - \underline{\alpha}_\ell' \mathbf{C}^{-1} \underline{\alpha}_\ell) \end{bmatrix}.$$

A numerical illustration of this remark is given in section 4.4.3

Theorem 4.6

The residual variances and the residual covariance of the estimated values of $\underline{x}_1$ and $\underline{x}_2$ are less than the corresponding estimates obtained from the actual values of $\underline{x}_1$ and $\underline{x}_2$. The difference between the two sets of estimates is given by the elements of $\mathbf{B}_{11}^{-1}$. This difference, in terms of expectations, gives the bias of the post-imputation estimates.

Proof

Let

$$\mathbf{B}_{11}^{-1} = \begin{bmatrix} z_{11} & z_{12} \\ z_{12} & z_{22} \end{bmatrix}. \tag{4.86}$$

Then from (4.85)

$$z_{11} = \frac{c_{22}}{c_{11}c_{22} - c_{12}^2}$$
$$= a_{11} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1, \tag{4.87}$$

which implies that

$$\underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1 = a_{11} - z_{11}.$$

But from (4.64)

$$V(\hat{x}_j) = \underline{\alpha}_j' \mathbf{C}^{-1} \underline{\alpha}_j.$$

Therefore

$$V(\hat{x}_1) = a_{11} - z_{11}. \tag{4.88}$$

Similarly,

$$z_{22} = \frac{c_{11}}{c_{11}c_{22} - c_{12}^2}$$
$$= a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2, \tag{4.89}$$

which implies that

$$V(\hat{x}_2) = a_{22} - z_{22}. \tag{4.90}$$

Next

$$z_{12} = \frac{-c_{12}}{c_{11}c_{22} - c_{12}^2}$$
$$= a_{12} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1, \tag{4.91}$$

which implies that

$$\underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_1 = a_{12} - z_{12}.$$

But from (4.66)

$$\text{Cov}(\hat{\underline{x}}_j, \hat{\underline{x}}_k) = \underline{\alpha}_j' \mathbf{C}^{-1} \underline{\alpha}_k$$

Therefore

$$\text{Cov}(\hat{\underline{x}}_1, \hat{\underline{x}}_2) = a_{12} - z_{12}. \tag{4.92}$$

From (4.88), (4.90) and (4.92) we note that the residual variances and the residual covariance of the estimated values of x_1 and x_2 are less than the corresponding estimates obtained from the actual values of x_1 and x_2. Moreover, from (4.87), (4.89) and (4.91), it follows that the differences between the corresponding elements of the two sets of estimates are given by the elements of $\mathbf{B}_{11}^{-1}$.

Remark

Note that the elements of $\mathbf{B}_{11}^{-1}$ give the maximum amount of bias in the post-imputation covariance matrix of x_1 and x_2. The maximum amount of bias, as mentioned earlier, corresponds to the case where all values of x_1 and x_2 are imputed. It therefore follows that if $\underline{x}_1$ and $\underline{x}_2$ are jointly missing on a proportion λ_{12} of all units where

$$\lambda_{12} = \frac{(n - n^{(12)})}{n - 1}$$

then their post-imputation variances (a_{11}^*, a_{22}^*) and covariance (a_{12}^*) obtained from the completed sample are to be adjusted to

$$a_{11}^{*} + \lambda_{12} z_{11}, \tag{4.93}$$

$$a_{22}^{*} + \lambda_{12} z_{22} \tag{4.94}$$

and

$$a_{12}^{*} + \lambda_{12} z_{12} \tag{4.95}$$

where z_{11}, z_{22} are the diagonal elements in $\mathbf{B}_{11}^{-1}$ and z_{12} is the off-diagonal element.

Corollary 4.6.1

The bias of the variance of x_1 is given by the inverse of the first element in $\mathbf{A}^{-1}$. And the covariance of x_1 and x_2 is unbiased, i.e.,

$$c_{11}^{-1} = a_{11} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_1$$
$$c_{22}^{-1} = a_{22} - \underline{\alpha}_2' \mathbf{C}^{-1} \underline{\alpha}_2$$
$$c_{12}^{-1} = a_{12} - \underline{\alpha}_1' \mathbf{C}^{-1} \underline{\alpha}_2 = 0.$$

Proof

The results immediately follow by substituting $c_{12} = 0$ in formulae (4.87), (4.89) and (4.91).

Note that these are the same results obtained by Buck (1960) for the case of units with one missing value given by formulae (3.18) and (3.26) of sections 3.5.1 and 3.5.2 respectively.

Theorem 4.7

$$\mathbf{B}_{11}^{-1} = (n_c - 1)^{-1}\mathbf{Q}_e,$$

where

$$\mathbf{Q}_e = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)}\hat{\mathbf{B}}$$

is the errors sum of squares and cross products matrix obtained from the multivariate regression of $\mathbf{Y} = (\underline{x}_j, \underline{x}_k)$ on the remaining $k - 2$ variables, and n_c is the number of complete cases.

Proof

Denote the sample variances and covariance of the residuals of the multivariate regression of the j-th and k-th variables on the remaining $k - 2$ variables by Φ_{jj}, Φ_{kk} and Φ_{jk}, respectively. Then we have

$$\begin{bmatrix} \Phi_{jj} & \\ \Phi_{jk} & \Phi_{kk} \end{bmatrix} = \frac{1}{n_c - 1}\mathbf{Q}_e$$

$$= \frac{1}{n_c - 1}\begin{bmatrix} \underline{x}_j'\underline{x}_j - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_j & \\ \underline{x}_j'\underline{x}_k - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_k & \underline{x}_k'\underline{x}_k - \underline{\hat{\beta}}_k'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_k \end{bmatrix}$$

or

$$\Phi_{jj} = (n_c - 1)^{-1}\left\{\underline{x}_j'\underline{x}_j - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_j\right\}, \tag{4.96}$$

where $\underline{x}_j'\underline{x}_j$ and $\underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_j$ are the corrected total sum of squares and the sum of squares due to regression respectively, $\mathbf{X}_{(jk)} = (\underline{x}_1, \ldots, \underline{x}_{j-1}, \underline{x}_{j+1}, \ldots, \underline{x}_{k-1})$.

The expression for Φ_{jj} given in (4.96) can be written as:

$$\Phi_{jj} = (n_c - 1)^{-1}\left\{\underline{x}_j'\underline{x}_j - (n_c - 1)\underline{\alpha}_j'\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\underline{\alpha}_j\right\}$$

since

$$\hat{\beta}'_j = \underline{\alpha}'_j \mathbf{C}^{-1}$$

and

$$\mathbf{C} = (\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})/(n_c - 1)$$

$$\therefore \quad \Phi_{jj} = (n_c - 1)^{-1}\left\{\underline{x}'_j\underline{x}_j - (n_c - 1)\underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_j\right\}$$

$$= a_{jj} - \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_j, \tag{4.97}$$

Replacing j by k we have

$$\Phi_{kk} = a_{kk} - \underline{\alpha}'_k\mathbf{C}^{-1}\underline{\alpha}_k. \tag{4.98}$$

Similarly,

$$\Phi_{jk} = (n_c - 1)^{-1}\left\{\underline{x}'_j\underline{x}_k - \hat{\underline{\beta}}'_j(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\hat{\underline{\beta}}_k\right\}$$

$$= (n_c - 1)^{-1}\left\{\underline{x}'_j\underline{x}_k - (n_c - 1)\underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_k\right\}$$

$$= a_{jk} - \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_k. \tag{4.99}$$

Therefore

$$\begin{bmatrix}\Phi_{jj} & \\ \Phi_{jk} & \Phi_{kk}\end{bmatrix} = \frac{1}{n_c - 1}\mathbf{Q}_e$$

$$= \frac{1}{n_c - 1}\begin{bmatrix}\underline{x}'_j\underline{x}_j - \hat{\underline{\beta}}'_j(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\hat{\underline{\beta}}_j & \\ \underline{x}'_j\underline{x}_k - \hat{\underline{\beta}}'_j(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\hat{\underline{\beta}}_k & \underline{x}'_k\underline{x}_k - \hat{\underline{\beta}}'_k(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\hat{\underline{\beta}}_k\end{bmatrix}$$

$$= \begin{bmatrix}(a_{jj} - \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_j) & (a_{jk} - \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_k) \\ (a_{jk} - \underline{\alpha}'_j\mathbf{C}^{-1}\underline{\alpha}_k) & (a_{kk} - \underline{\alpha}'_k\mathbf{C}^{-1}\underline{\alpha}_k)\end{bmatrix}$$

$$= \mathbf{B}_{11}^{-1}. \tag{4.100}$$

Hence

$$\mathbf{B}_{11}^{-1} = (n_c - 1)^{-1}\mathbf{Q}_e.$$

Therefore the elements of $\mathbf{B}_{11}^{-1}$ are actually the sample variances and covariance of the residuals of the multivariate regression of the j-th and k-th variables on the remaining $k-2$ variables. Note that the elements of $\mathbf{Q}_e$ are readily obtained from the MANOVA table of the multivariate regression of the j-th and k-th variables on the remaining $k-2$ variables.

The importance of this result is that it enables us to obtain the bias of the post-imputation covariance matrix directly from the MANOVA table of the multivariate regression of $\underline{x}_j, \underline{x}_k$ on the remaining $k-2$ variables. Thus the actual determination of bias does not require the computations of $\mathbf{A}^{-1}$ or $\mathbf{B}_{11}^{-1}$. These are only required for the theoretical derivation of the amount of bias.

Theorem 4.8

$$\Phi_{jk} = a_{jk}(1 - R_{jk}^2),$$

where a_{jk} is the covariance of the j-th and k-th variables obtained from the n_c complete cases. R_{jk}^2 is the jk-th element in the matrix $(\mathbf{R})$ of the coefficients of determinations of the multivariate regression equation $(\hat{\underline{x}}_j \hat{\underline{x}}_k) = \mathbf{X}_{(jk)}\hat{\mathbf{B}}$, where $\mathbf{X}_{(jk)} = (\underline{x}_{i1}, \ldots, \underline{x}_{i(j-1)}, \underline{x}_{i(j+1)}, \underline{x}_{i(j+2)}, \ldots, \underline{x}_{i(k-1)})$, $i = 1, \ldots, n_c$.

Proof

From the multivariate regression equation

$$(\underline{x}_j, \underline{x}_k) = \mathbf{X}_{(jk)}\hat{\mathbf{B}} + \mathbf{U}, \tag{4.101}$$

we have

$$R^2_{jk} = 1 - \frac{(\underline{x}_j'\underline{x}_k - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_k)}{\underline{x}_j'\underline{x}_k}, \tag{4.102}$$

and from (4.99) we have

$$\Phi_{jk} = \frac{(\underline{x}_j'\underline{x}_k - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_k)}{n_c - 1}, \tag{4.103}$$

Substituting the value of $(\underline{x}_j'\underline{x}_k - \underline{\hat{\beta}}_j'(\mathbf{X}'_{(jk)}\mathbf{X}_{(jk)})\underline{\hat{\beta}}_k)$ from (4.103) in (4.102) we get

$$R^2_{jk} = 1 - \frac{(n_c - 1)\Phi_{jk}}{\underline{x}_j'\underline{x}_k}, \tag{4.104}$$

which implies

$$\Phi_{jk} = \frac{\underline{x}_j'\underline{x}_k(1 - R^2_{jk})}{n_c - 1}.$$

We should note that $\underline{x}_j'\underline{x}_k$ is the corrected cross product term of the regression equation given by (4.101), hence

$$(\underline{x}_j'\underline{x}_k/n_c - 1) = a_{jk},$$

where a_{jk} is the covariance of the j-th and k-th variables obtained from the n_c-complete cases. Therefore,

$$\Phi_{jk} = a_{jk}(1 - R^2_{jk}). \tag{4.105}$$

Thus if the j-th and k-th variables are jointly missing on $m^{(jk)}$ units out of n units, then from (4.105), the correction for the bias of the post-imputation covariance of the j-th and k-th variables is given by

$$\lambda_{jk}\Phi_{jk} = \lambda_{jk}a_{jk}(1 - R_{jk}^2), \tag{4.106}$$

where $\lambda_{jk} = m^{(jk)}/n$ is the proportion of units with missing values on both the j-th and k-th variables.
from which the relative bias is given by

$$\lambda_{jk}\frac{\Phi_{jk}}{a_{jj}} = \lambda_{jk}(1 - R_{jk}^2). \tag{4.107}$$

Note that (4.105) gives the maximum amount of bias in the estimation of the covariance of the j-th and k-th variables as a function of two forces, namely, R_{jk}^2 and the covariance of the j-th and k-th variables (a_{jk}) computed from the complete cases . This formula together with (4.16) play an important role in the investigation of the missingness mechanism in the method of Buck which will be discussed in section 4.6.

Note that from (4.41) we have

$$0 \leq R_{jk}^2 \leq 1.$$

Therefore, and from (4.107) it follows that

1- For fixed λ_{jk}, the relative bias decreases as R_{jk}^2 increases, and

2- For fixed R_{jk}^2, the relative bias increases as λ_{jk} increases.

Corollary 4.8.1

$$\Phi_{jj} = a_{jj}(1 - R_j^2),$$

Proof

The proof follows immediately by putting k=j in formulae (4.101) through (4.105).

Note that corollary 4.8.1 gives the same result obtained by theorem 4.2 for the case of units with one missing value. Thus theorem 4.2 can be considered as a special case of theorem 4.8.

Theorem 4.9

If the j-th and k-th variables are jointly missing on $m^{(jk)}$ units, then the post-imputation estimates, $V(x_j)$ and $Cov(x_j, x_k)$ are inconsistent.

Proof

Noting that the amount of bias in the estimation of the variance of the j-th variable and the covariance of the j-th and k-th variables are given, respectively, by

$$\frac{m^{(j)}}{n} z_{jj} = \frac{n - n^{(j)}}{n} \Phi_{jj} \tag{4.108}$$

and

$$\frac{m^{(jk)}}{n} z_{jk} = \frac{n - n^{(jk)}}{n} \Phi_{jk} \tag{4.109}$$

it follows that

$$\lim_{n \to \infty} \left[\frac{n - n^{(j)}}{n} \Phi_{jj} \right] \neq 0, \quad n^{(j)} \neq n \tag{4.110}$$

and

$$\lim_{n \to \infty} \left[\frac{n - n^{(jk)}}{n} \Phi_{jk} \right] \neq 0, \quad n^{(jk)} \neq n. \tag{4.111}$$

Thus the estimates of the variances and covariances given by Buck's method are generally inconsistent.

4.4.3 REAL DATA ILLUSTRATIONS

In this section we shall be concerned with the numerical illustration and validation of theorem 4.5. This is because theorem 4.5 summarizes most of the findings of sections 4.4.1 and 4.4.2.

We shall consider two patterns of missingness. For the first pattern, 11 observations on both $\underset{\sim}{x}_1$ and x_2 were picked at random from the data of Bumpus (1898) and considered as missing. This shall be referred to as pattern (5). Pattern (6) is obtained by selecting at random 13 observations on $\underset{\sim}{x}_1, x_2$ and x_5 from the data of Bumpus (1898) and consider them as missing.

The resulting incomplete data (missing data patterns (5) and (6)) as well as the summary statistics (means and covariance matrix) for the complete cases of the two patterns are given in the appendix. Also displayed in the appendix are the results of the multivariate regression analysis, required for the estimation of missing values in the two patterns. The data are analyzed using SPSS and STATGRAPHICS.

ANALYSIS OF MISSING DATA PATTERN (5)

Using the notation of section 3.3 we have, n = 49, $n_c = n^{(jk)} = 38$ and $m^{(jk)} = (n - n^{(jk)}) = 11$. Further,

$$\mathbf{A} = \begin{array}{c} \begin{array}{ccccc} x_1 & x_2 & x_3 & x_4 & x_5 \end{array} \\ \begin{pmatrix} 13.7020 & & & & \\ 14.7866 & 26.9616 & & & \\ 2.0257 & 2.5919 & .6253 & & \\ 1.4102 & 2.1404 & .3285 & .3051 & \\ 2.2888 & 2.9774 & .5419 & .3907 & 1.0035 \end{pmatrix} \end{array}$$

$$\mathbf{A}^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} .2221390 & & & & \\ -.0821358 & .11646 & & & \\ -.2802510 & .0218064 & 4.5568100 & & \\ -.011659 & -.484985 & -2.68933 & 11.7524 & \\ -.1070820 & .0188463 & -.8391640 & -1.657850 & 2.2834500 \end{pmatrix} \end{array}.$$

And since only x_1 and x_2 have joint missing values, it follows that

$$\mathbf{B}_{11} = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{pmatrix} .2221390 & -.0821358 \\ -.0821358 & .1164600 \end{pmatrix} \end{array} \text{and } \mathbf{B}_{11}^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 \end{matrix} \\ \begin{pmatrix} 6.08972 & 4.2949 \\ 4.2949 & 11.6157 \end{pmatrix} \end{array}.$$

Using Woolf's procedure, the quantities required for the estimation of the multivariate regression of $\underline{x}_1, \underline{x}_2$ on $\underline{x}_3, \underline{x}_4, \underline{x}_5$ are

$$\overline{X}_h = (31.45, 18.4763, 20.7842), \quad h = 3, 4, 5$$

and from $\mathbf{A}$, we have

$$\underline{\alpha}_1' = (2.0257 \;\; 1.4102 \;\; 2.2888), \; \underline{\alpha}_2' = (2.5919 \;\; 2.1404 \;\; 2.9774),$$

$$\mathbf{C} = \begin{array}{c} \begin{matrix} x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} .6253 & .3285 & .5419 \\ .5419 & .3907 & 1.0035 \end{pmatrix} \end{array}$$

and

$$\mathbf{C}^{-1} = \begin{array}{c} \begin{matrix} x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 4.12549 & -3.16905 & -.993976 \\ -.993976 & -1.78139 & 2.226830 \end{pmatrix} \end{array}.$$

The estimated coefficients of the regression of $\underline{x}_1$ on $\underline{x}_3, \underline{x}_4, \underline{x}_5$ are given by

$$\hat{\beta}'_{h1} = \underline{\alpha}'_1 \mathbf{C}^{-1} = (1.612998514 \quad 2.153959061 \quad .571155143), \quad h = 3,4,5$$

so that

$$\hat{\beta}_{01} = \overline{X}_1 - \sum_h \hat{\beta}_{h1} \overline{X}_h = 55.577, \quad h = 3,4,5$$

and the estimated coefficients of the regression of $\underline{x}_2$ on $\underline{x}_3, \underline{x}_4, \underline{x}_5$ are given by

$$\hat{\beta}'_{h2} = \underline{\alpha}'_2 \mathbf{C}^{-1} = (.950358769 \quad 5.683521675 \quad .240990092), \quad h = 3,4,5$$

so that

$$\hat{\beta}_{02} = \overline{X}_2 - \sum_h \hat{\beta}_{h2} \overline{X}_h = 101.1973, \quad h = 3,4,5$$

then the missing values are estimated and a completed data matrix is created by imputing these estimates.

The quantities required for the estimation of the bias are

$$V(\hat{x}_1) = \underline{\alpha}'_1 \mathbf{C}^{-1} \underline{\alpha}_1 = 7.612224049$$

$$V(\hat{x}_2) = \underline{\alpha}'_2 \mathbf{C}^{-1} \underline{\alpha}_2 = 15.34576859$$

$$\mathrm{Cov}(\hat{x}_1, \hat{x}_2) = \underline{\alpha}'_1 \mathbf{C}^{-1} \underline{\alpha}_2 = 10.49162215$$

It follows that the maximum amount of bias in the estimation of $V(x_1), V(x_2)$, and $\mathrm{Cov}(x_1, x_2)$ are given, respectively, by

$$z_{11} = a_{11} - \underline{\alpha}'_1 \mathbf{C}^{-1} \underline{\alpha}_1 = 13.7020 - 7.612224049 = 6.0897$$

$$z_{22} = a_{22} - \underline{\alpha}'_2 \mathbf{C}^{-1} \underline{\alpha}_2 = 26.9616 - 15.34576859 = 11.6158$$

$$z_{12} = a_{12} - \underline{\alpha}'_1 \mathbf{C}^{-1} \underline{\alpha}_2 = 14.7866 - 10.49162215 = 4.2950$$

which are the same figures displayed in $\mathbf{B}_{11}^{-1}$.

Thus the amount of bias is actually a function of the elements of $\mathbf{B}_{11}^{-1}$. And since only $n - n^{(jk)} = 11$ observations are jointly missing on x_1 and x_2, it follows that the estimates obtained from the completed data for the variances a_{11}^* of x_1, a_{22}^* of x_2 and the covariance a_{12}^* of x_1 and x_2 have to be adjusted for bias using (4.93) through (4.95).

Now, from the results of the analysis of the incomplete data pattern (5) the MANOVA table of the multivariate regression of x_1, x_2 on the remaining $k-2$ variables is given by the following table:

Table 4(6): MANOVA TABLE FOR THE MULTIVARIATE REGRESSION OF x_1, x_2 on x_3, x_4, x_5

S.V.	d.f.	SSP.
Reg.	3	$\begin{bmatrix} 281.64428 & \\ 388.18 & 567.79093 \end{bmatrix} = \mathbf{Q}_h$
Resid.	34	$\begin{bmatrix} 225.32940 & \\ 158.92 & 429.78802 \end{bmatrix} = \mathbf{Q}_e$
Total (corr.)	37	$\begin{bmatrix} 506.97368 & \\ 547.1 & 997.57895 \end{bmatrix} = \mathbf{Q}_t$

thus

$$\frac{1}{37}\mathbf{Q}_e = \frac{1}{37}\begin{bmatrix} 225.32940 & 158.92 \\ & 429.78802 \end{bmatrix}$$

$$= \begin{bmatrix} 6.090 & 4.2951 \\ & 11.6160 \end{bmatrix}$$

$$= \mathbf{B}_{11}^{-1}.$$

Therefore the computation of the amount of bias from the MANOVA table

is straight forward. It does not require the computation of either $\mathbf{A}^{-1}$ or $\mathbf{B}_{11}^{-1}$.

ANALYSIS OF MISSING DATA PATTERN (6)

Here, we have n = 49, $n_c = n^{(jkl)} = n^{(125)} = 36$ and $m^{(jkl)} = m^{(125)} = (n - n^{(125)}) = 13$. Further, we re-arrange the elements of A so that those variables with jointly missing values belong to $\mathbf{D}_{11}$. Recall that $\mathbf{D}_{11}$ is the covariance matrix of those variables with jointly missing values. And since we have $\underline{x}_1, x_2$ and x_5 with jointly missing values then,

$$\mathbf{A} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 11.9714 & & & & \\ 12.1524 & 23.9111 & & & \\ 1.4510 & 2.3035 & .5395 & & \\ .9129 & 1.8600 & .2398 & .2231 & \\ 1.9848 & 2.6079 & .2777 & .1984 & .8739 \end{pmatrix} \end{array}$$

$$\mathbf{D}_{11} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_5 \end{matrix} \\ \begin{pmatrix} 11.9714 & 12.1524 & 1.9848 \\ 12.1524 & 23.9111 & 2.6079 \\ 1.9848 & 2.6079 & .8739 \end{pmatrix} \end{array},$$

so that the elements of $\mathbf{D}_{11}$ are the variances and covariances of those variables with jointly missing values. Also we have

$$\mathbf{A}^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 \end{matrix} \\ \begin{pmatrix} 0.209501 & & & & \\ -.0802344 & .175021 & & & \\ -.211635 & -.0320269 & 3.92829 & & \\ 0.237497 & -1.0062 & -3.12201 & 15.2638 & \\ -0.223049 & -0.101457 & .0367256 & -9.91235E-3 & 1.94423 \end{pmatrix} \end{array}$$

$$\mathbf{B}_{11} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_5 \end{matrix} \\ \begin{pmatrix} 0.209501 & -0.0802344 & -0.223049 \\ -0.0802344 & 0.175021 & -0.101457 \\ -0.223049 & -0.101457 & 1.94423 \end{pmatrix} \end{array}$$

and

$$\mathbf{B}_{11}^{-1} = \begin{array}{c} \begin{matrix} x_1 & x_2 & x_5 \end{matrix} \\ \begin{pmatrix} 7.45266 & 4.03417 & 1.06551 \\ 4.03417 & 8.07554 & 0.884225 \\ 1.06551 & 0.884225 & 0.682724 \end{pmatrix} \end{array}.$$

Thus multiplying the elements of $\mathbf{B}_{11}^{-1}$ by the proportion of missing values we get the amount of bias in the corresponding elements of $\mathbf{D}_{11}$.

Now, from the results of the analysis of the incomplete data pattern (6) given in the appendix, the MANOVA table of the multivariate regression of x_1, x_2 and x_5 on the remaining $k-3$ variables is given by the following table:

Table 4(7): MANOVA TABLE FOR THE MULTIVARIATE REGRESSION OF x_1, x_2, x_5 on x_3, x_4

S.V.	d.f.	SSP.
Reg.	2	$\begin{bmatrix} 158.15068 & & \\ 284.15000 & 554.3019 & \\ 32.18000 & 60.3400 & 6.69299 \end{bmatrix} = \mathbf{Q}_h$
Resid.	33	$\begin{bmatrix} 260.84932 & & \\ 141.19000 & 282.58699 & \\ 37.29000 & 30.94000 & 23.89257 \end{bmatrix} = \mathbf{Q}_e$
Total (corr.)	35	$\begin{bmatrix} 419.000 & & \\ 425.340 & 836.88889 & \\ 69.470 & 91.28000 & 30.58556 \end{bmatrix} = \mathbf{Q}_t$

thus

$$\frac{1}{35}\mathbf{Q}_e = \frac{1}{35}\begin{bmatrix} 260.84932 & & \\ 141.19000 & 282.58699 & \\ 37.29000 & 30.94000 & 23.89257 \end{bmatrix}$$

$$= \begin{bmatrix} 7.4528 & & \\ 4.0340 & 8.073914 & \\ 1.0654 & 0.884000 & 0.682645 \end{bmatrix}$$

$$= \mathbf{B}_{11}^{-1}.$$

Thus the elements of $\mathbf{B}_{11}^{-1}$ are actually the sample variances and covariances of the residuals of the multivariate regression of the elements of $\mathbf{D}_{11}$ on the remaining variables.

4.5 BUCK'S METHOD IN REGRESSION ANALYSIS

In this section we shall investigate the effect of imputation via Buck's method on the estimated regression coefficients, coefficient of determination, and the standard error of the estimated regression coefficients in regression analysis. The discussion will concentrate on the case of one variable subject to missingness. Specifically, we shall assume that an incomplete multivariate data set was completed via Buck's method and made available to the users. A user may wish to estimate regression functions on the basis of the completed data set. Our objective is to investigate the validity of the results and conclusions obtained by the data user.

Consider a sample $\mathbf{X} = (x_{ij})$, $(i = 1,\ldots,n; j = 1,\ldots,k)$ of n units (cases), on each of which it is desired to have measurements on k variables. Assume that each unit can have more than one missing value. Using the results of section 3.6, for each $j = 1\ldots,k$, we have

$$n = n_c + m^{(j)} + \sum_{\ell \neq j}^{k} m^{(\ell)} - \sum_{i=2}^{k-1}(i-1)U^{(i)}, \qquad (4.112)$$

where $U^{(i)}$ is the number of units with 'i' missing values, $i > 1$. That is, the n units on the j-th variable can be decomposed into: n_c-complete units, $m^{(j)}$ units with x_j only missing, and $\sum_{\ell \neq j}^{k} m^{(\ell)} - \sum_{i=2}^{k-1}(i-1)U^{(i)}$ units with x_j observed but any combination of the remaining $k-1$ variables missing.

Note that the decomposition given by (4.112) above corresponds to the case of units with more than one missing value. The corresponding decomposition for the case of units with one missing value is obtained from (4.112) by setting $\sum_{i=2}^{k-1}(i-1)U^{(i)} = 0$. Thus we get

$$n = n_c + m^{(j)} + \sum_{\ell \neq j}^{k} m^{(\ell)} \tag{4.113}$$

and for the case of one variable subject to missingness we set

$$\sum_{\ell \neq j}^{k} m^{(\ell)} = \sum_{i=2}^{k-1}(i-1)U^{(i)} = 0 \tag{4.114}$$

in (4.112) above. Thus the n units are decomposed as

$$n = n_c + m^{(j)} = n^{(j)} \tag{4.115}$$

Now, assume that the missing values of the j-th variable were imputed via Buck's method as outlined in section 3.3. That is, from the model

$$x_{ij} = \beta_o + \sum_{h \neq j}^{k} x_{ih}\beta_{hj} + \epsilon_{ij}, \; i = 1, \ldots, n_c \tag{4.116}$$

we estimate the complete-case regression coefficients $\underline{\beta}_c$ by

$$\hat{\underline{\beta}}_c = (X_c' X_c)^{-1} X_c' x_{j,obs} \tag{4.117}$$

by minimizing the complete-case residual sum of squares given by

$$\epsilon'\epsilon = \sum_{n_c}(x_{ij} - \hat{x}_{ij})^2, \tag{4.118}$$

so that

$$\hat{x}_{j,mis} = \mathbf{X}_{n-n_c}\underline{\hat{\beta}}_c, \tag{4.119}$$

where $\mathbf{X}_{n-n_c}$ is the matrix of the incomplete-cases.

Now, assume that a user of the completed data wishes to fit the model

$$x_{ij} = \beta_o^* + \sum_{h\neq j}^{k} x_{ih}\beta_{hj}^* + \zeta_{ij},\ i = 1,\ldots,n_c,\ldots,n. \tag{4.120}$$

Then we have the following results for the case of one variable subject to missingness:

Theorem 4.10

The pre– and post-imputation regression coefficients ($\underline{\hat{\beta}}_c$ and $\underline{\hat{\beta}}^*$) of the regression of the j-th variable on the remaining $k-1$ variables are the same, that is,

$$\underline{\hat{\beta}}_c = \underline{\hat{\beta}}^*$$

Proof

Let SSE_j be the residual sum of squares of the model given by (4.120). Using the decomposition given by (4.115), we can write

$$\begin{aligned} SSE_j = \zeta'\zeta &= \sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2 \\ &= \sum_{n_c}(x_{ij} - \hat{x}_{ij})^2 + \sum_{m(j)}(x_{ij} - \hat{x}_{ij})^2 \end{aligned} \tag{4.121}$$

the proof follows by noting that the second term of (4.121) is already minimized by imputations and the first term is minimized by $\hat{\beta}_c$ given in (4.117).

Thus the Least squares estimates of the regression coefficients obtained from the completed data are the same as those obtained from the complete cases.

It is interesting to note that the above result does not hold for the case of units with one missing value. In the latter case, using the decomposition given by (4.113), the residual sum of squares can be written as

$$\begin{aligned} \mathrm{SSE_j} = \epsilon'\epsilon &= \sum_{i=1}^{n}(x_{ij} - \hat{x}_{ij})^2 \\ &= \sum_{n_c}(x_{ij} - \hat{x}_{ij})^2 + \sum_{\star}(x_{ij} - \hat{x}_{ij})^2 + \sum_{m^{(j)}}(x_{ij} - \hat{x}_{ij})^2 \qquad (4.122) \end{aligned}$$

where $\star = \sum_{\ell \neq j} m^{(\ell)}$

$$= \sum_{n_c}(x_{ij} - \hat{x}_{ij})^2 + \sum_{\star}(x_{ij} - \hat{x}_{ij})^2 \qquad (4.123)$$

since the last term of (4.122) is minimized by imputations.

We should note that the first term of (4.123) is minimized by $\hat{\beta}_c$ given in (4.117). However, the second term of (4.123) is not minimized by imputations since it is independent of the imputed values. It therefore follows that

$$\underline{\hat{\beta}}_c \neq \underline{\hat{\beta}}^{\star}$$

Similarly, using the decomposition given by (4.112), the same result can be shown to be true for the case of units with more than one missing value.

Theorem 4.11

The estimated standard errors of the regression coefficients are deflated by imputations.

Proof

Let $S(\underline{\hat{\beta}}^*)$ be the dispersion matrix of the estimated regression coefficients of the model given by (4.120). Then we have

$$S(\underline{\hat{\beta}}^*) = (\mathbf{X}'\mathrm{X})^{-1}(n-1)^{-1}SSE_j \tag{4.124}$$

substituting the value of SSE_j from (4.121) we have

$$\begin{aligned} S(\underline{\hat{\beta}}^*) &= (\mathbf{X}'\mathrm{X})^{-1}(n-1)^{-1}SSE_j \\ &= (\mathbf{X}'\mathrm{X})^{-1}(n-1)^{-1}\left(\sum_{n_c}(x_{ij}-\hat{x}_{ij})^2 + \sum_{m^{(j)}}(x_{ij}-\hat{x}_{ij})^2\right) \\ &= (\mathbf{X}'\mathrm{X})^{-1}(n-1)^{-1}\left(\sum_{n_c}(x_{ij}-\hat{x}_{ij})^2\right) \end{aligned} \tag{4.125}$$

since $\sum_{m^{(j)}}(x_{ij}-\hat{x}_{ij})^2 = 0$ by imputations.

Thus the estimated standard errors of the regression coefficients will tend to be misleadingly smaller due to imputation. Consequently,

Corollary 4.11.1

The conventional t-statistic for testing the significance of the individual regression coefficients is inflated by imputations.

Proof

We have

$$t = \hat{\beta}_j^*/\sqrt{V(\hat{\beta}_j^*)} \tag{4.126}$$

The proof follows by noting that $\sqrt{V(\hat{\beta}_j^*)}$ will be misleadingly smaller due to imputations as shown in theorem 4.11 above. Hence type I error will be made more frequently. In other words, the individual regression coefficients will have more chances of being declared significantly different from zero.

Theorem 4.12

The coefficient of determination (R_j^2) of the model given by (4.120) above is unrealistically inflated by imputations.

Proof

R_j^2 can be written as

$$R_j^2 = 1 - \frac{SSE}{CTSS} = 1 - \frac{SSE}{SSE + SSR/\beta_o} \tag{4.127}$$

where SSR/β_o and CTSS are the corrected regression and total sums of squares respectively. The proof follows by noting that SSE is reduced due to imputation.

To avoid adopting the above mentioned misleading results, the sample variance of the residuals must be corrected for bias as outlined in section 3.5.1.

However, the crucial question is whether or not the data user is aware of the existence of imputed values. If the answer is 'no' then the user will 'unconsciously' adopt some or all of the above mentioned misleading conclusions. This situation is more likely to happen in cases where the data imputer and the data user are two distinct entities which is always true for public-use data bases which are shared by many users.

Hence the conclusion arrived at in this section is that the imputed observations should be clearly identified and the method of imputation must be well documented. This will minimize the possibilities of adopting misleading conclusions by the data user.

As a numerical verification of theorems 4.10–4.12 and corollary 4.11.1 given above, the following pattern of missingness is obtained from the data of Bumpus by considering X_4 and X_5 only. Then the missing data pattern (7), given in table A(8), is obtained by assuming that the last 15 observations

of X_5 are missing. The missing observations are then imputed via Buck's method. The results of the pre-imputation and post-imputation regression analyses are given in tables A(8.1)–A(8.4) of the appendix. Comparing tables A(8.2) and A(8.4) we note that:

1- The pre–and post-imputation regression coefficients are the same.

2- The post-imputation standard errors of the regression coefficients are deflated.

3- The post-imputation t-statistics are inflated.

4- The p-values at which $H_0 : \underline{\beta} = 0$ is rejected are decreasing.

And by comparing tables A(8.1) and A(8.3) we note that the coefficient of determination is inflated while its standard error is deflated. Moreover, comparing tables A(5.12) and A(5.14) we note that theorem 4.10 does not hold when more than one variable are subject to missingness.

4.6 BUCK'S METHOD AND THE MISSINGNESS MECHANISM

Most work on inference with missing data is based on the explicit or implicit assumption that the missing data are missing at random (MAR) or missing completely at random (MCAR). MCAR is the *strong* assumption of missingness which means that the process causing an observation to be missing is independent of the value of the observation (observed or missing) or the identity of the underlying population. On the other hand, the *weak* assumption of missingness MAR means that the missing observation is independent of its value but might depend on the value of an observed variable. A direct consequence of the MCAR assumption is that the joint distribution of any subset of the data is the same whether it contains missing values or not. Hence standard statistical techniques, under MCAR assumption, can be applied to the completely recorded data to obtain estimates of the missing

observations or sample estimates of the corresponding population parameters. While it is immaterial whether the MAR or MCAR assumption is used for imputation strategy, it makes a great difference in deletion-pairwise methods which require only MCAR for their validity (see Little and Rubin, 1987, chap. 3).

The method of Buck can be viewed as a two-stage process for the joint estimation of the sample missing observations and the corresponding population parameters. In the first stage, using deletion-pairwise strategy, a subset of complete cases is obtained by deleting any case with at least one missing observation. The second stage applies regression techniques to the subset of complete cases obtained in the first stage to impute the missing data. Thus the weaknesses of both imputation and deletion-pairwise strategies are reflected in the method. Afifi and Elashoff (1966) stated that: *"Buck carries out his calculations conditional upon the complete vector observations. The particular reasons for this conditioning are not clear"*. (See Afifi and Elashoff (1966), pp. 600, second paragraph). However, although Buck (1960) does not make any explicit assumptions about the missingness mechanism in his method, yet the application of the deletion-pairwise strategy in the first stage of the method is actually an implicit MCAR assumption for the missingness mechanism. In other words, the conditioning of Buck's calculations upon the complete data vector is an implicit MCAR assumption for the missingness mechanism. It is only under MCAR assumption that the complete cases are a random sub sample of the original population. It is worth mentioning that the literature about the missingness mechanism, MAR and MCAR assumptions have only been developed in 1976 by Rubin, D. B.

Note that (4.16) expresses the bias of the j-th variable's variance as a function of two forces, namely, R_j^2 and the variance of the j-th variable (a_{jj}^*)

obtained from the n_c-complete cases. Similarly, formula (4.105) expresses the bias of the covariance of the j-th and k-th variables as a function of R^2_{jk} and the covariance of the j-th and k-th variables (a^*_{jk}) obtained from the n_c-complete cases. The dependence of the bias in the covariance matrix on a^*_{jj}, R^2_j, a^*_{jk} and R^2_{jk} is actually a dependence on the n_c-complete cases of $\mathbf{X}$ since these quantities are computed from them. Furthermore, the n_c-complete cases are a result of a specific selection which is completely determined by the pattern (frequency distribution) of missing observations. For each pattern of missing data we will have a different set of n_c-complete cases. Each set of these n_c-complete cases can be considered as a random sub sample of the original population if and only if there is no selection bias in considering the n_c-complete cases. If the pattern of missing values, that leads to the specific selection of the n_c-complete cases, is not missing completely at random then the n_c-complete cases are not a random sub-sample of the original population. Thus all the estimates obtained, including a^*_{jj}, a^*_{jk} and the conclusions arrived at will be biased towards the non-representative sample of the n_c-complete cases. Thus formulae (4.16) and (4.105) actually state that the performance of Buck's method is only valid if the missing observations are missing completely at random (MCAR).

4.7 CONCLUSIONS

Various issues about Buck's method have been discussed in this chapter. The theoretical and/or numerical verification of some of the statements about the method given by Buck and other scholars have been the prime objective. However, in the process of this verification many results have come up. In most of the cases, the discussion of the method came under two subheadings, one for the case of units with one missing value and the other for the case of units with more than one missing value. This is because we

believe that the statistical properties of the estimates in the two cases are quite different. Hence the theoretical derivations of the statistical properties of the estimates in the two cases require different methods of analysis. For instance, only univariate regression and analysis of variance are required in the case of units with one missing value. The case of units with more than one missing value requires, as tools of analysis, the multivariate regression, multivariate analysis of variance and covariance analysis.

Issues that have been discussed in this chapter include, the biasedness in the covariance matrix, missingness mechanism in Buck's method and Buck's method in regression analysis.

In the next chapter we shall try to set the context for some connections between imputation techniques (Buck's method) and the maximum likelihood methods of estimation from incomplete data.

CHAPTER V

SOME RELATIONS BETWEEN IMPUTATION AND MAXIMUM LIKELIHOOD METHODS OF ESTIMATION

5.1 INTRODUCTION

In chapter II we have introduced three main strategies for dealing with the problem of missing data in multivariate analysis. These are: the deletion-pairwise strategy; the imputation strategy; and the maximum likelihood (ML) approach. For the ML approach, the factorization method of Anderson (1957) and the EM algorithm were briefly introduced in sections 2.4.1 and 2.4.2 respectively. We have already noted that the three areas are not mutually disjoint, in the sense that a single method can combine more than one strategy. For example Buck's method is a combination of deletion and imputation strategies.

Unlike imputation techniques, the statistical theory of ML estimation from incomplete data is relatively well developed. This is because the corresponding theory in the complete-data case is well developed in most of its aspects. Moreover, most of the statistical literature on missing data considers each of the above mentioned three strategies as a distinct category. There seems to be very little research on the possible interactions between the three strategies of handling missing data.

The objective of this chapter is to investigate more on the possible links between the three areas. Specifically, we shall investigate the relationship between Buck's method, as a deletion-imputation strategy, and

1- Anderson's (1957) factorization method

2- The EM algorithm

as ML methods of estimation from incomplete multivariate data.

5.2 RELATION BETWEEN ANDERSON'S AND BUCK'S METHODS: UNITS WITH ONE MISSING VALUE SUBJECT TO ONE VARIABLE

In section 2.4.1 we presented the theoretical aspects of Anderson's (1957) factorization approach. The objective of this section is to establish the equivalence between Anderson's and Buck's methods. To achieve this we apply the theory of Anderson's method, introduced in chapter II, to the incomplete bivariate, trivariate and multivariate normal distributions to estimate the unknown parameters. Then these estimates are compared and contrasted with those obtained via the corresponding special cases of Buck's method.

5.2.1 ANDERSON'S METHOD FOR THE BIVARIATE NORMAL DISTRIBUTION

Consider a bivariate normal sample (y_{i1}, y_{i2}), $i = 1, \ldots, n$; where (y_{i1}, y_{i2}), $i = 1, \ldots, n_c$ are n_c-complete bivariate observations and $y_{i1}, i = n_c + 1, \ldots, n$ are $(n - n_c)$ observations with missing y_{i2}. Thus we have n_c bivariate observations on both y_{i1} and y_{i2} and $(n-n_c)$ univariate observations on y_{i1}, that is,

$$f(y_{i1}, y_{i2}) = (2\pi)^{-n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})}, \tag{5.1}$$

and

$$f(y_{i1}) = (2\pi)^{-\frac{(n-n_c)}{2}} \sigma_{11}^{-\frac{(n-n_c)}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i-\mu)^2}{\sigma_{11}}}. \tag{5.2}$$

Thus the joint density of a bivariate normal sample with n_c bivariate (y_{i1}, y_{i2}) observations and $n - n_c$ univariate y_{i1} observations is given by

$$f(Y_{obs} \mid \underline{\mu}, \Sigma) = (2\pi)^{-n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})}$$
$$.(2\pi)^{-\frac{(n-n_c)}{2}} \sigma_{11}^{-\frac{(n-n_c)}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i-\underline{\mu})^2}{\sigma_{11}}}. \tag{5.3}$$

Therefore the loglikelihood function (ignoring the missing data mechanism) is

$$\ell(\underline{\mu}, \Sigma \mid Y_{obs}) = -\frac{n_c}{2}\ln \mid \Sigma \mid -\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu}) - \frac{1}{2}(n - n_c)\ln \sigma_{11}$$
$$-\frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i - \mu)^2}{\sigma_{11}}. \tag{5.4}$$

Thus the maximum likelihood estimates of $\underline{\mu}$ and Σ can be found by maximizing (5.4) with respect to these values. The likelihood equations, however, do not have an obvious solution. A simpler approach to this maximization was suggested by Anderson (1957).

Anderson (1957) noted that the joint density of y_{i1} and y_{i2} can be factorized into two terms, that is,

$$f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma) = f(y_{i1} \mid \mu_1, \sigma_{11}) f(y_{i2} \mid y_{i1}, \beta_o, \beta_1, \sigma^2) \tag{5.5}$$

where, by the properties of the bivariate normal distribution,

$$f(y_{i1} \mid \mu_1, \sigma_{11}) \sim N(\mu_1, \sigma_{11}),$$

$$f(y_{i2} \mid y_{i1}, \beta_o, \beta_1, \sigma^2) \sim N(\beta_o + \beta_1 y_{i1}, \sigma^2)$$

and

$$\beta_1 = \sigma_{12}/\sigma_{11}, \tag{5.6}$$

$$\beta_o = \mu_2 - \beta_1 \mu_1, \tag{5.7}$$

$$\sigma^2 = \sigma_{22} - \sigma_{12}^2/\sigma_{11}. \tag{5.8}$$

The likelihood function of the observed data Y_{obs} can be factorized as follows:

$$\begin{aligned} L(\mu, \Sigma \mid Y_{obs}) &= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}) \\ &= \prod_{i=1}^{n_c} f(y_{i1} \mid \mu_1, \sigma_{11}) f(y_{i2} \mid y_{i1}, \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}) \\ &= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}) \prod_{i=1}^{n_c} f(y_{i2} \mid y_{i1}, \beta_o, \beta_1, \sigma^2) \end{aligned} \tag{5.9}$$

Then the ML estimates of $\mu_1, \sigma_{11}, \beta_o, \beta_1$ and σ^2 are those values that maximize the individual terms of the RHS of (5.9).

Maximizing the first term of the RHS of (5.9) with respect to μ_1 and σ_{11} we have

$$L(\mu_1, \sigma_{11} \mid y_{i1}) = (2\pi)^{\frac{-n}{2}} (\sigma_{11})^{\frac{-n}{2}} \text{Exp}[-\frac{1}{2\sigma_{11}} \sum_{i=1}^{n} (y_{i1} - \mu_1)^2]$$

from which the loglikelihood is

$$\ell(\mu_1, \sigma_{11} \mid y_{i1}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_{11}) - \frac{1}{2\sigma_{11}} \sum_{i=1}^{n} (y_{i1} - \mu_1)^2. \tag{5.10}$$

Thus

$$\frac{\partial \ell}{\partial \mu_1} = \frac{1}{\sigma_{11}} \sum_{i=1}^{n} (y_{i1} - \mu_1) = 0$$

or

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{y_{i1}}{n} = \bar{y}_1 \tag{5.11}$$

and

$$\frac{\partial \ell}{\partial \sigma_{11}} = -\frac{n}{2\sigma_{11}} + \frac{2\sum_{i=1}^{n}(y_{i1}-\mu_1)^2}{4\sigma_{11}^2} = 0$$

or

$$\hat{\sigma}_{11} = \frac{\sum_{i=1}^{n}(y_{i1}-\mu_1)^2}{n} = \frac{\sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2}{n}. \tag{5.12}$$

Thus the ML estimates of the mean and variance of Y_1 are the usual ML estimates of the univariate normal distribution based on 'n' observations $(y_{11},\ldots,y_{n1})$.

Maximizing the second term of the RHS of (5.9) with respect to β_o , β_1 and σ^2 we have

$$L(\beta_o+\beta_1 y_{i1},\sigma^2 \mid y_{i1}) = (2\pi)^{\frac{-n_c}{2}}(\sigma^2)^{\frac{-n_c}{2}}\text{Exp}[-\frac{1}{2\sigma^2}\sum_{i=1}^{n_c}(y_{i2}-\beta_o-\beta_1 y_{i1})^2]$$

from which the loglikelihood is

$$\ell(.\mid y_{i1}) = -\frac{n_c}{2}\ln(2\pi) - \frac{n_c}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n_c}(y_{i2}-\beta_o-\beta_1 y_{i1})^2. \tag{5.13}$$

Thus

$$\frac{\partial \ell}{\partial \beta_o} = \frac{4\sigma^2\sum_{i=1}^{n_c}(y_{i2}-\beta_o-\beta_1 y_{i1})}{4\sigma^4} = 0$$

from which

$$\hat{\beta}_o = \bar{y}_2^c - \beta_1\bar{y}_1^c, \tag{5.14}$$

where $\bar{y}_1^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{i1}$ and $\bar{y}_2^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{i2}$. That is, $\bar{y}_1^c$ and $\bar{y}_2^c$ are the complete-case estimates of the means.

Similarly we have,

$$\frac{\partial \ell}{\partial \beta_1} = \frac{4\sigma^2\sum_{i=1}^{n_c} y_{i1}(y_{i2}-\beta_o-\beta_1 y_{i1})}{4\sigma^4} = 0.$$

On solving and substituting the value of $\hat{\beta}_o$ from (5.14) we get

$$\hat{\beta}_1 = S_{12}/S_{11} \tag{5.15}$$

where

$$S_{12} = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i1} - \bar{y}_1^c)(y_{i2} - \bar{y}_2^c)$$

and

$$S_{11} = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i1} - \bar{y}_1^c)^2.$$

That is, the ML estimates of the parameters of the conditional distribution $f(y_{i2} \mid y_{i1})$ are based on the n_c-complete cases.

Finally,

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{2\sum_{i=1}^{n_c}(y_{i2} - \beta_o - \beta_1 y_{i1})^2}{\sigma^2} = 0.$$

On solving we get

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n_c}(y_{i2} - \hat{\beta}_o - \hat{\beta}_1 y_{i1})^2}{n_c}$$

$$= \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i2} - \hat{y}_{i2})^2 = S_{22} - \hat{\beta}_1^2 S_{11}. \tag{5.16}$$

Using the invariance property of the ML method of estimation, estimates of other parameters can then be obtained. In particular, from (5.7) above we have

$$\begin{aligned}\hat{\mu}_2 &= \hat{\beta}_o + \hat{\beta}_1\hat{\mu}_1 \\ &= (\bar{y}_2^c - \hat{\beta}_1\bar{y}_1^c) + \hat{\beta}_1\hat{\mu}_1 \\ &= \bar{y}_2^c + \hat{\beta}_1(\hat{\mu}_1 - \bar{y}_1^c),\end{aligned} \tag{5.17}$$

and from (5.8) we have

$$\begin{aligned}\hat{\sigma}_{22} &= \hat{\sigma}^2 + \hat{\beta}_1^2\hat{\sigma}_{11} \\ &= S_{22} - S_{12}^2/S_{11} + \hat{\beta}_1^2\hat{\sigma}_{11} \\ &= S_{22} - \hat{\beta}_1^2 S_{11} + \hat{\beta}_1^2\hat{\sigma}_{11} \\ &= S_{22} + \hat{\beta}_1^2(\hat{\sigma}_{11} - S_{11}),\end{aligned} \tag{5.18}$$

and finally, from (5.6) we have

$$\hat{\sigma}_{12} = \hat{\beta}_1\hat{\sigma}_{11} = \frac{S_{12}}{S_{11}}\hat{\sigma}_{11}. \tag{5.19}$$

5.2.2 ANDERSON'S METHOD FOR THE TRIVARIATE NORMAL DISTRIBUTION

Consider a trivariate normal sample (y_{i1}, y_{i2}, y_{i3}), $i = 1, \ldots, n$; where (y_{i1}, y_{i2}, y_{i3}), $i = 1, \ldots, n_c$ are n_c-complete trivariate observations and (y_{i1}, y_{i2}), $i = n_c + 1, \ldots, n$ are $(n - n_c)$ bivariate observations with missing y_{i3}. Thus we have n_c trivariate observations on both y_{i1}, y_{i2} and y_{i3} and $(n - n_c)$ bivariate observations on y_{i1} and y_{i2}, that is,

$$f(y_{i1}, y_{i2}, y_{i3}) = (2\pi)^{-\frac{3}{2}n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})}, \tag{5.20}$$

and

$$f(y_{i1}, y_{i2}) = (2\pi)^{-n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})}. \tag{5.21}$$

Thus the joint density of a trivariate normal sample with n_c trivariate (y_{i1}, y_{i2}, y_{i3}) observations and $(n - n_c)$ bivariate (y_{i1}, y_{i2}) observations is given by

$$\begin{aligned}f(Y_{obs} \mid \mu, \Sigma) = (2\pi)^{-\frac{3}{2}n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})} \\ .(2\pi)^{-n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})}.\end{aligned} \tag{5.22}$$

Therefore the loglikelihood function (ignoring the missing data mechanism) is

$$\ell(\underline{\mu}, \Sigma \mid Y_{obs}) = -\frac{n_c}{2} \ln \mid \Sigma \mid -\frac{1}{2} \sum_{i=1}^{n_c} (y_i - \underline{\mu})' \Sigma^{-1} (y_i - \underline{\mu})$$

$$- \frac{n_c}{2} \ln \mid \Sigma \mid -\frac{1}{2} \sum_{i=n_c+1}^{n} (y_i - \underline{\mu})' \Sigma^{-1} (y_i - \underline{\mu}). \quad (5.23)$$

Thus the maximum likelihood estimates of μ and Σ can be found by maximizing (5.23) with respect to these values. A simpler approach to this maximization is Anderson's factorization method.

The joint density of y_{i1}, y_{i2} and y_{i3} can be factorized as follows:

$$f(y_{i1}, y_{i2}, y_{i3} \mid \underline{\mu}, \Sigma) = f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma).f(y_{i3} \mid y_{i1}, y_{i2}, \beta_o, \beta_1, \beta_2, \sigma^2) \quad (5.24)$$

where, by the properties of the trivariate normal distribution,

$$f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma) \sim N(\mu_1, \mu_2; \sigma_{11}, \sigma_{22}),$$

$$f(y_{i3} \mid y_{i1}, y_{i2}, \beta_0, \beta_1, \beta_2, \sigma^2) \sim N(\beta_o + \beta_1 y_{i1} + \beta_2 y_{i2}; \sigma^2)$$

and

$$\underline{\beta}_3' = \underline{\alpha}_3' \mathbf{C}^{-1}$$

or

$$[\sigma_{31} \quad \sigma_{32}] \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}^{-1} = [\beta_1 \quad \beta_2].$$

From which we have

$$\beta_1 = \frac{\sigma_{31}\sigma_{22} - \sigma_{32}\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \quad (5.25)$$

$$\beta_2 = \frac{\sigma_{32}\sigma_{11} - \sigma_{31}\sigma_{12}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}, \tag{5.26}$$

$$\beta_o = \mu_3 - \beta_1\mu_1 - \beta_2\mu_2 \tag{5.27}$$

$$\sigma_{11}\beta_1 + \sigma_{12}\beta_2 = \sigma_{13} \tag{5.28}$$

$$\sigma_{12}\beta_1 + \sigma_{22}\beta_2 = \sigma_{23}. \tag{5.29}$$

Note that the sum of squares due to regression $SS(\underline{\beta}_3)$ is given by

$$SS(\underline{\beta}_3) = \underline{\beta}_3'(\mathbf{X}_{(3)}'\mathbf{X}_{(3)})\underline{\beta}_3,$$

therefore, the variance of the estimated values of y_3 is given by

$$n_c^{-1}SS(\underline{\beta}_3) = [\beta_1 \quad \beta_2]\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix}$$

$$= \beta_1^2\sigma_{11} + 2\beta_1\beta_2\sigma_{12} + \beta_2^2\sigma_{22}.$$

Hence the model variance σ^2 is given by

$$\sigma^2 = \sigma_{33} - (\beta_1^2\sigma_{11} + 2\beta_1\beta_2\sigma_{12} + \beta_2^2\sigma_{22}). \tag{5.30}$$

The likelihood function of the observed data Y_{obs} can be factorized in the following:

$$L(\underline{\mu}, \Sigma \mid Y_{obs}) = \prod_{i=1}^{n_c} f(y_{i1}, y_{i2}, y_{i3} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma)$$

$$= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma) f(y_{i3} \mid y_{i1}, y_{i2}, \underline{\mu}, \Sigma). \prod_{i=n_c+1}^{n} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma)$$

$$= \prod_{i=1}^{n} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma). \prod_{i=1}^{n_c} f(y_{i3} \mid \beta_o + \beta_1 y_{i1} + \beta_2 y_{i2}, \sigma^2) \tag{5.31}$$

Then the ML estimates of $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}, \beta_o, \beta_1, \beta_2$ and σ^2 are those values that maximize the individual terms of the RHS of (5.31).

Maximizing the first term of the RHS of (5.31) with respect to μ_1 and σ_{11} we have the usual ML estimates of the mean and variance of the bivariate normal distribution, that is,

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{y_{i1}}{n} = \bar{y}_1 \tag{5.32}$$

$$\hat{\mu}_2 = \sum_{i=1}^{n} \frac{y_{i2}}{n} = \bar{y}_2, \tag{5.33}$$

and

$$\hat{\sigma}_{11} = \frac{\sum_{i=1}^{n}(y_{i1} - \mu_1)^2}{n} = \frac{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2}{n} \tag{5.34}$$

$$\hat{\sigma}_{22} = \frac{\sum_{i=1}^{n}(y_{i2} - \mu_2)^2}{n} = \frac{\sum_{i=1}^{n}(y_{i2} - \bar{y}_2)^2}{n}. \tag{5.35}$$

The likelihood function of the second term of the RHS of (5.31) is given by

$$L(\beta_o + \beta_1 y_{i1} + \beta_2 y_{i2}, \sigma^2 \mid y_{i1}, y_{i2}) = (2\pi)^{\frac{-n_c}{2}} (\sigma^2)^{\frac{-n_c}{2}}$$
$$.Exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n_c} (y_{i3} - \beta_o - \beta_1 y_{i1} - \beta_2 y_{i2})^2],$$

from which the loglikelihood is

$$\ell(. \mid y_{i1}, y_{i2}) = -\frac{n_c}{2} \ln(2\pi) - \frac{n_c}{2} \ln\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n_c} (y_{i3} - \beta_o - \beta_1 y_{i1} - \beta_2 y_{i2})^2. \tag{5.36}$$

Maximizing (5.36) with respect to β_o, β_1, β_2 and σ^2 we get

$$\hat{\beta}_o = \bar{y}_3^c - \beta_1 \bar{y}_1^c - \beta_2 \bar{y}_2^c, \tag{5.37}$$

where $\overline{y}_1^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{i1}$, $\overline{y}_2^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{i2}$ and $\overline{y}_3^c = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{i3}$. That is, $\overline{y}_1^c$, $\overline{y}_2^c$ and $\overline{y}_3^c$ are the complete-case estimates of the means. Also we have

$$\hat{\beta}_1 = \frac{S_{31}S_{22} - S_{32}S_{12}}{S_{11}S_{22} - S_{12}^2} \tag{5.38}$$

$$\hat{\beta}_2 = \frac{S_{32}S_{11} - S_{31}S_{12}}{S_{11}S_{22} - S_{12}^2}, \tag{5.39}$$

where

$$S_{jk} = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{ij} - \overline{y}_j^c)(y_{ik} - \overline{y}_k^c), \quad j \neq k = 1, 2, 3$$

and for j=k=1,2 we have

$$S_{11} = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i1} - \overline{y}_1^c)^2.$$

$$S_{22} = \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i2} - \overline{y}_2^c)^2.$$

That is, the ML estimates of the parameters of the conditional distribution $f(y_{i3} \mid y_{i1}, y_{i2})$ are based on the n_c-complete cases.

Finally,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{i=1}^{n_c}(y_{i3} - \hat{\beta}_o - \hat{\beta}_1 y_{i1} - \hat{\beta}_2 y_{i2})^2}{n_c} \\ &= \frac{1}{n_c}\sum_{i=1}^{n_c}(y_{i3} - \hat{y}_{i3})^2 \\ &= S_{33} - (\hat{\beta}_1^2 S_{11} + \hat{\beta}_2^2 S_{22} + 2\hat{\beta}_1\hat{\beta}_2 S_{12}), \; 1 < m = 1, \ldots k-1. \end{aligned} \tag{5.40}$$

Using the invariance property of the ML method of estimation, estimates of other parameters can then be obtained. In particular, from (5.27) above we

have

$$\begin{aligned}\hat{\mu}_3 &= \hat{\beta}_o + \hat{\beta}_1\hat{\mu}_1 + \hat{\beta}_2\hat{\mu}_2 \\ &= \overline{y}_3^c - \hat{\beta}_1\overline{y}_1^c - \hat{\beta}_2\overline{y}_2^c + \hat{\beta}_1\hat{\mu}_1 + \hat{\beta}_2\hat{\mu}_2 \\ &= \overline{y}_3^c + \hat{\beta}_1(\hat{\mu}_1 - \overline{y}_1^c) + \hat{\beta}_2(\hat{\mu}_2 - \overline{y}_2^c) \\ &= \overline{y}_3^c + \sum_{j=1}^{2}\hat{\beta}_j(\hat{\mu}_j - \overline{y}_j^c), \end{aligned} \tag{5.41}$$

and from (5.30) we have

$$\hat{\sigma}_{33} = \hat{\sigma}^2 + \hat{\beta}_1^2\hat{\sigma}_{11} + 2\hat{\beta}_1\hat{\beta}_2\hat{\sigma}_{12} + \hat{\beta}_2^2\hat{\sigma}_{22}.$$

Substituting the value of $\hat{\sigma}^2$ from (5.40), we have

$$\begin{aligned}\hat{\sigma}_{33} &= S_{33} - \hat{\beta}_1^2 S_{11} - 2\hat{\beta}_1\hat{\beta}_2 S_{12} - \hat{\beta}_2^2 S_{22} + \hat{\beta}_1^2\hat{\sigma}_{11} + 2\hat{\beta}_1\hat{\beta}_2\hat{\sigma}_{12} + \hat{\beta}_2^2\hat{\sigma}_{22} \\ &= S_{33} + \hat{\beta}_1^2(\hat{\sigma}_{11} - S_{11}) + 2\hat{\beta}_1\hat{\beta}_2(\hat{\sigma}_{12} - S_{12}) + \hat{\beta}_2^2(\hat{\sigma}_{22} - S_{22}). \end{aligned} \tag{5.42}$$

Finally, from (5.28) and (5.29) we get

$$\begin{aligned}\hat{\sigma}_{13} &= \hat{\beta}_1\hat{\sigma}_{11} + \hat{\beta}_2\hat{\sigma}_{12} \\ &= \frac{S_{31}S_{22} - S_{32}S_{12}}{S_{11}S_{22} - S_{12}^2}\hat{\sigma}_{11} + \frac{S_{32}S_{11} - S_{31}S_{12}}{S_{11}S_{22} - S_{12}^2}\hat{\sigma}_{12}. \end{aligned} \tag{5.43}$$

and

$$\begin{aligned}\hat{\sigma}_{23} &= \hat{\beta}_1\hat{\sigma}_{12} + \hat{\beta}_2\hat{\sigma}_{22} \\ &= \frac{S_{31}S_{22} - S_{32}S_{12}}{S_{11}S_{22} - S_{12}^2}\hat{\sigma}_{12} + \frac{S_{32}S_{11} - S_{31}S_{12}}{S_{11}S_{22} - S_{12}^2}\hat{\sigma}_{22}. \end{aligned} \tag{5.44}$$

5.2.3 GENERALIZATION OF ANDERSON'S METHOD TO THE MULTIVARIATE NORMAL DISTRIBUTION

The generalization of Anderson's factorization method to the multivariate normal distribution with one variable subject to missingness can be dealt with as follows:

Consider a k-variate multivariate normal sample $(y_{i1}, y_{i2}, \ldots, y_{ik})$, $i = 1, \ldots, n$; where $(y_{i1}, \ldots, y_{i,k-1})$, $i = 1, \ldots, n_c$ are n_c-complete k-variate observations and $(y_{i1}, \ldots, y_{i,k-1})$, $i = n_c + 1, \ldots, n$ are $(n - n_c)$ observations on $k-1$ variables with missing y_{ik}. Thus we have n_c k-variate observations on $(y_{i1}, \ldots, y_{ik})$ and $(n - n_c)$ observations on $k-1$ variables $(y_{i1} \ldots, y_{i,k-1})$, that is,

$$f(y_{i1}, \ldots, y_{i,k}) = (2\pi)^{-\frac{k}{2}n_c} | \Sigma |^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})}, \tag{5.45}$$

and

$$f(y_{i1}, \ldots, y_{i,k-1}) = (2\pi)^{-\frac{(k-1)}{2}n_c} | \Sigma |^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})}. \tag{5.46}$$

Thus the joint density of a multivariate normal sample with n_c k-variate $(y_{i1}, \ldots, y_{ik})$ observations and $(n - n_c)$ $k-1$-variate $(y_{i1}, \ldots, y_{i,k-1})$ observations is given by

$$\begin{aligned} f(Y_{obs} \mid \underline{\mu}, \Sigma) = {} & (2\pi)^{-\frac{k}{2}n_c} | \Sigma |^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})} \\ & .(2\pi)^{-\frac{(k-1)}{2}n_c} | \Sigma |^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})}. \end{aligned} \tag{5.47}$$

Therefore the loglikelihood function (ignoring the missing data mechanism) is

$$\begin{aligned} \ell(\underline{\mu}, \Sigma \mid Y_{obs}) = {} & -\frac{n_c}{2} \ln | \Sigma | -\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu}) \\ & -\frac{n_c}{2} \ln | \Sigma | -\frac{1}{2}\sum_{i=n_c+1}^{n}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu}). \end{aligned} \tag{5.48}$$

Thus the maximum likelihood estimates of $\underline{\mu}$ and Σ can be found by maximizing (5.48) with respect to these values. A simpler approach to this maximization is Anderson's factorization method.

The joint density of $(y_{i1} \ldots, y_{ik})$ can be factorized as follows:

$$f(y_{i1} \ldots, y_{ik} \mid \underline{\mu}, \Sigma) = f(y_{i1}, \ldots, y_{i,k-1} \mid \underline{\mu}, \Sigma).f(y_{ik} \mid y_{i1}, \ldots, y_{i,k-1}, \beta_o, \ldots \beta_{k-1}; \sigma^2), \quad (5.49)$$

where, by the properties of the multivariate normal distribution,

$$f(y_{i1}, \ldots, y_{i,k-1} \mid \underline{\mu}, \Sigma) \sim N(\mu_1, \ldots, \mu_{k-1}; \sigma_{11} \ldots, \sigma_{k-1,k-1}),$$

$$f(y_{ik} \mid y_{i1}, \ldots, y_{i,k-1}, \beta_0, \beta_1, \beta_2, \ldots, \beta_{k-1}, \sigma^2) \sim N(\beta_o + \sum_{j=1}^{k-1} \beta_j y_{ij}; \sigma^2)$$

and

$$\underline{\beta}'_k = \underline{\alpha}'_k \mathbf{C}^{-1}$$

or

$$[\beta_1 \quad \ldots \quad \beta_{k-1}] = [\sigma_{k1} \quad \ldots \quad \sigma_{k,k-1}] \begin{bmatrix} \sigma_{11} & \ldots & \sigma_{1,k-1} \\ \vdots & \vdots & \vdots \\ \sigma_{k-1,1} & \ldots & \sigma_{k-1,k-1} \end{bmatrix}^{-1}. \quad (5.50)$$

From which we have

$$\beta_o = \mu_3 - \beta_1 \mu_1 - \ldots - \beta_{k-1} \mu_{k-1}, \quad (5.51)$$

and

$$[\sigma_{k1} \quad \ldots \quad \sigma_{k,k-1}] = [\beta_1 \quad \ldots \quad \beta_{k-1}] \begin{bmatrix} \sigma_{11} & \ldots & \sigma_{1,k-1} \\ \vdots & \vdots & \vdots \\ \sigma_{k-1,1} & \ldots & \sigma_{k-1,k-1} \end{bmatrix}. \quad (5.52)$$

Note that the sum of squares due to regression $SS(\underline{\beta}_k)$ is given by

$$SS(\underline{\beta}_k) = \underline{\beta}'(\mathbf{X}'_{(k)}\mathbf{X}_{(k)})\underline{\beta}_k,$$

therefore, the variance of the estimated values is given by

$$n_c^{-1}SS(\underline{\beta}_k) = [\beta_1 \quad \dots \quad \beta_{k-1}] \begin{bmatrix} \sigma_{11} & \dots & \sigma_{1,k-1} \\ \vdots & \vdots & \vdots \\ \sigma_{k-1,1} & \dots & \sigma_{k-1,k-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}.$$

$$= \beta_1^2\sigma_{11} + \dots + \beta_{k-1}^2\sigma_{k-1,k-1} + 2\beta_m\beta_l\sigma_{lm},\ l < m = 1, \dots, k-1.$$

Hence the model variance σ^2 is given by

$$\sigma^2 = \sigma_{kk} - (\beta_1^2\sigma_{11} + \dots + \beta_{k-1}^2\sigma_{k-1,k-1} + 2\beta_l\beta_m\sigma_{lm}). \tag{5.53}$$

The likelihood function of the observed data Y_{obs} can be factorized in the following:

$$\begin{aligned} L(\underline{\mu}, \Sigma \mid Y_{obs}) &= \prod_{i=1}^{n_c} f(y_{i1}, \dots, y_{ik} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1}, \dots, y_{i,k-1} \mid \underline{\mu}, \Sigma) \\ &= \prod_{i=1}^{n_c} f(y_{i1}, \dots, y_{i,k-1} \mid \underline{\mu}, \Sigma).f(y_{ik} \mid y_{i1}, \dots, y_{i,k-1}, \underline{\mu}, \Sigma) \\ &\quad \cdot \prod_{i=n_c+1}^{n} f(y_{i1}, \dots, y_{i,k-1} \mid \underline{\mu}, \Sigma) \\ &= \prod_{i=1}^{n} f(y_{i1}, \dots, y_{i,k-1} \mid \underline{\mu}, \Sigma). \prod_{i=1}^{n_c} f(y_{ik} \mid \beta_0 + \sum_{j=1}^{k-1} \beta_j y_{ij}; \sigma^2). \end{aligned} \tag{5.54}$$

Then the ML estimates of $\mu_j, \sigma_{jj}, \beta_j, j = 1, \dots, k-1$ together with β_0 and σ^2 are those values that maximize the individual terms of the RHS of (5.54).

Maximizing the first term of the RHS of (5.54) with respect to μ_j and σ_{jj} we have the usual ML estimates of the mean and variance of the multivariate normal distribution, that is,

$$\hat{\mu}_j = \sum_{i=1}^{n} \frac{y_{ij}}{n} = \bar{y}_j, \quad j = 1, \ldots, k. \tag{5.55}$$

and

$$\hat{\sigma}_{jj} = \frac{\sum_{i=1}^{n}(y_{ij} - \mu_j)^2}{n} = \frac{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}{n}, \quad j = 1, \ldots, k. \tag{5.56}$$

The likelihood function of the second term of the RHS of (5.54) is given by

$$L(\beta_o + \sum_{j=1}^{k-1} \beta_j y_{ij}, \sigma^2 \mid y_{i1}, \ldots, y_{ik-1}) = (2\pi)^{\frac{-n_c}{2}} (\sigma^2)^{\frac{-n_c}{2}}$$

$$.Exp[-\frac{1}{2\sigma^2} \sum_{i=1}^{n_c} (y_{ik} - \beta_0 - \sum_{j=1}^{k-1} \beta_j y_{ij})^2],$$

from which the loglikelihood is

$$\ell(. \mid y_{i1} \ldots, y_{i,k-1}) = -\frac{n_c}{2} \ln(2\pi) - \frac{n_c}{2} \ln\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n_c} (y_{i2} - \beta_o - \sum_{j=1}^{k-1} \beta_j y_{ij})^2. \tag{5.57}$$

Maximizing (5.57) with respect to β_0, β_j, $j = 1, \ldots, k-1$ and σ^2 we get

$$\hat{\beta}_o = \bar{y}_k^c - \sum_{j=1}^{k-1} \beta_j \bar{y}_j^c, \tag{5.58}$$

where $\bar{y}_j^c = \frac{1}{n_c} \sum_{i=1}^{n_c} y_{ij}$, $j = 1, \ldots, k$. That is, $\bar{y}_j^c$, $j = 1, \ldots, k$ are the complete-case estimates of the means. Also we have

$$[\hat{\beta}_1 \quad \dots \quad \hat{\beta}_{k-1}] = [S_{k1} \quad \dots \quad S_{k,k-1}] \begin{bmatrix} S_{11} & \dots & S_{1,k-1} \\ \vdots & \vdots & \vdots \\ S_{k-1,1} & \dots & S_{k-1,k-1} \end{bmatrix}^{-1} . \quad (5.59)$$

where

$$S_{jm} = \frac{1}{n_c} \sum_{i=1}^{n_c} (y_{ij} - \overline{y}_j^c)(y_{im} - \overline{y}_m^c), \quad j \neq m = 1, \dots, k$$

and for $j = m = 1, \dots, k-1$ we have

$$S_{jj} = \frac{1}{n_c} \sum_{i=1}^{n_c} (y_{ij} - \overline{y}_j^c)^2, \; j = 1, \dots, k-1$$

That is, the ML estimates of the parameters of the conditional distribution $f(y_{ik} \mid y_{i1}, \dots, y_{i,k-1})$ are based on the n_c-complete cases.

Finally,

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n_c} (y_{ik} - \hat{\beta}_o - \sum_{j=1}^{k-1} \hat{\beta}_j y_{ij})^2}{n_c} \\ &= \frac{1}{n_c} \sum_{i=1}^{n_c} (y_{ik} - \hat{y}_{ik})^2 \\ &= S_{kk} - (\hat{\beta}_1^2 S_{11} + \dots, + \hat{\beta}_{k-1}^2 S_{k-1,k-1} + 2\hat{\beta}_l \hat{\beta}_m S_{lm}), \\ &\qquad l < m = 1, \dots k-1. \end{aligned} \quad (5.60)$$

Using the invariance property of the ML method of estimation, estimates of other parameters can then be obtained. In particular, from (5.58) above we have

$$\hat{\mu}_k = \hat{\beta}_o + \sum_{j=1}^{k-1} \hat{\beta}_j \hat{\mu}_j$$

$$= \overline{y}_k^c - \sum_{j=1}^{k-1} \hat{\beta}_j \overline{y}_j^c + \sum_{j=1}^{k-1} \hat{\beta}_j \hat{\mu}_j$$

$$= \overline{y}_k^c + \sum_{j=1}^{k-1} \hat{\beta}_j (\hat{\mu}_j - \overline{y}_j^c), \tag{5.61}$$

and from (5.53) we have

$$\hat{\sigma}_{kk} = \hat{\sigma}^2 + (\hat{\beta}_1^2 \hat{\sigma}_{11} + \ldots + \hat{\beta}_{k-1}^2 \hat{\sigma}_{k-1,k-1} + 2\hat{\beta}_l \hat{\beta}_m \hat{\sigma}_{lm}). \tag{5.62}$$

Substituting the value of $\hat{\sigma}^2$ from (5.60) we get

$$\hat{\sigma}_{kk} = S_{kk} + \sum_{j=1}^{k-1} \hat{\beta}_j (\hat{\sigma}_{jj} - S_{jj}) + 2\hat{\beta}_l \hat{\beta}_m (\hat{\sigma}_{lm} - S_{lm}),\ l < m = 1, \ldots, k-1. \tag{5.63}$$

Finally, from (5.52) we get

$$[\hat{\sigma}_{k1} \quad \ldots \quad \hat{\sigma}_{k,k-1}] = [\hat{\beta}_1 \quad \ldots \quad \hat{\beta}_{k-1}] \begin{bmatrix} S_{11} & \ldots & S_{1,k-1} \\ \vdots & \vdots & \vdots \\ S_{k-1,1} & \ldots & S_{k-1,k-1} \end{bmatrix}. \tag{5.64}$$

5.2.4 EQUIVALENCE OF BUCK'S AND ANDERSON'S METHODS

In this section we shall try to estabish the equivalence of Buck's and Anderson's methods. To achieve this we consider the following special cases of Buck's method:

a- k=2,

b- k=3,

c- The multivariate case.

Where k is the number of variables. Moreover, in each of the above cases, (a), (b) and (c) we shall consider the case of units with one missing value

subject to one variable. Simply, this is the case of one variable subject to missingness.

Buck's method for k=2

Let x_{ij}, $(i = 1, 2, \ldots, n;\ j = 1, 2.)$ represent a sample of n units, on each of which it is desired to have measurements on k=2 variables. Assume that y_{i1} is fully observed but y_{i2} is missing on $(n - n_c)$ units, where n_c is the number of complete cases, that is,

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} \\ \vdots & \vdots \\ y_{n_{c1}} & y_{n_{c2}} \\ y_{n_{c1}+1} & ? \\ \vdots & \vdots \\ y_{n1} & ? \end{bmatrix} \tag{5.65}$$

For the data matrix **X**, let **A** be a 2x2 matrix that denotes the covariance matrix of the n_c-complete cases, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \tag{5.66}$$

Then the regression coefficients of y_2 on y_1 are given by

$$\hat{\beta}_1 = a_{12}a_{11}^{-1} \tag{5.67}$$

and

$$\hat{\beta}_0 = \bar{y}_2^c - \hat{\beta}_1\bar{y}_1^c. \tag{5.68}$$

The missing values of $\underline{y}_2$ are estimated by

$$\hat{\underline{y}}_2 = \mathbf{X}_{(2)}\hat{\beta}_1.$$

or

$$\hat{y}_{i2} = \hat{\beta}_0 + \hat{\beta}_1 y_{i1}, \quad i = 1, \ldots, n \tag{5.69}$$

and a completed data set is created by imputing the missing values of y_{i2}.

Then, on the basis of the completed data set, we have the following results that establish the equivalence of the two methods.

Theorem 5.1

The estimated mean and variance of y_1 from the completed data are the same as the corresponding ML estimates obtained by (5.11) and (5.12). Specifically,

$$\hat{\mu}_1 = \sum_{i=1}^{n} \frac{y_{i1}}{n} = \bar{y}_1$$

and

$$\hat{\sigma}_{11} = \frac{\sum_{i=1}^{n} (y_{i1} - \mu_1)^2}{n} = \frac{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2}{n}.$$

Proof

The proof follows by noting that y_1 is independent of the missing values.

Theorem 5.2

The ML estimates of the parameters of the conditional distribution of y_{i2} given y_{i1}, namely $\hat{\beta}_o$ and $\hat{\beta}_1$ given by (5.14) and (5.15) are the same as those obtained from the completed data. That is,

$$\underline{\hat{\beta}}_c = \underline{\beta}^*$$

Proof

See the proof of theorem 4.10.

Theorem 5.3

The mean of y_2 obtained from the completed data is the same as the ML estimate given by (5.17) above, that is,

$$\bar{y}_2 = n^{-1}\left\{\sum_{i=1}^{n_c} y_{i2} + \sum_{i=n_c+1}^{n} \hat{y}_{i2}\right\} = \bar{y}_2^c + \hat{\beta}_1(\hat{\mu}_1 - \bar{y}_1^c)$$

Proof

The estimated mean ($\bar{y}_2$) from the completed data can be written as

$$\bar{y}_2 = n^{-1}\left\{\sum_{i=1}^{n_c} y_{i2} + \sum_{i=n_c+1}^{n} \hat{y}_{i2}\right\}, \tag{5.70}$$

where

$$\hat{y}_{i2} = \bar{y}_2^c + \sum_{j=1}^{k-1} \hat{\beta}_j(y_{ij} - \bar{y}_j^c). \tag{5.71}$$

Substituting the value of $\hat{y}_{i2}$ from (5.71) in (5.70) we have

$$\begin{aligned}
\bar{y}_2 &= n^{-1}\left\{n_c\bar{y}_2^c + \sum_{i=n_c+1}^{n}(\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j(y_{ij} - \bar{y}_j^c))\right\} \\
&= n^{-1}\left\{n_c\bar{y}_2^c + (n-n_c)\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j\sum_{i=n_c+1}^{n}(y_{ij} - \bar{y}_j^c)\right\} \\
&= n^{-1}\left\{n\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j\sum_{i=n_c+1}^{n}(y_{ij} - \bar{y}_j^c)\right\} \\
&= n^{-1}\left\{n\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j\sum_{i=n_c+1}^{n}y_{ij} - \sum_{j=1}^{k-1}\hat{\beta}_j(n-n_c)\bar{y}_j^c\right\}
\end{aligned}$$

$$= n^{-1}\left\{ n\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j\Big(\sum_{i=1}^{n} y_{ij} - \sum_{i=1}^{n_c} y_{ij}\Big) - (n-n_c)\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c \right\}$$

$$= n^{-1}\left\{ n\bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j(n\hat{\mu}_j - n_c\bar{y}_j^c) - n\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c + n_c\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c \right\}$$

$$= n^{-1}\left\{ n\bar{y}_2^c + n\sum_{j=1}^{k-1}\hat{\beta}_j\hat{\mu}_j - n_c\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c - n\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c + n_c\sum_{j=1}^{k-1}\hat{\beta}_j\bar{y}_j^c \right\}$$

$$= n^{-1}\left\{ n\bar{y}_2^c + n\sum_{j=1}^{k-1}\hat{\beta}_j(\hat{\mu}_j - \bar{y}_j^c) \right\}$$

$$= \bar{y}_2^c + \sum_{j=1}^{k-1}\hat{\beta}_j(\hat{\mu}_j - \bar{y}_j^c), \qquad (5.72)$$

and the proof follows by putting k=2 in (5.72).

Buck's method for k=3

Let x_{ij}, $(i = 1,2,\ldots,n;\ j = 1,2,3.)$ represent a sample of n units, on each of which it is desired to have measurements on k=3 variables. Assume that y_{i1} and y_{i2} are fully observed but y_{i3} is missing on $(n - n_c)$ units, where n_c is the number of complete cases, that is,

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ \vdots & \vdots & \vdots \\ y_{n_c1} & y_{n_c2} & y_{n_c3} \\ y_{n_c1+1} & y_{n_c2+1} & ? \\ \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & ? \end{bmatrix} \qquad (5.73)$$

For the data matrix **X**, let **A** be a 3x3 matrix that denotes the covariance matrix of the n_c-complete cases, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{5.74}$$

We can write $\mathbf{A}$ as

$$\mathbf{A} = (a_{sj}) = \begin{bmatrix} \mathbf{C} & \underline{\alpha}_3 \\ \underline{\alpha}_3' & a_{33} \end{bmatrix} \quad (s = 1, 2, 3) \tag{5.75}$$

where

$$a_{33} = V(\underline{y}_3)$$
$$\underline{\alpha}_3' = (a_{31}, a_{32})$$

and

$$\mathbf{C} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}. \tag{5.76}$$

Then the regression coefficients of y_3 on y_1, y_2 are given by

$$\underline{\hat{\beta}}_3' = \underline{\alpha}_3' \mathbf{C}^{-1}. \tag{5.77}$$

By writing $\mathbf{X}_{(3)} = (\underline{y}_1, \underline{y}_2)$, of order nx2, we can estimate the value of $\underline{y}_3$ for those cases which have $\underline{y}_3$ only missing by taking

$$\underline{\hat{y}}_3 = \mathbf{X}_{(3)} \underline{\hat{\beta}}_3,$$

or

$$\hat{y}_{i3} = \hat{\beta}_0 + \sum_{h=1}^{k-1} y_{ih} \hat{\beta}_{h1}, \quad i = 1, \ldots . n \tag{5.78}$$

Then the equivalence of Buck's and Anderson's methods for this case (k=3) can be established by the direct application of theorems 5.1, 5.2 and by putting k=3 in (5.70) of theorem 5.3.

Buck's method for the multivarjate case

Let y_{ij}, $(i = 1,\ldots,n;\ j = 1,\ldots,k.)$ represent a sample of n units, on each of which it is desired to have measurements on k variables. Assume that $(y_{i1},\ldots,y_{i(k-1)})$ are fully observed but y_{ik} is missing on $(n - n_c)$ units, where n_c is the number of complete cases, that is,

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1(k-1)} & y_{1k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n_{c1}} & y_{n_{c2}} & \cdots & y_{n_{c(k-1)}} & y_{n_{ck}} \\ y_{n_{c1}+1} & y_{n_{c2}+1} & \cdots & y_{n_{c(k-1)}+1} & ? \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{n(k-1)} & ? \end{bmatrix} \tag{5.79}$$

For the data matrix $\mathbf{X}$, let $\mathbf{A}$ be a kxk matrix that denotes the covariance matrix of the n_c-complete cases, i.e.,

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kk} \end{bmatrix} \tag{5.80}$$

We can write $\mathbf{A}$ as

$$\mathbf{A} = (a_{sj}) = \begin{bmatrix} \mathbf{C} & \underline{\alpha}_k \\ \underline{\alpha}_k' & a_{kk} \end{bmatrix} \quad (s = 1,\ldots,k) \tag{5.81}$$

where

$$a_{kk} = V(\underline{x}_k)$$

$$\underline{\alpha}_k' = (a_{k1}, a_{k2}, \ldots, a_{k(k-1)})$$

and

$$\mathbf{C} = \begin{bmatrix} a_{11} & \cdots & a_{1(k-1)} \\ \vdots & \vdots & \vdots \\ a_{(k-1)1} & \cdots & a_{(k-1)(k-1)} \end{bmatrix}. \tag{5.82}$$

Then the regression coefficients of x_k on $x_1, \ldots, x_{k-1}$ are given by

$$\underline{\hat{\beta}}'_k = \underline{\alpha}'_k \mathbf{C}^{-1}. \tag{5.83}$$

By writing $\mathbf{X}_{(k)} = (\underline{x}_1, \ldots, \underline{x}_{k-1})$, of order $nx(k-1)$, we can estimate the value of $\underline{x}_k$ for those cases which have $\underline{x}_k$ only missing by taking

$$\hat{\underline{x}}_k = \mathbf{X}_{(k)} \underline{\hat{\beta}}_k,$$

or

$$\hat{x}_{ik} = \hat{\beta}_0 + \sum_{h=1}^{k-1} x_{ih} \hat{\beta}_{h1}, \quad i = 1, \ldots, n. \tag{5.84}$$

The equivalence of Buck's and Anderson's methods for this multivariate case follows immediately from theorems 5.1, 5.2 and 5.3.

Remark 1

Note that the ML estimates of the parameters of the conditional distributions given by Anderson's factorization method (5.14–5.15; 5.37–5.39 and 5.58–5.59) for the case of one variable subject to missingness are exactly the same quantities required for Buck's imputations. Since the parameters of Buck's regression equations, required for imputation, are Least-squares-estimtes, it follows that the normality assumption is a necessary condition for the equivalence of Anderson's and Buck's methods. It also follows that the estimated conditional distributions via Anderson's method can be used

for imputation purposes. The resulting completed data set will be exactly the same as the one created via Buck's method.

Remark 2

It is interesting to note that the first terms of the right hand side of (5.16)–(5.19), (5.40)–(5.42), (5.60)–(5.63) and the second term of the right hand side of (5.64) are the estimates obtained from the complete-case analysis. Thus the remaining terms represent adjustments based on the additional information from the $(n - n_c)$ additional observations on y_{i1}, (y_{i1}, y_{i2}) and $(y_{i1}, \ldots, y_{i(k-1)})$, respectively. In other words, the complete-case analysis estimates can be obtained as a special case of the corresponding ML estimates by ignoring adjustment terms in the latter.

Remark 3

For patterns of missingness that do not conform for Anderson's factorization method, conformability can be created by deleting some few observations and consider them as missing values. The following simple example illustrates the point.

$$
\begin{array}{ccc}
y_{11} & y_{12} & y_{13} \\
y_{21} & y_{21} & y_{23} \\
y_{31} & y_{32} & y_{33} \\
y_{41} & y_{42} & y_{43} \\
y_{51} & y_{52} & ? \\
y_{61} & ? & ? \\
y_{71} & ? & ? \\
y_{81} & ? & y_{83} \\
y_{91} & ? & y_{93}.
\end{array}
$$

Clearly, there is no possible factorization for this pattern. However, a possible factorization can be obtained by deleting y_{83} and y_{93} and consider them as missing values.

5.3 RELATION BETWEEN ANDERSON'S AND BUCK'S METHODS: UNITS WITH MORE THAN ONE MISSING VALUE

In this section we shall study Anderson's factorization method for the case of units with more than one missing value. We start our study with the trivariate normal distribution where units have two missing values. For this case we consider two different cases of missingness, case (1) and case (2). These two cases will also be generalized to their corresponding multivariate versions. The objective is to investigate the equivalence of these two cases with the corresponding versions of Buck's method.

5.3.1 ANDERSON'S METHOD FOR THE TRIVARIATE NORMAL DISTRIBUTION: CASE (1)

Consider a trivariate normal sample $(y_{i1}, y_{i2}, y_{i3}), i = 1, \ldots, n$; where $(y_{i1}, y_{i2}, y_{i3}), i = 1, \ldots, n_c$ are n_c-complete trivariate observations and $y_{i1}, i = n_c + 1, \ldots, n$ are $(n - n_c)$ univariate observations with missing y_{i2} and y_{i3}, that is,

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ \vdots & \vdots & \vdots \\ y_{n_c 1} & y_{n_c 2} & y_{n_c 3} \\ y_{n_c+1} & ? & ? \\ \vdots & \vdots & \vdots \\ y_{n1} & ? & ? \end{bmatrix} \tag{5.85}$$

Thus we have n_c trivariate observations on both y_{i1}, y_{i2} and y_{i3} and $(n - n_c)$ univariate observations on y_{i1}, that is,

$$f(y_{i1}, y_{i2}, y_{i3}) = (2\pi)^{-\frac{3}{2}n_c} |\Sigma|^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})}, \tag{5.86}$$

and

$$f(y_{i1}) = (2\pi)^{-\frac{(n-n_c)}{2}} \sigma_{11}^{-\frac{(n-n_c)}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i - \underline{\mu})^2}{\sigma_{11}}}. \tag{5.87}$$

Thus the joint density of a trivariate normal sample with n_c trivariate (y_{i1}, y_{i2}, y_{i3}) observations and $(n - n_c)$ univariate y_{i1} observations is given by

$$f(Y_{obs} \mid \underline{\mu}, \Sigma) = (2\pi)^{-\frac{3}{2}n_c} \mid \Sigma \mid^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})}$$
$$.(2\pi)^{-\frac{(n-n_c)}{2}} \sigma_{11}^{-\frac{(n-n_c)}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i - \underline{\mu})^2}{\sigma_{11}}}. \qquad (5.88)$$

Therefore the loglikelihood function (ignoring the missing data mechanism) is

$$\ell(\underline{\mu}, \Sigma \mid Y_{obs}) = -\frac{n_c}{2} \ln \mid \Sigma \mid -\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu})$$
$$-\frac{1}{2}(n - n_c)\ln \sigma_{11} - \frac{1}{2}\sum_{i=n_c+1}^{n}\frac{(y_i - \underline{\mu})^2}{\sigma_{11}}. \qquad (5.89)$$

Thus the maximum likelihood estimates of $\underline{\mu}$ and Σ can be found by maximizing (5.89) with respect to these values. Using Anderson's factorization method, the likelihood function of the observed data Y_{obs} can be factorized in the following:

$$L(\underline{\mu}, \Sigma \mid Y_{obs}) = \prod_{i=1}^{n_c} f(y_{i1}, y_{i2}, y_{i3} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11})$$
$$= \prod_{i=1}^{n_c} f(y_{i1} \mid \mu_1, \sigma_{11}) f(y_{i2}, y_{i3} \mid y_{i1}, \mu_1, \sigma_{11})$$
$$. \prod_{i=n_c+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11})$$
$$= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}). \prod_{i=1}^{n_c} f(y_{i2}, y_{i3} \mid y_{i1}, \mu_1, \sigma_{11})$$
$$= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}). \prod_{i=1}^{n_c} f(y_{i2} \mid \beta_o + \beta_1 y_{i1}, \sigma^2)$$
$$. \prod_{i=1}^{n_c} f(y_{i3} \mid \beta_o + \beta_1 y_{i1} + \beta_2 y_{i2}, \sigma^2). \qquad (5.90)$$

Then the ML estimates of μ_j, σ_{jj}, j = 1, 2, 3; $\beta_0, \beta_1, \beta_2$ and the corresponding model variances are those values that maximize the individual terms of the RHS of (5.90).

Maximizing the first term of the RHS of (5.90) with respect to μ_1 and σ_{11} we have the usual ML estimates of the mean and variance of the univariate normal distribution obtained in section 5.2.1

Maximizing the second term of the RHS of (5.90) we obtain the ML estimates of the parameters of the conditional distribution $f(y_{i2} \mid y_{i1})$ as was also given in section 5.2.1.

Finally, the ML estimates of the parameters of the third term of the RHS of (5.90) are those given in section 5.2.2.

Remark 4

Case (1) above can be generalized to the multivariate normal distribution (units with more than one missing value). Since the approach is the same, we shall only give the possible factorization, which is as follows:

$$
\begin{aligned}
L(\underline{\mu}, \Sigma \mid Y_{obs}) &= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{ik}) \\
&= \prod_{i=1}^{n_c} f(y_{i1}, \ldots, y_{ik} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \\
&= \prod_{i=1}^{n_c} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma).f(y_{i,k-1}, y_{ik} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma) \\
&\quad \cdot \prod_{i=n_c+1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma). \prod_{i=1}^{n_c} f(y_{i,k-1}, y_{ik} \mid y_{i1}, \ldots, \\
&\qquad\qquad y_{i,k-2}, \underline{\mu}, \Sigma)
\end{aligned}
$$

$$= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \cdot \prod_{i=1}^{n_c} f(y_{i,k-1} \mid \beta_0 + \sum_{j=1}^{k-2} \beta_j y_{ij}; \sigma^2) \cdot$$

$$\cdot \prod_{i=1}^{n_c} f(y_{i,k} \mid \beta_0 + \sum_{j=1}^{k-1} \beta_j y_{ij}; \sigma^2). \qquad (5.91)$$

5.3.2 ANDERSON'S METHOD FOR THE TRIVARIATE NORMAL DISTRIBUTION: CASE (2)

Consider a trivariate normal sample $(y_{i1}, y_{i2}, y_{i3}), i = 1, \ldots, n$; where $(y_{i1}, y_{i2}, y_{i3}), i = 1, \ldots, n_c$ are n_c-complete trivariate observations; (y_{i1}, y_{i2}), $i = n_c + 1, \ldots, m$ are m bivariate observations with y_{i3} missing and $y_{i1}, i = m+1, \ldots, n$ are $(n - n_c - m)$ univariate observations with y_{i2} and y_{i3} missing. That is,

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ \vdots & \vdots & \vdots \\ y_{n_c 1} & y_{n_c 2} & y_{n_c 3} \\ y_{n_c 1+1} & y_{n_c 2+2} & ? \\ \vdots & \vdots & \vdots \\ y_{n_c 1+m} & y_{n_c 2+m} & ? \\ y_{n_c 1+m+1} & ? & ? \\ \vdots & \vdots & \vdots \\ y_{n1} & ? & ? \end{bmatrix} \qquad (5.92)$$

Thus we have n_c trivariate observations on both (y_{i1}, y_{i2}, y_{i3}); m bivariate observations on (y_{i1}, y_{i2}) and $(n - n_c - m)$ univariate observations on y_{i1}, that is,

$$f(y_{i1}, y_{i2}, y_{i3}) = (2\pi)^{-\frac{3}{2}n_c} \mid \Sigma \mid^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c} (y_i - \underline{\mu})' \Sigma^{-1} (y_i - \underline{\mu})}, \qquad (5.93)$$

$$f(y_{i1}, y_{i2}) = (2\pi)^{-m} \mid \Sigma \mid^{-\frac{m}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n_c+m} (y_i - \underline{\mu})' \Sigma^{-1} (y_i - \underline{\mu})}, \qquad (5.94)$$

and

$$f(y_{i1}) = (2\pi)^{-\frac{(n-n_c-m)}{2}} \sigma_{11}^{-\frac{(n-n_c-m)}{2}} e^{-\frac{1}{2}\sum_{i=n_c+m+1}^{n} \frac{(y_i-\mu)^2}{\sigma_{11}}}. \quad (5.95)$$

Thus the joint density of a trivariate normal sample with n_c trivariate (y_{i1}, y_{i2}, y_{i3}) observations; m bivariate (y_{i1}, y_{i2}) observations and $(n - n_c - m)$ univariate y_{i1} observations is given by

$$\begin{aligned} f(Y_{obs} \mid \underline{\mu}, \Sigma) = {} & (2\pi)^{-\frac{3}{2}n_c} \mid \Sigma \mid^{-\frac{n_c}{2}} e^{-\frac{1}{2}\sum_{i=1}^{n_c}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})} \\ & .(2\pi)^{-m} \mid \Sigma \mid^{-\frac{m}{2}} e^{-\frac{1}{2}\sum_{i=n_c+1}^{n_c+m}(y_i-\underline{\mu})'\Sigma^{-1}(y_i-\underline{\mu})} \\ & .(2\pi)^{-\frac{(n-n_c-m)}{2}} \sigma_{11}^{-\frac{(n-n_c-m)}{2}} e^{-\frac{1}{2}\sum_{n_c+m+1}^{n} \frac{(y_i-\mu)^2}{\sigma_{11}}}. \end{aligned} \quad (5.96)$$

Therefore the loglikelihood function (ignoring the missing data mechanism) is

$$\begin{aligned} \ell(\underline{\mu}, \Sigma \mid Y_{obs}) = {} & -\frac{n_c}{2} \ln \mid \Sigma \mid -\frac{1}{2}\sum_{i=1}^{n_c}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu}) \\ & -\frac{m}{2} \ln \mid \Sigma \mid -\frac{1}{2}\sum_{i=n_c+1}^{n_c+m}(y_i - \underline{\mu})'\Sigma^{-1}(y_i - \underline{\mu}) \\ & -\frac{1}{2}(n - n_c - m)\ln \sigma_{11} - \frac{1}{2}\sum_{i=n_c+m+1}^{n} \frac{(y_i - \underline{\mu})^2}{\sigma_{11}}. \end{aligned} \quad (5.97)$$

Thus the maximum likelihood estimates of $\underline{\mu}$ and Σ can be found by maximizing (5.97) with respect to these values. Using Anderson's factorization method, the likelihood function of the observed data Y_{obs} can be factorized in the following:

$$
\begin{aligned}
L(\underline{\mu}, \Sigma \mid Y_{obs}) &= \prod_{i=1}^{n_c} f(y_{i1}, y_{i2}, y_{i3} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n_c+m} f(y_{i1}, y_{i2} \mid \underline{\mu}, \Sigma) \\
&\quad \cdot \prod_{n_c+m+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}) \\
&= \prod_{i=1}^{n_c} f(y_{i1} \mid \mu_1, \sigma_{11}).f(y_{i2}, y_{i3} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&\quad \cdot \prod_{i=n_c+1}^{n_c+m} f(y_{i1} \mid \mu_1, \sigma_{11}).f(y_{i2} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&\quad \cdot \prod_{i=n_c+m+1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}) \\
&= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}). \prod_{i=1}^{n_c} f(y_{i2}, y_{i3} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&\quad \cdot \prod_{i=n_c+1}^{n_c+m} f(y_{i2} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}). \prod_{i=1}^{n_c} f(y_{i2} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&\quad .f(y_{i3} \mid y_{i1}, y_{i2}, \underline{\mu}, \Sigma). \prod_{i=n_c+1}^{n_c+m} f(y_{i2} \mid y_{i1}, \mu_1, \sigma_{11}) \\
&= \prod_{i=1}^{n} f(y_{i1} \mid \mu_1, \sigma_{11}). \prod_{i=1}^{n_c+m} f(y_{i2} \mid \beta_0 + \beta_1 y_{i1}, \sigma^2) \\
&\quad \cdot \prod_{i=1}^{n_c} f(y_{i3} \mid \beta_0 + \beta_1 y_{i1} + \beta_2 y_{i2}, \sigma^2). \qquad (5.98)
\end{aligned}
$$

Then the ML estimates of μ_j, σ_{jj}, $j = 1, 2, 3$; $\beta_0, \beta_1, \beta_2$ and the corresponding model variances are those values that maximize the individual terms of the RHS of (5.98). These estimates are the same as those obtained in sections

5.2.1 and 5.2.2.

Remark 5

The possible factorization for the multivariate generalization of this case is as follows:

$$
\begin{aligned}
L(\underline{\mu}, \Sigma \mid Y_{obs}) &= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{ik}) \\
&= \prod_{i=1}^{n_c} f(y_{i1}, \ldots, y_{ik} \mid \underline{\mu}, \Sigma) \prod_{i=n_c+1}^{n_c+m} f(y_{i1}, \ldots, y_{i,k-1} \mid \underline{\mu}, \Sigma) \\
&\quad \cdot \prod_{i=n_c+m+1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \\
&= \prod_{i=1}^{n_c} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma).f(y_{i,k-1}, y_{ik} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma) \\
&\quad \cdot \prod_{i=n_c+1}^{n_c+m} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma).f(y_{i,k-1} \mid y_{i1}, \ldots, y_{i,k-2}) \\
&\quad \cdot \prod_{i=n_c+m+1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \prod_{i=1}^{n_c} f(y_{i,k-1}, y_{ik} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma) \\
&\quad \cdot \prod_{i=n_c+1}^{n_c+m} f(y_{i,k-1} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \prod_{i=1}^{n_c} f(y_{i,k-1} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma) \\
&\quad .f(y_{ik} \mid y_{i1}, \ldots, y_{i,k-1}) \prod_{i=n_c+1}^{n_c+m} f(y_{i,k-1} \mid y_{i1}, \ldots, y_{i,k-2}, \underline{\mu}, \Sigma)
\end{aligned}
$$

$$= \prod_{i=1}^{n} f(y_{i1}, \ldots, y_{i,k-2} \mid \underline{\mu}, \Sigma) \prod_{i=1}^{n_c+m} f(y_{i,k-1} \mid \beta_0 + \sum_{j=1}^{k-2} \beta_j y_{ij}, \sigma^2)$$

$$\cdot \prod_{i=1}^{n_c} f(y_{i,k} \mid \beta_0 + \sum_{j=1}^{k-1} \beta_j y_{ij}, \sigma^2). \qquad (5.99)$$

5.3.3 EQUIVALENCE OF BUCK'S AND ANDERSON'S METHODS

In section 5.2.4 we have established the equivalence of Buck's and Anderson's methods for the case of units with one missing value subject to one variable. This result has been generalized to the multivariate normal distribution with missing values on one variable. In this section we shall try to study the same equivalence for the case of units with more than one missing value. Specifically, we shall consider Buck's method for case (1) and case (2) above. The equivalence shall then be studied by comparing the estimated parameters for case (1) and case (2) with their corresponding versions of Buck's method.

Buck's method for the Trivariate Normal Distribution: Case (1)

The data matrix for this case is given by

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ \vdots & \vdots & \vdots \\ y_{n_{c1}} & y_{n_{c2}} & y_{n_{c3}} \\ y_{n_{c1}+1} & ? & ? \\ y_{n1} & ? & ? \end{bmatrix},$$

and the variance-covariance matrix of the n_c-complete cases is given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Using our notation of section 4.4, we estimate the values of $\underset{\sim}{y}_2$ and $\underset{\sim}{y}_3$ for those cases which have y_2 and y_3 missing, by estimating the multivariate regression equation

$$\hat{\mathbf{Y}} = \mathbf{X}_{(23)}\hat{\mathbf{B}}, \tag{5.100}$$

where $\hat{\mathbf{Y}} = (\hat{y}_2, \hat{y}_3)$ and $\mathbf{X}_{(23)} = y_{i1}$, i.e.,

$$\hat{y}_{i2} = \hat{\beta}_{02} + \hat{\beta}_{12}y_{i1}, \quad i = 1, \ldots, n_c, \tag{5.101}$$

and

$$\hat{y}_{i3} = \hat{\beta}_{03} + \hat{\beta}_{13}y_{i1}, \quad i = 1, \ldots, n_c. \tag{5.102}$$

where $\hat{\beta}_{12} = a_{12}a_{11}^{-1}$ and $\hat{\beta}_{13} = a_{13}a_{11}^{-1}$.

Thus we have

$$\hat{\underset{\sim}{y}}_{i2} = \hat{\beta}_{02} + \hat{\beta}_{12}y_{i1}, \quad i = n_{c+1}, \ldots, n, \tag{5.103}$$

and

$$\hat{\underset{\sim}{y}}_{i3} = \hat{\beta}_{03} + \hat{\beta}_{13}y_{i1}, \quad i = n_{c+1}, \ldots, n. \tag{5.104}$$

Note that Buck's method estimates the jointly missing values on the second and third variables from the multivariate regression of those variables on the first variable. This is done on the basis of the n_c complete cases. However, the conditional distributions of Anderson's factorization for this case, given by (5.90), do not correspond to Buck's regression equations which are required for imputation. Specifically, the second term of the RHS of (5.90) allows for the imputation of the missing values of y_{i2}. This is because the second term of (5.90) corresponds exactly to Buck's regression equation required for the imputation of the missing values of y_{i2}. However, the missing values of y_{i3} can never be estimated on the basis of the factorization since

there is no corresponding term that allows for its estimation. In fact, the last term of the RHS of (5.90) does not allow for this imputation since it is conditioned on both y_{i1} and y_{i2}. Thus, unlike the case of missingness on one variable, Anderson's method for the case of units with more than one missing value cannot be used as an imputation method. It therefore follows that there is no equivalence relation between Anderson's and Buck's methods for this case of units with two missing values. By extension, the same conclusion follows for the corresponding multivariate normal case.

Buck's method for the Trivariate Normal Distribution: Case (2)

For this case, the data matrix is given by

$$\mathbf{X} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ \vdots & \vdots & \vdots \\ y_{n_{c1}} & y_{n_{c2}} & y_{n_{c3}} \\ y_{n_{c1}+1} & y_{n_{c2}+2} & ? \\ \vdots & \vdots & \vdots \\ y_{n_{c1}+m} & y_{n_{c2}+m} & ? \\ y_{n_{c1}+m+1} & ? & ? \\ \vdots & \vdots & \vdots \\ y_{n1} & ? & ? \end{bmatrix}.$$

Similarly, we estimate the values of $\underset{\sim}{y}_2$ and $\underset{\sim}{y}_3$ for those cases which have y_2 and y_3 missing, by estimating the multivariate regression equation

$$\hat{\mathbf{Y}} = \mathbf{X}_{(23)}\hat{\mathbf{B}}, \tag{5.105}$$

where $\hat{\mathbf{Y}} = (\hat{y}_2, \hat{y}_3)$ and $\mathbf{X}_{(23)} = y_{i1}$, i.e.,

$$\hat{y}_{i2} = \hat{\beta}_{02} + \hat{\beta}_{12} y_{i1}, \quad i = 1, \ldots, n_c, \tag{5.106}$$

and

$$\hat{y}_{i3} = \hat{\beta}_{03} + \hat{\beta}_{13} y_{i1}, \quad i = 1, \ldots, n_c. \tag{5.107}$$

where $\hat{\beta}_{12} = a_{12}a_{11}^{-1}$ and $\hat{\beta}_{13} = a_{13}a_{11}^{-1}$.

Thus we have

$$\hat{y}_{i2} = \hat{\beta}_{02} + \hat{\beta}_{12} y_{i1}, \quad i = n_{c+1}, \ldots, n \tag{5.109}$$

and

$$\hat{y}_{i3} = \hat{\beta}_{03} + \hat{\beta}_{13} y_{i1}, \quad i = n_{c+1}, \ldots, n. \tag{5.109}$$

Here, note that (5.108) and (5.109) impute the missing values of y_{i2} and y_{i3} from the multivariate regression of those variables on y_{i1} from the n_c-complete cases. However, the second term of the RHS of Anderson's factorization given by (5.98) allows for the estimation of the missing values of y_{i2} conditioned on y_{i1} from the $(n_c + m)$-complete cases. In other words, all-available-data on y_{i2} are used for the estimation of the missing values of y_{i2}. Similarly, the last term of the RHS of (5.98) allows for the estimation of missing y_{i3} only for those cases with y_{i1} and y_{i2} observed. It remains that the jointly missing values on both y_{i2} and y_{i3} cannot be estimated as there is no corresponding term in the factorization that can allow for their estimation. Hence, for this case, Anderson's method cannot be used as an imputation method for creating complete data sets. It therefore follows that there is no equivalence betweeen Anderson's and Buck's methods for this case of units with two missing values. By extension, the same conclusion follows for the corresponding multivariate normal case.

5.4 RELATION BETWEEN ITERATED BUCK'S METHOD AND THE EM ALGORITHM

Let x_{ij}, $(i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, k)$ represent the sample of n units, on each of which it is desired to have measurements on k variables. The

observations x_{ij} can be represented in the form of an nxk matrix, $\mathbf{X}$, in which some of the elements are missing. Without loss of generality, assume that the last $n - n_c$ units have missing entries. Thus we write

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n_{c1}} & x_{n_{c2}} & \cdots & x_{n_{ck}} \\ x_{n_{c1}+1} & ? & \cdots & x_{n_{ck}} \\ \vdots & \vdots & \vdots & \vdots \\ ? & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{5.110}$$

where ? denotes a missing value.

For the data matrix $\mathbf{X}$, let

$P_i \equiv$ The set of the observed variables in the i-th unit

$P_T \equiv$ The matrix of the complete cases

$\underline{\mu}_A \equiv$ The vector of means of the complete cases (P_T)

$\Sigma_A \equiv$ The variance-covariance matrix of the complete cases (P_T)

$\underline{\mu} \equiv$ The vector of means of the completed data (after imputation)

$\Sigma \equiv$ The variance-covariance matrix of the completed data (after imputation).

Then using Buck's method, for the case of units with one missing value, we estimate the value of x_j for those units which have x_j only missing by taking

$$\hat{x}_{ij} = E(x_{ij} \mid x_{i1}, \ldots, x_{i(j-1)}, x_{i(j+1)}, \ldots, x_{ik}; \underline{\mu}_A, \Sigma_A)$$

or

$$\hat{x}_{ij} = E(x_{ij} \mid P_i, \underline{\mu}_A, \Sigma_A). \tag{5.111}$$

For those units with more than one missing value (say, v< k missing values) we estimate the missing values by estimating the multivariate regression equation given by

$$(\hat{x}_{ij}, \ldots, \hat{x}_{iv}) = E(x_{ij}, \ldots, x_{iv} \mid x_{i(v+1)}, \ldots, x_{ik}; \underline{\mu}_A, \Sigma_A)$$

or

$$(\hat{x}_{ij}, \ldots, \hat{x}_{iv}) = E(x_{ij}, \ldots, x_{iv} \mid P_i, \underline{\mu}_A, \Sigma_A) \tag{5.112}$$

(5.111) and (5.112) are the usual regression equations required by Buck's method for the estimation of the missing values. The regression coefficients of these equations can be computed with relative ease using Woolf's procedure as outlined in section 3.4. The incomplete data matrix **X** can now be completed by imputing the missing values.

Using theorems 4.6 and 4.7, the elements of the estimated variance-covariance matrix from the completed data $(\sigma^*_{jj}, \sigma^*_{jk})$, are to be adjusted for bias as follows:

$$\hat{\sigma}_{jj} = \sigma^*_{jj} + \lambda_j \Phi_{jj} \tag{5.113}$$

and

$$\hat{\sigma}_{jk} = \sigma^*_{jk} + \lambda_{jk} \Phi_{jk}, \tag{5.114}$$

where Φ_{jj} is the sample variance of the residuals of the regression of the j-th variable on the remaining $(k-1)$ variables, and Φ_{jk} is the sample covariance of the residuals of the multivariate regression of the j-th and k-th variables on the remaining $(k-2)$ variables. λ_j and λ_{jk} are the proportions of missing values in the j-th variable and in the j-th and k-th variables respectively.

(5.113) and (5.114) can be rewritten as

$$\hat{\sigma}_{jj} = \frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{x}_{ij} - \hat{\mu}_j)^2 + \Phi_{jj,i}\right\}, \tag{5.115}$$

$$\hat{\sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left\{(\hat{x}_{ij} - \hat{\mu}_j)(\hat{x}_{ik} - \hat{\mu}_k) + \Phi_{jk,i}\right\}, \tag{5.116}$$

where

$$\Phi_{jj,i} = \begin{cases} \Phi_{jj}, & \text{if the j-th variable is missing on the i-th unit} \\ 0, & \text{Otherwise} \end{cases} \tag{5.117}$$

$$\Phi_{jk,i} = \begin{cases} \Phi_{jk}, & \text{if both the j-th and k-th variables are missing} \\ & \text{on the i-th unit} \\ 0, & \text{Otherwise} \end{cases} \tag{5.118}$$

and Φ_{jj}, Φ_{jk} are as defined above.

Now for the data completed via (5.111) and (5.112), let

$$\hat{x}_{ij} = \begin{cases} x_{ij}, \text{ if } x_{ij} \text{ is observed,} \\ \text{a linear combination of the variables observed in} \\ \text{the i-th unit } (P_i) \text{ if } x_{ij} \text{ is missing.} \end{cases} \tag{5.119}$$

Formula (5.119) reimputes (from the completed data) the previously imputed missing values. Thus (5.119) defines an iterated version of Buck's method which is similar to the one obtained by Kasap (1973).

At each iteration the data are completed by imputations and the variance-covariance matrix is computed. This matrix is then adjusted for bias using (5.115) and (5.116). Specifically, at each iteration we have, for those units with more than one missing value

$$(\hat{x}_{ij}, \ldots, \hat{x}_{iv}) = E(x_{ij}, \ldots, x_{iv} \mid P_i, \underline{\mu}, \Sigma) \tag{5.120}$$

which, for the case of units with one missing value (v=1), reduces to

$$\hat{x}_{ij} = E(x_{ij} \mid P_i, \underline{\mu}, \Sigma). \tag{5.121}$$

Then on the basis of the recompleted data, we have

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} \hat{x}_{ij}, \tag{5.122}$$

$$\hat{\sigma}_{jj} = \frac{1}{n} \sum_{i=1}^{n} \left\{ (\hat{x}_{ij} - \hat{\mu}_j)^2 + \Phi_{jj,i} \right\}, \tag{5.123}$$

$$\hat{\sigma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \left\{ (\hat{x}_{ij} - \hat{\mu}_j)(\hat{x}_{ik} - \hat{\mu}_k) + \Phi_{jk,i} \right\}, \tag{5.124}$$

where $\Phi_{jj,i}$ and $\Phi_{jk,i}$ are as defined by (5.117) and (5.118), respectively.

The method iterates between (5.120)–(5.124) until the estimated parameters in two successive iterations are not materially different.

Remark 6

Assuming multivariate normality, the above formulation of the *iterated Buck's method* is equivalent to the *Missing Information Principle* of Orchard and Woodbury (1972) and the *EM algorithm*, discussed in chapter II. In other words, the EM algorithm is equivalent to an iterated version of Buck's method.

5.5 CONCLUSIONS

In this chapter we have established some relations between imputation techniques and the ML methods of estimation from incomplete data. Specifically, some equivalence relations are established between Buck's method and

Anderson's (1957) fatorization method on the one hand and Buck's method and the EM algorithm on the other hand.

We have started by giving detailed generalizations of Anderson's method to the trivariate and multivariate normal distributions with one variable subject to missingness. On the basis of these generalizations, we have shown that Anderson's method is equivalent to the special case of Buck's method where units have one missing value subject to one variable. This equivalence holds under normality assumptions.

We have also studied, in detail, the generalizations of Anderson's method to the trivariate and multivariate normal distributions where units have more than one missing value. For this case, we have shown the non-equivalence of Anderson's and Buck's methods.

As for the EM algorithm, we have shown that it is equivalent to an iterated version of Buck's method under the multivariate normality assumptions.

CHAPTER VI

ESTIMATION FROM NON-RANDOMLY MISSING CATEGORICAL DATA

6.1 INTRODUCTION

Most work on inference with missing data is based on the explicit or implicit assumption that the missing data are missing at random (MAR) or missing completely at random (MCAR). In chapter II we have discussed the assumptions for the missingness mechanism (MAR/MCAR) and the consequences of each of these assumptions in relation to the required method of analysis as given by Rubin (1976). Specifically, we have seen that the MLE's of the population parameters from incomplete data require the MAR assumption for the missingness mechanism. On the other hand, the use of deletion strategy requires the MCAR assumption. A direct consequence of these assumptions is that they allow the ignorability of the missingness mechanism. Hence standard statistical techniques, under MAR/MCAR assumptions, can be applied to the completely recorded data to obtain estimates of the missing observations or sample estimates of the corresponding population parameters.

The process of collecting categorical data may lead to some observations that cannot be certainly classified to one of the underlying categories. Considerable attention has been devoted to the analysis of such partially classified data. Three pioneering papers dealing with estimation in this case are: Hartley (1958), Blumenthal (1968) and Hocking and Oxspring (1971). In these studies, the partially classified observations are considered to be missing from their respective categories. Moreover, all these studies make the explicit or implicit assumption that the partially classified (missing) ob-

servations are missing at random MAR. In this context of categorical data, the MAR assumption implies that the process leading to partial classification (missingness) does not depend on the identity of the true underlying category.

In some problems, however, the process leading to partial classification may be non-random, i.e., depends on the category identity of the observations. In these cases, estimates based on MAR assumption do not give satisfactory results. In particular, estimation of population proportions must take into account the non-random nature of the missingness mechanism. Pregibon (1977) provided a model for non-random missing data and applied it as a tool of imputation. Little (1980, 1982) indicated how prior odds can be used with non-randomly missing data and estimates can be obtained via the EM algorithm. Nordheim (1978) developed a model for the estimation of population proportions for two categories for data subject to some non-random partial categorization. In a subsequent paper, Nordheim (1984) extended his original work of (1978) by considering misclassification. He also extended the work to the multicategory problem with non-random partial classification and misclassification.

The objective of this chapter is to explore the area of non-ranodm missingness with special emphasis to categorical data. Specifically, we shall try to extend the work of Nordheim (1978, 1984) to the case of both non-random partial classification and non-random misclassification. To achieve this, in section 6.2 we give a review of the original work of Nordheim (1978, 1984). In section 6.3 we consider the extension of Nordheim's work of section 6.2 to the case of non-random misclassification.

6.2 ESTIMATION OF POPULATION PROPORTIONS FROM NON-RANDOMLY MISSING CATEGORICAL DATA WITH NON-RANDOM PARTIAL CLASSIFICATION

Suppose that 100 objects are to be classified to categories A and B. Suppose that 40 objects are classified to A, 40 are classified to B and 20 objects are uncertainly classified to A or B. This situation is shown below

Example 1

Classification	Number of Objects
A	40
B	40
A or B	20
Total	100

Under MAR assumption (for the uncertainly classified objects) the standard estimates of the proportions of the two categories are .5 for A and .5 for B. Furthermore, the missingness mechanism would be described as an entirely random process with probability .2 that an individual object is uncertainly classified.

If, however, the MAR assumption is inappropriate, the above standard estimates are quite misleading. Thus a new procedure for the estimation of category proportions with different probabilities of uncertain classification is required.

To facilitate the development of the procedure, the following notation is introduced:

π_A = Population proportion of category A

π_B = Population proportion of category B

P_{Aa} = Probability that an object of category A is classified to category A

P_{Aab} = Probability that An object of category A is classified to uncertain category A or B

P_{Bb} = Probability that an object of category B is classified to category B

P_{Bab} = Probability that an object of category B is classified to uncertain category A or B

N_a = Number of observations classified to category A

N_b = Number of observations classified to category B

N_{ab} = Number of observations classified to uncertain category A or B

In this notation, capital letters in the subscripts refer to true categories while small letters refer to observed categories. The objective is to estimate π_A and π_B.

The above quantities satisfy the following constraints

$$\pi_A + \pi_B = 1$$
$$P_{Aa} + P_{Aab} = P_{Bb} + P_{Bab} = 1$$
$$N_a + N_b + N_{ab} = N,$$

where N is the total number of observations.

The distribution of $N_a, N_b, N_{ab}, \pi_A, P_{Aab}$ and P_{Bab} is trinomial with probability distribution

$$L(N_a,\ N_b,\ N_{ab}; \pi_A, P_{Aab}, P_{Bab}) \propto \frac{N!}{N_a!\ N_b!\ N_{ab}!}(\pi_A P_{Aa})^{N_a}((1-\pi_A)P_{Bb})^{N_b} (\pi_A P_{Aab} + (1-\pi_A)P_{Bab})^{N_{ab}}. \quad (6.1)$$

From which the loglikelihood is given by

$$\log L(.) \propto N_a \log(\pi_A P_{Aa}) + N_b \log(\pi_B P_{Bb}) + N_{ab} \log(\pi_A P_{Aab} + \pi_b P_{Bab}). \tag{6.2}$$

In general, there are two degrees of freedom in the observation (e.g., N_a, N_b) and three independent parameters (π_A, P_{Aa} and P_{Bb}). Thus the estimation of π_A requires the imposition of at least one additional constraint on the parameters.

In considering the constraint to be imposed on the parameters, this procedure utilizes the knowledge of the experienced analyst. It is anticipated that researchers can develop estimates of P_{Aa} and P_{Bb} by experiment or by experienced judgement.

Introduce the constraint

$$R = \frac{P_{Bab}}{P_{Aab}}. \tag{6.3}$$

Assume that R is known. Note that R=1 corresponds to the MAR case where the probabilities of uncertain classification are equal.

Inserting R in the last term of the R.H.S. of (6.2), the third term becomes

$$\begin{aligned} N_{ab} \log[\pi_A P_{Aab} + (1-\pi_A) R P_{Aab}] &= N_{ab} \log[\pi_A P_{Aab} + R P_{Aab} - \pi_A R P_{Aab}] \\ &= N_{ab} \log[P_{Aab}(\pi_A + R - \pi_A R)] \\ &= N_{ab} \log[P_{Aab}(\pi_A + R(1-\pi_A))] \\ &= N_{ab} \log P_{Aab} + N_{ab} \log[\pi_A + R(1-\pi_A)] \end{aligned} \tag{6.4}$$

Hence the loglikelihood given by (6.2) now becomes

$$\log L(.) \propto N_a \log \pi_A + N_a \log P_{Aa} + N_b \log(1-\pi_A) + N_b \log P_{Bb} + N_{ab} \log P_{Aab} + N_{ab} \log [\pi_A + R(1-\pi_A)]. \quad (6.5)$$

Maximizing $\log L$ as a function of π_A, we have

$$\frac{d \log L}{d\pi_A} = \frac{N_a}{\pi_A} - \frac{N_b}{1-\pi_A} + \frac{N_{ab}(1-R)}{\pi_A + R(1-\pi_A)} = 0$$

or

$$\begin{aligned} &[N_a(1-\pi_A)(\pi_A + R(1-\pi_A)) - N_b \pi_A(\pi_A + R \\ &(1-\pi_A)) + N_{ab}(1-R)\pi_A(1-\pi_A)] \\ &\div [\pi_A(1-\pi_A)(\pi_A + R(1-\pi_A))] = 0. \end{aligned} \quad (6.6)$$

Manipulating (6.6) we get a quadratic equation

$$\pi_A^2 N(1-R) + \pi_A [R(N+N_a) - (N_a + N_{ab})] - RN_a = 0 \quad (6.7)$$

which is to be solved numerically to obtain an interpretable root (between 0 and 1) as an MLE of the population proportion π_A.

Remark 1

Note that

$$\frac{d^2 \log L}{d\pi_A^2} = \frac{-N_a}{\pi_A^2} - \frac{N_b}{(1-\pi_A)^2} - \frac{N_{ab}(1-R)^2}{(\pi_A + R(1-\pi_A))^2}$$

is negative for π_A between 0 and 1 which ensures that the interpretable root is a maximum.

Remark 2

Setting R=1 in equation (6.7), we get

$$\pi_A(N - N_{ab}) - N_a = 0$$

or

$$\pi_A(N_a + N_{ab}) - N_a = 0$$

since $N = N_a + N_b + N_{ab}$.

Therefore

$$\pi_A = \frac{N_a}{N_a + N_b}. \tag{6.8}$$

Which is the standard estimate obtained by ignoring the uncertainly classified objects (N_{ab}) since they are missing at random (MAR). This is what occurs in obtaining the standard estimates of example 1, i.e., $\pi_A = 40/80 = \pi_B = .5$ ignoring the 20 uncertainly classified elements.

Remark 3

The framework developed in section 6.2 can be extended to include misclassification. In this case some of the observations classified to category A(B) may actually belong to category B(A). The model to describe such data must include the conditional probabilities of misclassification, i.e., (P_{Ab} and P_{Ba}).

The quantities necessary for the two category problem with partial classification and misclassification satisfy

$$\pi_A + \pi_B = 1$$

$$P_{Aa} + P_{Ab} + P_{Aab} = P_{Bb} + P_{Ba} + P_{Bab} = 1$$

$$N_a + N_b + N_{ab} = N.$$

In obtaining the MLE of π_A, all the P's are assumed known. The likelihood is

$$L(.) \propto (\pi_A P_{Aa} + (1-\pi_A)P_{Ba})^{N_a} . (\pi_A P_{Ab} + (1-\pi_A)P_{Bb})^{N_b}$$
$$. ((1 - P_{Aa} - P_{Ab})\pi_A + (1-\pi_A)(1 - P_{Ba} - P_{Bb}))^{N_{ab}}.$$

From which the loglikelihood is

$$\log L(.) \propto N_a \log(\pi_A P_{Aa} + (1-\pi_A)P_{Ba}) + N_b \log(\pi_A P_{Ab} + (1-\pi_A)P_{Bb})$$
$$+ N_{ab} \log((1 - P_{Aa} - P_{Ab})\pi_A + (1-\pi_A)(1 - P_{Ba} - P_{Bb})). \tag{6.9}$$

Solving dlogL/dπ_A=0 leads to a quadratic equation for the MLE of π_A.

Remark 4

Putting $P_{Ab} = P_{Ba} = 0$ in equation (6.9) we get

$$\log L(.) \propto N_a \log(\pi_A P_{Aa}) + N_b \log((1-\pi_A)P_{Bb})$$
$$+ N_{ab} \log(\pi_A P_{Aab} + \pi_B P_{Bab}) \tag{6.10}$$

since $1 - P_{Ab} = P_{Aab}$ and $1 - P_{Ba} = P_{Bab}$.
Which is the case of partial classification without misclassification given by equation (6.2).

Remark 5

Putting $1 - P_{Aa} - P_{Ab} = 1 - P_{Bb} - P_{Ba} = 0$ in equation (6.9) we get

$$\log L(.) \propto N_a \log(\pi_A P_{Aa} + (1-\pi_A)P_{Ba}) + N_b \log(\pi_A P_{Ab} + (1-\pi_A)P_{Bb}) \tag{6.11}$$

Which is the case of misclassification without partial classification.

The framework developed above, for the two category problem, can be extended to the multicategory problem with non-random partial classification and misclassification. Nordheim (1984) noted that the multicategory problem can become very involved in practice owing to the difficulty of specifying values for the various probabilities involved in the formulation. Hence the utility of the model is limited to particular restricted situations.

6.3 ESTIMATION OF POPULATION PROPORTIONS FROM NON-RANDOMLY MISSING CATEGORICAL DATA WITH NON-RANDOM PARTIAL CLASSIFICATION AND NON-RANDOM MISCLASSIFICATION

In this section we shall extend Nordheim's work of section 6.2 by considering the case of non-random misclassification. The extension was motivated by noting that if the probability of misclassification differs according to the true category then the standard statistical techniques are inappropriate to allow correct estimation of the population proportion of the two categories. In this extension, an object which is misclassified from either of the two categories is considered to be missing from both categories. Moreover, a separate category (M) for those individuals who are more likely to be misclassified is incorporated in the model. It is anticipated that the experienced analyst can help identifying the elements of (M). This is similar to the idea of Press (1968) who allowed the probabilities of misclassification to vary with observation–such as, for example, the assessment of an interviewer of the probability that an interviewee is telling the truth.

To illustrate the point consider the following simple example. Suppose that 100 objects are to be classified to categories A and B. Suppose that 30 objects are classified to A, 30 are classified to B, 25 are uncertainly classified to A or B, and 15 are misclassified. This situation is shown below

Example 2

Classification	Number of Objects
A	30
B	30
A or B	25
Misclassified	15
Total	100

Under MAR assumption (for both uncertainly classified and misclassified objects) the standard estimates of the proportions of the two categories are .5 for A and .5 for B. Furthermore, the missing data process would be described as an entirely random process with probabilities .25 and .15 that an individual object is uncertainly classified or misclassified respectively.

If, however, the MAR assumption is inappropriate, the above standard estimates are quite misleading. Thus a new procedure for the estimation of category proportions with different probabilities of uncertain classification and misclassification is required. To enhance the development of the procedure we adopt the notation of section 6.2 and let

N_m = Number of misclassified observations.

Then we have the following constraints

$$\pi_A + \pi_B = 1$$
$$P_{Aa} + P_{Ab} + P_{Aab} = P_{Bb} + P_{Ba} + P_{Bab} = 1$$
$$N_a + N_b + N_{ab} + N_m = N,$$

where N is the total number of observations.

The distribution of N_a, N_b, $N_{ab}, N_m; \pi_A, P_{Aab}, P_{Bab}, P_{Ab}$ and P_{Ba} is multinomial of order 4, i.e.,

$$L(.) \propto \frac{N!}{N_a!\, N_b!\, N_{ab}!\, N_m!}(\pi_A P_{Aa})^{N_a}((1-\pi_A)P_{Bb})^{N_b}$$
$$.(\pi_A P_{Aab} + (1-\pi_A)P_{Bab})^{N_{ab}}$$
$$.(\pi_A P_{Ab} + (1-\pi_A)P_{Ba})^{N_m}. \tag{6.12}$$

And the loglikelihood is obtained from (6.12) as

$$\log L \propto N_a \log(\pi_A P_{Aa}) + N_b \log(\pi_B P_{Bb}) + N_{ab} \log(\pi_A P_{Aab} + \pi_b P_{Bab})$$
$$+ N_m \log(\pi_A P_{Ab} + \pi_B P_{Ba}). \tag{6.13}$$

In general, there are three degrees of freedom in the observation (e.g., N_a, N_b, and N_{ab}) and five independent parameters (π_A, P_{Aa}, P_{Aab}, P_{Bb} and P_{Bab}). Thus the estimation of π_A requires the imposition of at least two additional constraints on the parameters.

In considering the constraints to be imposed on the parameters, this procedure utilizes the knowledge of the experienced analyst to provide reasonable estimates of P_{Aab}, P_{Ab}, P_{Bb}, and P_{Bab}.

Introduce the constraints

$$\mathrm{R} = \frac{\mathrm{P_{Bab}}}{\mathrm{P_{Aab}}} \tag{6.14}$$

and

$$\mathrm{Q} = \frac{\mathrm{P_{Ba}}}{\mathrm{P_{Ab}}}. \tag{6.15}$$

Assume that R and Q are known. Note that R=Q=1 corresponds to the MAR case where the probabilities of uncertain classification are equal and those of misclassification are also equal.

Inserting R and Q in the third and fourth terms of the R.H.S. of equation (6.13) respectively, the third term becomes

$$\begin{aligned} N_{ab}\log[\pi_A P_{Aab}+(1-\pi_A)RP_{Aab}] &= N_{ab}\log[\pi_A P_{Aab}+RP_{Aab}-\pi_A RP_{Aab}] \\ &= N_{ab}\log[P_{Aab}(\pi_A+R-\pi_A R)] \\ &= N_{ab}\log[P_{Aab}(\pi_A+R(1-\pi_A))] \\ &= N_{ab}\log P_{Aab}+N_{ab}\log[\pi_A+ \\ &\quad R(1-\pi_A)] \end{aligned} \tag{6.16}$$

Similarly the fourth term of equation (6.13) can be written as

$$\begin{aligned} N_m\log[\pi_A P_{AB}+(1-\pi_A)QP_{Ab}] &= N_m\log[\pi_A P_{Ab}+QP_{Ab}-\pi_A QP_{Ab}] \\ &= N_m\log[P_{Ab}(\pi_A+Q-Q\pi_A)] \\ &= N_m\log[P_{Ab}(\pi_A+Q(1-\pi_A))] \\ &= N_m\log P_{Ab}+N_m\log[\pi_A+ \\ &\quad Q(1-\pi_A)] \quad . \end{aligned} \tag{6.17}$$

Hence the loglikelihood for π_A given by equation (6.13) now becomes

$$\begin{aligned} \log L \propto\ & N_a\log\pi_A+N_a\log P_{Aa}+N_b\log(1-\pi_A)+N_b\log P_{Bb}+N_{ab}\log P_{Aab} \\ & +N_{ab}\log[\pi_A+R(1-\pi_A)]+N_m\log P_{Ab}+N_m\log[\pi_A+Q(1-\pi_A)]\,. \end{aligned} \tag{6.18}$$

Maximizing $\log L$ as a function of π_A, we have

$$\frac{d\log L}{d\pi_A}=\frac{N_a}{\pi_A}-\frac{N_b}{1-\pi_A}+\frac{N_{ab}(1-R)}{\pi_A+R(1-\pi_A)}+\frac{N_m(1-Q)}{\pi_A+Q(1-\pi_A)}=0$$

or

$$\Big[N_a(1-\pi_A)(\pi_A+R(1-\pi_A))(\pi_A+Q(1-\pi_A))-N_b\pi_A(\pi_A+R(1-\pi_A))(\pi_A+Q(1-\pi_A))+N_{ab}(1-R)\pi_A(1-\pi_A)(\pi_A+Q(1-\pi_A))+N_m(1-Q)\pi_A(1-\pi_A)(\pi_A+R(1-\pi_A))\Big]$$
$$\div\Big[\pi_A(1-\pi_A)(\pi_A+R(1-\pi_A))(\pi_A+Q(1-\pi_A))\Big]=0.$$

That is,

$$N_a\Big[\pi_A^2+Q\pi_A(1-\pi_A)+R\pi_A+QR(1-\pi_A)-2R\pi_A^2-2R\pi_AQ(1-\pi_A)-\pi_A^3(1-R)-Q\pi_A^2(1-\pi_A)(1-R)\Big]$$
$$=N_a\Big[\pi_A^2+Q\pi_A-Q\pi_A^2+R\pi_A+QR-QR\pi_A-2R\pi_A^2-2R\pi_AQ+2R\pi_A^2Q-\pi_A^3+\pi_A^3R-Q\pi_A^2+RQ\pi_A^2+Q\pi_A^3-RQ\pi_A^3\Big]$$
$$=N_a\Big[\pi_A^2+Q\pi_A-2Q\pi_A^2+R\pi_A+QR-3QR\pi_A-2R\pi_A^2+3RQ\pi_A^2\pi_A^3+\pi_A^3R+Q\pi_A^3-RQ\pi_A^3\Big]$$
$$=N_a\Big[\pi_A(Q+R-3QR)+\pi_A^2(1-2Q-2R+3RQ)+\pi_A^3(R+Q-RQ-1)+QR\Big], \tag{6.19}$$

and

$$-N_b\Big[\pi_A^2+R\pi_A-R\pi_A^2)(\pi_A+Q(1-\pi_A))\Big]$$
$$=-N_b\Big[(\pi_A^2+R\pi_A-R\pi_A^2)(\pi_A+Q-Q\pi_A)\Big]$$

$$= -N_b\Big[\pi_A^3 + \pi_A^2 Q - \pi_A^3 Q + R\pi_A^2 + R\pi_A Q - R\pi_A^2 Q - R\pi_A^3 - RQ\pi_A^2 + RQ\pi_A^3\Big]$$

$$= -N_b\Big[RQ\pi_A + \pi_A^2(Q + R - 2RQ) + \pi_A^3(1 - Q - R + RQ)\Big] \tag{6.20}$$

and

$$+ N_{ab}\Big[(\pi_A + \pi_A R + \pi_A^2 R - \pi_A^2)(\pi_A + Q - \pi_A Q)\Big]$$

$$= N_{ab}\Big[\pi_A^2 + Q\pi_A - Q\pi_A^2 - R\pi_A^2 - RQ\pi_A + RQ\pi_A^2 + R\pi_A^3 + RQ\pi_A^2 - RQ\pi_A^3 - \pi_A^3 - Q\pi_A^2 + Q\pi_A^3\Big]$$

$$= N_{ab}\Big[Q\pi_A(1 - R) + \pi_A^2(1 - 2Q - R + 2RQ) + \pi_A^3(R - RQ + Q - 1)\Big] \tag{6.21}$$

and

$$+ N_m\Big[(\pi_A - \pi_A Q)(1 - \pi_A)(\pi_A + R - R\pi_A)\Big]$$

$$= N_m\Big[(\pi_A - \pi_A^2 - \pi_A Q + \pi_A^2 Q)(\pi_A + R - \pi_A R)\Big]$$

$$= N_m\Big[\pi_A^2 + R\pi_A - R\pi_A^2 - \pi_A^3 - R\pi_A^2 + R\pi_A^3 - \pi_A^2 Q - \pi_A RQ + RQ\pi_A^2 + \pi_A^3 Q + \pi_A^2 RQ - \pi_A^3 RQ\Big]$$

$$= N_m\Big[R\pi_A(1 - Q) + \pi_A^2(1 - 2R - Q + 2RQ) + \pi_A^3(-1 + R + Q - RQ)\Big] \tag{6.22}$$

(6.19) + (6.20) + (6.21) + (6.22) = 0. Hence

$$-\pi_A^3(1-Q-R+RQ)(N_a+N_b+N_{ab}+N_m)$$

$$+N_a\pi_A^2(1-2Q-2R+3RQ)-N_b\pi_A^2(Q+R-2RQ)+N_{ab}\pi_A^2(1-2Q-R$$

$$+2RQ)+N_m\pi_A^2(1-2R-Q+2RQ)$$

$$+\pi_A\Big[QN_a+RN_a+3RQN_a-RQN_b+QN_{ab}-RQN_{ab}+RN_m-RQN_m\Big]$$

$$+RQN_a=0$$

or

$$-\pi_A^3(1-Q-R+RQ)(N_a+N_b+N_{ab}+N_m)$$

$$+\pi_A^2\Big[N_a-2QN_a-2RN_a+3RQN_a-QN_b-RN_b+2RQN_b+N_{ab}$$

$$-2QN_{ab}-RN_{ab}+2RQN_{ab}+N_m-2RN_m-QN_m+2RQN_m\Big]$$

$$+\pi_A\Big[Q(N_a+N_{ab})+R(N_a+N_m)-RQ(3N_a+N_b+N_{ab}+N_m)\Big]$$

$$+RQN_a=0.$$

That is,

$$-\pi_A^3(1-Q-R+RQ)(N_a+N_b+N_{ab}+N_m)$$

$$+\pi_A^2\Big[RQ(3N_a+2N_b+2N_{ab}+2N_m)-R(2N_a+N_b+N_{ab}+2N_m)-Q(2N_a$$

$$+N_b+2N_{ab}+N_m)\Big]$$

$$+\pi_A\Big[Q(N_a+N_{ab})+R(N_a+N_m)-RQ(N+2N_a)\Big]+RQN_a=0$$

or

$$-\pi_A^3(1-Q-R+RQ)N+\pi_A^2\Big[RQ(2N+N_a)-R(N+N_a+N_m)-Q(N+$$
$$+N_a+N_{ab})+N_a+N_{ab}+N_m\Big]$$
$$+\pi_A\Big[Q(N_a+N_{ab})+R(N_a+N_m)-RQ(N+2N_a)\Big]+RQN_a=0. \tag{6.23}$$

Thus, maximizing $\log L$ as a function of π_A leads to a cubic equation for the maximum likelihood estimate of π_A, i.e,

$$\pi_A^3(1-Q-R+RQ)N-\pi_A^2\Big[RQ(2N+N_a)-R(N+N_a+N_m)$$
$$-Q(N+N_a+N_{ab})+N_a+N_{ab}+N_m\Big]-\pi_A\Big[Q(N_a+N_{ab})+R(N_a+N_m)$$
$$-RQ(N+2N_a)\Big]-RQN_a=0 \tag{6.24}$$

The above equation is to be solved numerically to obtain an interpretable root (between 0 and 1) as an estimate of the population proportion π_A.

Remark 6

Note that

$$\frac{d^2\log L}{d\pi_A^2}=\frac{-N_a}{\pi_A^2}-\frac{N_b}{(1-\pi_A)^2}-\frac{N_{ab}(1-R)^2}{(\pi_A+R(1-\pi_A))^2}-\frac{N_m(1-Q)^2}{(\pi_A+Q(1-Q))^2}$$

is negative for π_A between 0 and 1 which ensures that the interpretable root is a maximum.

Special Case (1)

Setting Q = 1 (MAR case) in equation (6.24), we get

$$
\begin{aligned}
&-\pi_A^2\left[RN - RN_m - N + N_m\right] - \pi_A\left[N_a + N_{ab} - RN_a + RN_m - RN\right] \\
&\qquad - RN_a = 0 \\
&= -\pi_A^2\Big[N(R-1) - N_m(R-1)\Big] - \pi_A\Big[-R(N + N_a - N_m) + (N_a + N_{ab})\Big] \\
&\qquad - RN_a = 0 \\
&= -\pi_A^2\Big[(R-1)(N - N_m)\Big] - \pi_A\Big[-R(N + N_a - N_m) + (N_a + N_{ab})\Big] \\
&\qquad - RN_a = 0 \\
&= \pi_A^2\Big[(1-R)(N_a + N_b + N_{ab})\Big] + \pi_A\Big[R(N_a + N_b + N_{ab} + N_m + N_a - N_m) \\
&\qquad - (N_a + N_{ab})\Big] - RN_a = 0.
\end{aligned}
$$

Since $N - N_m = N_a + N_b + N_{ab}$.

$$
\begin{aligned}
&= \pi_A^2\Big[(1-R)(N_a + N_b + N_{ab})\Big] + \pi_A\Big[R(N_a + N_b + N_{ab} + N_a) - (N_a \\
&\qquad + N_{ab})\Big] - RN_a = 0 \\
&= \pi_A^2 N(1-R) + \pi_A\big[R(N + N_a) - (N_a + N_{ab})\big] - RN_a = 0 \qquad (6.25)
\end{aligned}
$$

Thus for Q=1 equation (6.24) reduces to equation (6.7) which gives Nordheim's MLE of π_A for the case of non-random partial classification without misclassification.

Special Case (2)

Putting $N_m = 0$ (which implies that $P_{Ab} = P_{Ba} = 0$), in equation (6.13) we get

$$\log L(.) \propto N_a \log(\pi_A P_{Aa}) + N_b \log((1-\pi_A)P_{Bb}) + N_{ab} \log(\pi_A P_{Aab} + \pi_B P_{Bab}). \quad (6.26)$$

Which is Nordheim's case of partial classification without misclassification given by equation (6.10).

Remark 7

Putting $P_{Aab} = P_{Bab} = 0$ in equation (6.13) we get

$$\log L(.) \propto N_a \log(\pi_A P_{Aa}) + N_b \log(\pi_B P_{Bb}) + N_m \log(\pi_A P_{Ab} + \pi_B P_{Ba}). \quad (6.27)$$

Which is the case of misclassification without partial classification.

Remark 8

Setting R=Q=1 in equation (6.23), we get

$$-\pi_A [N_{ab} - (N - N_m)] - N_a = 0$$

or

$$-\pi_A [N_{ab} - N_a - N_b - N_{ab}] - N_a = 0$$

since $N - N_m = N_a + N_b + N_{ab}$.

Therefore

$$\pi_A [N_a + N_b] = N_a$$

or

$$\pi_A = \frac{N_a}{N_a + N_b}. \quad (6.28)$$

Which is the standard estimate obtained by ignoring both the uncertainly classified objects (N_{ab}) and the misclassified objects (N_m) since they are both missing at random (MAR).

6.4 CONCLUSIONS

In this chapter we have been concerned with the problem of non-random missingness with special emphasis to categorical data. Specifically, we have extended the work of Nordheim (1978, 1984) to the case of non-random misclassification. Our extension can further be extended to the multicategory problem with non-random misclassification. As we can see, the developed procedure is heavily dependent upon the choice of the parameters R and Q. However, Nordheim (1984) performed a sensitivity analysis by determining a range of plausible values for $\hat{\pi}_A$ depending on the range of R deemed reasonable. The conclusion arrived at is that when non-randomly missing data exist, even a rough estimate of R can result in improved estimates for π_A compared to those estimates obtained by assuming MAR, (R=1). We expect the same conclusion to hold for our extension. A final decision on this can be reached by performing a sensitivity analysis similar to the one performed by Nordheim (1984) for the case of non-random partial classification.

It is our hope that the underlying ideas of this chapter can help in developing more general non-random missingness models.

CHAPTER VII

CONCLUDING REMARKS

In this thesis we have been concerned with the study of imputation techniques in multivariate analysis with emphasis to the method of Buck (1960). The importance of Buck's method as an imputation technique stems from its pioneering nature and its extensive use in the literature of statistical analysis with missing data. Moreover, the method combines two of the three major strategies for handling missing data, namely deletion strategy and imputation strategy. For the remaining third strategy (ML methods), we have shown that, under certain conditions, Anderson's (1957) factorization approach is equivalent to Buck's method. We have also shown that, under multivariate normality assumptions, the EM algorithm is equivalent to an iterated version of Buck's method. The method of Buck is therefore central to all strategies for handling missing data in multivariate analysis.

We have started by a critical review of the method in which the use of Woolf's (1951) procedure for the construction of the regression equations, required for the application of the method, is illustrated. The utilization of the available data by the method of Buck is then studied. The various issues raised in the literature about the method are enumerated and briefly described.

A simplified procedure for the estimation of the bias of the variances for the case of units with one missing value is given. A particular property of the procedure is that it does not require the tedious computation of the inverse of the covariance matrix of the complete cases. Apart from its relative ease of computations, the developed procedure has the advantage of giving

a functional relationship between the relative bias and the coefficient of determination. We have shown that the bias in the estimation of the variances is a function of the sample variance of the residuals resulting from the regression of that variable on the remaining variables. On the other hand, the relative bias is found to be an increasing function of the proportion of missing values and a decreasing function of R^2. However, the resulting estimate is shown to be inconsistent.

The fact that the bias in Buck's method is a decreasing function of R^2 has led us to suggest a simple modification to the method. In this modification, the missing values of the j-th variable are estimated from a subset of the remaining $k-1$ variables rather than the whole set of the $k-1$ variables. The choice of this subset can make use of a procedure developed by Beale *et al* (1967) for the determination of the best subset of variables that maximizes the coefficient of determination R^2.

Given the fact that the amount of bias in Buck's method is a decreasing function of R^2, one would expect the method to give better performance under linearity assumptions. However, The effect of multivariate normality assumptions on the performance of Buck's method needs further investigation. This might require a simulation study where the amount of bias is compared for normal and non-normal incomplete samples. This simulation study may also be needed for the comparison of the method's results with other imputation strategies. The latter has already been done by Haitovsky (1968). Therefore, the statement of Kim and Curry (1977), that the results of Buck cannot be taken seriously because it is based on the examination of a single data set, is lacking a bit of literature review.

We have extended Buck's method to the case of units with more than one missing value. Here, we disagree with Buck's conclusion about the unbiasedness of the covariances. On the contrary, we have shown that this is only true for the case of units with one missing value. In fact the conclusions of Buck (1960) about the biasedness are obtained as special cases of our formulations. For the case of units with more than one missing value, we have shown that the bias of the variances is a function of their respective sample variances of the residuals obtained from the multivariate regression of those variables with missing values on the remaining other variables. Similarly, the bias of the covariances is a function of their respective sample covariances of the residuals obtained from the multivariate regression of those variables with missing values on the remaining other variables. These estimates of the variances and covariances via Buck's method are shown to be inconsistent.

We should note that Buck has dealt explicitly with the case of units with one missing value; the case of units with more than one missing value was left as a generalization. From the application point of view this generalization is straightforward. However, the statistical properties of the resulting estimate of the covariance matrix in the two cases are quite different as we have shown.

The use of the completed data via Buck's method in regression analysis is also studied. This is done for the special case of only one variable subject to missingness. The effect of imputations on the estimated regression coefficients and their precision, the coefficient of determination, and the conventional t-test are investigated. A practical conclusion of this investigation is that the imputed values as well as the method of imputation should be clearly identified. The existence of unrecognized imputed values in a multivariate data set is shown to have caused very misleading conclusions. This is particularly important in public-use data bases which are shared by many

users who may not be aware of the presence of imputed values. In this concern, the study of the effect of imputed values in regression analysis for the cases of units with one missing value and units with more than one missing value is perhaps worthwhile.

More importantly, we have seen that the mechanism that leads to missing data is an important factor that determines the validity of the use of the method. This result agrees with the statement of Little and Rubin (1987) that says: "*Buck's method is only valid under MCAR assumption for the missing values*". The practical implication of this result is that practitioners are advised against the blind application of the method without studying the missingness mechanism. Applying the method when the missing observations are not missing at random might give quite misleading results. Here, the important remark of Afifi and Elashoff (1966), that "*...the specific reasons behind the conditioning of computations on the complete cases are not clear*" has been explained by relating it to the concept of missingness mechanism.

Some relations between the maximum likelihood strategy and imputation strategy for handling missing data in multivariate analysis have been established. Specifically, we have established some relations between Anderson's (1957) factorization method and the method of Buck, on the one hand, and Buck's method and the EM algorithm on the other hand. To establish these relations we have started by elaborating on the generalizations of Anderson's (1957) factorization method to the trivariate and multivariate normal distributions with one variable subject to missingness. These generalizations are then used to show that the special case of Buck's method where units have one missing value subject to one variable is equivalent to Anderson's factorization method under the normality assumptions. This equivalence leads to the interesting result that Anderson's method, as a ML

method of estimation, can as well be used as an imputation technique. This is because the conditional distributions of the factorization method correspond exactly to the regression equations required for imputation by Buck's method.

The above mentioned equivalence between Anderson's and Buck's methods for the case of units with one missing value has motivated us to study the same for the case of units with more than one missing value. To achieve this we have studied, in detail, the generalization of Anderson's method to the case of units with more than one missing value. Unlike the case of one variable subject to missingness, no equivalence relation is obtained between Anderson's and Buck's methods for the case of units with more than one missing value. In fact, in this case Anderson's method can be viewed as a partial imputation technique.

As for Buck's method and the EM algorithm, we have shown that, under multivariate normality assumptions, the EM algorithm due to Dempster *et al* (1977) and the Missing Information Principle of Orchard and Woodbury (1972) are equivalent to an iterated version of Buck's method.

The various relations that have been established above make it clear that the various strategies for handling missing data are not mutually exclusive. For example, we have shown that Anderson's method, as a ML method of estimation, can also be used as an imputation technique. Such type of relations might make it worthwhile to conduct further research on the possible relations between some other imputation techniques and ML methods of estimation from incomplete data. In particular, one can investigate the existence of possible relations between Dear's principal component method and the singular value decomposition method, as imputation techniques, and Anderson's method and the EM algorithm as ML methods of estimation from

incomplete data.

The work of Nordheim (1978, 1984) has been extended to include the case of non-random misclassification where the probabilities of misclassification are not equal. This extesion is achieved by viewing the misclassified elements as missing from their respective categories. The extension can be viewed as a general case that, under certain conditions, reduces to Nordheim's (1978, 1984) procedures. Our extension of Nordheim's procedure can further be generalized to the multicategory problem with non-random misclassification.

In fact, the estimation of non-randomly missing data is an important area of research. An extension of the method of Buck in such a way that it can handle the estimation of non-randomly missing data is perhaps a good suggestion for future research. The work of Nordheim (1978, 1984) and Pregibon (1977), who developed a similar method and employed it as a tool for imputation, might be a good starting point for this suggested extension. Moreover, the statement given by Haitovsky (1968) might also help in this extension. The statement goes as follows:

"*Albert E. Beaton, in conversation, has suggested assigning dummy variables for the missing values, and adding interactions between the dummy and the explanatory variables, to account for different slopes for the different groups of non-random missing observations or categories.*"

Other potential areas for future research include:

a- Development of hybrid methods where two or more strategies could be combined to overcome the limitations of an individual strategy (see for example the EM algorithm and Buck's method).

b- Comparison of the performance of various methods dealing with missing

data would also be worthwhile. Specifically, one can study and compare the performance of Federspiel (1959), Buck (1960) and Kasap (1973) methods. This proposed study may also include an additional method: A version of Buck's method that uses all-available-data for imputation (details of this aditional method are found in Buck, 1960).

In this thesis we have been confined to the derivation of point estimators and their statistical properties. Indeed, a complementary part of the story is the interval estimation and test of hypothesis. The latter has had very little development in the whole area of statistical analysis with missing data, e.g., Rao (1956), Li *et al* (1991) and Alvo and Cabilio (1995).

We conclude this thesis by giving the following general remarks:

i- Despite the enormous literature in the area of statistical analysis with missing data, there is little indication that survey researchers have paid much attention to the literature. When faced with missing data problems, most survey researchers are likely to choose one of the historical approaches (deletion strategy), and then proceed to interpret the resulting statistics as usual. The resulting estimates often require ad hoc adjustments to yield satisfactory estimates.

ii- Most of the literature concerns the derivation of point estimates of parameters, with interval estimation and testing based on large-sample theory. Tests and interval estimates from small samples with missing values have had very little development (Little and Rubin, 1987).

iii- Often the complete data have a distribution belonging to the regular exponential family. Consequently, the complete data ML estimate is the unique solution of the likelihood equations and has the desired asymptotic properties. In contrast, the incomplete data often have a distribution outside

the regular exponential family, and it is possible for the likelihood function to have multiple stationary values. Hence, one cannot always be certain that a given solution of the likelihood equation is the ML estimate (Murray, 1977). In fact, very little work has been done on diagnostic tests on the validity of the assumed probability models when data are incomplete, or on the robustness of estimates derived from them, see for example Liu (1996).

iv- Asymptotic theory for incomplete data patterns is not highly developed. One complicating issue is how to generalize the notion of letting the sample size tend to infinity in the context of incomplete data patterns. A weak condition is to let the proportion of incomplete units tend to zero as the sample size increases (Press and Scott, 1976). A more appropriate asymptotic theory is obtained by allowing the proportion of units with each observed pattern of response to remain constant as the sample size increases (Little, 1979).

v- We feel that there is a need for computer programmes for analyzing general incomplete data problems. These programs should be clearly documented with regard to robustness under distributional and missingness assumptions. This documentation will safeguard the practicing statisticians against the blind application of those programmes without knowing their underlying theoretical assumptions that may not be valid for the case under consideration.

LIST OF REFERENCES

Afifi, A. A, and Elashoff, R. M. (1966). Missing observations in multivariate statistics I: Review of the literature. *Journal of the American Statistical Association*, **61**, 595–604.

Afifi, A. A, and Elashoff, R. M. (1967). Missing observations in multivariate statistics II: Point estimation in simple linear regression. *Journal of the American Statistical Association*, **62**, 10–29.

Afifi, A. A, and Elashoff, R. M. (1969a). Missing observations in multivariate statistics III: Large sample analysis in simple linear regression. *Journal of the American Statistical Association*, **64**, 337–358.

Afifi, A. A, and Elashoff, R. M. (1969b). Missing observations in multivariate statistics IV: A note on simple linear regression. *Journal of the American Statistical Association*, **64**, 359–365.

Allan, F. E. and Wishart, J. (1930). A method of estimating the yield of a missing plot in field experimental work. *Journal of Agricultural Science*, **20**, 399–406.

Alvo, M. and Cabilio, P. (1991). On the balanced incomplete block design for rankings. *Annals of Statistics*, **19**, 1597–1613.

Alvo, M. and Cabilio, P. (1995). Testing ordered alternatives in the presence of incomplete data. *Journal of the American Statistical Association*, **90**, 1015–1024.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, **52**, 200–203.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied botany. *Journal of the Royal Statistical Society*, **B4**, 137–170.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate statistical analysis. *Journal of the Royal Statistical Society*, **B37**, 129–146.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Bello, A. L. (1992). A study of some aspects of imputation techniques in discriminant and regression analyses. *Unpublished Ph.D. thesis*, Department of Statistics, University of Oxford.

Blumenthal, S. (1968). Multinomial sampling with partially categorized data. *Journal of the American Statistical Association*, **63**, 542–551.

Brothwell, D. and Krzanowski, W. J. (1974). Evidence of biological differences between early British populations from the neolithic to medieval times, as revealed by eleven commonly available cranial vault measurements. *Journal of Archaeological Science*, **1**, 249–260.

Brownstone, D. (1991). Multiple imputations for linear regression models. *Working Paper* **MBS**, 1–37. University of California, Irvine, Institute for Mathematical Behavioral Sciences.

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*, **B22**, 302–307.

Chan, L. S. and Dunn, O. J. (1972). The treatment of missing values in discriminant analysis—I: The sampling experiment. *Journal of the American Statistical Association*, **67**, 473–477.

Cochran, W. G. (1977). *Sampling techniques.* Wiley, New York.

Cook, N. R. (1997). An imputation method for non-ignorable missing data in studies of blood pressure. *Statistics in Medicine*, **16**, 2713–2728.

Dear, R. E. (1959). A principal component missing data method for multiple regression models. *Report SP-86*, System Development Corporation, Santa Monica, CA.

Dempster, A. P. Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, **B39**, 1–38.

Dodge, Y. (1985) *Analysis of experiments with missing data.* Wiley, New York.

Dong, H. K. (1985). Non-Germian and singular matrices in maximum likelihood factor analysis, *Applied psychological Measurement*, **9**, 363–366.

Edgett, G. L. (1956). Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association*, **51**, 122–131.

Efron, B. (1994). Missing data, imputation, and the bootstrap (with discussion). *Journal of the American Statistical Association*, **89**, 463–478.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, **91** , 490–517.

Federspiel, C. F. (1959). An investigation of some multiple regression methods for incomplete samples. *Unpublished Ph.D. thesis*, North Carolina State College.

Frane, J. W. (1978). Missing data and BMDP: Some pragmatic approaches. In *proceedings of the statistical computing section, American Statistical Association*, Washington, D.C., 27–33.

Garrett, M. F., Nan, M. L., and Gwendolyn, E. P. Z. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, **91**, 99–108.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, **59**, 834–844.

Good, I. J. (1969). Some applications of the singular value decomposition of a matrix. *Technometrics*, **27**, 823–831.

Greenlees, J. S., Reece, W. S., and Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association*, **77**, 251–261.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society*, **B30**, 67–82.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, **14**, 174–194.

Hartley, H. O. and Hocking, R. R. (1971). The analysis of incomplete data. *Biometrics*, **27**, 783–823.

Hocking, R. R. and Oxspring, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data. *Journal of the American Statistical Association*, **66**, 65–70.

Hocking, R. R. and Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, **63**, 159–173.

Herzog, T., and Rubin, D. B. (1983). Using multiple imputations to handle nonresponse in sample surveys. *Incomplete data in sample surveys.* Volume II, New York, Academic Press, 209–245.

Jackson, E. C. (1968). Missing values in linear multiple discriminant analysis. *Biometrics*, **24**, 835–844.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, **91**, 222–230.

Kasap, H. S. (1973). The problems of missing observations in medical surveys. *Unpublished M.Sc. thesis*, Department of Statistics, University of Oxford.

Kim, J. O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, **6**, 215–240.

Kish, L. (1965). *Survey sampling.* Wiley, New York.

Knol, D. L. and Ten Berge, J. M. F. (1989). Least-squares approximation of an improper correlation matrix by a proper one. *Psychometrika*, **54**, 53–61.

Krzanowski, W. J. (1987). Cross-validation in principal component analysis. *Biometrics*, **43**, 575–584.

Krzanowski, W. J. (1988). Missing value imputation in multivariate data using the singular value decomposition of a matrix. *Biometrical Letters*, **25**, 31–39.

Li, K. H. Merg, X. L., Raghunathan, T. E. and Rubin, D. B. (1991). Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica*, **1**, 65–92.

Little, R. J. A. (1979). Maximum likelihood inference for multiple regression with missing values. *Journal of the Royal Statistical Society*, **B41**, 76–87.

Little, R. J. A. (1980). Superpopulation models for non-response. I: The ignorable case; II: The non-ignorable case. In *Non-Response in Sample Surveys: The Theory of Current Practice, Part V.* Washington, D.C., National Academy of Sciences, Panel on Incomplete Data.

Little, R. J. A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, **77**, 237–250.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, **83**, 1198–1202.

Little, R. J. A. and Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *American Statistician*, **37**, 218–220.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York.

Liu, C. (1996). Bayesian robust multivariate linear regression with incomplete data. *Journal of the American Statistical Association*, **91**, 1219–1227.

Lord, F. M. (1955). Estimation of parameters from incomplete data. *Journal of the American Statistical Association*, **50**, 870–876.

Madow, W. G., Olkin, I., Nisselson. H. and Rubin, D. B. (eds.). (1983). *Incomplete data in sample surveys.* Volume II, New York, Academic Press.

Manly, B. F. J. (1986). *Multivariate statistical methods: A primer.* Chapman and Hall, London.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate analysis.* Academic press, London.

Matthai, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhya*, **11**, 145–152.

McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions.* Wiley, New York.

Murray, G. O. (1977). Discussion of paper by Dempster, Laird and Rubin. *Journal of the Royal Statistical Society,* **B39**, 27.

Nordheim, E. V. (1978). Obtaining information from non-randomly missing data. In *Proceedings of the Statistical Computing Section, American Statistical Association,* Washington, D. C., 34–39.

Nordheim, E. V. (1984). Inference from non-randomly missing categorical data: An example from a genetic study on Turner's Syndrome. *Journal of the American Statistical Association,* **79**, 772–780.

Orchard, T., and Woodbury, M. A. (1972). Missing information principle: Theory and applications. In *proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability, Th-*

eory of statistics, University of California Press, **1**, 697–715.

Pregibon, D. (1977). Typical survey data: Estimation and imputation. *Survey Methodology,* **2**, 70–102.

Press, S. J. (1968). Estimating from misclassified data. *Journal of the American Statistical Association,* **63**, 123–133.

Press, S. J. and Scott, A. J. (1974). Missing variables in Bayesian regression II. *Journal of the American Statistical Association,* **71**, 366–369.

Rao, C. R. (1952). *Advanced statistical methods in biometric research.* New York, Wiley.

Rao, C. R. (1956). Analysis of dispersion with incomplete observations on one of the characters. *Journal of the Royal Statistical Society,* **B18**, 259–264.

Rao, C. R. and Toutenburg, H. (1995). *Linear models: Least squares and alternatives.* Springer Verlag, New York.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association,* **90**, 106–121.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association,* **90**, 122–129.

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to nonresponse, In *proceedings of the survey research methods section of the American Statistical Association,* 20–34.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of non-respondents in sample surveys. *Journal of the American Statistical Association*, **72**, 538–543.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley, New York.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, **91**, 473–489.

Rubin, D. B. and Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. In *Proceedings of the Statistical Computing Section, American Statistical Association*, 83–88.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, **91**, 1278-1288.

Tocher, K. D. (1952). The design and analysis of block experiments. *Journal of the Royal Statistical Society*, **B14**, 45–100.

Toutenburg, H., Heumann, C., Fieger, A., and Park, S. (1995). Missing values in regression: mixed and weighted mixed estimation. *A paper presented to the Symposia Gaussiana, Conference B, Germany*, 289–301.

Trawinski, I. M. and Bargmann, R. W. (1964). Maximum likelihood estimation with incomplete multivariate data. *Annals of Mathematical Statistics*, **35**, 647–657.

Troxel, A. B., Lipsitz, S. R. and Brennan, T. A. (1997). Weighted estimated equations with nonignorably missing response data. *Biometrics*, **53**, 857-869.

Wilks, S. S. (1932). Moments and distribution of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, **3**, 163–195.

Woolf, B. (1951). Computation and interpretation of multiple regressions. *Journal of the Royal Statistical Society*, **B13**, 100–119.

Yates, F. (1933). The analysis of replicated experiments when field results are incomplete, *The Empire Journal of Experimental Agriculture*, **1**, 129–142.

Zhao, L. P., Lipsitz, S. R., and Lew, D. (1996). Regression Analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.

APPENDIX

Table A(1): THE DATA OF BUMPUS (1898)

Case#	x_1	x_2	x_3	x_4	x_5
1	156	245	31.6	18.5	20.5
2	154	240	30.4	17.9	19.6
3	153	240	31.0	18.4	20.6
4	153	236	30.9	17.7	20.2
5	155	243	31.5	18.6	20.3
6	163	247	32.0	19.0	20.9
7	157	238	30.9	18.4	20.2
8	155	239	32.8	18.6	21.2
9	164	248	32.7	19.1	21.1
10	158	238	31.0	18.8	22.0
11	158	240	31.3	18.6	22.0
12	160	244	31.1	18.6	20.5
13	161	246	32.3	19.3	21.8
14	157	245	32.0	19.1	20.0
15	157	235	31.5	18.1	19.8
16	156	237	30.9	18.0	20.3
17	158	244	31.4	18.5	21.6
18	153	238	30.5	18.2	20.9
19	155	236	30.3	18.5	20.1
20	163	246	32.5	18.6	21.9
21	159	236	31.5	18.0	21.5
22	155	240	31.4	18.0	20.7
23	156	240	31.5	18.2	20.6
24	160	242	32.6	18.8	21.7
25	152	232	30.3	17.2	19.8
26	160	250	31.7	18.8	22.5
27	155	237	31.0	18.5	20.0
28	157	245	32.2	19.5	21.4
29	165	245	33.1	19.8	22.7
30	153	231	30.1	17.3	19.8
31	162	239	30.3	18.0	23.1
32	162	243	31.6	18.8	21.3
33	159	245	31.8	18.5	21.7
34	159	247	30.9	18.1	19.0
35	155	243	30.9	18.5	21.3
36	162	252	31.9	19.1	22.2
37	152	230	30.4	17.3	18.6
38	159	242	30.8	18.2	20.5
39	155	238	31.2	17.9	19.3
40	163	249	33.4	19.5	22.8
41	163	242	31.0	18.1	20.7
42	156	237	31.7	18.2	20.3
43	159	238	31.5	18.4	20.3
44	161	245	32.1	19.1	20.8
45	155	235	30.7	17.7	19.6
46	162	247	31.9	19.1	20.4

Case#	x_1	x_2	x_3	x_4	x_5
47	153	237	30.6	18.6	20.4
48	162	245	32.5	18.5	21.1
49	164	248	32.3	18.8	20.9

Number of cases=49

Source: Manly, B.F.J. (1986). Multivariate Statistical Methods: A primer. Chapman and Hall, London, pp. 2–3.

Table A(2): MISSING DATA PATTERN (1)

Case#	x_1	x_2	x_3	x_4	x_5
1	156.00000	245.00000	31.60000	18.50000	20.50000
2	154.00000	240.00000	30.40000	????????	19.60000
3	153.00000	240.00000	31.00000	18.40000	20.60000
4	?????????	236.00000	30.90000	17.70000	20.20000
5	155.00000	243.00000	31.50000	18.60000	20.30000
6	163.00000	?????????	32.00000	19.00000	20.90000
7	157.00000	238.00000	30.90000	18.40000	20.20000
8	155.00000	239.00000	32.80000	18.60000	21.20000
9	164.00000	248.00000	????????	19.10000	21.10000
10	158.00000	238.00000	31.00000	18.80000	????????
11	158.00000	240.00000	31.30000	????????	22.00000
12	160.00000	?????????	31.10000	18.60000	20.50000
13	?????????	246.00000	32.30000	19.30000	21.80000
14	157.00000	245.00000	32.00000	19.10000	????????
15	?????????	235.00000	31.50000	18.10000	19.80000
16	156.00000	237.00000	30.90000	????????	20.30000
17	158.00000	244.00000	31.40000	18.50000	21.60000
18	153.00000	238.00000	30.50000	18.20000	20.90000
19	155.00000	236.00000	30.30000	18.50000	20.10000
20	163.00000	246.00000	????????	18.60000	21.90000
21	159.00000	236.00000	31.50000	18.00000	21.50000
22	155.00000	240.00000	31.40000	18.00000	20.70000
23	156.00000	240.00000	31.50000	18.20000	20.60000
24	160.00000	?????????	32.60000	18.80000	21.70000
25	152.00000	232.00000	30.30000	????????	19.80000
26	?????????	250.00000	31.70000	18.80000	22.50000
27	155.00000	237.00000	31.00000	18.50000	20.00000
28	157.00000	245.00000	32.20000	19.50000	21.40000
29	165.00000	245.00000	????????	19.80000	22.70000
30	153.00000	231.00000	30.10000	17.30000	????????
31	162.00000	?????????	30.30000	18.00000	23.10000
32	162.00000	243.00000	31.60000	18.80000	21.30000
33	159.00000	245.00000	31.80000	18.50000	21.70000
34	?????????	247.00000	30.90000	18.10000	19.00000
35	155.00000	243.00000	30.90000	18.50000	????????
36	162.00000	?????????	31.90000	19.10000	22.20000
37	152.00000	?????????	30.40000	17.30000	18.60000
38	159.00000	242.00000	30.80000	18.20000	20.50000
39	155.00000	238.00000	????????	17.90000	19.30000

Case#	x_1	x_2	x_3	x_4	x_5
40	163.00000	249.00000	33.40000	????????	22.80000
41	163.00000	242.00000	31.00000	18.10000	20.70000
42	156.00000	?????????	31.70000	18.20000	20.30000
43	159.00000	238.00000	31.50000	18.40000	20.30000
44	161.00000	245.00000	32.10000	????????	20.80000
45	155.00000	235.00000	30.70000	17.70000	19.60000
46	162.00000	247.00000	31.90000	19.10000	20.40000
47	153.00000	237.00000	????????	18.60000	20.40000
48	?????????	245.00000	32.50000	18.50000	21.10000
49	164.00000	248.00000	32.30000	18.80000	????????

????????? and ???????? indicate missing values

Tables A(2.1)–A(2.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (1)

Table A(2.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	157.1500	2.9069	20
X_2	240.6500	3.5285	20
X_3	31.3450	.5925	20
X_4	18.4350	.3964	20
X_5	20.7050	.5808	20

Table A(2.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 8.4500 & & & & \\ 4.6868 & 12.4500 & & & \\ .3982 & 1.0534 & .3510 & & \\ .2418 & .8813 & .1268 & .1571 & \\ .4887 & .8861 & .1856 & .0698 & .3373 \end{bmatrix}$$

Number of cases = 20

Tables A(2.3)–A(2.13): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (1)

Table A(2.3):

Dependent Variable X_1: total length

Multiple R	.47991
R Square	.23032
Adjusted R Square	.02507
Standard Error	2.87022

Table A(2.4):
Variables in the Equation

Variable	β	SE(β)	T–Value	Sig T
X_5	.615546	1.388361	.443	.6638
X_4	-.960966	2.280632	-.421	.6795
X_3	-.062308	1.504425	-.041	.9675
X_2	.405943	.255991	1.586	.1336
(Constant)	66.383457	47.1483885	1.408	.1795

Table A(2.5):
Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	36.97755	9.24439
Residual	15	123.57245	8.23816
Total		160.55000	

F = 1.12214, Signif F = .3829

Table A(2.6):
Dependent Variable X_2: alar extent

Multiple R	.73814
R Square	.54486
Adjusted R Square	.42348
Standard Error	2.67911

Table A(2.7):
Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	128.88575	32.2214
Residual	15	107.66425	7.17762
Total		236.55	

F = 4.48916, Signif F = .0139

Table A(2.8):
Dependent Variable X_3: length of beak & head

Multiple R	.67402
R Square	.45431
Adjusted R Square	.30879
Standard Error	.49258

Table A(2.9):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	3.03002	.75750
Residual	15	3.63948	.24263
Total		6.6695	

F = 3.12203, Signif F = .0469

Table A(2.10):

Dependent Variable X_4: length of humerus

Multiple R	.68970
R Square	.47568
Adjusted R Square	.33587
Standard Error	.32304

Table A(2.11):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	1.42016	.35504
Residual	15	1.56534	.10436
Total		2.98550	

F = 3.40218, Signif F = .0360

Table A(2.12):

Dependent Variable X_8: length of keel of sternum

Multiple R	.58465
R Square	.34182
Adjusted R Square	.16630
Standard Error	.53032

Table A(2.13):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	2.19087	.54772
Residual	15	4.21863	.28124
Total		6.40950	

F = 1.94749, Signif F = .1547

Table A(3): MISSING DATA PATTERN (2)

Case#	x_1	x_2	x_3	x_4	x_5
1	?????????	245.00000	31.60000	18.50000	20.50000
2	154.00000	240.00000	30.40000	17.90000	19.60000
3	153.00000	?????????	31.00000	18.40000	20.60000
4	153.00000	236.00000	30.90000	17.70000	20.20000
5	155.00000	243.00000	31.50000	18.60000	20.30000
6	?????????	247.00000	32.00000	19.00000	20.90000
7	157.00000	238.00000	30.90000	18.40000	????????
8	155.00000	239.00000	32.80000	18.60000	21.20000
9	164.00000	248.00000	32.70000	19.10000	21.10000
10	158.00000	238.00000	31.00000	18.80000	22.00000
11	158.00000	240.00000	31.30000	18.60000	22.00000
12	160.00000	244.00000	31.10000	18.60000	20.50000
13	161.00000	246.00000	32.30000	????????	21.80000
14	157.00000	?????????	32.00000	19.10000	20.00000
15	157.00000	235.00000	31.50000	18.10000	19.80000
16	156.00000	?????????	30.90000	18.00000	20.30000
17	158.00000	244.00000	31.40000	18.50000	21.60000
18	153.00000	?????????	30.50000	18.20000	20.90000
19	155.00000	236.00000	30.30000	18.50000	20.10000
20	?????????	246.00000	32.50000	18.60000	21.90000
21	159.00000	236.00000	31.50000	18.00000	21.50000
22	155.00000	240.00000	31.40000	18.00000	20.70000
23	156.00000	240.00000	31.50000	18.20000	20.60000
24	160.00000	242.00000	????????	18.80000	21.70000
25	152.00000	232.00000	30.30000	17.20000	19.80000
26	160.00000	250.00000	31.70000	18.80000	22.50000
27	155.00000	237.00000	31.00000	18.50000	20.00000
28	157.00000	245.00000	32.20000	19.50000	????????
29	?????????	245.00000	33.10000	19.80000	22.70000
30	153.00000	231.00000	30.10000	????????	19.80000
31	162.00000	239.00000	30.30000	18.00000	23.10000
32	162.00000	243.00000	31.60000	18.80000	21.30000
33	159.00000	245.00000	31.80000	????????	21.70000
34	159.00000	247.00000	30.90000	18.10000	19.00000
35	155.00000	243.00000	30.90000	????????	21.30000
36	162.00000	252.00000	31.90000	19.10000	22.20000
37	152.00000	230.00000	30.40000	17.30000	????????
38	159.00000	242.00000	30.80000	18.20000	20.50000
39	155.00000	238.00000	????????	17.90000	19.30000
40	163.00000	249.00000	????????	19.50000	22.80000
41	163.00000	242.00000	????????	18.10000	20.70000
42	156.00000	237.00000	31.70000	18.20000	20.30000
43	159.00000	238.00000	31.50000	18.40000	20.30000
44	161.00000	245.00000	32.10000	19.10000	20.80000
45	155.00000	235.00000	30.70000	17.70000	19.60000
46	162.00000	247.00000	31.90000	19.10000	20.40000
47	153.00000	237.00000	30.60000	18.60000	20.40000
48	162.00000	245.00000	32.50000	18.50000	????????
49	164.00000	248.00000	32.30000	18.80000	20.90000

????????? and ???????? indicate missing values

Tables A(3.1)–A(3.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (2)

Table A(3.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	157.8621	2.9069	29
X_2	240.9655	5.0743	29
X_3	31.3310	.6714	29
X_4	18.4069	.4735	29
X_5	20.7690	.9551	29

Table A(3.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 11.4803 & & & & \\ 12.6379 & 25.7488 & & & \\ 1.1723 & 1.8333 & .4508 & & \\ 1.0331 & 1.7002 & .2016 & .2242 & \\ 1.6170 & 1.6525 & .1671 & .1881 & .9122 \end{bmatrix}$$

Number of cases = 29

Tables A(3.3)–A(3.12): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (2)

Table A(3.3):

Dependent Variable X_1: total length

Multiple R	.79286
R Square	.62862
Adjusted R Square	.56673
Standard Error	2.23027

Table A(3.4):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	202.06949	50.51737
Residual	24	119.37878	4.97412
Total		321.44827	

F = 10.15605, Signif F = .0001

Table A(3.5):

Dependent Variable X_2: alar extent

Multiple R	.80106
R Square	.64170
Adjusted R Square	.58198
Standard Error	3.28078

Table A(3.6):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	462.64179	115.66045
Residual	24	258.32373	10.76349
Total		720.96552	

F = 10.74563, Signif F = .0000

Table A(3.7):

Dependent Variable X_3: length of beak & head

Multiple R	.65393
R Square	.42763
Adjusted R Square	.33223
Standard Error	.54866

Table A(3.8):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	5.39752	1.34938
Residual	24	7.22455	.30102
Total		12.62207	

F = 4.48264, Signif F = .0076

Table A(3.9):

Dependent Variable X_4: length of humerus

Multiple R	.78755
R Square	.62023
Adjusted R Square	.55694
Standard Error	.31520

Table A(3.10):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	3.89420	.97355
Residual	24	2.38442	.09935
Total		6.27862	

F = 9.79911, Signif F = .0001

Table A(3.11):

Dependent Variable X_5: length of keel of sternum

Multiple R	.52784
R Square	.27861
Adjusted R Square	.15838
Standard Error	.87621

Table A(3.12):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	7.11634	1.77909
Residual	24	18.42573	.76774
Total		25.54207	

F = 2.31731, Signif F = .0863

Table A(4): MISSING DATA PATTERN (3)

Case#	x_1	x_2	x_3	x_4	x_5
1	156.00000	245.00000	????????	18.50000	20.50000
2	154.00000	?????????	30.40000	17.90000	19.60000
3	153.00000	240.00000	31.00000	18.40000	20.60000
4	153.00000	236.00000	30.90000	17.70000	20.20000
5	155.00000	?????????	31.50000	18.60000	20.30000
6	163.00000	247.00000	32.00000	19.00000	20.90000
7	?????????	238.00000	30.90000	18.40000	20.20000
8	155.00000	239.00000	32.80000	18.60000	21.20000
9	164.00000	?????????	32.70000	19.10000	21.10000
10	158.00000	238.00000	31.00000	18.80000	22.00000
11	158.00000	240.00000	31.30000	18.60000	22.00000
12	160.00000	244.00000	31.10000	18.60000	20.50000

Case#	x_1	x_2	x_3	x_4	x_5
13	161.00000	246.00000	32.30000	????????	21.80000
14	?????????	245.00000	32.00000	19.10000	20.00000
15	157.00000	235.00000	31.50000	18.10000	19.80000
16	156.00000	237.00000	30.90000	18.00000	????????
17	158.00000	244.00000	31.40000	18.50000	21.60000
18	?????????	238.00000	30.50000	18.20000	20.90000
19	155.00000	236.00000	30.30000	18.50000	????????
20	163.00000	246.00000	32.50000	????????	21.90000
21	159.00000	236.00000	31.50000	18.00000	21.50000
22	155.00000	240.00000	????????	18.00000	20.70000
23	156.00000	240.00000	????????	18.20000	20.60000
24	160.00000	242.00000	????????	18.80000	21.70000
25	152.00000	232.00000	30.30000	17.20000	19.80000
26	160.00000	250.00000	31.70000	18.80000	22.50000
27	155.00000	237.00000	31.00000	18.50000	????????
28	157.00000	245.00000	32.20000	????????	21.40000
29	165.00000	245.00000	33.10000	????????	22.70000
30	153.00000	231.00000	30.10000	17.30000	19.80000
31	162.00000	239.00000	????????	18.00000	23.10000
32	162.00000	?????????	31.60000	18.80000	21.30000
33	159.00000	245.00000	31.80000	18.50000	21.70000
34	159.00000	247.00000	30.90000	18.10000	????????
35	155.00000	243.00000	30.90000	????????	21.30000
36	?????????	252.00000	31.90000	19.10000	22.20000
37	152.00000	230.00000	30.40000	17.30000	18.60000
38	?????????	242.00000	30.80000	18.20000	20.50000
39	155.00000	238.00000	31.20000	17.90000	19.30000
40	163.00000	249.00000	33.40000	19.50000	22.80000
41	163.00000	242.00000	31.00000	18.10000	20.70000
42	156.00000	237.00000	31.70000	18.20000	20.30000
43	159.00000	238.00000	31.50000	18.40000	20.30000
44	161.00000	245.00000	32.10000	19.10000	20.80000
45	155.00000	?????????	30.70000	17.70000	19.60000
46	162.00000	247.00000	31.90000	19.10000	20.40000
47	153.00000	237.00000	30.60000	18.60000	????????
48	162.00000	245.00000	32.50000	18.50000	21.10000
49	164.00000	248.00000	32.30000	18.80000	20.90000

????????? and ???????? indicate missing values

Tables A(4.1)–A(4.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (3)

Table A(4.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	158.1250	3.8029	24
X_2	240.6667	5.7760	24
X_3	31.5167	.7772	24
X_4	18.3750	.5951	24
X_5	20.8042	1.0050	24

Table A(4.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 14.4620 & & & & \\ 18.2174 & 33.3623 & & & \\ 1.9326 & 3.1884 & .6041 & & \\ 1.7424 & 2.9870 & .3665 & .3541 & \\ 2.0690 & 3.8145 & .4469 & .4210 & 1.0100 \end{bmatrix}$$

Number of cases = 24

Tables A(4.3)–A(4.12): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (3)

Table A(4.3):

Dependent Variable X_1: total length

Multiple R	.83789
R Square	.70206
Adjusted R Square	.63934
Standard Error	2.28384

Table A(4.4):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Square
Regression	4	233.52249	58.38062
Residual	19	99.10251	5.21592
Total		332.62500	

F = 10.15605, Signif F = .0001

Table A(4.5):

Dependent Variable X_2: alar extent

Multiple R	.90687
R Square	.82242
Adjusted R Square	.78503
Standard Error	2.67803

Table A(4.6):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Square
Regression	4	631.06798	157.76700
Residual	19	136.26535	7.17186
Total		767.33333	

F = 21.99806, Signif F = .0000

Table A(4.7):

Dependent Variable X_3: length of beak & head

Multiple R	.79571
R Square	.63316
Adjusted R Square	.55593
Standard Error	.51793

Table A(4.8):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Square
Regression	4	8.79665	2.19916
Residual	19	5.09669	2.6825
Total		13.89334	

F = 8.19828, Signif F = .0005

Table A(4.9):

Dependent Variable X_4: length of humerus

Multiple R	.91473
R Square	.83673
Adjusted R Square	.80235
Standard Error	.26456

Table A(4.10):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	6.81514	1.70378
Residual	19	1.32986	.06999
Total		8.14500	

F = 24.34227, Signif F = .0000

Table A(4.11):

Dependent Variable X_5: length of keel of sternum

Multiple R	.71236
R Square	.50745
Adjusted R Square	.40376
Standard Error	.77601

Table A(4.12):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	11.78789	2.94697
Residual	19	11.44169	.60219
Total		23.22958	

F = 4.89373, Signif F = .0070

Table A(5): MISSING DATA PATTERN (4)

Case#	x_1	x_2	x_3	x_4	x_5
1	156.00000	245.00000	31.60000	18.50000	20.50000
2	154.00000	240.00000	30.40000	17.90000	????????
3	153.00000	240.00000	31.00000	18.40000	????????
4	153.00000	236.00000	30.90000	????????	20.20000
5	?????????	243.00000	31.50000	18.60000	20.30000
6	163.00000	247.00000	????????	19.00000	20.90000
7	157.00000	238.00000	30.90000	????????	20.20000
8	155.00000	239.00000	32.80000	?????????	21.20000
9	?????????	248.00000	32.70000	19.10000	21.10000
10	158.00000	238.00000	31.00000	18.80000	22.00000
11	158.00000	240.00000	31.30000	18.60000	????????
12	160.00000	244.00000	31.10000	18.60000	20.50000
13	161.00000	246.00000	32.30000	????????	21.80000
14	157.00000	245.00000	????????	19.10000	20.00000
15	157.00000	235.00000	31.50000	18.10000	19.80000
16	156.00000	237.00000	30.90000	18.00000	20.30000
17	158.00000	244.00000	31.40000	18.50000	21.60000
18	153.00000	238.00000	30.50000	18.20000	20.90000
19	155.00000	?????????	30.30000	18.50000	20.10000
20	163.00000	?????????	32.50000	18.60000	21.90000
21	159.00000	236.00000	????????	18.00000	21.50000
22	155.00000	240.00000	31.40000	18.00000	20.70000
23	156.00000	240.00000	31.50000	18.20000	20.60000
24	160.00000	242.00000	32.60000	18.80000	21.70000

Case#	x_1	x_2	x_3	x_4	x_5
25	152.00000	232.00000	30.30000	17.20000	19.80000
26	160.00000	?????????	31.70000	18.80000	22.50000
27	?????????	237.00000	31.00000	18.50000	20.00000
28	157.00000	245.00000	32.20000	19.50000	21.40000
29	165.00000	?????????	33.10000	19.80000	22.70000
30	153.00000	231.00000	????????	17.30000	19.80000
31	?????????	239.00000	30.30000	18.00000	23.10000
32	162.00000	243.00000	31.60000	????????	21.30000
33	159.00000	245.00000	31.80000	18.50000	21.70000
34	159.00000	247.00000	30.90000	18.10000	19.00000
35	155.00000	243.00000	30.90000	????????	21.30000
36	162.00000	252.00000	31.90000	????????	22.20000
37	?????????	230.00000	30.40000	17.30000	18.60000
38	159.00000	242.00000	????????	18.20000	20.50000
39	155.00000	238.00000	31.20000	17.90000	????????
40	163.00000	249.00000	33.40000	19.50000	????????
41	163.00000	242.00000	31.00000	18.10000	????????
42	?????????	237.00000	31.70000	18.20000	20.30000
43	159.00000	?????????	31.50000	18.40000	20.30000
44	161.00000	245.00000	????????	19.10000	20.80000
45	155.00000	?????????	30.70000	17.70000	19.60000
46	162.00000	247.00000	31.90000	19.10000	????????
47	153.00000	?????????	30.60000	18.60000	20.40000
48	?????????	245.00000	32.50000	18.50000	21.10000
49	164.00000	248.00000	????????	18.80000	20.90000

????????? and ???????? indicate missing values

Tables A(5.1)-A(5.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (4)

Table A(5.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	156.8571	2.4133	14
X_2	240.8571	4.4351	14
X_3	31.3357	.6209	14
X_4	18.3571	.5258	14
X_5	20.7500	.8662	14

Table A(5.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 5.8242 & & & & \\ 7.0549 & 19.6703 & & & \\ .8670 & 1.3747 & .3855 & & \\ .7626 & 1.4473 & .2278 & .2765 & \\ .5923 & .7462 & .2665 & .2962 & .7504 \end{bmatrix}$$

Number of cases = 14

Tables A(5.3)–A(5.13): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (4)

Table A(5.3):

Dependent Variable X_1: total length

Multiple R	.72690
R Square	.52838
Adjusted R Square	.31877
Standard Error	1.99188

Table A(5.4):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	40.00588	10.00147
Residual	9	35.70841	3.96760
Total		75.71429	

F = 2.52079, Signif F = .1148

Table A(5.5):

Dependent Variable X_2: alar extent

Multiple R	.74848
R Square	.56023
Adjusted R Square	..36478
Standard Error	3.53484

Table A(5.6):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	143.25861	35.81465
Residual	9	112.45567	12.49507
Total		255.71428	

F = 2.86630, Signif F = .0873

Table A(5.7):

Dependent Variable X_3: length of beak & head

Multiple R	.73174
R Square	.53545
Adjusted R Square	.32898
Standard Error	.50864

Table A(5.8):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	2.68375	.67094
Residual	9	2.32839	.25871
Total		5.01214	

F = 2.59340, Signif F = .1082

Table A(5.9):

Dependent Variable X_4: length of humerus

Multiple R	.85500
R Square	.73103
Adjusted R Square	.61149
Standard Error	.32775

Table A(5.10):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	2.62753	.65688
Residual	9	.96676	.10742
Total		3.59429	

F = 6.11521, Signif F = .0116

Table A(5.11):

Dependent Variable X_5: length of keel of sternum

Multiple R	.70969
R Square	.50366
Adjusted R Square	.28306
Standard Error	.73347

Table A(5.12):

Variables in the Equation

Variable	β	SE(β)	T–Value	Sig T
X_4	1.300797	.607019	2.143	.0607
X_1	-.020018	.122562	-.163	.8739
X_2	-.064385	.065752	-.979	.3561
X_3	.197336	.476158	.414	.6883
(Constant)	9.335119	14.510802	.643	.5361

Table A(5.13):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	4	4.91316	1.22829
Residual	9	4.84184	.53798
Total		9.75500	

F = 2.28314, Signif F = .1397

Table A(5.14): POST-IMPUTATION REGRESSION COEFFICIENTS OF X_5 ON X_1, X_2, X_3, X_4 FOR THE MISSING DATA PATTERN (4)

Variable	β	SE(β)	T-Value	Sig T
X_4	.952540	.390821	2.437	.0189
X_1	.033641	.053748	.626	.5346
X_3	.098438	.233346	.422	.6752
X_2	-.026106	.046045	-.567	.5736
(Constant)	1.147089	6.044673	.190	.8504

Table A(6): MISSING DATA PATTERN (5)

Case#	x_1	x_2	x_3	x_4	x_5
1	?????????	?????????	31.60000	18.50000	20.50000
2	154.00000	240.00000	30.40000	17.90000	19.60000
3	153.00000	240.00000	31.00000	18.40000	20.60000
4	153.00000	236.00000	30.90000	17.70000	20.20000
5	155.00000	243.00000	31.50000	18.60000	20.30000
6	163.00000	247.00000	32.00000	19.00000	20.90000
7	157.00000	238.00000	30.90000	18.40000	20.20000
8	155.00000	239.00000	32.80000	18.60000	21.20000
9	?????????	?????????	32.70000	19.10000	21.10000
10	158.00000	238.00000	31.00000	18.80000	22.00000
11	158.00000	240.00000	31.30000	18.60000	22.00000
12	160.00000	244.00000	31.10000	18.60000	20.50000
13	161.00000	246.00000	32.30000	19.30000	21.80000
14	157.00000	245.00000	32.00000	19.10000	20.00000
15	157.00000	235.00000	31.50000	18.10000	19.80000
16	156.00000	237.00000	30.90000	18.00000	20.30000
17	?????????	?????????	31.40000	18.50000	21.60000
18	153.00000	238.00000	30.50000	18.20000	20.90000
19	155.00000	236.00000	30.30000	18.50000	20.10000
20	163.00000	246.00000	32.50000	18.60000	21.90000
21	159.00000	236.00000	31.50000	18.00000	21.50000
22	?????????	?????????	31.40000	18.00000	20.70000
23	156.00000	240.00000	31.50000	18.20000	20.60000
24	160.00000	242.00000	32.60000	18.80000	21.70000
25	152.00000	232.00000	30.30000	17.20000	19.80000
26	160.00000	250.00000	31.70000	18.80000	22.50000
27	155.00000	237.00000	31.00000	18.50000	20.00000

Case#	x_1	x_2	x_3	x_4	x_5
28	?????????	?????????	32.20000	19.50000	21.40000
29	165.00000	245.00000	33.10000	19.80000	22.70000
30	?????????	?????????	30.10000	17.30000	19.80000
31	?????????	?????????	30.30000	18.00000	23.10000
32	162.00000	243.00000	31.60000	18.80000	21.30000
33	159.00000	245.00000	31.80000	18.50000	21.70000
34	159.00000	247.00000	30.90000	18.10000	19.00000
35	?????????	?????????	30.90000	18.50000	21.30000
36	162.00000	252.00000	31.90000	19.10000	22.20000
37	152.00000	230.00000	30.40000	17.30000	18.60000
38	159.00000	242.00000	30.80000	18.20000	20.50000
39	?????????	?????????	31.20000	17.90000	19.30000
40	163.00000	249.00000	33.40000	19.50000	22.80000
41	163.00000	242.00000	31.00000	18.10000	20.70000
42	156.00000	237.00000	31.70000	18.20000	20.30000
43	159.00000	238.00000	31.50000	18.40000	20.30000
44	?????????	?????????	32.10000	19.10000	20.80000
45	155.00000	235.00000	30.70000	17.70000	19.60000
46	162.00000	247.00000	31.90000	19.10000	20.40000
47	153.00000	237.00000	30.60000	18.60000	20.40000
48	?????????	?????????	32.50000	18.50000	21.10000
49	164.00000	248.00000	32.30000	18.80000	20.90000

????????? indicates missing values.

Tables A(6.1)-A(6.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (5)

Table A(6.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	157.9737	3.7016	38
X_2	241.1053	5.1925	38
X_3	31.4500	.7907	38
X_4	18.4763	.5524	38
X_5	20.7842	1.0018	38

Table A(6.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 13.7020 & & & & \\ 14.7866 & 26.9616 & & & \\ 2.0257 & 2.5919 & .6253 & & \\ 1.4102 & 2.1404 & .3285 & .3051 & \\ 2.2888 & 2.9774 & .5419 & .3907 & 1.0035 \end{bmatrix}$$

Number of cases = 38

Tables A(6.3)-A(6.6): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (5)

Table A(6.3):

Dependent Variable X_1: total length

Multiple R	.74535
R Square	.55554
Adjusted R Square	.51632
Standard Error	2.57436

Table A(6.4):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	3	281.64428	93.88143
Residual	34	225.32940	6.62734
Total		506.97368	

F = 14.16579, Signif F = .0000

Table A(6.5):

Dependent Variable X_2: alar extent

Multiple R	.75443
R Square	.56917
Adjusted R Square	.53115
Standard Error	3.55539

Table A(6.6):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	3	567.79093	189.26364
Residual	34	429.78802	12.64082
Total		997.57895	

F = 14.97241, Signif F = .0000

Table A(7): MISSING DATA PATTERN (6)

Case#	x_1	x_2	x_3	x_4	x_5
1	?????????	?????????	31.60000	18.50000	????????
2	154.00000	240.00000	30.40000	17.90000	19.60000
3	153.00000	240.00000	31.00000	18.40000	20.60000
4	?????????	?????????	30.90000	17.70000	????????
5	155.00000	243.00000	31.50000	18.60000	20.30000
6	163.00000	247.00000	32.00000	19.00000	20.90000
7	157.00000	238.00000	30.90000	18.40000	20.20000
8	155.00000	239.00000	32.80000	18.60000	21.20000
9	164.00000	248.00000	32.70000	19.10000	21.10000
10	?????????	?????????	31.00000	18.80000	????????
11	158.00000	240.00000	31.30000	18.60000	22.00000
12	160.00000	244.00000	31.10000	18.60000	20.50000
13	161.00000	246.00000	32.30000	19.30000	21.80000
14	157.00000	245.00000	32.00000	19.10000	20.00000
15	?????????	?????????	31.50000	18.10000	????????
16	156.00000	237.00000	30.90000	18.00000	20.30000
17	158.00000	244.00000	31.40000	18.50000	21.60000
18	?????????	?????????	30.50000	18.20000	????????
19	155.00000	236.00000	30.30000	18.50000	20.10000
20	163.00000	246.00000	32.50000	18.60000	21.90000
21	159.00000	236.00000	31.50000	18.00000	21.50000
22	?????????	?????????	31.40000	18.00000	????????
23	156.00000	240.00000	31.50000	18.20000	20.60000
24	160.00000	242.00000	32.60000	18.80000	21.70000
25	?????????	?????????	30.30000	17.20000	????????
26	160.00000	250.00000	31.70000	18.80000	22.50000
27	155.00000	237.00000	31.00000	18.50000	20.00000
28	?????????	?????????	32.20000	19.50000	????????
29	?????????	?????????	33.10000	19.80000	????????
30	153.00000	231.00000	30.10000	17.30000	19.80000
31	162.00000	239.00000	30.30000	18.00000	23.10000
32	162.00000	243.00000	31.60000	18.80000	21.30000
33	159.00000	245.00000	31.80000	18.50000	21.70000
34	?????????	?????????	30.90000	18.10000	????????
35	155.00000	243.00000	30.90000	18.50000	21.30000
36	162.00000	252.00000	31.90000	19.10000	22.20000
37	152.00000	230.00000	30.40000	17.30000	18.60000
38	159.00000	242.00000	30.80000	18.20000	20.50000
39	155.00000	238.00000	31.20000	17.90000	19.30000
40	?????????	?????????	33.40000	19.50000	????????
41	163.00000	242.00000	31.00000	18.10000	20.70000
42	156.00000	237.00000	31.70000	18.20000	20.30000
43	159.00000	238.00000	31.50000	18.40000	20.30000
44	161.00000	245.00000	32.10000	19.10000	20.80000
45	?????????	?????????	30.70000	17.70000	????????
46	162.00000	247.00000	31.90000	19.10000	20.40000
47	153.00000	237.00000	30.60000	18.60000	20.40000
48	162.00000	245.00000	32.50000	18.50000	21.10000
49	?????????	?????????	32.30000	18.80000	????????

????????? and ???????? indicate missing values

Tables A(7.1)–A(7.2): SUMMARY STATISTICS FOR THE COMPLETE CASES OF THE MISSING DATA PATTERN (6)

Table A(7.1):

VARIABLE	MEAN	STD DEV	CASES
X_1	158.1667	3.4600	36
X_2	241.4444	4.8899	36
X_3	31.4361	.7345	36
X_4	18.4750	.4723	36
X_5	20.8389	.9348	36

Table A(7.2):

COVARIANCE MATRIX

$$\begin{bmatrix} 11.9714 & & & & \\ 12.1524 & 23.9111 & & & \\ 1.4510 & 2.3035 & .5395 & & \\ .9129 & 1.8600 & .2398 & .2231 & \\ 1.9848 & 2.6079 & .2777 & .1984 & .8739 \end{bmatrix}$$

Number of cases = 36

Tables A(7.3)–A(7.8): RESULTS OF THE MULTIPLE REGRESSIONS FOR THE MISSING DATA PATTERN (6)

Table A(7.3):

Dependent Variable X_1: total length

Multiple R	.61437
R Square	.37745
Adjusted R Square	.33972
Standard Error	2.81150

Table A(7.4):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	2	158.15068	79.07534
Residual	33	260.84932	7.90452
Total		419	

F = 10.00381, Signif F = .0004

Table A(7.5):

Dependent Variable X_2: alar extent

Multiple R	.81384
R Square	.66234
Adjusted R Square	.64187
Standard Error	2.92630

Table A(7.6):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	2	554.30190	277.15095
Residual	33	282.58699	8.56324
Total		836.88889	

F = 32.36519, Signif F = .0000

Table A(7.7):

Dependent Variable X_5: length of keel of sternum

Multiple R	.46779
R Square	.21883
Adjusted R Square	.17148
Standard Error	.85089

Table A(7.8):

Analysis of Variance

S.V	DF	Sum of Squares	Mean Squares
Regression	2	6.69299	3.34649
Residual	33	23.89257	.72402
Total		30.58556	

F = 4.62212, Signif F = .0170

Table A(8): MISSING DATA PATTERN (7)

Case#	x_4	x_5	x_5 Estimated
1	18.50000	20.50000	20.94093
2	17.90000	19.60000	20.41826
3	18.40000	20.60000	20.85382
4	17.70000	20.20000	20.24403
5	18.60000	20.30000	21.02804
6	19.00000	20.90000	21.37649
7	18.40000	20.20000	20.85382
8	18.60000	21.20000	21.02804
9	19.10000	21.10000	21.46360
10	18.80000	22.00000	21.20226

Case#	x_4	x_5	x_5 Estimated
11	18.60000	22.00000	21.02804
12	18.60000	20.50000	21.02804
13	19.30000	21.80000	21.63782
14	19.10000	20.00000	21.46360
15	18.10000	19.80000	20.59248
16	18.00000	20.30000	20.50537
17	18.50000	21.60000	20.94093
18	18.20000	20.90000	20.67959
19	18.50000	20.10000	20.94093
20	18.60000	21.90000	21.02804
21	18.00000	21.50000	20.50537
22	18.00000	20.70000	20.50537
23	18.20000	20.60000	20.67959
24	18.80000	21.70000	21.20226
25	17.20000	19.80000	19.80847
26	18.80000	22.50000	21.20226
27	18.50000	20.00000	20.94093
28	19.50000	21.40000	21.81205
29	19.80000	22.70000	22.07338
30	17.30000	19.80000	19.89558
31	18.00000	23.10000	20.50537
32	18.80000	21.30000	21.20226
33	18.50000	21.70000	20.94093
34	18.10000	19.00000	20.59248
35	18.50000	????????	20.94093
36	19.10000	????????	21.46360
37	17.30000	????????	19.89558
38	18.20000	????????	20.67959
39	17.90000	????????	20.41826
40	19.50000	????????	21.81205
41	18.10000	????????	20.59248
42	18.20000	????????	20.67959
43	18.40000	????????	20.85382
44	19.10000	????????	21.46360
45	17.70000	????????	20.24403
46	19.10000	????????	21.46360
47	18.60000	????????	21.02804
48	18.50000	????????	20.94093
49	18.80000	????????	21.20226

Number of cases read = 49
???????? indicates missing values

Tables A(8.1)–A(8.2): RESULTS OF THE REGRESSION OF X_5 on X_4 FOR THE MISSING DATA PATTERN (7)

Table A(8.1):

Multiple R	.50357
R Square	.25358
Adjusted R Square	.23096
Standard Error	.83908

Table A(8.2):
Variables in the Equation

Variable	β	SE(β)	T–Value	Sig T
X_4	.871120	.260168	3.348	.0020
(Constant)	4.825205	4.805532	1.004	.3226

Tables A(8.3)–A(8.4): RESULTS OF THE POST–IMPUTATION REGRESSION OF X_5 on X_4 FOR THE MISSING DATA PATTERN (7)

Table A(8.3):

Multiple R	.57704
R Square	.33298
Adjusted R Square	.31878
Standard Error	.70309

Table A(8.4):
Variables in the Equation

Variable	β	SE(β)	T–Value	Sig T
X_4	.871120	.179843	4.844	.0000
(Constant)	4.825205	3.323113	1.452	.1531