Chapter 1

Introduction

1-1 BACKGROUND TO THE RESEARCH

Researchers who utilise factor analyses frequently encounter the problem of what to do when confronted with missing values. There have been numerous techniques proposed in the literature to deal with the problem of missing values. This thesis examines the following techniques: complete cases only, imputing means, all available cases, an iterative principal component method, singular value decomposition and the EM algorithm. It is not surprising that several papers have been written that compare the performance of these techniques in various multivariate situations (Bello 1993; Little 1988; Frane 1976; Gleason and Staelin 1975; Beale and Little 1975; Timm 1970; Haitovsky 1968; Buck 1960). These papers used either real life or simulated data and Monte Carlo methods to investigate the ability of the various techniques to estimate the missing values themselves, the correlation matrix or the covariance matrix. Detailed discussions of these comparisons are found in chapter 5.
Unfortunately, there are specific problems that have to be considered when investigating the various techniques’ abilities to handle missing values in a factor analysis. There are several parameters in a factor analysis that need to be estimated when there are missing values present. There are also several methods used to estimate factor loadings and to determine the number of factors to include. Researchers also use various methods to rotate original factor loadings. These problems are specific to a factor analysis and have not been addressed in detail in the literature.

Two authors (Finkbeiner 1979; Hendricks Brown 1983) have compared some of the techniques to handle missing values when they are applied to missing values in maximum likelihood factor analysis. Finkbeiner worked with 50 replicates of samples of size 64. Two factors, 6 variables and 2 ‘missingness’ patterns were used. Hendricks Brown examined a single factor model and used asymptotic methods of comparison as opposed to Monte Carlo methods. He was particularly interested in the influence of patterns of missing data on the covariance structure. Both of these authors were more interested in sampling properties than covering as many conditions as possible. This thesis uses Monte Carlo methods and varies the number of cases n, the number of variables p, the average intercorrelation among variables φ and the proportion of missing values k. The performances of six techniques that handle missing values are investigated over a wide range of conditions. This will inform researchers about the conditions under which certain techniques outperform others and when certain techniques should not be used.
1-2 RESEARCH PROBLEM

The main problem addressed in this research is:

*When performing a factor analysis on a data set in which some of the data are missing, which technique for dealing with missing values best estimates the factor loadings and specific variances? If there is not one technique that is always the best, then under what conditions will a particular technique outperform the others?*

This thesis concludes that there is no one technique that is best to use for all of the conditions studied. The EM algorithm is generally the strongest technique to estimate both the factor loadings and specific variances. Unfortunately, the EM algorithm is much more costly in terms of the computing time than the other methods. The conditions when the EM algorithm is outperformed by all available cases and an iterative principal component method are discussed. Some key areas for future research are also identified.

To obtain the above results, some intermediary and associated problems need to be considered. As well as the main research problem, the following research questions are also addressed:

- How are random multivariate data sets generated that are typical of that encountered when a factor analysis is applied?
- Do all of the missing value techniques work for all possible combinations when data are randomly deleted?
- How is the difference between two correlation matrices compared and tested?
• Which technique is the best estimator of the correlation matrix of a data set used for a factor analysis?
• How are two factor analyses compared?
• How do the various techniques compare in terms of their costs in computing time?
• What effect does a change in the correlation matrix have on the factor loadings?

1-3 JUSTIFICATION FOR THE RESEARCH

A factor analysis usually involves a large-scale collection of data with large numbers of variables. Researchers that perform factor analyses frequently encounter the problem of missing data from such a collection. Most of the literature on missing values (outlined in Section 1-1) examines the abilities of various techniques to estimate the missing values themselves or to estimate the covariance or correlation matrix of the complete data set. These general approaches to analyse the techniques cannot address the specific problems inherent when conducting a factor analysis with missing values.

The articles written about missing values in factor analysis (Finkbeiner 1979; Hendricks Brown 1983) have concentrated on the influence of patterns of missing data and sampling properties for maximum likelihood factor analysis. There has not been an extensive simulation study into missing values in factor analysis. This thesis attempts to fill this gap by using Monte Carlo methods to cover as many conditions as possible for principal component factor analysis.
1-4 METHODOLOGY

The six techniques previously proposed in the literature to handle missing values are examined for their ability to handle missing values specifically in factor analyses. Data that are typical of that encountered in a factor analysis are simulated and some of them are randomly removed to represent missing values. The number of cases n, the number of variables p, the average intercorrelation among the variables φ and the proportion of missing values k are all varied to investigate their effect on these techniques.

Graphical comparisons and regression analyses are used to compare the six techniques’ abilities to estimate original factor loadings and specific variances in a factor analysis. The original factor loadings and specific variances obtained from complete data sets are compared to the factor loadings and specific variances obtained when some of the data are randomly deleted using root-mean-square measures. Regression models are also used to predict the error of each technique and investigate what parameters are important in predicting the factor loadings and specific variances. Some of the techniques are iterative and may provide better estimates of parameters than the other techniques; however, this must be weighed against the extra time it takes to calculate their estimates. The costs of each technique are compared using the computing time required to estimate the correlation matrix that is to be input into a factor analysis. Finally, the effect that changes in the correlation matrix have on the factor loadings are analysed theoretically using partial derivatives.
1-5 OUTLINE OF THIS THESIS

Chapter 2 introduces the terminology and basic multivariate concepts that are used throughout this thesis. The factor analysis model and its underlying assumptions that are used in this thesis are described. The importance of examining the effect of missing values in a factor analysis is discussed. Many of the techniques that handle missing values are based on the premise that these values are missing at random. The implications of this issue are explored.

The techniques that are used in this thesis to handle the problem of missing values are discussed in chapter 3. When data are deleted randomly, some of the techniques might produce non positive-definite matrices or they may not work at all. For example, if there is at least one missing value in each case, then using only the complete cases will not be able to produce a correlation matrix. These problems are also addressed in chapter 3. Chapter 4 describes the procedure used to simulate data that are representative of that found in situations where factor analyses are conducted. In particular, multivariate data are generated that have a correlation matrix with a specified average intercorrelation.

The iterative principal component method and the EM algorithm require an initial estimate of the correlation matrix. In Chapter 5, the techniques: complete cases only, all available cases and imputing means are investigated in respect to their ability to estimate the original correlation matrix. Chapter 6 analyses all six techniques regarding their ability to handle missing values in a factor analysis. The techniques are compared as to how well they estimate loadings and specific variances. The
computing time taken by each of the techniques to calculate the correlation matrix is also compared.

A theoretical analysis of the derivatives of the factor loadings with respect to changes in the correlation matrix is conducted in Chapter 7. An example is used to illustrate the complicated form of the derivative expression, the effects on loadings and which loadings are more influenced by changes in the correlation matrix. The conclusions about each of the research questions and the main research problem posed in section 1-2 are set out in chapter 8. The limitations of the research in this thesis are discussed. The implications of the conclusions for researchers are discussed and some areas of future research are also identified.

1-6 MAJOR CONTRIBUTIONS

This thesis makes several contributions to the current knowledge about missing values in factor analyses. Existing techniques to handle missing values and established Monte Carlo methods are used to investigate how missing values influence factor loadings and specific variances. The conditions when a particular technique outperforms the others are established with respect to correlation matrices, factor loadings and specific variances. The effects of missing value techniques on correlation matrices have been previously addressed. This thesis explores these effects for a wider range of intercorrelations.

Changes are introduced to the usual methods of generating multivariate data. To ease interpretability, the correlation structure is based on the average intercorrelation rather
than the usual geometric progression of eigenvalues. The pattern of observed non-positive definite matrices that occur for some missing value techniques is explained using perturbation theory.

When data are randomly deleted, there is a chance that complete cases only and all available cases will not be able to produce an estimate of the correlation or covariance matrix. Formulae are derived to calculate the probability that these techniques will fail. Finally, the effect that small changes in the correlation matrix (possibly as a result of a missing value technique) will have on the factor loadings is algebraically analysed using derivatives.

1-7 CONCLUSION

This chapter has laid the foundation for this thesis. It has introduced the main research problem as well as the associated research questions. The reasons for conducting this research were justified by referring to the relative neglect in the literature of a detailed simulation study. The methodology was briefly described in general terms and the thesis was outlined. Based on these fundamentals, the thesis proceeds with a detailed description of the research conducted.
Chapter 2

Basic Concepts

2-1 INTRODUCTION

Before describing the research, this chapter introduces the terminology and explains the basic multivariate concepts that are used throughout this thesis. The factor analysis model and assumptions used in this thesis are described. Comparisons are made of the various methods proposed in the literature that estimate factor loadings and the number of factors to include in a factor analysis. The importance of examining the effect of missing values in a factor analysis is also discussed. Many of the techniques that handle missing values are based on the premise that all missing values are missing at random. The implications of this premise are explored.

2-2 MULTIVARIATE ANALYSIS

Scientists and researchers who examine more than one variable in their studies have realised the importance of examining the simultaneous relationships among the variables using multivariate techniques. When dealing with several variables, an
uninformed researcher might analyse each variable ignoring the other variables. Unfortunately, this univariate approach for each variable does not provide vital information about a complex intertwined system. A simultaneous examination of all of the variables using multivariate techniques is required to reveal more aspects of the system being analysed.

The greatest impetus for the use of multivariate techniques has been the introduction of high-speed computers with large storage capacities and easy-to-use software packages. These developments have made it easier to implement the intensive calculations involved when using multivariate techniques. The realisation of the need for multivariate analysis and the assistance of computer technology have resulted in multivariate techniques being accepted more widely and used in many scientific fields. As well as university research, multivariate techniques are also being utilised in industry and government. There are few fields of study or research that have failed to incorporate multivariate techniques.

A wide variety of multivariate techniques are available. The technique used depends on the analysis objectives, the type of data and the type of problem. Many of the techniques focus on summarising a large number of variables and data using fewer parameters. Factor analysis is an example of this type of technique and is examined throughout this thesis. This type of analysis attempts to account for the variation in a number of original variables using a smaller number of factors. Some important multivariate concepts are discussed before describing factor analysis in detail.
2-2.1 Multivariate Concepts

Multivariate data contain observations on several variables for a number of individuals, objects or cases. Throughout this thesis, the term cases is used to refer to the things, individuals or objects on which the measurements are taken. These measurements (or data) relate to the attributes of the cases that are being recorded and can be arranged and displayed in matrix form.

The notation $x_{ij}$ refers to the particular measurement of the $j^{th}$ variable on the $i^{th}$ case.

The measurements are arranged in the data matrix $X$ $(n \times p)$ as shown below. The $p$ variables are contained in the columns and the rows represent the $n$ cases.

$$X = \begin{bmatrix}
    x_{11} & x_{12} & \cdots & x_{1p} \\
    x_{21} & x_{22} & \cdots & x_{2p} \\
    \vdots & \vdots & \ddots & \vdots \\
    x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix} = \begin{bmatrix}
    X_1 & X_2 & \cdots & X_p
\end{bmatrix}$$  \hspace{1cm} (2-1)

Not only is it very convenient to put the data into matrix form, it also allows many statistical calculations to be performed quickly and efficiently. A vector can be established that contains the sample means for each of the $p$ variables as follows:

$$\bar{X} = \begin{bmatrix}
    \bar{x}_1 \\
    \bar{x}_2 \\
    \vdots \\
    \bar{x}_p
\end{bmatrix}$$  \hspace{1cm} (2-2)

Each of the elements of the above vector can be calculated using the following equation:

$$\left\{ \bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \quad j=1,2,\ldots,p \right\}$$  \hspace{1cm} (2-3)
The brackets \{\} in the previous equation are used to represent each of the elements in the matrix shown in equation 2-2.

The sample variance for each of the \(p\) variables is given by:

\[
s_{ji}^2 = s_{ji} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2 \quad j = 1, 2, \ldots, p
\]  

(2-4)

It is customary to use the notation \(s_i^2\) for the sample variance. In matrix notation, the variances lie along the main diagonal and it is easier to introduce the notation \(s_{ii}\) to denote the sample variance of the \(i^{th}\) variable.

The symmetric matrix of variances and covariances, shown below, is called the covariance matrix.

\[
S = \begin{bmatrix}
    s_{11} & s_{12} & \cdots & s_{1p} \\
    s_{21} & s_{22} & \cdots & s_{2p} \\
    \vdots & \vdots & & \vdots \\
    s_{p1} & s_{p2} & \cdots & s_{pp}
\end{bmatrix}
\]  

(2-5)

Each element of the sample covariance matrix is calculated using the following equation:

\[
S = \left\{ s_{ik} = \frac{1}{n-1} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)(x_{kj} - \bar{x}_k) \quad i = 1, 2, \ldots, p \quad k = 1, 2, \ldots, p \right\}
\]  

(2-6)
The sample correlation matrix is defined as:

\[ R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \]  
(2-7)

Each element of the sample correlation matrix is given by:

\[ R = \left\{ r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii} S_{kk}}} \right\} \]  
(2-8)

To relate \( R \) to \( S \), a diagonal matrix \( D \) that contains the standard deviations for each of the \( p \) variables is defined as follows:

\[ D = [\text{diag}(S)]^{1/2} = \text{diag}(\sqrt{S_{11}}, \sqrt{S_{22}}, \ldots, \sqrt{S_{pp}}) \]  
(2-9)

The relationship between the matrices \( R \) and \( S \) is shown in the following equations:

\[ R = D^{-1} S D^{-1} \]  
(2-10)
\[ S = D R D \]  
(2-11)

The population mean, covariance and correlation matrices that correspond to equations 2-3, 2-6 and 2-8 respectively are given by the following:

\[ \mu = \{ \mu_i \} = \text{E}(X) \]  
(2-12)
\[ \Sigma = \{ \sigma_{ij} \} = \text{E}(X - \mu)(X - \mu)' \]  
(2-13)
\[ \rho = \begin{cases} \rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii} \sigma_{jj}}} \end{cases} \]  
(2-14)
2-3 FACTOR ANALYSIS

Factor analysis attempts to identify the structure within a set of observed variables. The main purpose of a factor analysis is to develop new variables (unobservable factors) that will enable an easier understanding of the original data. By studying the interrelationships among the original variables, factor analysis attempts to describe the original variables in terms of a smaller number of factors. These factors express what is common among the original variables.

The development of factor analysis was originated by Spearman (1904) from his work with intelligence tests. Typically, these tests contain a large number of questions that depend on verbal ability, mathematical ability, memory, logic etc. He examined the correlations between test scores of various types to determine if intelligence is made up of a single factor (general intelligence) or of several factors such as mathematical ability. He suggested that these intelligence tests contain two factors: one that is common to all tests (general intelligence) and another that is specific to the particular test. The results are still debatable and have later been modified to consist of several common factors as well as a part that is specific to the test.

Psychology and sociology are not the only subjects that utilise factor analysis as a statistical tool. There has been tremendous variety and complexity in the applications of factor analyses. A thorough summary of the range of research is beyond the scope of this thesis; however, a brief overview of the areas in which factor analysis is used has been included.
Before the 1960’s, factor analysis was used in a variety of contexts. Harman (1967) discusses a number of research areas in the United States including: politics, sociology, economics, accident research, taxonomy, biology, medicine and geology. Malinowski and Howery (1980) discuss some uses for factor analysis in chemistry. They discuss how factor analysis is used in component analysis, nuclear magnetic resonance and chromatography. Over the last couple of decades, factor analysis has also been utilised by business and financial analysts. Cooper (1983) provides an elementary overview of factor analysis and describes the usefulness of factor analysis in classifying a large number of ratios from financial statements into four main categories: liquidity, leverage, profitability and turnover ratios. Kline (1994) discusses the modern use of factor analysis in personality testing. Hair, Anderson, Tatham & Black (1998) show how factor analysis is used to reduce the number of items in large questionnaires. They also illustrate how factor analysis quantifies and measures the structure in market research.

2-3.1 Model and Assumptions

Factor analysis begins with the random vector \( X' = [X_1, X_2 \ldots X_p] \)' which represents the observations of \( p \) variables. It is assumed that each of these variables \( X_i \) can be represented by a linear combination of \( m \) unobservable common factors \( F_j \), as well as a specific factor \( \varepsilon_i \). The following orthogonal factor model has been taken from Johnson and Wichern (1998).
Orthogonal Factor Model with m Common Factors:

\[ \mathbf{X} = \mu + \mathbf{LF} + \mathbf{\epsilon} \]

\[ \begin{align*}
X_1 &= \mu_1 + \ell_{11}F_1 + \ell_{12}F_2 + \cdots + \ell_{1m}F_m + \epsilon_1 \\
X_2 &= \mu_2 + \ell_{21}F_1 + \ell_{22}F_2 + \cdots + \ell_{2m}F_m + \epsilon_2 \\
&\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\
X_p &= \mu_p + \ell_{p1}F_1 + \ell_{p2}F_2 + \cdots + \ell_{pm}F_m + \epsilon_p 
\end{align*} \]  \hspace{1cm} (2-15)

where: \( \mathbf{L} \) = matrix of factor loadings

\( \mathbf{F} \) = vector of common factors

\( \mathbf{\epsilon} \) = vector of specific factors

\( \mu_i \) = mean of the \( i^{th} \) variable

\( \epsilon_i \) = \( i^{th} \) specific (or unique) factor

\( F_j \) = \( j^{th} \) common factor (unobservable)

\( \ell_{ij} \) = loading of the \( i^{th} \) variable on the \( j^{th} \) factor

Assumptions used for the Orthogonal Factor Model:

\[ \mathbf{E}(\mathbf{F}) = \mathbf{0} \quad \text{Cov}(\mathbf{F}) = \mathbf{E}[\mathbf{FF}'] = \mathbf{I} \]  \hspace{1cm} (2-16)

\[ \mathbf{E}(\mathbf{\epsilon}) = \mathbf{0} \quad \text{Cov}(\mathbf{\epsilon}) = \mathbf{E}[\mathbf{\epsilon \epsilon}'] = \mathbf{\Psi} = \begin{bmatrix}
\psi_1 & 0 & \cdots & 0 \\
0 & \psi_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \psi_p
\end{bmatrix} \]  \hspace{1cm} (2-17)

\( \mathbf{F} \) and \( \mathbf{\epsilon} \) are also assumed to be independent, so

\[ \text{Cov}(\mathbf{\epsilon}, \mathbf{F}) = \mathbf{E}(\mathbf{\epsilon F}') = \mathbf{0} \]  \hspace{1cm} (2-18)

The assumptions about \( \mathbf{F} \) and \( \mathbf{\epsilon} \) imply a covariance structure for \( \mathbf{X} \). The interested reader is referred to Johnson and Wichern (1998) for the derivation of the following covariance structure:
Covariance Structure for the Orthogonal Factor Model:

\[ \Sigma = \text{Cov}(X) = LL' + \Psi \]  \hspace{1cm} (2-19)

\[ \text{ie} \quad \text{Var}(X_i) = \sigma_{ii} = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \psi_i \]  \hspace{1cm} (2-20)

\[ \text{Cov}(X_i, X_j) = \ell_{i1} \ell_{j1} + \ell_{i2} \ell_{j2} + \ldots + \ell_{im} \ell_{jm} \]  \hspace{1cm} (2-21)

\[ \text{Cov}(X, F) = L \]  \hspace{1cm} (2-22)

\[ \text{ie} \quad \text{Cov}(X_i, F_j) = \ell_{ij} \]  \hspace{1cm} (2-23)

There are two components in the variance given in equation 2-20. The part of the variance \( \sigma_{ii} \) explained by the \( m \) common factors, that is \( \sum_{k=1}^{m} \ell_{ik}^2 \), is called the \textit{communality} of the \( i \)-th variable. This communality is given by:

\[ h_i^2 = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 \]  \hspace{1cm} (2-24)

The portion of the variance due to the specific factor, that is \( \psi_i \), is known as the \textit{specific variance}. The variance from equation 2-20 can be partitioned as follows:

\[ \text{Var}(X_i) = \sigma_{ii} = h_i^2 + \psi_i \]  \hspace{1cm} (2-25)

Equation 2-19 demonstrates that the factors \( LL' \) express all of the covariances (off-diagonal terms of \( \Sigma \)) exactly. A factor analysis is most useful when the number of factors \( m \) is small relative to the number of variables \( p \). Unfortunately, some covariance matrices cannot be factored as \( LL' + \Psi \) when \( m \) is much smaller than \( p \). Lawley and Maxwell (1971) provide two examples of correlation matrices that cannot be factored for a one factor model.
2-3.2 Factor Rotation

If there is a solution when the number of factors \( m \) equals one, then it will usually be unique. However, if \( m > 1 \) and a solution exists, then it is easy to show that this solution is not necessarily unique. Let \( T \) be an orthogonal matrix. Then,

\[
LT(LT)' = LTT'L' = LL'
\] (2-26)

Even though the loadings for \( L \) and \( LT \) are, in general, different from one another, they both generate the same covariance matrix \( \Sigma \). Since multiplication of a matrix by an orthogonal matrix is equivalent to a rotation of axes, this procedure is commonly referred to as rotating the loadings. This particular procedure is known as an orthogonal rotation because the loadings are rotated using an orthogonal matrix. Oblique rotations are also possible, but they are not discussed in this thesis.

Rotating factor loadings to simplify the structure is mathematically feasible. The ideal rotation is to obtain a pattern where each variable has a high loading on a single factor and has small or moderate loadings on the remaining factors. These rotations attempt to make each factor define a separate cluster of intercorrelated variables. There are three main methods for carrying out a rotation; namely, varimax, quartimax and equimax rotations. It is not the purpose of this thesis to investigate in any detail the various orthogonal rotations of loadings. The reader is referred to Tabachnick and Fidell (1996) for further descriptions of these procedures.

Orthogonal rotations can improve the interpretability of the factors but they are also potentially dangerous for the untrained researcher. The non-uniqueness of loadings
means that different researchers could use different rotation methods to get a variety of results using the same data.

2.3.3 Methods of Estimating Loadings

To perform a factor analysis, the factor loadings $\ell_i$ and specific variances $\psi_i$ initially have to be estimated. Various iterative methods were used in the early days of factor analysis to estimate the factor loadings. Two of the more popular methods used today to estimate these loadings are the principal component and maximum likelihood methods.

Both methods have particular strengths and weaknesses. The maximum likelihood method is based on maximum likelihood equations and the assumption that the common factors $F$ and the specific factors $\varepsilon$ are normally distributed. The reader is referred to Lawley and Maxwell (1971) to get a broad coverage of this method. This method must be solved iteratively and sometimes fails to converge (Rencher 1998). It may also produce a communality estimate greater than one when factoring the correlation matrix $R$. The loadings are scale invariant, which means that a change of scale to any variable merely produces proportional changes to the loadings. This method also leads to a significance test for the adequacy of the factor model with $m$ factors.

The principal component method is used in this thesis. Unlike the maximum likelihood method, this method does not make any distributional assumptions and
does not require $\mathbf{R}$ to be non-singular. The loadings of the principal component method are not scale invariant as is the case for the maximum likelihood method.

The principal component method extracts factors such that each factor accounts for the maximum amount of remaining variance. Referring to equation 2-19 and using $\mathbf{S}$ instead of $\Sigma$, an estimate of $\mathbf{L}$ needs to be found such that:

$$\mathbf{S} = \hat{\mathbf{L}} \hat{\mathbf{L}}' + \hat{\Psi} \quad (2-27)$$

Let $\mathbf{S}$ have the eigenvalue-eigenvector pairs $(\lambda_i, \mathbf{e}_i)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Then,

$$\mathbf{S} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2' + \ldots + \lambda_p \mathbf{e}_p \mathbf{e}_p'$$

$$= \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1' \\ \sqrt{\lambda_2} \mathbf{e}_2' \\ \ldots \\ \sqrt{\lambda_p} \mathbf{e}_p' \end{bmatrix} = \hat{\mathbf{L}} \hat{\mathbf{L}}' \quad (2-28)$$

For the situation where the number of factors $m$ equals the number of variables $p$, the above model explains the covariance exactly. Unfortunately, this is not particularly useful because factor analysis aims to have fewer factors than variables. Neglecting the effects of the last $p - m$ eigenvalues and eigenvectors and allowing for specific factors, the approximation becomes:
\[ S = \hat{L}_m \hat{L}_m' + \hat{\Psi} \]

where:

\[ \hat{\Psi}_1 = s_{ii} - \sum_{j=1}^{n} \hat{\xi}_{ij}^2 \quad i = 1,2,\ldots,p \]  
\[ \hat{\lambda}_1 = \hat{\xi}_{i1}^2 + \hat{\xi}_{i2}^2 + \cdots + \hat{\xi}_{im}^2 \]  

In practice, \( R \) is often factored instead of \( S \) in equation 2-27 to obtain the loadings.

Throughout this thesis, \( R \) is used initially to estimate the factor loadings and specific variances. For the factor analysis based on \( R \), the proportion of total sample variance due to the \( j \)th factor is given by:

\[ \frac{\hat{\lambda}_j}{p} \]  

2-3.4 Methods that Estimate the Number of Factors

The number of factors \( m \) to keep in the analysis is an important question that has to be answered by a researcher. There have been several approaches suggested to determine the optimal number of factors to include. The first technique examines the percentage of variance (equation 2-32) explained by each factor. Researchers that use this technique specify a certain percentage of the total variation that needs to be explained by the factors. The factors are then ordered from the factor that explains the most variation (largest eigenvalue) to the factor that explains the least. The process involves accumulating the variation explained by each of the \( p \) factors. Once
the specified percentage of variation is reached, only those $m$ factors are retained in
the study.

Unfortunately, the specified percentage is arbitrary and depends on the type of
research being conducted. For some studies, it may be appropriate to include factors
that account for 95% of the variance. In other less precise studies, the included
factors might only account for 70% of the variance. The nature of this ad hoc
measure makes it very difficult to manage in a simulation study that is meant for a
broad range of researchers and research interests.

Equation 2-25 shows that factors contain both common and specific variance. There
is more specific variance in later factors. A scree plot is used as a visual aid to specify
the number of factors such that the specific variance does not become dominant in the
factor. The eigenvalues are ordered from largest to smallest and are plotted against
the number of factors. The curve that is obtained from such a process is used to
evaluate the number of factors to retain. If the curve shows a steep drop followed by
a linear trend, the number of factors should equal the number of eigenvalues in the
steeper section. Kaiser (1970) mentions the concern that this approach becomes
subjective when there is more than one major bend. The subjective nature of this
technique makes it awkward to work with in a simulation study.

The procedure used in this thesis was first proposed by Guttman (1954) and sets $m$
equal to the number of eigenvalues of $R$ that are greater than one. This eigenvalue
criteria is based on the rationale that any single factor that is to be retained should
account for the variance of at least (on average) a single variable. Each variable is
considered to contribute a value of one to the total eigenvalues. Therefore, the significant factors that should be retained are those that have eigenvalues greater than one. Comparing this technique to the scree plot, Cattell (1966) states that the scree plot results in at least one more factor and as much as three more factors being included in the study compared to the eigenvalue criteria. This eigenvalue criteria seems to work well in practice and is convenient to use in a simulation study.

2-4 MISSING VALUES

A factor analysis typically involves a large-scale collection of data with a large number of variables. Researchers that perform factor analyses frequently encounter the problem that some of the data are missing. The factor analysis procedures discussed previously in this chapter are based on the data matrix having measurements recorded for each variable on all cases.

Unfortunately, a complete data matrix may not be possible to obtain for a variety of reasons. If the experimental units are animals or plants, they may die before all of the variables have been measured. Archaeological studies often have post-mortem damage or erosion that makes it impossible to record some measurements. An industrial experiment may result in missing values as a result of mechanical breakdowns. In surveys, respondents may accidentally or purposely omit answers. All of these examples illustrate that even in well-designed experiments, it is quite common to have missing values for a wide variety of reasons.
There have been numerous proposals in the literature to estimate missing values. The procedures discussed in this thesis (refer to chapter 3) are intended for data matrices whose measurements are missing by a process that is unrelated to any of the relationships among the variables. This is defined as missing completely at random (MCAR) by Little and Rubin (1987). These authors also define a weaker assumption of missing at random (MAR), which means that the missing values are related to the observed data but not to the missing data. All of the procedures discussed to handle missing values in chapter 3 are based on MCAR, unless otherwise stated. There may be many real-life cases of factor analyses where it is not possible to assume MCAR. However, to enable comparisons of several techniques over a wide range of situations and to get some initial results and conclusions, this thesis assumes that the missing values are MCAR.

Researchers must be very careful to examine the underlying processes that result in missing values in their data. It is not always appropriate to use the procedures discussed in chapter 3 unless the data are MCAR. The untrained researcher could seriously bias their results if they assume MCAR and the data are not missing at random. The fact that a variable is missing may indicate that there is something unusual with this variable. This is evident in surveys where respondents may accidentally or deliberately omit answers. If a question deals with a particularly sensitive topic, then it may be found that the nonresponse would be higher for certain groups of respondents than for others. For example, high-income earners may be more unwilling than others to disclose their income. It would be inappropriate in this situation to assume MCAR. However, if the respondent accidentally skipped a question, then it would seem reasonable to assume MCAR.
Readers who are interested in more details about the difference between MAR and MCAR are referred to Rubin (1976). Rubin discussed when it was appropriate to assume that missing values are MCAR. He formulated the weakest conditions on a process that causes missing values such that the process can be ignored. Again, this thesis assumes that missing values are MCAR unless otherwise stated.
Chapter 3

Techniques Investigated

3-1 INTRODUCTION

This chapter discusses a variety of techniques that have been proposed in the literature to handle the problem of missing values. Computer programs have been written by the author using Mathematica for each of the techniques discussed. The program listings, together with detailed explanations of their working, are contained in appendix A. Further computing considerations are discussed at the end this chapter. For example, some of the techniques might not work for particular patterns of randomly deleted data. Non-positive definite matrices are also a problem that is considered for some of the techniques.

3-2 COMPLETE CASES ONLY

As its name suggests, this method simply involves excluding any cases that contain missing values for any of the measured variables. That is, only the cases that have all of the variables present are used with this technique. There is some inconsistency in
the literature in regards to the name used for this technique. It is also known as the case-wise or list-wise deletion method.

The removal of any cases that have missing values results in the original data matrix being reduced in size. However, a correlation matrix can be determined or a factor analysis performed from the remaining cases as measurements are present for every variable. Assuming MCAR, this technique produces a matrix that is a sub-sample of the original data matrix and therefore does not bias any sample estimates. It results in a positive definite covariance matrix as long as the number of cases \( n \) still remaining is greater than the number of variables \( p \).

Complete cases only is very easy to use and is available on most popular statistical software programs (often as the default option). However, the user should be aware of some of the disadvantages in using this method. Deleting an entire case because one of its variables is unrecorded is potentially wasting important information that could be obtained from the remaining observations. A missing value can sometimes represent an important element in a study. For example, the fact that a person did not fill in a question about their level of income may reflect some information that the researcher might want to investigate further.

Removing entire cases can also be drastic if the missing values are spread throughout the data matrix. It could potentially result in the removal of large amounts of valuable information. A poor correlation matrix estimate would be expected if the reduction in the number of cases is large. This technique also has the potential to break down
altogether if every case has at least one variable missing. It seems reasonable that this method should only be applied if the percentage of missing values is small.

Under the assumption of MCAR, it is possible to determine the expected number of cases that will remain after the removal of those that contain missing values (refer to table 3-1). The calculations are a simple application of the binomial distribution. For example, consider the situation when there are 15 variables and 5 percent of the data are missing at random. The probability that any one case has no missing values is given by:

\[
X \sim \text{Bin}(15, 0.05) \\
\Pr(X = 0) = (1 - 0.05)^{15} = 0.463
\]

The other interpretation of the above calculation is that complete cases only will include, on average, only 46.3% of the original cases.

**Table 3-1 Percentage of Cases Remaining Using Complete Cases Only**

<table>
<thead>
<tr>
<th>Percentage Missing</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 %</td>
<td>95.1</td>
<td>90.4</td>
<td>86.0</td>
<td>81.8</td>
<td>77.8</td>
<td>60.5</td>
</tr>
<tr>
<td>2 %</td>
<td>90.4</td>
<td>81.7</td>
<td>73.9</td>
<td>66.8</td>
<td>60.4</td>
<td>36.4</td>
</tr>
<tr>
<td>5 %</td>
<td>77.4</td>
<td>59.9</td>
<td>46.3</td>
<td>35.9</td>
<td>27.7</td>
<td>7.7</td>
</tr>
<tr>
<td>10%</td>
<td>59.1</td>
<td>34.9</td>
<td>20.6</td>
<td>12.2</td>
<td>7.2</td>
<td>0.5</td>
</tr>
<tr>
<td>15%</td>
<td>44.4</td>
<td>19.7</td>
<td>8.7</td>
<td>3.9</td>
<td>1.7</td>
<td>0.0</td>
</tr>
<tr>
<td>20%</td>
<td>32.8</td>
<td>10.7</td>
<td>3.5</td>
<td>1.1</td>
<td>0.4</td>
<td>0.0</td>
</tr>
</tbody>
</table>
It is evident from table 3-1 that more of the original data set is eliminated when there are a larger number of variables. This thesis simulates factor analyses (refer to chapters 5 and 6) with a large number of variables \((p = 15, 25, 50)\). Therefore, it is expected that complete cases only will perform poorly in these situations.

Appendix A-1 contains the Mathematica code for the implementation of the technique complete cases only. As was explained in section 2-2.3, a factor analysis will be performed starting with the correlation matrix \(R\) of the data. Hence, the program shown in appendix A-1 finishes by calculating the correlation matrix for the complete data matrix after the removal of cases with missing values.

The technique of complete cases only may not be able to be performed for certain situations. For example, if there are no cases or only one case remaining after data are randomly deleted, then a correlation matrix cannot be calculated. As random deletion patterns are used in the simulation study in chapters 5 and 6, this situation has to be considered. A program is written in Mathematica (refer to Appendix C) to check if complete cases only will not work for a matrix with random deletions. If complete cases only will not work for a particular deletion pattern, then this is simply ignored and another one is used.

It is possible to determine the probability that complete cases only cannot be used on a matrix with random deletions. The probability depends on the number of variables \(p\), the number of cases \(n\) and the proportion of missing values deleted \(k\) from a data matrix.
A case will not be deleted if there are no deletions in that case. The probability that there are no deletions in a case is given by:

\[ X = \text{number of deletions in a case} \]
\[ X \sim \text{Bin}(p, k) \]
\[ \Pr(X = 0) = (1 - k)^p \]

Complete cases only will fail when there are no cases or only one case remaining. This probability is determined in the next expression:

\[ Y = \text{number of cases remaining} \]
\[ Y \sim \text{Bin}(n, (1 - k)^p) \]
\[ \Pr(Y \leq 1) = \Pr(Y = 0) + \Pr(Y = 1) \]
\[ = \binom{n}{0}[(1 - k)^p]^0\left[1 - (1 - k)^p\right]^n + \binom{n}{1}[(1 - k)^p]^1\left[1 - (1 - k)^p\right]^{n-1} \]
\[ = \left[1 - (1 - k)^p\right]^n + n\left[1 - (1 - k)^p\right]\left[1 - (1 - k)^p\right]^{n-1} \]
\[ = \left[1 - (1 - k)^p\right]^{n-1}\left[1 + (n - 1)(1 - k)^p\right] \]

Hence, the formula to determine the probability of complete cases only failing completely (none or one case remaining after random deletions) is given by:

\[ \left[1 - (1 - k)^p\right]^{n-1}\left[1 + (n - 1)(1 - k)^p\right] \quad (3-1) \]

Table 3-2 shows the probability that complete cases only will fail for a data matrix with randomly deleted data. This table contains the values of n, p and k that are used in the simulation study (refer to chapter 4).
Table 3-2  Percentage of Situations when Complete Cases Only Fails

<table>
<thead>
<tr>
<th>k</th>
<th>p = 15</th>
<th>p = 25</th>
<th>p = 50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 50</td>
<td>n = 100</td>
<td>n = 200</td>
</tr>
<tr>
<td>1 %</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2 %</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>5 %</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>10 %</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The above table highlights some important concerns for the simulation study in chapter 4. It should be noted that when running a simulation of a data matrix that has 50 variables and 10% of the data are deleted, complete cases only fails between 45% and 97% of the time.

3-3 IMPUTING MEANS

Originally developed by Wilks (1932), this technique simply replaces (imputes) any missing values of a variable with the mean of all existing cases of that particular variable. The correlation matrix can then be determined or a factor analysis performed from the full data matrix. This technique does not take into account any of the intercorrelations that might exist among the variables. Mean imputation has the advantage of producing consistent estimates, but it can give an inaccurate representation of the relationships among the variables as the missing values are simply approximated by the means of the variables.

Imputing means is a popular technique for handling missing values; however, the effect that it has on the covariance and correlation matrices must be considered. This
technique produces a positive semi-definite covariance matrix but the variances and covariances are generally underestimated. If there are several missing values for a variable, then mean imputation results in more values being closer to the mean. Therefore, this results in a reduced variance and covariance. It is possible to determine analytically the extent of this reduction. Consider the variance of a variable $X_j$ that has some of its values missing. Using the terms proposed by Little and Rubin (1987), the variance from the $j$ cases which are present is given by:

$$s_{jj}^{(i)} = \frac{\sum (x_{ij} - \bar{x}_{j}^{(i)})^2}{n^{(i)} - 1}$$ (3-2)

where: $n^{(i)}$ refers to the number of cases that are observed and the mean $\bar{x}_{j}^{(i)}$ and summation in 3-2 are calculated over those $n^{(i)}$ cases.

The variance of the variable $X_j$ with all of the missing values replaced by the mean $\bar{x}_{j}^{(i)}$, $s_{jj}$ (means), is given by:

$$s_{jj}^{(\text{means})} = \frac{\sum (x_{ij} - \bar{x}_{j}^{(i)})^2 + \sum (\bar{x}_{j}^{(i)} - \bar{x}_{j}^{(i)})^2}{n - 1}$$

$$= \frac{\sum (x_{ij} - \bar{x}_{j}^{(i)})^2}{n - 1}$$ (3-3)

Comparing the variance of the variable excluding any missing values (equation 3-2) to the variance of that variable containing means for the missing values (equation 3-3), it is easily seen that:

$$s_{jj}^{(\text{means})} = \frac{s_{jj}^{(i)}(n^{(i)} - 1)}{n - 1}$$ (3-4)
The above equation shows that the variance of a variable as a result of imputing the mean for every missing value is underestimated by a factor of \( \frac{(n^{(i)} - 1)}{(n - 1)} \).

Using a similar argument, the covariance between two variables \((j\) and \(k)\) that contain missing values is underestimated by a factor of \( \frac{(n^{(jk)} - 1)}{(n - 1)} \). The superscript \((jk)\) refers to the observations where both variables \(j\) and \(k\) are present.

The covariance obtained by using the means from all available cases of both variables is given by:

\[
\begin{align*}
S_{jk}^{(jk)} &= \frac{\sum \left( x_{ij} - \bar{x}_{j}^{(i)} \right) \left( x_{ik} - \bar{x}_{k}^{(k)} \right)}{(n^{(jk)} - 1)} \\
&= S_{jk}^{(jk)}
\end{align*}
\]  
(3-5)

The comparison of the covariance with means imputed \(S_{jk}\) (means) and the covariance using all available cases \(S_{jk}^{(jk)}\) is shown below:

\[
S_{jk} (\text{means}) = \frac{n^{(jk)} - 1}{n - 1} S_{jk}^{(jk)}
\]  
(3-6)

An obvious approach to compensate for the reductions in variance and covariance is to simply multiply both of these measures (equations 3-4 and 3-6) by the reciprocal of the factor of reduction. That is, these terms are defined as follows:

\[
\begin{align*}
S_{ij} (\text{means}) &= \frac{S_{ij}^{(i)} (n^{(i)} - 1)}{n - 1} \times \frac{n - 1}{n^{(i)} - 1} \\
&= S_{ij}^{(i)} \\
S_{jk} (\text{means}) &= \frac{n^{(jk)} - 1}{n - 1} S_{jk}^{(jk)} \times \frac{n - 1}{n^{(jk)} - 1} \\
&= S_{jk}^{(jk)}
\end{align*}
\]  
(3-7)
This procedure is commonly referred to as adjusting the degrees of freedom. Throughout this thesis, this adjustment is used whenever imputing means is mentioned. The correlation matrix also has terms that are underestimated. The correlation based on the variance from all available observations is given by:

\[
    r_{jk} = \frac{S^{(jk)}_{ji}}{\sqrt{S^{(j)}_{ii}} \sqrt{S^{(k)}_{kk}}}
\]  

(3-9)

Using the variance and covariance factors of reduction (equations 3-4 and 3-6), the reduction factor of the above correlation is given by:

\[
    r_{jk} = \frac{(n^{(jk)} - 1)/(n - 1)}{\sqrt{(n^{(j)} - 1)/(n - 1) \sqrt{(n^{(k)} - 1)/(n - 1)}}} r_{jk} \text{(means)}
\]

\[
    = \frac{(n^{(jk)} - 1)/(n - 1)}{\sqrt{(n^{(j)} - 1)(n^{(k)} - 1)/(n - 1)^2}} r_{jk} \text{(means)}
\]

\[
    = \frac{(n^{(jk)} - 1)}{\sqrt{(n^{(j)} - 1)(n^{(k)} - 1)}} r_{jk} \text{(means)}
\]  

(3-10)

When using imputing means, the correlation between two variables must be multiplied by the reciprocal of the factor of reduction. That is, each off diagonal term of the correlation matrix determined from imputing means \( r_{jk} \text{(means)} \) must be multiplied by \( \sqrt{(n^{(j)} - 1)(n^{(k)} - 1)} / (n^{(jk)} - 1) \).

Imputing means with the adjustment for degrees of freedom yields consistent estimates of the covariances and correlations being estimated. Unfortunately, the estimated correlation and covariance matrices may not be positive semi-definite and
some correlations may lie outside the range (-1,1). The complete Mathematica program for mean imputation is shown in appendix A-2.

3-4 ALL AVAILABLE CASES

The technique of complete cases only (discussed in section 3-2) does not make use of an entire case if an observation is missing for any variable. This is potentially wasteful of important information that has been collected. The technique of all available cases attempts to utilise all of the information collected. There are several methods that can be used to estimate correlations among variables that contain some missing values. Three of the more popular methods are discussed below. The difference among them lies in which values are used to determine the mean that is used for the variance and covariance. All three methods produce consistent estimates of the covariances and correlations being estimated.

3-4.1 Covariance mean from all available cases; Variance mean from all available cases.

This first method uses as much information from the incomplete data matrix as possible to estimate parameters. When computing the variance, this method uses the means of each variable for all of the available cases. The variance of a variable is the same as shown in equation 3-2. The covariance of two variables must use only the data where both observations are present. However, the means used to determine the covariance for this method comes from the means determined above. Hence, the covariance is the same as shown in equation 3-5.
Notice that using imputing means and adjusting the degrees of freedom gives the same results as this procedure. The advantages and disadvantages of this method have been previously addressed in section 3-3.

### 3-4.2 Covariance mean from paired cases; Variance mean from all available cases.

This method uses the means from the available cases when computing the variance as in equation 3-2. However, to determine the covariance between two variables, the mean of each variable is calculated only from those observations where both observations are present for both variables. That is,

$$
s_{jk}^{(jk)} = \frac{\sum_{(jk)} (x_{ij} - \bar{x}_j^{(jk)})(x_{ik} - \bar{x}_k^{(jk)})}{(n_{(jk)}^{(jk)} - 1)} \quad (3-11)
$$

Again, these equations may result in the correlation and covariance matrices not being positive semi-definite and the correlations lying outside the range (-1,1).

### 3-4.3 Covariance mean from paired cases; Variance mean from paired cases.

This third method of dealing with available cases looks at two variables at a time. The mean of each variable is determined from those data where both observations are available for both variables. Hence, the variance will have the same base as the covariance.
\[ s_{ij}^{(jk)} \text{ (avail)} = \frac{\sum (x_{ij} - \bar{x}_j^{(jk)})^2}{n^{(jk)} - 1} \] (3-12)

\[ s_{jk}^{(rk)} \text{ (avail)} = \frac{\sum (x_{ij} - \bar{x}_j^{(jk)})(x_{ij} - \bar{x}_k^{(rk)})}{n^{(jk)} - 1} \] (3-13)

\[ r_{jk}^{(rk)} \text{ (avail)} = \frac{s_{jk}^{(rk)}}{\sqrt{s_{ij}^{(jk)} s_{ik}^{(rk)}}} \] (3-14)

This method has the advantage that all correlations are between (-1,1). However, the correlation matrix may not be positive semi-definite. Throughout the thesis, this procedure is used whenever the technique of all available cases is mentioned. The Mathematica program for this procedure is shown in appendix A-3.

As is the situation for complete cases only, the technique of all available cases does not work for certain patterns of random deletions of the data matrix. If, for any pair of variables, there are no cases or only one case where data are present for both variables, then the correlation between this pair of variables cannot be determined. For these situations, all available cases does not work.

The derivation of the equation to determine the probability that this technique fails is shown below. Consider any pair of variables, each containing \( n \) observations with a probability \( k \) that any of the observations are missing. The probability for a pair of variables to have no cases or only one case where data are present for both variables can be found by substituting \( p = 2 \) into equation 3-1. That is,

\[ \left[ 1 - (1-k)^2 \right]^{n-1} \left[ 1 + (n-1)(1-k)^2 \right] \]
For a pair of variables, this is the expression to determine the probability that their correlation cannot be determined. The above probability to occur for any 2 pairs among p variables is given by:

\[
\binom{p}{2} \left[ 1 - (1 - k)^2 \right]^{n-1} \left[ 1 + (n - 1)(1 - k)^2 \right] \tag{3-15}
\]

For the various levels of n, p and k that are used for the simulation study in chapter 4; available cases will never fail. However, researchers that use other levels for their simulation studies would be advised to keep equation 3-15 in mind.

3.5 ITERATIVE PRINCIPAL COMPONENT METHOD

Dear (1959) devised a method for estimating missing values based on principal components. This method breaks the original data matrix into its known and unknown parts and uses the first principal component to determine the missing values. One of the main advantages of this method is that it does not require any assumptions about the distribution of its variables.

Dear defined an indicator matrix M as follows:

\[
M = \{m_{ij}\} \tag{3-16}
\]

where: 
\[
m_{ij} = 0 \text{ if } x_{ij} \text{ is missing}
\]
\[
m_{ij} = 1 \text{ if } x_{ij} \text{ is observed}
\]
Let $X$ be an $n \times p$ data matrix consisting of $n$ cases with $p$ variables. A standardized matrix $Z$ is formed from the data matrix $X$. Dear used the technique of complete cases only to determine an initial estimation of the correlation matrix $R$. The largest eigenvalue $\lambda_1$ of $R$ is determined along with its associated eigenvector $e_{ik}$ where $k$ refers to the variable. The first principal component for the $i$th case is given by:

$$
\gamma_i = \sum_{k=1}^{p} e_{ik} z_{ik} m_{ik} \tag{3-17}
$$

Any missing values are replaced by the point on the first principal component line as follows:

$$
\hat{z}_{ik} = \begin{cases} 
  z_{ik} & \text{if } m_{ik} = 1 \\
  e_{ik} \gamma_i & \text{if } m_{ik} = 0
\end{cases} \tag{3-18}
$$

The above steps are repeated for all cases that contain missing values. Since this method uses complete cases only to estimate the original correlation matrix $R$, problems can arise when much of the data are missing (as discussed in section 3-2). Bello (1993) used mean imputation to avoid this problem. He also suggested using an iterative principal component method to improve the estimation of the missing values. This iterative version still estimated the first principal component and the missing values as shown in equations 3-17 and 3-18. After the missing values are estimated, the estimated correlation matrix $\hat{R}$ is then recalculated along with its first principal component. The missing values are then re-estimated using equation 3-18. This procedure continues to be iterated until the imputed missing values satisfied a convergence criteria.
Bello's improvements are used in this thesis with the possible exception of the technique used to estimate the original correlation matrix $R$. There are three possible methods that can be used to provide an initial estimate of $R$ (complete cases only, imputing means and available cases). Simulations are performed in chapter 5 to determine which technique provides a better initial estimate of the original correlation matrix $R$ (with no deleted values). The iterative procedure continues until the average difference among the imputed values is less than $10^{-6}$. The Mathematica program for the iterative principal component method is found in appendix A-4.

3-6 SINGULAR VALUE DECOMPOSITION METHOD

A simple and inventive method to impute missing values using the singular value decomposition (SVD) of a matrix was first suggested by Krzanowski (1988). Any data matrix $X$ $(n \times p)$ can be decomposed using the singular value decomposition (Good 1969).

The decomposition has the following form:

$$X = UDV'$$  \hspace{1cm} (3-19)

where $U'U = I_p$ \hspace{1cm} (3-20)

$V'V = VV' = I_p$ \hspace{1cm} (3-21)

$D = \text{diag} (d_1, d_2, \ldots, d_p)$ \hspace{1cm} (3-22)

The $d_i$'s are the square roots of the eigenvalues of $XX'$ and $XX'$. The $i^{\text{th}}$ column of the $(n \times p)$ matrix $U$ is the eigenvector corresponding with the $i^{\text{th}}$ largest eigenvalue.
$d_i^2$ of $XX'$. The $i^{th}$ column of the $(p \times p)$ matrix $V$ is the eigenvector corresponding to the $i^{th}$ largest eigenvalue $d_i^2$ of $XX'$.

Equation 3-19 can also be written in the form:

$$x_{ij} = \sum_{t=1}^{p} u_{it} \cdot d_i, v_{jt} \quad (3-23)$$

Krzanowski (1988) noted that the roles of $n$ and $p$ can be interchanged if $X$ is such that $p > n$. For any missing value $\hat{x}_{ij}$, the following procedure can be used. The $i^{th}$ row is deleted from the data matrix $X$ and a SVD is calculated from the remaining $((n - 1) \times p)$ data matrix. This is denoted by:

$$^iX = \bar{U}\bar{D}\bar{V}' \quad (3-24)$$

The $j^{th}$ variable or column is deleted from $X$ and a SVD is performed on the remaining $(n \times (p - 1))$ data matrix. This is given by:

$$^jX = \bar{U}\bar{D}\bar{V}' \quad (3-25)$$

Referring to equation 3-23, the maximum data estimates for $u_{it}$ and $v_{jt}$ are given by $\bar{u}_i$ and $\bar{v}_j$. Which estimate to use for $d_i$ is a little harder to determine. It can either be estimated by $\bar{d}_i$, $\bar{d}_i$, or some combination of both of them. Krzanowski (1988) used the compromise of the geometric mean $\sqrt[3]{\bar{d}_i} \sqrt[3]{\bar{d}_i}$. 

41
Combining all of the above information, the estimate for a missing value \( \hat{x}_{ij} \) is given by:

\[
\hat{x}_{ij} = \sum_{t=1}^{T} \left( \bar{u}_t \sqrt{d_t} \right) \left( \bar{v}_t \sqrt{d_t} \right)
\]  

(3-26)

Krzanowski (1988) argued that the above equation should be summed over (p-1) using maximum data precepts. When the data contained more than one missing value, Krzanowski established an iterative system. Initially, the values that were present in the data matrix were standardised. To obtain the initial estimates for the missing values, Krzanowski used the technique of imputing means. Each missing value was then re-estimated using equation 3-26 until the imputed values converged to a satisfactory criteria. The average squared difference in imputed values between successive iterations was used as the convergence criteria. When this value was less than \( 10^6 \), the iterative procedure was terminated. The final estimates were then de-standardised to obtain the final estimates of the missing values.

Experimentation with this iterative procedure for the SVD revealed that it tends not to converge when there are a large number of missing values. The imputed values tend to cycle rather than converge. Correspondence with Krzanowski in July 1998, gave some insight into the problem. For each iteration, two SVD's must be performed for each missing value (equations 3-24 and 3-25). The columns of both SVD's for each missing value contain eigenvectors that are susceptible to arbitrary sign changes. This can make the successive imputed values swing around wildly and not converge. In order to overcome this, at the beginning of each iteration all the signs of both \( \bar{u}_n \) and
\( \tilde{v}_{ji} \) are compared with the original SVD obtained from the estimated complete data matrix \( X \). The signs of \( \tilde{u}_i \) and \( \tilde{v}_{ji} \) are all changed if there are more than 80% differences in signs to allow for legitimate changes to the signs.

Equation 3-26 is summed over p-1 singular values. The estimated complete data matrix \( X \) at the beginning of each iteration has p singular values. Likewise, \( \tilde{d} \) has p singular values. However, there are only p-1 singular values in \( \tilde{d} \) when a variable is deleted from \( X \). Krzanowski (1988) assumed that when a variable was deleted from \( X \), the smallest singular value and its corresponding column of \( V \) should be eliminated (to produce p-1 singular values). Further experimentation after the paper was published, led Krzanowski to believe that deleting the smallest singular value was not always the most appropriate one to delete.

From equation 3-26, it is not just a matter of taking the first p-1 elements of \( \tilde{d} \) and the corresponding columns of \( V \). A matching process needs to be established to compare which one of the p singular values from the original matrix \( X \) from each iteration least matches the p-1 singular values from \( \tilde{d} \). Once this is found, the corresponding singular value of \( \tilde{d} \) and the corresponding column of \( V \) are deleted. Krzanowski assumed that the largest singular value had to be included. He compared the remaining singular values using a ratio measure to determine when one singular value was closer than another to the original singular value.

To compare which singular value is closer to another value, this study used an absolute measure of the difference. The following procedure was used to determine
which singular value from $\tilde{d}$ had to be dropped. As the first singular value was automatically kept, the procedure started comparing the second singular values. The absolute difference between the second singular value from $X$ (as it stands at the beginning of each iteration) and $\tilde{d}$ is determined. If the absolute difference of the third singular value from $\tilde{d}$ and the second singular value from $X$ is smaller than the previous value, then this is an indication that the third singular value is closer to the singular value of the original matrix $X$. Hence, the second singular value of $\tilde{d}$ would be removed along with the corresponding column of $V$. This matching process continues until one of the singular values is removed. It should be noted that the last singular value is usually the one removed, but this is certainly not always the case. The Mathematica program for SVD can be found in appendix A-5.

3.7 REgression Method

Buck (1960) developed a method to estimate any missing values by implementing well-known regression techniques. The technique of complete cases only was used to estimate the initial mean matrix $\bar{X}$ and the covariance matrix $S$. Regression equations were established with the missing values as the dependent variables. The regression coefficients were then calculated using the known measurements as the independent variables.

The terminology used to explain this technique is taken from Gleason and Staelin (1975). The original data matrix $X$ is standardised to the matrix $Z$. The correlation
matrix $\mathbf{R}$ is simply given by $\mathbf{R} = \frac{1}{n} \mathbf{Z}'\mathbf{Z}$. $\mathbf{Z}$ can now be partitioned into two submatrices representing the missing $\mathbf{Z}_1$ and the observed data $\mathbf{Z}_2$.

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}$$  \hspace{1cm} (3-27)

where: $\mathbf{Z}_1$ is $n \times q$ and $\mathbf{Z}_2$ is $n \times (p-q)$.

$\mathbf{Z}_1$ can now be reconstructed from $\mathbf{Z}_2$ and the intercorrelations between $\mathbf{Z}_1$ and $\mathbf{Z}_2$. The approach used here is to reconstruct $\mathbf{Z}_1$ using the regressions of $\mathbf{Z}_1$ on $\mathbf{Z}_2$. A partition of the correlation matrix $\mathbf{R}$ can be formed from the above partition of $\mathbf{Z}$.

$$\mathbf{R} = \frac{1}{n} \mathbf{Z}'\mathbf{Z}$$

$$= \frac{1}{n} \begin{bmatrix} \mathbf{Z}_1' \mathbf{Z}_1 & \mathbf{Z}_1' \mathbf{Z}_2 \\ \mathbf{Z}_2' \mathbf{Z}_1 & \mathbf{Z}_2' \mathbf{Z}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{bmatrix}$$  \hspace{1cm} (3-28)

From regression theory, it is well known that $\mathbf{Z}_1$ can be reconstructed from the above information using the following equation. For example, refer to Johnson and Wichern (1998).

$$\hat{\mathbf{Z}}_1 = \mathbf{Z}_2 \mathbf{R}_{22}^{-1} \mathbf{R}_{21}$$  \hspace{1cm} (3-29)

An initial estimate of the correlation matrix $\mathbf{R}$ must be made to be able to use the above method. This can be accomplished using complete cases only, imputing means or all available data. Gleason and Staclin (1975) suggested using imputing means; whereas, Buck (1960) suggested using complete cases only. None of the authors
made any mention about the technique of using all available cases. It has been argued
previously in section 3-2 that complete cases only is not very effective for matrices
with a large number of variables as most of the original data matrix could be
eliminated. Timm (1970) found that using complete cases only was inferior to all
other procedures when $p \geq 5$. A simulation study of the three techniques is used to
investigate which method actually provides a better estimate of the correlation matrix.
This is discussed in more detail in chapter 5.

The regression technique described in this section is not used in this thesis. Iterating
this procedure has considerably improved this technique. The details of this iterative
procedure are discussed in the next section.

3-8 EM ALGORITHM

In most multivariate analyses, it is very common for some of the variables to be
dependent on others. This is particularly the situation with the variables that are
analysed using a factor analysis. It seems reasonable then that the missing values
could be estimated from the observed values by using the dependencies among the
variables. This is the fundamental premise of the EM algorithm.

The EM algorithm essentially consists of two steps – expectation (E-step) and
maximisation (M-step) - that are used iteratively until a convergence criteria is
satisfied. Initially, the EM algorithm estimates the missing values. The parameters
are then estimated and the missing values are then re-estimated assuming that the
estimated parameters are correct. The parameters are then re-estimated and the cycle
continues until they converge. As the EM algorithm simply involves filling in missing values and iterating, the algorithm has been proposed in various contexts throughout the literature. Orchard and Woodbury (1972) appear to be the first to mention the general applications of the idea. They called it the 'missing information principle'. The term EM algorithm was first proposed by Dempster, Laird and Rubin (1977). They examined the full generality of the algorithm and proved that each iteration increases the likelihood and other general results. They also provide a wide range of examples and uses of the algorithm. For an overview of the history, popularity and bibliographical review of the EM algorithm, the interested reader is referred to Meng and Van Dyk (1997).

The approach used in this thesis is to estimate the sufficient statistics rather than the individual missing observations. The sufficient statistics are given by:

\[
T_1 = \sum_{i=1}^{n} X_i = n\bar{X} \tag{3-30}
\]

\[
T_2 = \sum_{i=1}^{n} X_iX'_i = (n-1)S + n\bar{X}\bar{X}' \tag{3-31}
\]

Initial estimates are required for the missing values. Chapter 5 investigates which of the techniques provides the better initial estimates. Once this technique is found, the initial estimates for \(\bar{\mu}\) and \(\bar{\Sigma}\) can be determined.

The E-step of the algorithm proceeds as follows. Each data vector \(X_i\) can be split into the missing \(X_1\) and observed \(X_2\) components. Thus, \(X_i = [X_{i,1} X_{i,2}]\). Given the estimates for \(\bar{\mu}\) and \(\bar{\Sigma}\) and given the observed values \(X_{i,2}\), the mean of the conditional
normal distribution of $X_i$ can be used to estimate the missing components $X_{i1}$. That is,

$$X_{i1} = E(X_{i1} | X_{i2}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{\mu}_1 + \tilde{\Sigma}_{12} \tilde{\Sigma}_{22}^{-1} (\tilde{X}_{i2} - \tilde{\mu}_2)$$

Equation 3-32 estimates the contribution of $X_{i1}$ to the sufficient statistic $T_1$. The estimation of the contribution of $X_{i1}$ to $T_2$ is given by the following two equations:

$$X_{i1}X_{i1}' = E(X_{i1}X_{i1}' | X_{i2}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{\Sigma}_{11} - \tilde{\Sigma}_{12} \tilde{\Sigma}_{22}^{-1} \tilde{\Sigma}_{21} + \tilde{X}_{i1}\tilde{X}_{i1}' \quad (3-33)$$

$$X_{i1}X_{i2}' = E(X_{i1}X_{i2}' | X_{i2}; \tilde{\mu}, \tilde{\Sigma}) = \tilde{X}_{i1}\tilde{X}_{i2}' \quad (3-34)$$

The estimators from equations 3-33 and 3-34 are determined for all $X_i$ that contain missing values. These results are combined with the sample data to yield $T_1$ and $T_2$. The $M$-step of the algorithm determines the maximum likelihood estimates assuming that the data are normally distributed. That is,

$$\tilde{\mu} = \frac{T_1}{n} \quad (3-35)$$

$$\tilde{\Sigma} = \frac{T_2}{n} - \tilde{\mu}\tilde{\mu}' \quad (3-36)$$

Once the new estimates for $\tilde{\mu}$ and $\tilde{\Sigma}$ are determined, the algorithm revises the estimates by iterating through the $E$ and $M$-steps. The process continues in this way until the estimates for $\tilde{\mu}$ and $\tilde{\Sigma}$ converge. The EM algorithm has two convergence criteria for the parameters $\tilde{\mu}$ and $\tilde{\Sigma}$. The process used in this thesis is terminated if the average squared difference in the elements of $\tilde{\mu}$ is less than $10^{-4}$ and the average
squared difference in the elements of \( \bar{\Sigma} \) is less than \( 10^{-2} \). The Mathematica program for the EM algorithm can be found in appendix A-6.

3-9 OTHER COMPUTING CONSIDERATIONS

In the presence of missing values, some of the techniques discussed in this chapter may produce non-positive definite (NPD) correlation matrices (i.e. correlation matrices with one or more negative eigenvalues). The techniques of all available cases and imputing means both have the potential to produce these NPD matrices as discussed previously. Unfortunately, a factor analysis cannot be conducted if the correlation matrix is NPD.

Several procedures have been proposed in the literature to modify an NPD matrix to achieve a positive semi-definite estimate of the matrix. If \( T \) is a \( p \times p \) matrix, then it is well known that \( T'T \) is positive semi-definite. Schwertman and Allen (1979) smooth the covariance matrix \( S \) by finding \( T \) such that the weighted square difference between each element of \( S \) and \( T'T \) is minimized. The expression to be minimized is:

\[
L = \sum_{i<j}^{n} W_{ij} (S_{ij} - \sum_{l=1}^{n} t_{il} t_{lj})^2
\]

(3-37)

where: \( W \) refers to a matrix of weights.
The situation where each squared difference is given equal weight is shown below:

\[ W_{ij} = \begin{cases} 
1, & i = j \\
2, & i < j 
\end{cases} \]  
(3-38)

Using equal weights, equation 3-37 becomes:

\[ L = \sum_{i=1}^{p} \sum_{j=1}^{p} (S_{ij} - \sum_{l=1}^{p} t_{il} t_{lj})^2 = \text{tr}(S - T'T)^2 \]  
(3-39)

Schwertman and Allen (1973), Rao (1973), Kristof (1970) and Golub (1968) all prove the following theorem for the weights used in equation 3-39.

**THEOREM 1** If \( S \) is a symmetric real matrix and \( \text{tr}(S - T'T)^2 \) is minimised with respect to a real matrix \( T \), then

\[ T'T = \sum_{\lambda_i \geq 0} \lambda_i e_i e_i' \]

where \( \lambda_i, e_i \) are the positive eigenvalues and corresponding eigenvectors of \( S \).

Hence, the estimated covariance matrix \( \hat{S} \) is given by:

\[ \hat{S} = T'T = \sum_{\lambda_i \geq 0} \lambda_i e_i e_i' \]  
(3-40)

Huseby, Schwertman and Allen (1980) provide an algorithm to smooth \( S \) using equation 3-40. Intuitively, the elements of an indefinite covariance matrix \( S \) that are estimated from the least amount of information (that is, more missing values) should be most altered. Using the degrees of freedom to estimate \( S_{ij} \) as weights in equation 3-37, the \( S_{ij} \) with the most information will change the least. In the literature, there does not appear to be an in-depth study about the performance of the two weighting
patterns (that is, equal weights and degrees of freedom for weights). The simpler method shown in equation 3-40 is used in this thesis.

The above procedures refer to smoothing the covariance matrix $S$. A slight adjustment is made to this procedure to smooth a correlation matrix $R$. The method used is based on Frane (1978). Define the correlation matrix $R$, the matrix of eigenvectors corresponding to the positive eigenvalues $V$ and the diagonal matrix of positive eigenvalues $E$. For the eigenvectors and eigenvalues of $R$, equation 3-40 is written in matrix form below:

$$
C = V^T E V
$$

(3-41)

$C$ has the form of a positive semi-definite symmetric matrix. Let $F$ be the diagonal matrix of square roots of the diagonal elements of $C$. The correlation matrix $R$ is smoothed using the following equation.

$$
\hat{R} = F^{-1} V^T E V F^{-1}
$$

(3-42)

Gabriel and Zamir (1979) examined iterative techniques for obtaining reduced rank approximation of matrices with weighted least squares. They investigated techniques involving criss-cross regressions with an application to the fitting of missing values. These techniques could be worth further investigation (refer to further research in section 8-6).

Dong (1985) discussed the use of adding a small ridge constant to the diagonal of $R$ for maximum likelihood factor analysis. This is equivalent to increasing the
uniqueness and forces $R$ to be positive definite. Dong did not recommend this procedure when an NPD matrix was due to missing data, as it will not necessarily make the solution substantively appropriate. Adding a small ridge constant may still be worth further investigation (refer to further research in section 8-6).
Chapter 4

Data Simulation

4-1 INTRODUCTION

Multivariate data need to be generated or collected to simulate the effects that missing values have on factor analyses. This chapter describes the procedure that is used to generate multivariate data that have a correlation matrix with a specified average intercorrelation $\phi$. After examining previous methods proposed in the literature for generating multivariate data, the methodology used in this thesis is described. This method combines and extends the models proposed by Bello (1993), Bryce and Maynes (1979) and Bendel and Mickey (1978).

4-2 PREVIOUS ATTEMPTS TO GENERATE RANDOM MULTIVARIATE DATA

In the literature, there have been several approaches used to produce multivariate data to examine the effects of missing data on correlation or covariance matrices. Buck (1960) and Krzanowski (1988) simply selected data from the literature to analyse the ability of their techniques to handle missing values. Timm (1970) selected several
correlation matrices from the literature and then generated sample data matrices using the Kaiser and Dickman (1962) method. This method utilises principal components to generate random sample data given a population correlation matrix. To enable easier comparisons among the large number of simulations used in this thesis, data need to be generated with strict parameters. The parameters needed for all of the simulations may not have been found in the matrices obtained from the literature, publications and institutions. For this reason, the above methods are not appropriate to use in this thesis.

Another approach that is used for generating multivariate data involves specifying all of the correlations within the correlation matrix. This means that $p(p-1)/2$ correlations have to be defined. Apart from the difficulty in specifying so many correlations, it is also very hard to categorise and group the data so that comparisons can be made. Herzberg (1969) used simulations with multiple regression with several dependent variables. The covariance structures were specified using two sets of parameters - the eigenvalues of the covariance matrix and parameters involving the dependence of the independent variables with the principal components. Little (1988), Beale and Little (1975) and Haitovsky (1968) constructed regression data with correlated independent variables specifying the means, variances and correlations. These procedures are useful for regression studies, but are not appropriate for the factor analyses used in this thesis. Tucker, Koopman and Linn (1969) generated correlation matrices by using the factor loadings as some of the input parameters. Unfortunately, these input parameters are too awkward and complex to use.
Chalmers (1975) and Bendel and Mickey (1978) have provided similar ways for generating correlation matrices with a given set of eigenvalues. The eigenvalues were represented by a single parameter. Costanza and Afifi (1978) also used Bendel and Mickey's method to generate data for discriminant analysis. Bryce and Maynes (1979) provided a method for generating sets of variables with a given eigenvalue structure. This method is combined with Bendel and Mickey's (1978) results to produce random data with a correlation matrix that have a specified eigenvalue structure. Bello (1993) used the methods described by Bendel and Mickey and Bryce and Maynes to generate data with covariance matrices containing a given eigenvalue structure. This model is extended by specifying the average intercorrelation among the variables.

4-3 OUTLINE OF METHOD

Several steps are required to generate random data that produce a correlation matrix with a specified average intercorrelation. Initially, data are generated that have a covariance matrix with specified eigenvalues. A series of orthogonal matrix transformations are applied to the covariance matrix to produce a correlation matrix with the given eigenvalues. A measure of the interdependencies among variables $v$ is defined as a single parameter to represent the eigenvalue structure. This measure of the interdependency among variables is difficult to work with, so the more familiar root-mean-square off diagonal correlation $\varphi$ is used instead. The relationship between $v$ and $\varphi$ is shown so that the eigenvalue structure can be specified using the parameter $\varphi$ rather than $v$. 
4-4 GENERATING DATA BY SPECIFYING ALL OF THE EIGENVALUES OF THE CORRELATION MATRIX

Initially random data are generated that have a covariance matrix $S$ with specified eigenvalues. Some notation is needed. Let $D$ be a diagonal matrix with the desired eigenvalues on its diagonal. That is,

$$D = \text{diag}[\lambda_1, \lambda_2, \ldots, \lambda_p]$$  \hspace{1cm} (4-1)

Let $S$ be defined as a random covariance matrix that has the above eigenvalues. It is well known that pre- and post-multiplying $D$ by an orthogonal matrix produces a symmetric matrix containing the eigenvalues of $D$. If a random orthonormal matrix $Q$ can be generated, then $S$ can be defined as:

$$S = QDQ'$$  \hspace{1cm} (4-2)

The method of Bryce and Maynes (1979) is used to generate random data $X$ that produce the covariance matrix $S$. Their procedure is discussed briefly. Any random $n \times p$ array of numbers can be generated as long as it has rank $p$. This matrix is denoted as $X^*$. Let $Z$ represent the results of the Gram-Schmidt orthogonalisation process to the columns of $X^*$. As $X^*$ is a random matrix, $Z$ must be a random orthonormal matrix.

The above steps to generate $Z$ are repeated to form another random orthonormal matrix $Q$ (as used in equation 4-2). The objective from here is to determine the data matrix $X$ for which the covariance matrix $X'X$ has the given set of eigenvalues.
That is,

\[ \mathbf{X}'\mathbf{X} = \mathbf{QDQ}' = \mathbf{S} \quad (4-3) \]

The Cholesky decomposition is now performed on \( \mathbf{QDQ}' \). The result of the decomposition is an upper triangular matrix \( \mathbf{T} \) where:

\[ \mathbf{T}'\mathbf{T} = \mathbf{QDQ}' \quad (4-4) \]

Finally, the data matrix \( \mathbf{X} \) is generated using the following equation:

\[ \mathbf{X} = \mathbf{ZT} \quad (4-5) \]

Now, the data matrix \( \mathbf{X} \) has the form such that:

\[
\begin{align*}
\mathbf{X}'\mathbf{X} &= (\mathbf{ZT})'(\mathbf{ZT}) \\
&= \mathbf{T}'\mathbf{Z}'\mathbf{ZT} \\
&= \mathbf{T}'\mathbf{T} \\
&= \mathbf{QDQ}' \\
\end{align*}
\]

Hence, the covariance matrix \( \mathbf{X}'\mathbf{X} \) has the eigenvalues contained in \( \mathbf{D} \) as desired. In order to generate a sample covariance matrix, \( \mathbf{X} \) should have the form such that:

\[
\frac{\mathbf{(X - \overline{X})'(X - \overline{X})}}{n - 1} = \mathbf{QDQ}'
\]

\[ (4-7) \]

\( \mathbf{X} \) is purposely generated with \( \overline{\mathbf{X}} = \mathbf{0} \) in the simulation study in chapters 5 and 6.

Therefore, equation 4-7 can be written as:

\[
\frac{n\mathbf{X}'\mathbf{X}}{n - 1} = \mathbf{QDQ}'
\]

i.e. \( \mathbf{X}'\mathbf{X} = (n - 1)\mathbf{QDQ}' \)

\[ (4-8) \]
To make this change in the above procedure described in this section, all that needs to be done is to apply the Cholesky decomposition on \((n - 1) QDQ'\) instead of \(QDQ'\) from equation 4-4. The result of this change means that \(X\) contains random data such that it produces a sample covariance matrix \(S\) that has the desired eigenvalue structure.

4-5 GENERATING DATA WITH A CORRELATION MATRIX OF SPECIFIED EIGENVALUE STRUCTURE

It still remains to generate a correlation matrix \(R\) that will have the appropriate eigenvalue structure established in section 4-4. The covariance matrix \(S\), produced in the previous section, cannot simply be standardised to produce the correlation matrix \(R\). Unfortuantly, standardisation does not preserve the eigenvalues as is desired. Bendel and Mickey (1978) suggested choosing another orthogonal matrix \(P\) such that:

\[
R = PQDQ'P' = PSP'
\]  (4-9)

In actual fact, \(P\) is a series of orthogonal matrices designed to transform \(S\) such that all of its diagonal elements equal one and it keeps the same eigenvalue structure. Each \(P_i\) is constructed from an identity matrix with the same dimensions as \(D\). \(\cos \theta\) replaces the two ones in the positions associated with the largest and smallest values of \(D\). One of the zeros in \(P\) is replaced by \(\sin \theta\). Its position is found using the row of \(D\) containing the largest value and the column with the smallest value. A zero in \(P\) corresponding with the row of the diagonal of \(D\) that contains the smallest value and the column that has the largest value is replaced with \((\sin \theta)\).
For example, if

$$D = \begin{bmatrix} 1.5 & 0 & 0 & 0 \\ 0 & 1.2 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$$

then

$$P = \begin{bmatrix} \cos \theta & 0 & 0 & \sin \theta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sin \theta & 0 & 0 & \cos \theta \end{bmatrix}$$

The correlation matrix $R$ is then found as follows:

$$S_1 = P_1SP_1'$$

$$S_2 = P_2S_1P_2'$$

$$\vdots$$

$$S_{p-1} = P_{p-1}S_{p-2}P_{p-1}' = R$$

(4-10)

The angle of rotation $\theta$ is chosen at each step above so that one diagonal element of $S_i$ is set to equal one. This procedure is repeated such that at $S_{p-1} = R$ all diagonal elements are equal to one. As each $P_i$ is an orthogonal matrix, $R$ will keep the desired eigenvalue structure, that is, the eigenvalue structure established in $D$.

It is now possible to generate a data matrix $X$ that produces a correlation matrix $R$ that has the specified eigenvalue structure. Now $X$ is simply formed using $X = ZT$ from equation 4-5. Remember that $Z$ is a random orthonormal matrix and $T$ is now the upper triangular matrix obtained from the Cholesky decomposition on $(n-1)R$.

That is,

$$T'T = (n - 1)R$$
Now $X$ has the form such that,

$$\frac{XX}{n-1} = \frac{1}{n-1} [T'Z'ZT]$$

$$= \frac{1}{n-1} [T'T]$$

$$= \frac{1}{n-1} [(n-1)R]$$

$$= R \quad (4-11)$$

Will there always be a real solution for the angle of rotation $\theta$ in equation 4-10 such that one of the diagonal elements can be set equal to one? Bendel and Mickey (1978) attempted to answer this question but omitted some of the details. Let the original covariance matrix $S$ be of any $p \times p$ dimension. Out of this matrix, consider only the elements of the matrix that are to be operated on:

$$\begin{bmatrix}
\alpha & \beta \\
\beta & \gamma
\end{bmatrix}$$

$\alpha$ refers to the largest diagonal element of $S$. Similarly, $\gamma$ is the smallest diagonal element. $\beta$ is the corresponding row and column element of $\alpha$ and $\gamma$. Because all of the diagonal elements of $S$ must add up to the number of variables $p$, it follows that $\alpha > 1$ and $\gamma < 1$.

In the steps above (refer to equation 4-10), the matrix to be operated on was pre- and post-multiplied by the rotational matrix as shown on the following page.
\[ P_i P'_i = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \]

\[ = \begin{bmatrix} \alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta & \beta (\cos^2 \theta - \sin^2 \theta) + (\gamma - \alpha) (\sin \theta \cos \theta) \\ \beta (\cos^2 \theta - \sin^2 \theta) + (\gamma - \alpha) (\sin \theta \cos \theta) & \gamma \cos^2 \theta + \alpha \sin^2 \theta - 2\beta \sin \theta \cos \theta \end{bmatrix} \]

(4-12)

From the above matrix, the element in the first row and column corresponds to the highest element of \( S \) and is set equal to one. To solve \( \alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta = 1 \), it is important to examine the function:

\[ f(\theta) = \alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta \]

(4-13)

where: \( f(0) = f(\pi) = \alpha \)

\[ f(\pi/2) = \gamma \]

It is evident that the above function is continuous with period \( \pi \). Referring to equation 4-13, a solution to \( f(\theta) = 1 \) must always exist as \( \alpha > 1 \) and \( \gamma < 1 \). In fact, there are two solutions. One lies between 0 and \( \pi/2 \) and the other lies between \( \pi/2 \) and \( \pi \). To get a representation of what is happening, one plot of equation 4-13 is done below using Mathematica. This plot was done for the values: \( \alpha = 1.5, \beta = 1 \) and \( \gamma = 0.5 \) as \( \theta \) varies from 0 to \( 2\pi \). The Mathematica commands to do this plot follow:

\( \alpha = 1.5; \)
\( \beta = 1; \)
\( \gamma = 0.5; \)
\( \text{Plot}[\alpha \cos[x]^2 + \gamma \sin[x]^2 + 2\beta \sin[x] \cos[x], \{x, 0, 2\pi\}] \)
It is a long process to solve $\alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta = 1$ algebraically, so the details and solution are shown in appendix E-1.

In summary, section 4-4 showed how to generate random data $X$, such that the covariance matrix would contain specified eigenvalues. It is desired that $X$ have a correlation matrix with these specified eigenvalues. Simply standardising the covariance matrix does not preserve the eigenvalues. Orthogonal transformations are used in this section to obtain $R$ that keeps the desired eigenvalues.

The steps to generate random data that have a correlation matrix with specified eigenvalues follows:

1. Specify the desired eigenvalues.
2. Generate the random orthonormal matrix $Z$.
3. Generate the random orthonormal matrix $Q$.
4. Perform orthogonal transformations on $QDQ'$ to obtain $R$.
5. Calculate $T$ by performing the Cholesky decomposition on $(n-1)R$.
6. Obtain the random data matrix $X$ by using $X = ZT$. 
4-6 GENERATING EIGENVALUES BY SPECIFYING THE NUMBER OF VARIABLES AND AN INDEX OF DEGREE OF INTERDEPENDENCE

The previous sections discussed the techniques necessary to produce random data with a correlation matrix that contains specified eigenvalues. It is not a trivial task to specify all of the eigenvalues to represent real life data. It involves specifying $p$ eigenvalues that satisfy $\sum \lambda_i = p$. For this study, it would be more convenient to specify the eigenvalues by one or two parameters. Bendel and Mickey (1978) expressed the eigenvalues using two parameters - the number of variables $p$ and an index of the degree of interdependence among the variables $v$. They supposed the ordered eigenvalues followed a geometric progression where:

$$\lambda_i = bv^{i-1} \quad (4-14)$$

i.e. 

$$\lambda_1 = b$$

$$\lambda_2 = bv$$

$$\lambda_3 = bv^2$$

$$\vdots$$

$$\lambda_p = bv^{p-1}$$

where $0 < v \leq 1$

By enforcing the condition $\sum \lambda_i = p$, the value of $b$ in the above equation becomes:

$$b = \begin{cases} 
\frac{p(1-v)}{(1 - v^p)}, & 0 < v < 1 \\
1, & v = 1 
\end{cases} \quad (4-15)$$

This model arose from the observed behaviour of the eigenvalues in both principal component and factor analyses of correlation matrices. In a factor analysis, there are
only a few large eigenvalues when there is a high intercorrelation among the variables. This is represented by a value of \( v \) close to zero and produces a steep decline in the values of the eigenvalues. Refer to equations 4-14 and 4-15. On the other hand, if \( v \) equals one, then all of the eigenvalues also equal one. This is representative of independent variables. Hence, \( v \) is simply a measure of the degree of interdependence among the variables. Values of \( v \) close to zero represent high intercorrelation and values of \( v \) close to one represent essentially independent variables.

Bendel and Mickey (1978) pointed out that this model was not particularly satisfactory when \( p \) was large i.e. \( p \geq 10 \). The example they used was for \( p = 15 \) and \( v = 0.4 \). Using equations 4-14 and 4-15, the eigenvalues for this situation are:

\[
9, 3.6, 1.44, 0.576, 0.230, 0.092, 0.037, 0.015, 0.0059, 0.0024, 0.00094, 0.0004, 0.0002, 0.00006, 0.00002.
\]

The correlation matrix that results from these eigenvalues is nearly singular due to some of the small eigenvalues. The number of variables is greater than 10 for the factor analyses studied in this thesis. A lower bound \( \delta = 0.1 \) is introduced to avoid singularity problems; however, these problems can still result (refer to section 5-3.5). Using this lower bound, the new model becomes:

\[
\lambda_i = bv^{i-1} + \delta \quad (4-16)
\]
Again, the condition \( \sum \lambda_i = p \) results in the following expression for \( b \):

\[
b = \begin{cases} 
\frac{p(1-\delta)(1-v)}{(1-v^p)}, & 0 < v < 1 \\
1-\delta, & v = 1 
\end{cases}
\]  

(4-17)

Using the previous example with \( p = 15, v = 0.4 \) and \( \delta = 0.1 \), the eigenvalues become:

8.20, 3.34, 1.40, 0.62, 0.31, 0.18, 0.13, 0.11, 0.11, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10.

Bendel (1973) stated that the correlation matrices for \( 10 \leq p \leq 30 \) were realistic examples of correlation matrices found in practice.

### 4.7 Generating Eigenvalues by Specifying the Number of Variables and Average Intercorrelation

The arbitrary choice of \( v \) to determine the level of interdependence among the variables makes the above model a little awkward. It is difficult for a researcher to understand what the various values of \( v \) between 0 and 1 actually represent other than some sort of index of the interdependencies among the variables. It is also impossible to convert the eigenvalues of a correlation matrix obtained from real life data into a single value for \( v \). This makes it difficult to compare data matrices obtained using \( v \) and those from real life. Therefore, it would be hard to extrapolate any conclusions from this simulation study to what occurs in real life. To enable easier comparisons between the results from this study and real life situations, another more interpretable measure of the interdependence among variables is used.
Gleason and Staelin (1975) used the root-mean-square, off diagonal correlation \( \varphi \) to measure the average level of correlation among the variables. This is given by:

\[
\varphi = \sqrt{\frac{\| \mathbf{R} \|^2 - p}{p(p - 1)}}
\]  

(4-18)

where: \( \| \mathbf{R} \|^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} R_{ij}^2 \)  

(4-19)

A value of \( \varphi \) close to zero represents no intercorrelation among the variables whilst a value of \( \varphi \) close to one represents high intercorrelation. The measure \( \varphi \) of the interdependencies among variables is much more interpretable and meaningful than the arbitrary choice of the measure \( \upsilon \). This thesis extends the models used previously in the literature (refer to section 4-6) by specifying the parameter \( \varphi \) instead of \( \upsilon \). A relationship needs to be established between \( \varphi \) and \( \upsilon \), so that the value of \( \upsilon \) that corresponds to \( \varphi \) can be placed into the model (equations 4-16 and 4-17) to determine the eigenvalues.

### 4.7.1 Determining the Relationship Between the Average Intercorrelation and Index of Interdependence

It has been established that:

\[
\lambda_i = b \upsilon^{i-1} + \delta
\]

\[
b = \begin{cases} 
\frac{p(1-\delta)(1-\upsilon)}{(1-\upsilon^p)}, & 0 < \upsilon < 1 \\
1-\delta, & \upsilon = 1
\end{cases}
\]
It is well known that \( \|R\|^2 = \sum_{i=1}^{p} \lambda_i^2 \). Hence, from equation 4-18:

\[
\varphi = \sqrt{\frac{\sum \lambda_i^2 - p}{p(p-1)}} \\
\phi^2 = \frac{\sum \lambda_i^2 - p}{p(p-1)} \\
\sum \lambda_i^2 = \phi^2 p(p-1) + p
\] (4-20)

Using equation 4-16, the ordered squared eigenvalues are given by:

\[
\lambda_1^2 = (b + \delta)^2 = b^2 + \delta^2 + 2b\delta \\
\lambda_2^2 = (bv + \delta)^2 = b^2v^2 + \delta^2 + 2b\delta v \\
\lambda_3^2 = (bv^2 + \delta)^2 = b^2v^4 + \delta^2 + 2b\delta v^2 \\
\vdots \\
\lambda_p^2 = (bv^{p-1} + \delta)^2 = b^2v^{2(p-1)} + \delta^2 + 2b\delta v^{p-1}
\] (4-21)

From the above information and using geometric series, it is possible to derive a second expression for \( \sum \lambda_i^2 \).

\[
\sum \lambda_i^2 = \frac{b^2(1 - v^{2p})}{1 - v^2} + \frac{2b\delta(1 - v^p)}{1 - v} + p\delta^2
\] (4-22)

It is now possible to derive the relationship between \( \varphi \) and \( v \). Equating equations 4-20 and 4-22 gives:

\[
\frac{b^2(1 - v^{2p})}{1 - v^2} + \frac{2b\delta(1 - v^p)}{1 - v} + p\delta^2 = p + \varphi^2 p(p-1)
\] (4-23)
Substituting in \( b = \frac{p(1 - \delta)(1 - v)}{1 - v^p} \) (equation 4-17) into the above expression gives:

\[
\frac{p^2(1 - \delta)^2(1 - v)^2(1 - v^2p)}{(1 - v^p)^2(1 - v^2)} + \frac{2p(1 - \delta)(1 - v)\delta(1 - v^p)}{(1 - v^p)(1 - v)} + p\delta^2 = p + \varphi^2p(p - 1) \tag{4-24}
\]

Simplifying the above expression gives:

\[
\frac{p^2(1 - \delta)^2(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} + 2p\delta(1 - \delta) = p[1 + \varphi^2(p - 1) - \delta^2]
\]

\[
\frac{p(1 - \delta)^2(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} + 2\delta(1 - \delta) = 1 + \varphi^2(p - 1) - \delta^2
\]

\[
\frac{p(1 - \delta)^2(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} = 1 - \delta^2 - 2\delta(1 - \delta) + \varphi^2(p - 1)
\]

\[
\frac{(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} = \frac{1 - \delta^2 - 2\delta + 2\delta^2 + \varphi^2(p - 1)}{p(1 - \delta)^2}
\]

\[
\frac{(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} = \frac{(1 - \delta)^2 + \varphi^2(p - 1)}{p(1 - \delta)^2} \tag{4-25}
\]

Re-arranging equation 4-25 in terms of \( \varphi \):

\[
\varphi^2 = \frac{(1 - v)(1 + v^p)p(1 - \delta)^2}{(1 - v^p)(1 + v)(p - 1)} - \frac{(1 - \delta)^2}{p - 1}
\]

\[
\varphi = \sqrt{\frac{(1 - \delta)^2}{p - 1} \left[ \frac{p(1 - v)(1 + v^p)}{(1 - v^p)(1 + v)} - 1 \right]} \tag{4-26}
\]

The three-dimensional plot for the above equation is done using Mathematica. The figures below (figures 4-2 to 4-4) show the relationship between \( \varphi \) and \( v \) with \( \delta = 0.1 \).

They are plotted for a number of variables between 2 and 75 as this is what is used in the simulated factor analyses (chapters 5 and 6).
The Mathematica program to plot equation 4-26 follows:

\[ \delta = 0.1; \]
\[ f[v_] = Sqrt[((1-\delta)^2/(p-1)) (p (1-v) (1+v p)/(1-v p) (1+v) - 1)); \]
\[ Plot3D[f[v], (v, 0.001, 0.999), (p, 2, 75), AxesLabel -> {"v", "p", "\(\varphi\)"}] \]

Figure 4-2 Three Dimensional Plot Showing the Relationship Between \(\varphi\) and \(v\)

This plot was rotated to give different perspectives of the same surface.

\[ Show[%, ViewPoint -> (0, -2, 0)] \]

Figure 4-3 Rotated Three Dimensional Plot of Figure 4-2
Show[\%\%, ViewPoint -> {2, -2, 0}]

![Graph](image)

**Figure 4-4 Rotated Three Dimensional Plot of Figure 4-2**

It is still necessary to rearrange equation 4-26 to write \( v \) as a function of \( \varphi \), that is \( v = f(\varphi) \). The expression for \( v \) could then be substituted into the proposed model (equations 4-16 and 4-17). Unfortunately, writing \( v \) as a function of \( \varphi \) is a very difficult task. Yet, it is still preferable to stipulate the parameter \( \varphi \) rather than \( v \).

Mathematica can be used to determine the value of \( v \) from equation 4-26 that corresponds to an input value of \( \varphi \). The value of \( v \) that corresponds to \( \varphi \) can then be substituted into the model developed so far. The commands to determine \( v \) from an input value of \( \varphi \) are shown below:

```mathematica
\[ \delta = 0.1; \]
\[ p = 15; \]
\[ f[v_] = \sqrt{(1-\delta)^2/(p-1)(p(1-v)(1+v*p)/(1-v*p)(1+v))-1}]\]
\[ \text{FindRoot}[f[v]==0.7,\{v,0.3\}] \]

\( \{v \rightarrow 0.226034\} \)
```

The first two lines specify values for \( \delta \) and \( p \). Equation 4-26 is designated as \( f[v_] \) on the third line. The command \text{FindRoot} uses a damped Newton's method, the secant method and Brent's method to find the roots of the equation. In the example above, the value of \( v \) that corresponds to \( \varphi = 0.7 \) is determined. The value of 0.3 on the last line is an initial guess of the root. Throughout this model, the initial root will always be estimated as 1-\( \varphi \). Figures 4-2 to 4-4 show that 1 - \( \varphi \) should be an acceptable initial
estimate. The value of v that corresponds to $\varphi = 0.7$, $p = 15$ and $\delta = 0.1$ is output as 0.226034. This value of v can be confirmed as correct by substituting these values into equation 4-26. Hence, it is now possible to input $\varphi$ as a parameter, convert it to the corresponding value of v and use the model as given by equations 4-16 and 4-17.

4-8 AN EXAMPLE TO PRODUCE RANDOM DATA BY SPECIFYING THE AVERAGE INTERCORRELATION OF THE CORRELATION MATRIX

All of the previous sections have been used to establish a model to simulate data that have a correlation matrix with a specified average intercorrelation $\varphi$. The complete Mathematica program for generating random data with a correlation matrix that has a specified average intercorrelation $\varphi$ is provided in appendix B. In this section, a random data matrix is produced with $n = 10$, $p = 5$ and $\varphi = 0.7$ using this program. The procedure GenData generates the $n \times p$ data matrix with the specified value for $\varphi$.

\[ x = \text{GenData}[10, 5, 0.7] \]

\[
\begin{align*}
(1.70065, -2.17183, -1.6333, -1.52227, -2.57265), \\
(0.16665, 0.495005, -0.673789, 0.185481, 0.182034), \\
(-1.19443, 1.54834, 1.74487, 1.51794, 0.250249), \\
(0.142797, 0.416278, 0.475446, 0.361742, 0.122746), \\
(0.299186, -0.799443, -0.444128, 0.378615, -0.581274), \\
(-0.684905, 0.858607, 0.190488, 0.79402, 0.846574), \\
(1.4587, -0.119164, -1.10714, -1.06194, 0.897992), \\
(-1.14638, 0.0112859, 0.981466, 1.1055, 0.208941), \\
(0.0512719, -0.274869, 0.42868, -0.945267, 0.603567), \\
(-0.793544, 0.0357933, 0.0374047, -0.442853, 0.0418188))
\]
These data can be rearranged into matrix form:

\[
\text{MatrixForm}[\mathbf{x}]
\]

\[
\begin{bmatrix}
1.70065 & -2.17183 & -1.6333 & -1.52227 & -2.57265 \\
0.16665 & 0.495005 & -0.673789 & -0.185481 & 0.182034 \\
-1.19443 & 1.54834 & 1.74487 & 1.51794 & 0.250249 \\
0.142797 & 0.416278 & 0.475446 & 0.361742 & 0.122746 \\
0.299186 & -0.799443 & -0.444128 & 0.378615 & -0.581274 \\
-0.684905 & 0.858607 & 0.190488 & 0.79402 & 0.846574 \\
1.4587 & -0.199164 & -1.10714 & -1.06194 & 0.897992 \\
-1.14638 & 0.0112859 & 0.981466 & 1.1055 & 0.208941 \\
0.0512719 & -0.274869 & 0.42868 & 0.945267 & 0.603567 \\
-0.793544 & 0.0357933 & 0.0374047 & -0.442853 & 0.0418188
\end{bmatrix}
\]

The correlation matrix for the above data is produced in Mathematica and written in matrix form using the following commands:

\[
\text{Corr} = \text{MatrixForm}[\text{CorrelationMatrix}[\mathbf{x}]]
\]

\[
\begin{bmatrix}
1 & -0.717498 & -0.879706 & -0.813905 & -0.47909 \\
-0.717498 & 1 & 0.739823 & 0.719692 & 0.782061 \\
-0.879706 & 0.739823 & 1 & 0.805991 & 0.496094 \\
-0.813905 & 0.719692 & 0.805991 & 1 & 0.383022 \\
-0.47909 & 0.782061 & 0.496094 & 0.383022 & 1
\end{bmatrix}
\]

The eigenvalues of the above correlation matrix are shown below:

\[
\text{Eigenvalues}[\text{Corr}]
\]

\{3.75527, 0.786857, 0.229066, 0.124253, 0.104557\}

To check that the Mathematica program actually produces a random data matrix with \(\varphi = 0.7\), equation 4-18 is used with the above correlation matrix.
\[
\varphi = \sqrt{\frac{\|R\|^2 - p}{p(p - 1)}} \\
= \sqrt{\frac{14.80001052 - 5}{5(5 - 1)}} \\
= 0.700000375
\]

The Mathematica program produced a random correlation matrix with \( \varphi \) accurate to six decimal places. This is deemed acceptable for this simulation study.
Chapter 5

Analysis of Techniques to Estimate the Correlation Matrix

5-1 INTRODUCTION

The techniques complete cases only, all available cases and imputing means are investigated as to their ability to estimate the original correlation matrix. Data are simulated as described in chapter 4. Appropriate levels are chosen for the number of variables, the number of cases, the average intercorrelation and the proportion of missing values to simulate data that are representative of factor analyses. Data are randomly deleted to represent missing values and to examine the conditions when each of the techniques performs well or inadequately. A root-mean-square error measure is used to measure the difference between the estimated and original correlation matrices. A test is used to examine for any significant differences between the estimated and original correlation matrices. Regression analysis is used to predict
the error of the techniques and provide information as to when one technique performs better than the others do.

5-2 MONTE CARLO DESIGN

When dealing with missing values, the iterative principal component method and the EM algorithm (discussed in chapter 3) require an initial estimate of the correlation matrix. The first part of the Monte Carlo design involves investigating which technique produces the best estimate of the correlation matrix from the complete data set. Only the three techniques of complete cases only, imputing means and all available cases are examined for their ability to estimate the correlation matrix. The other techniques are considered too expensive to run for an initial estimate of the correlation matrix.

5-2.1 Study Design

Initially, data with a specified correlation structure are randomly simulated as described in detail in chapter 4. The number of cases n, the number of variables p, the average intercorrelation among the variables \( \varphi \) and the proportion of missing values \( k \) are varied. The various levels of these parameters will assist to determine the conditions when one technique performs better than the other techniques. The different values used for these parameters are shown below:

\[
\begin{align*}
n &= 50, 100, 200 \\
p &= 15, 25, 50 \\
\varphi &= 0.1, 0.3, 0.5, 0.7, 0.9 \\
k &= 1 \%, 2 \%, 5 \%, 10 \%
\end{align*}
\]
The above levels of parameters are chosen to simulate a factor analysis that will resemble those encountered in real life situations. A large number of variables $p$ are used in this study. In the literature, none of the previous authors that examined simulations for missing values considered the effect of more than 30 variables. Gleason and Staelin (1975) analysed 10, 15 and 30 variables. Bello (1993) and Timm (1970) analysed a maximum of 10 variables. Beale and Little (1975), Haitovsky (1968) and Buck (1960) analysed 5 variables or less. As this thesis is concerned with the effect that missing values have on a factor analysis, it is very important to consider a large number of variables.

Five levels for the average intercorrelation among variables are chosen. It is suspected that this factor will have a large influence on the three techniques’ abilities to handle missing values. Four levels between 1% and 10% are chosen for the proportion of missing values. A maximum of 10% of the data are deleted to test the robustness of each of the three techniques.

For each value of $n$, $p$, $\phi$ and $k$, five data matrices are simulated as discussed in chapter 4. For each of the five data matrices, data are randomly deleted 20 times to provide missing values. A simple procedure is used to randomly delete data from the complete data matrix. Two random numbers are used to determine the row and the column of the position in the original data matrix that is to be deleted. This procedure is repeated until the desired number of deletions is reached (defined by $k$). If the two random numbers indicate a position that has already been deleted, then this position is simply ignored and another deletion position is used. Altogether, 100 simulations are run for each combination of $n$, $p$, $\phi$ and $k$ ($5 \times 20$). It would be preferable to generate
random data 100 times for each factor combination as opposed to the five that are used. Unfortunately, this proved too expensive to run in computer time, particularly for the larger number of variables. The comparison of the differences for both of these simulation patterns is analysed in section 5-3.1.

5.2.2 Measurements to Compare the Original and Estimated Correlation Matrices

The three techniques (complete cases only, all available cases and imputing means) are applied to the data matrix with missing values to estimate the correlation matrix \( \mathbf{R} \). Several criteria have been used in the literature for comparing the estimated correlation matrix \( \hat{\mathbf{R}} \) with the original correlation matrix \( \mathbf{R} \) (based on the complete data set). Timm (1970) discussed the Euclidean norm \( \| \mathbf{R} - \hat{\mathbf{R}} \|^2 \) to measure the difference between the estimated and the original correlation matrix. As the distribution of this measure is unknown, Timm used an efficiency ratio to compare two techniques.

Gleason and Staelin (1975) used the Euclidean norm in a similar fashion to Timm. They used the root-mean-square measure \( \sqrt{\frac{\| \mathbf{R} - \hat{\mathbf{R}} \|^2}{p(p-1)}} \). Bello (1993) also adopted the Euclidean norm, but measured the difference between the estimated and original covariance matrices.

As correlation matrices are used in this thesis and a single measure of the difference between the two correlation matrices is preferred to an efficiency measure, Gleason
and Staelin’s measure is used. The equation used to measure the difference between
two correlation matrices is given by:

\[ E_\alpha = \sqrt{\frac{\hat{R}_\alpha - R}{p(p - 1)}} \quad \text{(5-1)} \]

where the subscript \( \alpha \) represents the technique being used.

5.2.3 Tests for the Equality of Correlation Matrices

For each simulation conducted in this chapter, two correlation matrices result: one for
the original data and another after data are deleted and a technique is used to estimate
the correlation matrix. Is there a significant difference between the original and the
estimated correlation matrices? Whilst this seems a logical and important question to
answer, there is no evidence of a study for this analysis in the literature.

It is suspected that the main reason for the lack of work in this area lies in the fact that
little is known about a statistical test to assess whether two correlation matrices are
equal. A likelihood ratio test for the equality of two covariance matrices is well
known. The reader is referred to Layard (1972) for a discussion on the tests proposed
for the equality of two covariance matrices.

Kullback (1967) suggested a test statistic for the equality of several correlation
matrices. It had the following form:

\[ \chi^2_i = \sum_{i=1}^{k} N_i \log \frac{|R|}{|R_i|} \quad \text{(5-2)} \]
where: $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_k$ are the sample correlation matrices from independent samples of size $n_1, n_2, \ldots, n_k$ from multivariate normal populations. Each $N_i = n_i - 1$ and

$$
\mathbf{R} = (N_1 \mathbf{R}_1 + N_2 \mathbf{R}_2 + \ldots + N_k \mathbf{R}_k) / N,
$$

where $N = N_1 + N_2 + \ldots + N_k$.

Kullback proposed that if all of the $k$ populations had the same non-singular correlation matrix, then the above statistic would have an asymptotic $\chi^2$ distribution with $p(k-1)(p-1)/2$ degrees of freedom. Jennrich (1970) was able to show that in general, this statistic did not follow an asymptotic $\chi^2$ distribution.

Pearson and Wilks (1933) suggested an asymptotic $\chi^2$ test for the equality of population correlation matrices for the bivariate situation. Kullback (1967) showed that this test does not generalise to larger correlation matrices. Cole (1968a, 1968b) developed the likelihood ratio test for the equality of two correlation matrices. This is given by the following statistic:

$$
\lambda = \prod_{i=1}^{k} \left[ \frac{D_i \mathbf{R}_i D_i}{\hat{D}_i \hat{P} \hat{D}_i} \right]^{n_i/2}
$$

(5-3)

where: $D_i = \text{diag}[s_{ii}^{1/2}(i)]$ and $s_{ii}(i)$ is a diagonal element of $S_i$. $\mathbf{R}_i$ and $S_i$ are the sample correlation and covariance matrices of the $i^{th}$ population respectively.

$\hat{P}$ and $\hat{D}_i$ are the solutions of:

$$
\hat{P} = \sum_{i=1}^{k} (n_i / N) \hat{D}_i^{-1} S_i \hat{D}_i^{-1}
$$

(5-4)

$$
(\hat{P}^{-1} \hat{D}_i^{-1} S_i \hat{D}_i^{-1} - I)_{i,j} = 0
$$

(5-5)

where: $N = n_1 + n_2 + \ldots + n_k$. 

79
The null distribution of $-2\log \lambda$ has a $\chi^2$ distribution with $(k-1)p(p-1)/2$ degrees of freedom. It is computationally very difficult and arduous to maximise the likelihood function particularly when using the equal correlation assumption. Modarres (1990) showed that this likelihood ratio test is very sensitive to departures from the assumption of multivariate normality.

Jennrich (1970) suggested looking at all the elements of the correlation difference matrix $R_1 - R_2$. The two sample correlation matrices are independent and non-singular from multivariate normal populations. Let $Vech(R_i)$ denote the correlation vector obtained from the $i^{th}$ correlation matrix by placing its subdiagonal elements under each other. It is well known that $\sqrt{n_i}Vech(R_1 - R_2)$ is asymptotically multivariate normal with covariance matrix $\Gamma$ (Steiger and Hakstian 1982). Hence, in testing whether several correlation matrices are equal against the alternative that at least one is not equal to another, the following statistic can be used:

$$Q = \sum_{i=1}^{k} n_i Vech(R_i - \bar{R})' \Gamma^{-1} Vech(R_i - \bar{R})$$  \hspace{1cm} (5-6)

This test has an asymptotic $\chi^2$ distribution with $(k-1)p(p-1)/2$ degrees of freedom. $\bar{R}$ is the weighted average of the correlation matrices. $\Gamma$ is a square $p(p-1)/2$ matrix. Even though it is easy to find a consistent estimator $\hat{\Gamma}$ for $\Gamma$ (pooling either the covariance or correlation matrices), the above statistic requires the inversion of $\hat{\Gamma}$. Throughout the simulations in this thesis, the minimum number of variables $p$ will be 15. A $105 \times 105$ matrix would be the smallest sized matrix to invert. Even this places enormous demands on computer storage and time. Jennrich (1970) has proposed a
simpler representation of equation 5-6 that only involves the inversion of a \( p \times p \) matrix.

The summary of his results is shown below for two correlation matrices:

\[
\chi^2 = \frac{1}{2} \text{tr}(Z^2) - \text{diag}'(Z)S^{-1}\text{diag}(Z)
\]  

(5-7)

where

\[
S = \delta_{ij} + \tilde{r}_{ij}\tilde{r}^{ji}
\]

\[
c = n_1 n_2 / (n_1 + n_2)
\]

\[
Z = c^{\frac{1}{2}}R^{-1}(R_1 - R_2)
\]

\[
\bar{R} = (R_1 + R_2) / 2
\]

\( \delta_{ij} \) is the Kronecker delta and the superscripts \( ij \) refer to the elements of the inverse matrix.

Modarres and Jernigan (1993) show that this test is much more robust compared to Cole's maximum likelihood ratio test. Equation 5-7 will be used throughout this thesis whenever two correlation matrices are compared for their equality. The computer program written in Mathematica to generate data; delete data to represent missing values; perform all available cases, complete cases only and imputing means; as well as performing Jennrich's test, is contained in appendix C.

5.3 RESULTS

5.3.1 Comparison of Simulation Patterns

A decision has to be made about the number of data matrices and deletion patterns to use for each combination of \( n, p, \varphi \) and \( k \). One hundred replicates in total are used for
each parameter combination. It would be preferable to generate 100 different data matrices and randomly delete data from each of these. Unfortunately, it is very expensive in computer running time to randomly generate a data matrix. Table 5-1 shows the time in CPU seconds for a Pentium 100 MHz computer to generate various sized matrices that are required for this thesis.

Table 5-1 Computing Time Required to Generate Random Data Matrices

<table>
<thead>
<tr>
<th>Number of variables (p)</th>
<th>Number of cases (n)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>100</td>
<td>200</td>
</tr>
<tr>
<td>15</td>
<td>14.8</td>
<td>15.7</td>
<td>17.2</td>
</tr>
<tr>
<td>25</td>
<td>29.8</td>
<td>31.9</td>
<td>38.3</td>
</tr>
<tr>
<td>50</td>
<td>106.8</td>
<td>114.8</td>
<td>128.3</td>
</tr>
</tbody>
</table>

According to the above table, it takes 128.3 seconds to generate a data matrix that contains 50 variables and 200 cases. Performing one hundred such simulations would then take 3.6 hours. It would take too long to generate 100 simulations for all of the 180 (3×3×5×4) combinations of parameters simulated in this thesis. To generate 100 simulations for each combination of parameters, it is easier and much quicker to generate 5 data matrices and randomly delete the data 20 times for each of these matrices. The effect that this faster simulation pattern has on the averages and standard deviations of the root-mean-square error E (equation 5-1) is examined in this section. The root-mean-square error term is the criteria used to decide which technique better estimates the correlation matrix.

A comparison is made on five data matrices with twenty deletion patterns (5×20) and the 100 data matrices with a single deletion pattern (100×1). Tables 5-2 and 5-3
show the averages and standard deviations of the error terms for the combination of \( n = 100, \ p = 15, \ k = 1\% \). The averages and standard deviations of \( E \) are determined for 100 data matrices using both simulation patterns.

### Table 5-2 Comparing Average Errors for Both Simulation Patterns for \( n = 100, \ p = 15, \ k = 1\% \)

<table>
<thead>
<tr>
<th>( \varphi )</th>
<th>Simulation Pattern</th>
<th>Complete Cases Only</th>
<th>Imputing Means</th>
<th>Available Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5 x 20</td>
<td>4.05</td>
<td>1.45</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>4.06</td>
<td>1.41</td>
<td>1.41</td>
</tr>
<tr>
<td>0.3</td>
<td>5 x 20</td>
<td>3.61</td>
<td>1.32</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>3.80</td>
<td>1.38</td>
<td>1.32</td>
</tr>
<tr>
<td>0.5</td>
<td>5 x 20</td>
<td>2.91</td>
<td>1.21</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>2.96</td>
<td>1.30</td>
<td>1.11</td>
</tr>
<tr>
<td>0.7</td>
<td>5 x 20</td>
<td>2.10</td>
<td>1.10</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>2.06</td>
<td>1.11</td>
<td>0.75</td>
</tr>
<tr>
<td>0.9</td>
<td>5 x 20</td>
<td>0.64</td>
<td>0.85</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>0.66</td>
<td>0.84</td>
<td>0.23</td>
</tr>
</tbody>
</table>

### Table 5-3 Comparing Standard Deviations of Errors for Both Simulation Patterns for \( n = 100, \ p = 15, \ k = 1\% \)

<table>
<thead>
<tr>
<th>( \varphi )</th>
<th>Simulation Pattern</th>
<th>Complete Cases Only</th>
<th>Imputing Means</th>
<th>Available Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5 x 20</td>
<td>4.20</td>
<td>1.90</td>
<td>1.90</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>4.38</td>
<td>2.39</td>
<td>2.38</td>
</tr>
<tr>
<td>0.3</td>
<td>5 x 20</td>
<td>6.30</td>
<td>2.30</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>6.57</td>
<td>2.62</td>
<td>2.44</td>
</tr>
<tr>
<td>0.5</td>
<td>5 x 20</td>
<td>6.70</td>
<td>2.30</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>7.06</td>
<td>2.55</td>
<td>2.26</td>
</tr>
<tr>
<td>0.7</td>
<td>5 x 20</td>
<td>7.30</td>
<td>3.00</td>
<td>2.00</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>8.58</td>
<td>2.61</td>
<td>1.86</td>
</tr>
<tr>
<td>0.9</td>
<td>5 x 20</td>
<td>1.60</td>
<td>1.90</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>2.14</td>
<td>1.78</td>
<td>0.49</td>
</tr>
</tbody>
</table>
According to table 5-2, there is little overall difference between the average errors for both simulation patterns for all three techniques. The largest disparity between the two simulation patterns is $0.19 \times 10^{-2}$. This situation occurs for complete cases only when the average level of intercorrelation $\varphi$ is 0.3. For all techniques, the largest differences between the two simulation patterns occurs when $\varphi = 0.3$ and 0.5. Table 5-3 reveals that the $5 \times 20$ simulation pattern results in smaller standard deviations of the errors as would be expected. However, the differences in standard deviations and averages are small enough to warrant the use of the $5 \times 20$ simulation pattern.

More extreme fluctuations between the two simulation patterns are expected when there are larger numbers of missing values. Both simulation patterns are run again for the same parameter combination as above, that is, $n = 100$, $p = 15$. Ten percent of the data are deleted to examine if the differences between the two simulation patterns becomes too large for more missing values. The results are shown below in tables 5-4 and 5-5.

**Table 5-4 Comparing Average Errors for Both Simulation Patterns for $n = 100$, $p = 15$, $k = 10\%$**

<table>
<thead>
<tr>
<th>$\varphi$</th>
<th>Simulation Pattern</th>
<th>Complete Cases Only</th>
<th>Imputing Means</th>
<th>Available Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$5 \times 20$</td>
<td>20.56</td>
<td>4.94</td>
<td>4.95</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>20.58</td>
<td>4.93</td>
<td>4.93</td>
</tr>
<tr>
<td>0.3</td>
<td>$5 \times 20$</td>
<td>18.76</td>
<td>4.74</td>
<td>4.52</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>18.93</td>
<td>4.71</td>
<td>4.47</td>
</tr>
<tr>
<td>0.5</td>
<td>$5 \times 20$</td>
<td>14.56</td>
<td>4.32</td>
<td>3.63</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>15.23</td>
<td>4.36</td>
<td>3.68</td>
</tr>
<tr>
<td>0.7</td>
<td>$5 \times 20$</td>
<td>10.81</td>
<td>3.83</td>
<td>2.54</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>10.55</td>
<td>3.84</td>
<td>2.49</td>
</tr>
<tr>
<td>0.9</td>
<td>$5 \times 20$</td>
<td>3.13</td>
<td>2.79</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>3.33</td>
<td>3.08</td>
<td>0.83</td>
</tr>
</tbody>
</table>
Table 5.5 Comparing Standard Deviations of Errors for Both Simulation Patterns for n = 100, p = 15, k = 10%

<table>
<thead>
<tr>
<th>( \phi )</th>
<th>Simulation Pattern</th>
<th>Complete Cases Only</th>
<th>Imputing Means</th>
<th>Available Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5 x 20</td>
<td>22.60</td>
<td>3.80</td>
<td>3.90</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>25.81</td>
<td>3.60</td>
<td>3.64</td>
</tr>
<tr>
<td>0.3</td>
<td>5 x 20</td>
<td>32.90</td>
<td>4.30</td>
<td>4.20</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>33.79</td>
<td>4.76</td>
<td>4.55</td>
</tr>
<tr>
<td>0.5</td>
<td>5 x 20</td>
<td>36.30</td>
<td>4.70</td>
<td>4.30</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>37.44</td>
<td>5.37</td>
<td>4.69</td>
</tr>
<tr>
<td>0.7</td>
<td>5 x 20</td>
<td>42.70</td>
<td>4.80</td>
<td>3.60</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>39.67</td>
<td>5.77</td>
<td>4.78</td>
</tr>
<tr>
<td>0.9</td>
<td>5 x 20</td>
<td>10.50</td>
<td>3.90</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>100 x 1</td>
<td>17.70</td>
<td>4.93</td>
<td>1.45</td>
</tr>
</tbody>
</table>

When 10% of the data are missing, the averages and standard deviations of the root-mean-square errors are much larger than when there are 1% missing values. Fortunately, there is still little difference in the means and standard deviations for both simulation patterns. Generally, the 5 x 20 simulation pattern produces smaller standard deviations of the error estimate. In comparison to the running time saved (refer to table 5.1), the difference in standard deviations of the two patterns are considered small enough to warrant the use of the 5 x 20 pattern for the simulation study.

5.3.2 Graphical Comparison of Three Techniques to Estimate the Correlation Matrix

The three techniques: complete cases only, imputing means and all available cases are compared for their ability to estimate the correlation matrix of the original data when some of the data are deleted. The root-mean-square error E (equation 5-1) is used to
measure the difference between the estimated and the original correlation matrices. The root-mean-square errors are initially plotted versus all the different criteria n, p, ϕ and k to visualise what is occurring and to investigate any interactions.

Previous research from Timm (1970) indicated that complete cases only should perform worse than the other techniques when p > 5. As this research deals with 15, 25 and 50 variables in the simulation of a factor analysis, it is suspected that complete cases only will perform poorly. To examine what happens for the smallest number of variables investigated in this study (p = 15), the three techniques are plotted to investigate the performance of complete cases only compared to the other two techniques. Plots of the errors E of the three techniques for p = 15 and k = 1%, 2% are shown in figures 5-1 and 5-2 for n = 50, 100 and 200.

![Figure 5-1](image)

**Figure 5-1** Comparison of E for Imputing Means, All Available Cases and Complete Cases Only (p = 15, k = 1%)
Figure 5-2   Comparison of E for Imputing Means, All Available Cases and Complete Cases Only (p = 15, k = 2%)

It is clear from the above figures that the technique of complete cases only performs poorly compared to the other two techniques. It is only when there is a very high level of intercorrelation among the variables (φ = 0.9) that complete cases only performs slightly better than imputing means. Out of the three levels selected for the number of variables in this thesis; namely, p = 15, 25 and 50, complete cases only should estimate the original correlation matrix R best when p = 15. Even at this level, complete cases only is outperformed by the other techniques. Theoretically, complete cases only will deteriorate as the number of variables increases. Some preliminary graphs at p = 25 supported this notion (refer to figures 5-3 and 5-4). Complete cases only performs worse than the other two techniques for all levels of factors examined at 25 variables. As complete cases only will continue to deteriorate for higher levels of p, this technique is disregarded in further investigations.
Figure 5-3  Comparison of E for Imputing Means, All Available Cases and Complete Cases Only (p = 25, k = 1%)

Figure 5-4  Comparison of E for Imputing Means, All Available Cases and Complete Cases Only (p = 25, k = 2%)

The complete range of parameters is plotted for imputing means and all available cases in figures 5-5 to 5-8. Each figure has 9 plots with 3 rows and 3 columns. The sample size increases in each column moving from left to right. The number of variables increases moving from top to bottom. Each plot displays the root-mean-
square error \( E \) for both all available cases and imputing means versus \( \varphi \). The four figures differ in the percentage of missing values.

The plots are very consistent in their shape for all of the parameters studied. All available cases performs better than imputing means for higher levels of intercorrelations among the variables as would be expected. Imputing means performs better for lower levels of intercorrelations. The lowest level of intercorrelation where all available cases continually outperforms imputing means lies between 0.1 and 0.3 for all plots.

Both techniques improve as \( \varphi \) and \( n \) increase. It is interesting that the graphs of imputing means consistently show that it improves as \( \varphi \) increases. It might be expected that the performance of imputing means would be largely unaffected by changes in \( \varphi \). However, it was shown in section 3-4.1 that imputing means for missing values and adjusting for the degrees of freedom produced the same results as one of the methods of all available cases. This available case method uses the mean from all available cases when determining the variance and covariance. Since imputing means and adjusting the degrees of freedom is one form of an available case method, it would be expected to improve as \( \varphi \) increases.
Figure 5-5 Graphical Comparison of E for All Available Cases and Imputing Means (k = 1%)
Figure 5-6 Graphical Comparison of E for All Available Cases and Imputing Means (k = 2%)
Figure 5-7 Graphical Comparison of E for All Available Cases and Imputing Means (k = 5%)
Figure 5-8 Graphical Comparison of E for All Available Cases and Imputing Means (k = 10%)
5.3.3 Regression Analysis of Errors

The consistency of the shape of the plots suggests that regression analysis might prove useful in predicting the root-mean-square error \( E \) of the two techniques. It might also provide some information on the situations when one technique performs better than the other. A linear regression model is initially fitted to the data. The independent variables are \( n, p, k \) and \( \varphi \), and the dependent variable is \( E \). Using SPSS, a stepwise regression procedure is conducted and a summary of the results is shown below. The details are found in appendices F-1A and F-2A. The reader should be aware of some of the terms in the SPSS output in the appendices. The term \( j \) refers to \( j \) and \( 1j \) refers to \( 1 - \varphi \). These terms had to be used in the appendices to be compatible with SPSS.

\[
\begin{align*}
E_{im} &= 0.032 - 1.00 \times 10^{-4} n + 0.340k - 0.015\varphi \quad (R^2 = 0.838) \quad (5-8) \\
E_{ac} &= 0.035 - 9.43 \times 10^{-5} n + 0.271k - 0.031\varphi \quad (R^2 = 0.790) \quad (5-9)
\end{align*}
\]

The subscripts \( im \) and \( ac \) in the above equations refer to imputing means and all available cases respectively. For both of the techniques shown in equations 5-8 and 5-9, all of the independent variables are very significant (p-value = 0.000) with the exception of the number of variables \( p \). This suggests that the ability of imputing means and all available cases to predict the original correlation matrix is not influenced by the number of variables. The number of variables studied in this analysis ranged from 15 to 50. It would be interesting to investigate whether the number of variables is significant for the two techniques when there are only a small numbers of variables present. This investigation is not done here, as this thesis is only concerned with a large number of variables that are present in a factor analysis.
Residual plots for both regression models are found in appendices F-1A and F-2A. The residuals appear to be normally distributed for both models. However, both of the residuals versus predicted values plots suggest that the assumption of homogeneity of variance may not be satisfied. There seems to be greater variation in the values of the standardised residuals corresponding to large values of predicted values than for small values. Plots of the residuals versus the number of cases n and the percentage of deletions k also seem to violate the assumption of homogeneity of variance. The standardised residuals seem to vary over a wider range for small values of n and large values of k.

Natural logarithm transformations are applied to the dependent variable only as well as both the dependent and independent variables in an attempt to overcome the problems identified with the linear models. Refer to appendices F-1B, F-1C, F-2B and F-2C for the complete details of these regression models. A summary of the regression models after the two transformations on both imputing means and all available cases is shown in tables 5-6 and 5-7.

Table 5-6 Summary of Regression Transformations for Imputing Means

<table>
<thead>
<tr>
<th>Natural Log Transformation</th>
<th>$R^2$</th>
<th>Comments on Residual Plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variables</td>
<td>0.878</td>
<td>Normally distributed; equal variance present.</td>
</tr>
<tr>
<td>Dependent &amp; Independent Variables</td>
<td>0.944</td>
<td>Normally distributed; equal variance present.</td>
</tr>
</tbody>
</table>
Table 5-7 Summary of Regression Transformations for All Available Cases

<table>
<thead>
<tr>
<th>Natural Log Transformation</th>
<th>R²</th>
<th>Comments on Residual Plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variables</td>
<td>0.835</td>
<td>Not normally distributed; resid vs fits shows unequal variance.</td>
</tr>
<tr>
<td>Dependent &amp; Independent Variables</td>
<td>0.961</td>
<td>Normally distributed; equal variance present.</td>
</tr>
</tbody>
</table>

From both of the above tables, the transformation applied to both the dependent and independent variables (multiplicative model) is the better regression technique for both imputing means and all available cases. The multiplicative models provide significant increases in the $R^2$ values over the linear models and they also satisfy the regression assumptions much better. The two multiplicative regression models are shown below:

\[
E_{lm} = 1.855n^{-0.513}k^{0.539}(1 - \varphi)^{0.238} \quad (R^2 = 0.944) \quad (5-10)
\]

\[
E_{ac} = 2.396n^{-0.531}k^{0.537}(1 - \varphi)^{0.845} \quad (R^2 = 0.961) \quad (5-11)
\]

The term $(1-\varphi)$ is used as an independent variable for both of the above models. This term is used to satisfy the theory for all available cases that as $\varphi \to 1$, $E_{ac} \to 0$. The plots shown in figures 5-5 to 5-8 suggest that the imputing means error also reduces as $\varphi$ approaches 1. The reason why imputing means decreases as $\varphi$ increases has been discussed previously in section 5-3.2

Equations 5-10 and 5-11 can be used to determine which technique is better to use for any factor analysis situation. The average level of intercorrelation $\varphi$ can be used as a criteria to identify when one technique performs better than the other. All available
cases performs better when its root-mean-square error is smaller than that from imputing means. That is,

\[
E_{sc} < E_{im} \\
2.396n^{-0.534}k^{0.537}(1 - \varphi)^{0.845} < 1.855n^{-0.513}k^{0.539}(1 - \varphi)^{0.238}
\]  

(5-12)

The coefficients of \(k\) for both techniques are almost equal. This suggests that the percentage of deletions has the same effect on both errors. Assuming that the coefficients of \(k\) are equal, equation 5-12 can now be solved for \(\varphi\).

\[
2.396n^{-0.534}(1 - \varphi)^{0.845} < 1.855n^{-0.513}(1 - \varphi)^{0.238} \\
(1 - \varphi)^{0.007} < 0.774n^{0.018} \\
1 - \varphi < \left(0.774n^{0.018}\right)^{0.607} \\
1 - \varphi < 0.656n^{0.030} \\
\varphi > 1 - 0.656n^{0.030}
\]

(5-13)

The above equation shows that all available cases performs better than imputing means when \(\varphi > 1 - 0.656n^{0.030}\). In terms of the values of \(n\) used in this chapter, all available cases is the better technique as follows:

\[
\begin{align*}
n &= 50 & \varphi &> 0.262 \\
n &= 100 & \varphi &> 0.247 \\
n &= 200 & \varphi &> 0.231
\end{align*}
\]

5-3.4 Comparison with Other Proposed Models

Gleason and Staelin (1975) conducted a similar study to that shown in this section. They used the methods of Wilks (1932) for imputing means and Glasser (1964) for all available cases. A total of 42 matrices of 7 basic types were constructed with various
values for n, p and φ. Three deletion patterns (k = 4%, 12% and 20%) were used on each of the 42 matrices, resulting in 126 matrices that were analysed. Table 5-8 summarises their design. It shows the lowest φ_L and highest φ_H levels of intercorrelation as well as the number of matrices N for each combination. Note that not all combinations of n and p were used to reduce the magnitude of the design.

### Table 5-8  Summary of Gleason & Staelin's Study Design

<table>
<thead>
<tr>
<th>Number of Cases (n)</th>
<th>Number of Variables (p)</th>
<th>φ_H</th>
<th>φ_L</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>15</td>
<td>.424</td>
<td>.170</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>15</td>
<td>.371</td>
<td>.115</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>.552</td>
<td>.118</td>
<td>54</td>
</tr>
<tr>
<td>200</td>
<td>15</td>
<td>.380</td>
<td>.111</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Apart from the volume of matrices studied, there are other differences between Gleason and Staelin’s study and the one presented in this thesis. Most of their matrices had 15 variables and they only studied a maximum of 30 variables. The intercorrelations ranged only from 0.11 to 0.55 as opposed to 0.1 to 0.9 used in this thesis. Some interesting comparisons can still be made. Gleason and Staelin also used multiplicative regression models for the errors from Wilks (W) and Glasser’s (G) methods. The two root-mean-square error equations are shown below:

\[
E_W = 1.362φ^{0.325}n^{-0.181}p^{-0.125}k^{0.288} \quad (R^2 = 0.91)
\]

\[
E_G = 1.194(1-φ)^{-0.202}n^{-0.278}p^{-0.128}k^{0.300} \quad (R^2 = 0.92)
\]
Compare these equations to equations 5-10 and 5.11. Gleason and Staelin indicated that the coefficient for \((1-\varphi)\) for the Glasser method is negative which is contrary to the theory. Their study did not have any matrices with \(\varphi > 0.55\). They suspected that the relationship over the total range of \(\varphi\) might be a second-degree polynomial. This thesis investigates larger values of \(\varphi\). Using these values, the coefficient of \(1-\varphi\) in equation 5-11 is positive, which agrees with the theory. The number of variables \(p\) is also significant in the above equations at a confidence level of 5%, whereas \(p\) was not included from the results in this thesis (refer to section 5.3.3). This difference could be due to the fact that Gleason and Staelin investigated smaller values of \(p\).

5.3.5 Testing the Equality of the Original and Estimated Correlation Matrices

For each combination of \(n, p, k\) and \(\varphi\), 100 matrices are generated. Data are removed to represent missing values and all available cases and imputing means are applied to these matrices. One hundred correlation matrices are estimated for both techniques. Equation 5-7 is then used to test for the equality of these 100 estimated correlation matrices \(\hat{R}\) with the original correlation matrix \(R\) (obtained from the data matrix with no missing values). There is a concern with using 100 \(\chi^2\) tests on the original and estimated correlation matrices. It is very difficult to determine the overall risk of making a Type I error. Obviously, the more tests that are conducted, the greater is the risk of making a Type I error. However, it is possible to take this into account. For example, at a 5% significance level, it is expected that 5 out of the 100 comparisons would be found to be significantly different when in fact they should not have been.
Jenrich’s test of two correlation matrices is also based on the assumption that the two sample correlation matrices are independent (refer to section 5-2.3). As most of the data are similar for the calculation of the original and estimated correlation matrices, the two matrices would not be independent. However, Jenrich’s test can still be used to compare which technique (imputing means or all available cases) performs better for various situations. The lack of independence can be accounted for by taking into account that there would be fewer significant results.

Tables 5-9 to 5-11 summarise the number of estimated correlation matrices \( \hat{R} \) that are significantly different to the original correlation matrix \( R \) for imputing means. Table 5-9 shows the number \( \hat{R} \) that are significantly different to \( R \) for \( n = 50, p = 15 \) and \( k = 1\%, 2\%, 5\% \) and \( 10\% \). The results are obtained from 100 replicates. For example, 62 \( \hat{R} \) are significantly different to \( R \) at \( \alpha = 0.05 \) for \( n = 50, p = 15 \) and \( k = 10\% \). As 5 significantly different correlation matrices are expected, imputing means performs poorly for this situation. Tables 5-10 to 5-11 show the same conditions for \( n = 100 \) and 200.

Overall, the number of significant results reduces as \( n \) increases. The number of significant results increases as \( k \) increases as is expected. Imputing means performs worse for higher levels of \( \varphi \). This is contradictory to the results found in section 5-3.2 where higher levels of \( \varphi \) results in estimates that are closer to the original correlation matrix. Before examining this apparent contradiction further, the tables 5-9 to 5-11 are repeated for all available cases to investigate whether the results are also contradictory for this technique.
Table 5-9  Number of $\hat{R}$ Significantly Different from R for Imputing Means  
(n = 50, p = 15)

<table>
<thead>
<tr>
<th>(a) k = 1%</th>
<th>(b) k = 2%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>φ</strong></td>
<td><strong>Significance Level</strong></td>
</tr>
<tr>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) k = 5%</th>
<th>(d) k = 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>φ</strong></td>
<td><strong>Significance Level</strong></td>
</tr>
<tr>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 5-10  Number of $\hat{R}$ Significantly Different from R for Imputing Means  
(n = 100, p = 15)

<table>
<thead>
<tr>
<th>(a) k = 1%</th>
<th>(b) k = 2%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>φ</strong></td>
<td><strong>Significance Level</strong></td>
</tr>
<tr>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) k = 5%</th>
<th>(d) k = 10%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>φ</strong></td>
<td><strong>Significance Level</strong></td>
</tr>
<tr>
<td></td>
<td>10%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 5-11  Number of $\hat{R}$ Significantly Different from R for Imputing Means  
(n = 200, p = 15)

<table>
<thead>
<tr>
<th>(a) $k = 1%$</th>
<th>(b) $k = 2%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) $k = 5%$</th>
<th>(d) $k = 10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-12  Number of $\hat{R}$ Significantly Different from R for All Available Cases  
(n = 50, p = 15)

<table>
<thead>
<tr>
<th>(a) $k = 1%$</th>
<th>(b) $k = 2%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) $k = 5%$</th>
<th>(d) $k = 10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(a) $k = 1%$</th>
<th>(b) $k = 2%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(c) $k = 5%$</th>
<th>(d) $k = 10%$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>Significance Level</td>
</tr>
<tr>
<td></td>
<td>$10%$</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 5-13  Number of $\hat{R}$ Significantly Different from R for All Available Cases (n = 100, p = 15)

<table>
<thead>
<tr>
<th></th>
<th>Significance Level</th>
<th></th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>10%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) $k = 5\%$

<table>
<thead>
<tr>
<th></th>
<th>Significance Level</th>
<th></th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>10%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(d) $k = 10\%$

Table 5-14  Number of $\hat{R}$ Significantly Different from R for All Available Cases (n = 200, p = 15)

<table>
<thead>
<tr>
<th></th>
<th>Significance Level</th>
<th></th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>10%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(c) $k = 5\%$

<table>
<thead>
<tr>
<th></th>
<th>Significance Level</th>
<th></th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varphi$</td>
<td>10%</td>
<td>5%</td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(d) $k = 10\%$
When comparing all sets of tables for both techniques, all available cases generally performs better than imputing means at estimating $\mathbf{R}$. Like imputing means, all available cases had fewer significant results as $n$ increases and more significant results as $k$ increases. For intermediate values of $\varphi$ (0.5 and 0.7) there are more significant results. Interestingly, there are no significant results at $\varphi = 0.9$ for all available cases. Again, this is contradictory to the results in section 5-3.2. As the results in this previous section are quite conclusive, this suggests that either something is awry in the test shown in equation 5-7 or there is another underlying factor biasing the results.

As Jennrich's test produces large numbers of unexpected significant results as $\varphi$ increases, the problem could be multicollinearity and ill-conditioning among the variables. It is beyond the scope of this thesis to discuss all of the issues involved with multicollinearity and ill-conditioned matrices. As equation 5-7 involves elements of the inverse of the average of the two correlation matrices, non-positive definite NPD and singular matrices could adversely affect the test statistic. Tabachnick and Fidell (1996) discuss the fact that division by a near-zero determinant (obtained from multicollinear and ill-conditioned matrices) produces very large and unstable numbers in the inverted matrix. The sizes of the numbers in inverted matrices fluctuate wildly with only minor changes. Myers (1990) shows that the number of small eigenvalues of $\mathbf{R}$ relate to the number of multicollinearities. Unfortunately, information was not collected on the eigenvalues of these matrices. However, the number of NPD matrices that had to be smoothed when all available cases and imputing means were applied to missing values was recorded. These are shown in the following tables.
### Table 5-15  Number of NPD matrices for Imputing Means  \((n = 50, p = 15)\)

<table>
<thead>
<tr>
<th>(\varphi)</th>
<th>(1%)</th>
<th>(2%)</th>
<th>(5%)</th>
<th>(10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>96</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>82</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>2</td>
<td>72</td>
<td>100</td>
</tr>
</tbody>
</table>

### Table 5-16  Number of NPD matrices for Imputing Means  \((n = 100, p = 15)\)

<table>
<thead>
<tr>
<th>(\varphi)</th>
<th>(1%)</th>
<th>(2%)</th>
<th>(5%)</th>
<th>(10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>66</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>88</td>
</tr>
</tbody>
</table>

### Table 5-17  Number of NPD matrices for Imputing Means  \((n = 200, p = 15)\)

<table>
<thead>
<tr>
<th>(\varphi)</th>
<th>(1%)</th>
<th>(2%)</th>
<th>(5%)</th>
<th>(10%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 5-18  Number of NPD matrices for All Available Cases  (n = 50, p = 15)

<table>
<thead>
<tr>
<th>φ</th>
<th>Percentage of Deletions (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-19  Number of NPD matrices for Imputing Means  (n = 100, p = 15)

<table>
<thead>
<tr>
<th>φ</th>
<th>Percentage of Deletions (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-20  Number of NPD matrices for Imputing Means  (n = 200, p = 15)

<table>
<thead>
<tr>
<th>φ</th>
<th>Percentage of Deletions (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>0.1</td>
<td>0</td>
</tr>
<tr>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>0.9</td>
<td>0</td>
</tr>
</tbody>
</table>

Tables 5-15 to 5-20 indicate that more NPD matrices result when more of the data are missing. The number of NPD’s that had to be smoothed should be an indication of
the presence of ill-conditioned matrices used in equation 5-7. Imputing means generally produces more NPD matrices as $\phi$ increases. All available cases produces more NPD matrices when $\phi = 0.5$ and 0.7. The same patterns but fewer NPD’s are observed for higher levels of $n$ as would be expected. The levels of $\phi$ where there are increases in NPD’s is almost identical to where there are large numbers of significant differences in correlation matrices. It appears as though there is a strong relationship between NPD’s and significant results in the test of the difference of two correlation matrices.

According to the model developed in chapter 4, more eigenvalues would be closer to 0.1 for higher levels of $\phi$. As NPD matrices result when one of these eigenvalues is less than zero, this might explain why imputing means produces more NPD matrices as $\phi$ increases. However, all available cases produces more NPD matrices and significant differences at $\phi = 0.5$ and 0.7 and none at $\phi = 0.9$. The apparent link between the number of NPD’s and ill-conditioned matrices with $\phi$ is explored using perturbation theory.

Let $R$ be a correlation matrix and $\lambda$, $e$ an associated eigenvalue and eigenvector of $R$. $C$ is a symmetric matrix. Under the assumption that a technique that handles missing values produces a symmetric perturbation, Taylor-series expansions are used to express the effect of these perturbations. Sibson (1979) developed the following expression for the effect of the perturbation on the eigenvalue.
Let \( \mathbf{R} \) be perturbed to:

\[
\mathbf{R}(\varepsilon) = \mathbf{R} + \varepsilon \mathbf{C} + O(\varepsilon^2)
\]  

(5-14)

Assume that the corresponding perturbation of \( \lambda \) is:

\[
\lambda(\varepsilon) = \lambda + \varepsilon \mu + O(\varepsilon^2)
\]  

(5-15)

Sibson was able to show that:

\[
\mu = \mathbf{e}^T \mathbf{C} \mathbf{e}
\]  

(5-16)

Therefore, \( \varepsilon(\mathbf{e}^T \mathbf{C} \mathbf{e}) \) expresses the effect of the perturbation on the eigenvalue.

Consider the simplistic assumption that the error produced from either imputing means or all available cases produces the same sized perturbation \( \varepsilon \) to all off-diagonal elements of \( \mathbf{R} \). The elements of \( \mathbf{C} \) have the following form:

\[
C_{ij} = \begin{cases} 
0 & \text{i = j} \\
1 & \text{i \neq j} 
\end{cases}
\]  

(5-17)

Using equations 5-14, 5-16 and 5-17, the corresponding change to the eigenvalue is given by:

\[
\varepsilon(\mathbf{e}^T \mathbf{C} \mathbf{e}) = 2\varepsilon \sum_{i \neq j} e_i e_j
\]  

(5-18)

where \( e_i = i^{th} \) component of \( \mathbf{e} \)
If the change in an eigenvalue given by equation 5-18 is greater than the eigenvalue itself (assuming that the eigenvalue is reducing), then an NPD matrix will result. Hence, an NPD matrix results when:

\[
2\varepsilon \sum_{i \neq j} e_i e_j > \lambda \\
\sum_{i \neq j} e_i e_j > \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon > 0 \quad \text{or} \quad \sum_{i \neq j} e_i e_j < \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon < 0
\]  

(5-19)

It is very difficult to get a numerical approximation for the left hand side of equation 5-19 as it involves the elements of the eigenvector. The right-hand side provides some useful information as to why the pattern of observed NPD matrices were observed for all available cases and imputing means (tables 5-15 and 5-20). Examining equation 5-19, more NPD matrices result for smaller values of \( \lambda \) and larger values of \( \varepsilon \). According to the model developed in chapter 4, there will be more eigenvalues around the lower bound of 0.1 as \( \phi \) increases. Table 5-21 shows the 10 smallest eigenvalues for each combination of \( p \) and \( \phi \). It is evident from this table as \( \phi \) increases, more eigenvalues are around 0.1 and are therefore more susceptible to becoming NPD.

Equation 5-19 shows that the number of NPD matrices will increase for smaller values of \( \lambda \). As the number of eigenvalues that are close to 0.1 increase as \( \phi \) increases, it is expected that the number of NPD matrices will increase as \( \phi \) increases. The exception to this pattern in increasing numbers of NPD matrices is for all available cases at \( \phi = 0.9 \). No NPD matrices result for this situation. The graphs in figures 5-5 to 5-8 might show a reason for this. According to these graphs, the error of this technique (defined in this section as \( \varepsilon \)) is very small at \( \phi = 0.9 \). According to
equation 5-19, there will be less NPD matrices for smaller values of \( \varepsilon \). It appears as though the error becomes small enough at \( \varphi = 0.9 \) to reduce the influence of the small eigenvalues. Hence, less NPD matrices result at this level.

Table 5-21  Last Ten Eigenvalues for the Model Developed in Section 4-7

<table>
<thead>
<tr>
<th>Avg Intercorrelation ( \varphi )</th>
<th>Number of Variables</th>
<th>Number of Variables</th>
<th>Number of Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( p = 15 )</td>
<td>( p = 25 )</td>
<td>( p = 50 )</td>
</tr>
<tr>
<td>0.1</td>
<td>1.10, 1.00, 0.91, 0.83, 0.76, 0.69, 0.63, 0.58, 0.53, 0.49</td>
<td>0.60, 0.55, 0.50, 0.46, 0.43, 0.39, 0.37, 0.34, 0.31, 0.29</td>
<td>0.17, 0.16, 0.15, 0.15, 0.14, 0.14, 0.14, 0.13, 0.13, 0.13</td>
</tr>
<tr>
<td>0.3</td>
<td>0.78, 0.57, 0.44, 0.33, 0.26, 0.21, 0.18, 0.16, 0.14, 0.13</td>
<td>0.13, 0.12, 0.12, 0.11, 0.11, 0.11, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
</tr>
<tr>
<td>0.5</td>
<td>0.31, 0.21, 0.15, 0.13, 0.11, 0.11, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
</tr>
<tr>
<td>0.7</td>
<td>0.12, 0.11, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
</tr>
<tr>
<td>0.9</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
<td>0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10, 0.10</td>
</tr>
</tbody>
</table>

It would be difficult to get meaningful results from the correlation difference test as the presence of NPD's places doubt on the validity of the results. Hence, Jennrich's test is abandoned for higher levels of \( n \) and \( p \).

5-4 Summary of Study and Results

This chapter has been concerned with a technique that best estimates a correlation matrix when there are missing values present. Two of the methods discussed in chapter 3 (the iterative principal component method and the EM algorithm) require an initial estimate of the correlation matrix. Complete cases only, all available cases and
imputing means were investigated as to their abilities to estimate the correlation matrix.

Data matrices were randomly simulated as described in chapter 4. Each matrix varied with respect to the levels of the parameters n, p, φ and k. Five data matrices were generated for each combination of n, p and φ. Data were randomly deleted 20 times according to the level of k to represent missing values. This resulted in 100 (5×20) data matrices for each combination of n, p, φ and k. There was some concern that deleting data 20 times for 5 data matrices would have a large reduction in the standard deviation of the root-mean-square errors $E$, compared to a design that simply generated 100 (100×1) matrices. The standard deviations of $E$ for the 5×20 design were smaller than the 100×1 design; however, the time saved to generate these matrices far outweighed this reduction.

Gleason and Staelin’s (1975) measure $E_\alpha = \sqrt{\frac{R_\alpha - R^2}{p(p-1)}}$ was used to measure the difference between the estimated and the original (no missing values) correlation matrices. Plots of the errors quickly revealed that complete cases only was performing poorly compared to the other two techniques and was removed from the study. This agreed with the work done by Timm (1970) which indicated that complete cases only performed worse than the other techniques when $p > 5$. Complete cases only was disregarded from further investigations.
The root-mean-square error $E$ was regressed against the variables $n$, $p$, $\varphi$ and $k$ to investigate the conditions where the techniques performed best. The final multiplicative regression models are shown below:

$$E_{im} = 1.855n^{-0.513}k^{0.539}(1 - \varphi)^{0.238}$$
$$E_{ac} = 2.396n^{-0.531}k^{0.537}(1 - \varphi)^{0.845}$$

From the above equations, the criteria as to when one technique outperforms the other was determined. All available cases performed better than imputing means to predict $R$ when $\varphi > 1 - 0.656n^{0.030}$ (refer to equation 5-13).

Jennrich’s (1970) test of correlation matrices produced more significant results when there were ill-conditioned matrices being examined. More ill-conditioned matrices were observed for imputing means as $\varphi$ increased. All available cases produced more ill-conditioned matrices at $\varphi = 0.5$ and 0.7. Hence, more significant differences were observed using Jennrich’s test at these levels. These findings were explained by assuming that the error produced from a technique produced the same size perturbation $\varepsilon$ to all off-diagonal elements of the correlation matrix $R$.

It was shown that NPD matrices result when:

$$\sum_{i \neq j} e_i e_j > \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon > 0 \text{ or } \sum_{i \neq j} e_i e_j < \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon < 0$$

where: $e_i = i^{th}$ component of the eigenvector $e$ of $R$.

This equation assisted in explaining the observed pattern in ill-conditioned matrices for all available cases and imputing means. The equation also explained the unusual amount of significant results from Jennrich’s test.
Chapter 6

Analysis of Techniques to Estimate Loadings and Specific Variances

6-1 INTRODUCTION

The main purpose of this chapter is to analyse the abilities of various techniques to estimate factor loadings and specific variances. The six techniques discussed in chapter 3: namely, complete cases only, all available cases, imputing means, singular value decomposition, the iterative principal component method and the EM algorithm are investigated. Root-mean-square measures are introduced to compare the original and estimated factor loadings and specific variances. All of the techniques are compared using graphs and multiplicative regression models. As well as the estimation of factor loadings and specific variances, the computing time it takes to perform each technique is also considered.
6-2 STUDY DESIGN

Only a couple of authors have designed studies to investigate missing values in a factor analysis. Finkbeiner (1979) studied maximum likelihood factor analysis and investigated complete cases only, imputing means, all available cases, a maximum likelihood approach and iterative regression and principal component methods. He worked with 50 replicates of samples of size 64. Two factors, 6 variables and 2 'missingness' patterns were used. Hendricks Brown (1983) examined a single factor model and used asymptotic methods of comparison as opposed to Monte Carlo methods. He was particularly interested in a few patterns of missing data and the covariance structure. Both of these authors were more interested in sampling properties than examining various levels of parameters.

This thesis generates random multivariate data that are typical of that obtained by researchers using factor analyses and examines as many conditions as are feasible. The procedures to perform this random data generation are described in chapter 4. The number of cases n, the number of variables p, the average intercorrelation \( \varphi \) and the percentage of deletions \( k \) are varied according to the levels shown below:

\[
\begin{align*}
n &= 50, 100, 200 \\
p &= 15, 25, 50 \\
\varphi &= 0.3, 0.5, 0.7 \\
k &= 1\%, 5\%, 10\%
\end{align*}
\]

These levels are similar to those used in chapter 5 to investigate which technique is better at estimating the initial correlation matrix. It is very time-consuming to perform all 6 techniques for each level of the four parameters above (refer to section
6-3.4). As a result, the number of levels of \( \phi \) and \( k \) are reduced from those used in chapter 5. It would be more unlikely to have real life data with \( \phi = 0.1 \) or 0.9 than the other levels, so these levels are removed. Section 5-3.2 shows that there is not much difference in the root-mean-square errors when there are 1% and 2% missing values. The root-mean-square errors in that section refer to the estimation of correlation matrices, but should still be a good indication of the errors in factor loadings and specific variances. Hence, \( k = 2\% \) is also removed from the levels used in chapter 5 to reduce the number of simulations.

For each level of \( n \), \( p \), \( \phi \) and \( k \), five data matrices are simulated as shown in chapter 4. Rather than generating more data matrices, the data from chapter 5 are simply re-used for the levels shown above. For each of the 5 data matrices, data are randomly deleted 10 times as described in section 5-2.1. The EM algorithm is costly to run in computing time (refer to section 6-3.4) and 10 deletion patterns are not always possible to complete in a reasonable amount of time. The number of deletion patterns used for the EM algorithm for the various values of \( n \), \( p \) and \( k \) are shown in the table below.

**Table 6-1  Number of Deletion Patterns used for the EM Algorithm**

<table>
<thead>
<tr>
<th></th>
<th>( n = 50 )</th>
<th></th>
<th>( n = 100 )</th>
<th></th>
<th>( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>( k = 1% )</td>
<td>( k = 5% )</td>
<td>( k = 10% )</td>
<td>( k = 1% )</td>
<td>( k = 5% )</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>50</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>
Chapter 6 – Analysis of Techniques to Estimate Loadings and Specific Variances

An initial estimate of $\mu$ and $S$ is required for the EM algorithm. The matrix of means $\mu$ is initially estimated by simply using whatever measurements are observed for each variable. The results from chapter 5 indicate that all available cases is the better technique at estimating $R$ when $\phi > 1 - 0.656n^{0.030}$. The smallest value of $\phi$ used in this section is 0.3. Comparing this rule with the values of $n$ (50, 100 and 200) used in this chapter, all available cases is always better at estimating $R$ than imputing means. As this technique is a better estimator of $R$, it seems reasonable to believe that generally it is a better estimator of $S$. Hence, all available cases is used for the initial estimate of $S$.

For smaller data matrices and small numbers of missing values, the singular value decomposition works very well. Unfortunately, for the large levels used for $n$, $p$ and $k$ in this thesis, the singular value decomposition did not converge for all situations. For a particular level of $n$ and $p$, this technique converges for some deletion patterns and not for others. Even after implementing the changes suggested by Krzanowski (personal communication in July 1998), the imputed values still tend to cycle rather than converge. The lack of consistent results makes it difficult to work with this technique in a large simulation study. Therefore, this technique is not used for the remainder of this thesis.

6-3 COMPARING TWO FACTOR ANALYSES

There is still the question of how to compare the original factor analysis based on a complete data set with the factor analyses obtained from the various techniques on data with missing values. To compare the results from different factor analyses,
Finkbeiner (1979) used three pieces of information. The sampling means and dispersions of $\mu$ and $\Psi$ were used as the first comparison criteria. The large number of simulations used in this thesis would mean that the number of means and dispersions to compare would become unwieldy. The mean square error of the specific variance $\Psi$ was also used. He did not use the mean square error of factor loadings $L$ due to the fact that most researchers are willing to rotate common factors.

Finkbeiner also used the average of the squared multiple coefficient of congruence defined by Tucker, Koopman and Linn (1969). This congruence coefficient is defined below:

$$\text{Coefficient of Congruence} = \frac{\sum_i a_{im} f_{ip}}{\sqrt{\sum_i a_{im}^2} \sqrt{\sum_i f_{ip}^2}} \quad (6-1)$$

where: $a_{im} = m^{th}$ column vector of factor loadings of $a$

$f_{ip} = p^{th}$ column vector of factor loadings of $f$

$a_{im} = i^{th}$ factor loading of $a_{m}$

$f_{ip} = i^{th}$ factor loading of $f_{p}$

According to Finkbeiner (1979), the sampling means and standard deviations of congruence coefficients reflect the distributions they summarise. The coefficient of congruence compares two column vectors of factor loadings of the original and estimated factor analyses. Tucker, Koopman and Linn (1969) indicate that equation 6-1 is analogous to a coefficient of correlation, but not identical, as the loadings are not converted to deviations from their means.
In this thesis, equation 6-1 is difficult to use as a comparison of a large number of simulations over a range of levels for n, p, φ and k as well as 50 replications. For example, consider one such simulation of a factor analysis that contains five factors and 50 replicates. For each replicate, five coefficients of congruence would have to be determined (one for each factor). Each of the five coefficients could be averaged over the 50 replicates. It is theoretically unappealing to then average the five coefficients for each factor to represent the situation for any combination of n, p, φ and k. This averaging would have to be done to compare factor analyses with differing numbers of factors. Equation 6-1 is useful for small simulation studies to get a feel for the similarity between vectors of factors, but is very difficult to work with in this large simulation study.

To compare the original and estimated factor analyses in this thesis, only two measures are used. The root-mean-square errors of both the factor loadings L and specific variances Ψ are used as the comparative measures. The root-mean-square error of Ψ is probably of more interest to researchers as specific variances are unaltered by rotations. The unrotated factor loadings are also of interest in this thesis as chapter 7 examines the theoretical nature of the changes in these loadings. Hence, the root-mean-square error of L is also considered. Examining the rotated loadings is beyond the scope of this thesis.
The root-mean-square measures of factor loadings and specific variances are shown below:

\[
FL_\alpha = \sqrt{\frac{\|\hat{L}_\alpha - L\|^2}{p^2 f^2}} \quad (6-2)
\]

where: \( \hat{L}_\alpha \) = estimated factor loading matrix from the technique \( \alpha \)

\( L \) = original factor loading matrix based on complete data set

\( p \) = number of variables

\( f \) = number of factors

\[
SV_\alpha = \sqrt{\frac{\|\hat{\Psi}_\alpha - \Psi\|^2}{p^2}} \quad (6-3)
\]

where: \( \hat{\Psi}_\alpha \) = estimated specific variances from the technique \( \alpha \)

\( \Psi \) = original specific variances

\( p \) = number of variables

Appendix D contains the computer program that determines the original and estimated factor loadings and specific variances. The root-mean-square measures above are also determined in this program.
RESULTS

6.4.1 Graphical Comparison of Factor Loadings and Specific Variances

All of the techniques discussed in this chapter are compared as to their ability to estimate the original factor loadings from the complete data set. The indices FL and SV are initially plotted versus all of the different parameters n, p, φ and k to visualise what is occurring and to investigate any interactions. The plots of FL are shown in figures 6-1 to 6-3. Note that the abbreviations im, em, pc and ac refer to imputing means, EM algorithm, iterative principal component method and all available cases respectively. Each figure has nine plots with three rows and three columns. The sample size increases in each column moving from left to right. The number of variables increases moving from top to bottom. Each plot displays FL for four techniques versus the average intercorrelation φ. The three figures differ in the percentage of deletions. Similar plots of SV are shown in figures 6-4 to 6-6.

Once again, the technique of complete cases only is not displayed on these plots as it performs far worse than all of the other techniques. Complete cases only produces root-mean-square FL that are between 5 and 50 times larger than the other techniques. Because of the scales involved on the plots, if complete cases only is included then it is hard to distinguish among the other techniques. The scale is consistent on each plot for each figure, but differs within figures. The exception in each figure is the bottom left plot (g). The scale for this plot is changed to include all of the four techniques.
Figure 6.1 Comparison of Four Techniques to Estimate Factor Loadings (k = 1%).
Figure 6-2  Comparison of Four Techniques to Estimate Factor Loadings (k = 5%).
Figure 6-3 Comparison of Four Techniques to Estimate Factor Loadings (k = 10%).
Figure 6-4  Comparison of Four Techniques to Estimate Specific Variances (k = 1%).
Figure 6-5  Comparison of Four Techniques to Estimate Specific Variances 
(k = 5%).
Figure 6-6  Comparison of Four Techniques to Estimate Factor Loadings
(k = 10%).
Generally, for all plots and techniques shown in figures 6-1 to 6-3, FL decreases as $n$ and $p$ increase. FL also increases as $k$ increases. The iterative principal component method deteriorates more than the other techniques do as $k$ increases. Imputing means is usually the worst technique at estimating the original factor loadings. All available cases and the EM algorithm outperform imputing means for most situations studied.

The EM algorithm performs better than all the other techniques except when $n = 50$, $p = 25$; $n = 50$, $p = 50$; and $n = 100$, $p = 50$. In fact, when $n = 50$ and $p = 50$ (plot $g$ in figures 6-1 to 6-3), all of the other techniques perform far better than the EM algorithm. The matrices at these levels are not of full rank and are susceptible to ill-conditioning when some techniques are applied to handle missing values. Refer to section 5-3.5 for the discussion of ill-conditioned and singular matrices.

When data are removed from the matrices with $n = 50$ and $p = 50$, imputing means and all available cases produce estimated correlation matrices that are NPD. As the EM algorithm and the iterative principal component method both use all available cases to estimate the initial correlation or covariance matrix, these techniques also produce NPD matrices. The EM algorithm involves the inverse of the partitioned covariance matrix $\Sigma_{22}$ that contains the variables with all observations present. It is, therefore, particularly susceptible to ill-conditioned matrices. All of the matrices with $n = 50$ and $p = 50$ result in NPD matrices that have to be smoothed as discussed in section 3-9. As these matrices will contain several eigenvalues that are close to zero, the term 'very ill-conditioned' is used to distinguish these matrices from the rest.
For the data matrices with $n = 50$, $p = 25$ and $n = 100$, $p = 50$, NPD correlation matrices result when $k \geq 5\%$. This explains why the EM algorithm performs well in figures 6-1d and 6-1h and deteriorates in figures 6-2d, 6-2h, 6-3d and 6-3h. Similar patterns are seen in figures 6-4d, 6-4h, 6-5d, 6-5h, 6-6d and 6-6h. The ill-conditioned matrices present for the matrices with $n = 50$, $p = 25$ and $n = 100$, $p = 50$ are referred to as ‘moderately ill-conditioned’. There are not as many NPD’s present compared to $n = 50$ and $p = 50$, but there are more NPD’s than the other matrices where the EM algorithm performed well. Very and moderately ill-conditioned are not quantified with respect to a criterion to distinguish between these two situations (although this is acknowledged as being important in section 8-6). These terms are used loosely to represent the two situations where there are NPD and ill-conditioned matrices present.

The researcher is probably very interested in detecting situations where the EM algorithm performs very poorly compared to the other techniques. Researchers would also be interested in detecting the situations where the EM algorithm is just starting to perform slightly poorer than the other techniques. Unfortunately, no information was recorded from this study on the eigenvalues for these situations. These might have revealed some information about a suitable criterion to identify when the EM algorithm performs very well, moderately or very poorly. Refer to section 8-6 for the discussion on further research about this criterion.

The plots of FL in figures 6-1 to 6-3 reveal that the EM algorithm performs better than the other techniques except for very or moderately ill-conditioned matrices (plots g, d and h). For both of these situations, all available cases is the best technique when there are more than 5\% missing values. For very ill-conditioned matrices, the
principal component method performs best when there are 1% missing values. The EM algorithm performs better than all the other techniques for moderately ill-conditioned matrices when there are 1% missing values.

The plots of SV are shown in figures 6-4 to 6-6. The plots of SV have similar appearances to the plots of FL. The performance of the iterative principal components method deteriorates as $k$ increases compared to the other techniques. Imputing means is usually the worst at 1% deletions, while the iterative principal component method and imputing means are the poorest performing techniques at 5% and 10% deletions. The EM algorithm again performs better than all the other techniques except when there are very or moderately ill-conditioned matrices. For very ill-conditioned matrices, all available cases is the best technique to use. The EM algorithm still performs best for moderately ill-conditioned matrices when there are 1% and 5% missing values. All available cases is the best technique for 10% missing values.

6-4.2 Regression Analysis of Factor Loadings

The root-mean-square of the factor loadings FL for each technique is regressed against the predictor variables $n$, $p$, $\phi$ and $k$. The data are analysed using stepwise regression on SPSS. As in section 5-3.3, linear models are initially fitted to the data. Again, the residuals versus predicted values plots suggest that the assumption of homogeneity of variance may not be satisfied. Multiplicative regression models (natural logarithm transformations) are used to overcome these problems. The details of these regression models are found in appendix F-3A to F-3F.
The final regression equations are shown below for each of the techniques:

\[
FL_{pc} = 1.530n^{-0.399}p^{-0.729}(1-\varphi)^{0.521}k^{0.622} \quad (R^2 = 0.915) \quad (6-4)
\]
\[
FL_{nm} = 1.806n^{-0.530}p^{-0.697}(1-\varphi)^{0.255}k^{0.541} \quad (R^2 = 0.922) \quad (6-5)
\]
\[
FL_{de} = 2.199n^{-0.535}p^{-0.726}(1-\varphi)^{0.613}k^{0.542} \quad (R^2 = 0.929) \quad (6-6)
\]
\[
FL_{ee} = 1.551n^{-0.492}p^{0.231}(1-\varphi)^{0.846}k^{0.830} \quad (R^2 = 0.859) \quad (6-7)
\]
\[
FL_{en} = 3.924n^{-0.809}p^{-0.518}(1-\varphi)^{0.835}k^{0.580} \quad (R^2 = 0.892) \quad (6-8)
\]

Each of the variables for all five techniques are significant in the above regression models. The residual plots for each of the models in equations 6-4 to 6-8 do not reveal any violations of assumptions. The results for FL_{en} do not include any information when n = 50 and p = 50. Figures 6-1 to 6-3, reveal that this technique is unduly influenced by very ill-conditioned matrices at these levels. Appendix F-3E shows the regression analysis when these levels are included. The \( R^2 \) value is only 0.792 and the residuals versus predicted values plot suggests a violation of the homogeneity of variance. The problems with these residuals are removed when these levels are not used in the regression analysis. The violations of regression assumptions and the fact that the EM algorithm performs unusually poorly at n = 50 and p = 50, seem to justify the removal of the observations at these levels to get a clearer regression analysis for the remaining situations.

An examination of equations 6-4 to 6-8 reveal some interesting facts. All techniques improve as n increases with the EM algorithm being the most responsive to changes in n. Complete cases only is the only technique to be adversely affected by increases in p. All of the other techniques improve for larger values of p. Complete cases only and the EM algorithm are more responsive to changes in 1-\( \varphi \) than the other
techniques. Complete cases only is the most adversely affected technique by increases in k, followed by the iterative principal component method.

6-4.3 Regression Analysis of Specific Variances

Multiplicative regression models again seem to be the most appropriate to use for modelling the root-mean-square specific variances SV. The details of the regression analysis of each technique versus n, p, \( \varphi \) and k are found in appendix F-4A to F-4E.

The final regression equations for each technique are shown below:

\[
\begin{align*}
SV_{pc} &= 0.433 n^{-0.236} p^{-0.243} (1-\varphi)^{0.942} k^{0.710} & (R^2 = 0.886) \\
SV_{im} &= 0.819 n^{-0.477} p^{-0.424} (1-\varphi)^{0.052} k^{0.554} & (R^2 = 0.910) \\
SV_{sc} &= 1.309 n^{-0.503} p^{-0.483} (1-\varphi)^{0.800} k^{0.533} & (R^2 = 0.891) \\
SV_{cc} &= 2.071 n^{-0.503} p^{0.052} (1-\varphi)^{0.975} k^{0.794} & (R^2 = 0.903) \\
SV_{em} &= 3.121 n^{-0.576} p^{-0.543} (1-\varphi)^{0.961} k^{0.712} & (R^2 = 0.885)
\end{align*}
\]

As in the previous section, the regression analysis for the EM algorithm does not include n = 50 and p = 50. The regression analysis when these levels are included is found in appendix F-4E. The \( R^2 \) value is only 0.712 as compared to 0.885 in equation 6-13. The residuals are not normally distributed and the residuals versus predicted values plot suggests a violation of the homogeneity of variance. The standardized residuals seem to vary over a wider range for small values of n and \( \varphi \) and large values of p. The residual problems are removed when these levels are not used in the regression analysis.
The regression equations 6-9 to 6-13 indicate some interesting points. The iterative principal component method is the technique that is least influenced by n when examining specific variances. Complete cases only is the only technique that is unduly influenced by increases in p. The increase to SV as a result of increases in p appears to be less than the corresponding increases to FL for complete cases only. The coefficients of imputing means and all available cases are very similar with the exception of φ. Imputing means is only slightly affected by changes in φ. The iterative principal component method, complete cases only and the EM algorithm are most affected by increases in k.

6-4.4 Analysis of Computing Time

The final criterion used to evaluate the performance of all the techniques to handle missing values in factor analysis is to compare the computer time used by these techniques. The techniques are timed on how long it takes for them to produce a correlation matrix that will be input into a factor analysis. The following tables summarise the time taken to perform each of these techniques. Each value in the tables are an average of 150 times. There are 50 replications and the times are also averaged over the three levels of φ. Consider table 6-2. The top left-hand time of 17 seconds refers to the average computer time taken by the iterative principal component method for n = 50, p = 15 and k = 1%. This is averaged over all levels of φ and the 50 replications.
Table 6-2  Computing Time Required for Iterative Principal Component Method

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>k = 1%</td>
<td>k = 5%</td>
<td>k = 10%</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>17</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>44</td>
<td>47</td>
<td>69</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>157</td>
<td>164</td>
<td>168</td>
</tr>
</tbody>
</table>

Table 6-3  Computing Time Required for Imputing Means

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>k = 1%</td>
<td>k = 5%</td>
<td>k = 10%</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>15</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>62</td>
<td>63</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 6-4  Computing Time Required for All Available Cases

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>k = 1%</td>
<td>k = 5%</td>
<td>k = 10%</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>27</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>107</td>
<td>104</td>
<td>98</td>
</tr>
</tbody>
</table>

Table 6-5  Computing Time Required for Complete Cases Only

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>k = 1%</td>
<td>k = 5%</td>
<td>k = 10%</td>
</tr>
<tr>
<td>15</td>
<td></td>
<td>0.8</td>
<td>0.6</td>
<td>0.5</td>
</tr>
<tr>
<td>25</td>
<td></td>
<td>1.7</td>
<td>1.1</td>
<td>0.8</td>
</tr>
<tr>
<td>50</td>
<td></td>
<td>4.9</td>
<td>2.8</td>
<td>2.6</td>
</tr>
</tbody>
</table>
Table 6-6  Computing Time Required for EM Algorithm

<table>
<thead>
<tr>
<th>p</th>
<th>n = 50</th>
<th>n = 100</th>
<th>n = 200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>k = 1%</td>
<td>k = 5%</td>
<td>k = 10%</td>
</tr>
<tr>
<td>15</td>
<td>21</td>
<td>73</td>
<td>224</td>
</tr>
<tr>
<td>25</td>
<td>96</td>
<td>1071</td>
<td>1737</td>
</tr>
<tr>
<td>50</td>
<td>2403</td>
<td>7317</td>
<td>7376</td>
</tr>
</tbody>
</table>

The non-iterative techniques of complete cases only, imputing means and all available cases take the least time in that order. The iterative principal component method is the next fastest and the EM algorithm takes much more time than the other techniques, particularly for large values of n, p and k.

Many articles have been written about methods to accelerate the slow convergence of the EM algorithm. It is beyond the scope of this thesis to explore faster methods except to recognise that they exist. The reader is referred to Jamshidian and Jennrich (1997) and Meng and Van Dyk (1997) for a summary and comparison of these quicker methods. Makuha, Pegram, Sparks and Zucchini (1997a,b) have also developed a pseudo-EM algorithm which excludes a correction term in the EM algorithm. This method is heuristic and has not been proven to always converge. The authors claim that the pseudo-EM algorithm is, on average, 21 times faster than the EM algorithm for the examples that they examined.

Multiplicative regression models are again used to model the computing time that it takes to complete each technique. Only the average time for the 50 replicates of each combination of n, p, φ and k is recorded. The regression analysis is still used realising that the variation in time for each combination is reduced.
The regression equations for the time to complete each technique are shown below.

The regression analyses are found in appendix F-5A to F-5E.

\[
T_{pc} = 0.008 n^{0.767} p^{1.807} (1 - \varphi)^{0.206} k^{0.038} \quad (R^2 = 0.996) \quad (6-14)
\]

\[
T_{im} = 0.001 n^{0.895} p^{2.004} \quad (R^2 = 0.999) \quad (6-15)
\]

\[
T_{ne} = 0.002 n^{0.763} p^{2.046} k^{-0.023} \quad (R^2 = 0.998) \quad (6-16)
\]

\[
T_{ce} = 0.001 n^{0.542} p^{1.265} (1 - \varphi)^{-0.234} k^{-0.257} \quad (R^2 = 0.896) \quad (6-17)
\]

\[
T_{em} = 0.107 p^{3.332} (1 - \varphi)^{0.603} k^{0.594} \quad (R^2 = 0.912) \quad (6-18)
\]

There are some concerns about the residual plots in appendix F-5A to F-5E. Some of the techniques have residuals that are not normally distributed and the plots of k versus standardised residuals suggest that the homogeneity of variance may be violated. While taking account of the residual problems, the time regressions are still used as exploratory rather than explanatory tools.

The computer time taken for the iterative principal component method increases for larger values of n, p and k. The only variables that are significant for the time taken for imputing means are n and p. The time for all available cases is not influenced by \( \varphi \). For larger values of k, the time to perform complete cases only reduces as would be expected. The time for the EM algorithm is not affected by n.

6-5 Summary of Results

The ability of various techniques to handle missing values in a factor analysis situation has been examined in this chapter. Complete cases only, all available cases,
imputing means, singular value decomposition, the iterative principal component method and the EM algorithm were all investigated for their ability to reproduce the original factor loadings and specific variances. The computing time that the techniques used to compute the correlation matrices that will be input into a factor analysis were also compared. Unfortunately, the singular value decomposition did not converge for all of the deletion patterns used for the levels of n and p. This technique was removed because of the inconsistent convergence.

Root-mean-square measures (FL and SV) were used to measure the difference between original and estimated factor loadings and specific variances. Plots of FL and SV versus φ were drawn over the various values for n, p and k to get a visual representation of what was happening. Multiplicative regression models were used to investigate the influence that the parameters had on each of the techniques.

Complete cases only was the worst technique for all criteria tested except the time it took to determine the correlation matrix. The maximum time to perform this technique was only about 13 seconds. However, the short duration and ease of this technique should be seriously weighed against the fact that FL and SV were between 5 and 50 times larger than the other techniques. Both FL and SV increased with increases in p and k and decreased with increases in φ and n.

The plots of the iterative principal component method revealed that it did better at estimating factor loadings than specific variances. It performed well for both FL and SV when there were 1% of the data missing, but deteriorated more than other
techniques for higher proportions of missing data. Comparing the coefficients of k in
the regression analyses with those of other techniques supported this notion.

The plots of imputing means and all available cases for both FL and SV revealed that
for every combination of n, p, \( \varphi \) and k, all available cases performed better than
imputing means. The regression analyses of FL and SV reveal that both sets of
coefficients for n, p and k are very similar for these two techniques. The constants
and coefficients of 1-\( \varphi \) differ for both techniques. The average intercorrelation \( \varphi \)
affects all available cases more than imputing means. The time taken for all available
cases takes about 50% more time than imputing means. However, for the largest
matrices studied in this thesis, the maximum time to perform all available cases was
only 5.37 minutes.

Plots of FL and SV reveal that the EM algorithm outperformed all other techniques
for most situations. The notable exception was when \( n = 50 \) and \( p = 50 \), where the
EM algorithm performed far worse than all other techniques. The EM algorithm was
very susceptible to ill-conditioned matrices at these levels. For these levels, the better
techniques to estimate factor loadings were the iterative principal component method
and all available cases. All available cases was better at estimating the specific
variances when ill-conditioned matrices were present. For larger sized matrices, the
EM algorithm was very time-consuming to perform. For example, it would take
approximately 4.5 hours for the EM algorithm to determine the correlation matrix for
\( n = 100, p = 50 \) and \( k = 10\% \).
Overall, the EM algorithm should be used for missing values in a factor analysis when time is not of concern to a researcher. This technique performed best for most conditions studied. If there is concern about the data producing a correlation matrix that is ill-conditioned or time is of concern, then the best choice is either all available cases or the iterative principal component method. All available cases will generally provide a better estimate of specific variances. The iterative principal component method better estimates the original factor loadings when 1% of the data are missing and all available cases is the better estimator when 5% and 10% of the data are missing.

The above recommendations are very general based on the conditions studied. If there was not much difference between two techniques, then the technique that applies to most situations is suggested. A researcher is recommended to use equations 6-4, 6-6, 6-8, 6-9, 6-11 and 6-13 for the conditions studied and select the technique that produces the smallest values for SV and FL (refer to the recommendations for researchers in section 8-5).
Chapter 7

Derivatives of Factor Loadings

7-1 INTRODUCTION

Some theoretical considerations regarding the effects that changes in the correlation matrix have on the loading matrix are examined in this chapter. All of the techniques discussed to handle missing values produce changes in the correlations among all the variables. For example, it was shown in section 3-3 that imputing means for missing values reduces the variances and covariances. As a result of imputing means, the correlations were shown to be altered by a factor of \( \frac{\sqrt{(n^{(j)} - 1)(n^{(k)} - 1)}}{n^{(j^2) - 1}} \). The effect on factor loadings as a result of multiple changes to the correlation matrix is very difficult to determine. The problem can be greatly simplified by using derivatives and considering the effect on factor loadings due to a small change in the correlation between two variables. This is simplistic in terms of the effect that missing value techniques have on the correlation matrix. However, it is still a useful exercise to
examine the complicated form of the expression for this simple situation. Analysing factor loadings using derivatives may also provide some information on the effect that changes in the correlation matrix have on loadings as well as which loadings are more influenced.

7-2 CHANGES IN EIGENVALUES WITH RESPECT TO CHANGES IN THE CORRELATION MATRIX

The next three sections are concerned with the changes in factor loadings as a result of a change in the correlation between two variables. Let \( R \) be the correlation matrix and \( r_{ij} \) and \( r_{ji} \) be the elements of \( R \) that change by a small amount. The change in loadings as a result of the change in the correlation between two variables is given by

\[
\frac{\partial L}{\partial r_{ij}}
\]

Define \( \lambda \) to be an eigenvalue of \( R \) and \( e \) an associated eigenvector of unit length.

As \( L = f(\lambda, e) \), the first step in defining \( \frac{\partial L}{\partial r_{ij}} \) is to determine the expression \( \frac{\partial \lambda}{\partial r_{ij}} \).

Differentiating \( Re = \lambda e \) gives:

\[
\frac{\partial}{\partial r_{ij}} (Re) = \frac{\partial}{\partial r_{ij}} (\lambda e)
\]

\[
R \frac{\partial e}{\partial r_{ij}} + \frac{\partial R}{\partial r_{ij}} e = \lambda \frac{\partial e}{\partial r_{ij}} + \frac{\partial \lambda}{\partial r_{ij}} e
\]

(7-1)

Pre-multiplying the above expression by \( e' \),

\[
e' R \frac{\partial e}{\partial r_{ij}} + e' \frac{\partial R}{\partial r_{ij}} e = e' \lambda \frac{\partial e}{\partial r_{ij}} + e' \frac{\partial \lambda}{\partial r_{ij}} e
\]

(7-2)
The eigenvectors have unit length, that is \( \mathbf{e}' \mathbf{e} = 1 \). Substituting this equation into equation 7-3 gives:

\[
\mathbf{e}' \mathbf{R} \frac{\partial \mathbf{e}}{\partial r_{ij}} + \mathbf{e}' \frac{\partial \mathbf{R}}{\partial r_{ij}} \mathbf{e} = \mathbf{e}' \lambda \frac{\partial \mathbf{e}}{\partial r_{ij}} + \frac{\partial \lambda}{\partial r_{ij}} \quad (7-4)
\]

Differentiating \( \mathbf{e}' \mathbf{e} = 1 \) gives:

\[
\frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} + \mathbf{e}' \frac{\partial \mathbf{e}}{\partial r_{ij}} = 0 \quad (7-5)
\]

But,

\[
\left( \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} \right)' = \mathbf{e}' \left( \frac{\partial \mathbf{e}'}{\partial r_{ij}} \right)' = \mathbf{e}' \frac{\partial \mathbf{e}}{\partial r_{ij}} \quad (7-6)
\]

Both of the terms in equation 7-5 are scalars. Hence, the transpose of the first term in equation 7-5 gives:

\[
\left( \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} \right)' = \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e}
\]

\[
\Rightarrow \mathbf{e}' \frac{\partial \mathbf{e}}{\partial r_{ij}} = \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} \quad (7-7)
\]

Substituting the above equation into equation 7-5 gives:

\[
2 \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} = 0
\]

\[
\Rightarrow \frac{\partial \mathbf{e}'}{\partial r_{ij}} \mathbf{e} = 0 \quad (7-8)
\]
Similarly,

$$e' \frac{\partial e}{\partial r_{ij}} = 0 \tag{7-9}$$

Substituting equation 7-9 into equation 7-4 gives:

$$e' R \frac{\partial e}{\partial r_{ij}} + e' \frac{\partial R}{\partial r_{ij}} e = \frac{\partial \lambda}{\partial r_{ij}} \tag{7-10}$$

The right-hand term of the above equation is a scalar. Using the fact that $R$ is symmetric gives:

$$\frac{\partial \lambda}{\partial r_{ij}} = \left( e' R \frac{\partial e}{\partial r_{ij}} + e' \frac{\partial R}{\partial r_{ij}} e \right)'$$

$$= \frac{\partial e'}{\partial r_{ij}} R e + e' \frac{\partial R}{\partial r_{ij}} e \tag{7-11}$$

Substituting equation 7-1 into the above equation gives:

$$\frac{\partial \lambda}{\partial r_{ij}} = \frac{\partial e'}{\partial r_{ij}} \lambda e + e' \frac{\partial R}{\partial r_{ij}} e$$

$$= \lambda \frac{\partial e'}{\partial r_{ij}} e + e' \frac{\partial R}{\partial r_{ij}} e \tag{7-12}$$

Substituting equation 7-8 into the above equation gives:

$$\frac{\partial \lambda}{\partial r_{ij}} = e' \frac{\partial R}{\partial r_{ij}} e \tag{7-13}$$
Now $\frac{\partial R}{\partial r_{ij}}$ is a matrix with 1 in the \{i,j\} and \{j,i\} positions and zero elsewhere.

Hence,

$$\frac{\partial \lambda}{\partial r_{ij}} = 2e_i e_j \quad (7-14)$$

It is important in the next sections to distinguish between the eigenvalue $\lambda$ and eigenvector $e$ that are being examined for changes and the other eigenvalues and eigenvectors. Denote the eigenvalue of particular concern $\lambda_k$ and its corresponding eigenvector $e_k$. Equation 7-14 becomes:

$$\frac{\partial \lambda_k}{\partial r_{ij}} = 2e_{ki} e_{kj} \quad (7-15)$$

where: $e_{ki} = i^{th}$ component of $e_k$

7-3 \text{ CHANGES IN EIGENVECTORS WITH RESPECT TO CHANGES IN THE CORRELATION MATRIX}

Before determining $\frac{\partial L}{\partial r_{ij}}$, an expression needs to be derived for $\frac{\partial e_k}{\partial r_{ij}}$. Again, consider the eigenvalue $\lambda_k$ and its corresponding eigenvector $e_k$.

From equation 7-2:

$$R \frac{\partial e_k}{\partial r_{ij}} + R \frac{\partial R}{\partial r_{ij}} e_k = \lambda_k \frac{\partial e_k}{\partial r_{ij}} + \lambda_k e_k$$

$$\left(R - \lambda_k I\right) \frac{\partial e_k}{\partial r_{ij}} = \frac{\partial \lambda_k}{\partial r_{ij}} e_k - \frac{\partial R}{\partial r_{ij}} e_k \quad (7-16)$$
Unfortunately, the determinant of $R - \lambda_k I$ is zero by its very definition and cannot be inverted. $R$ is a symmetric matrix with the spectral decomposition $\sum \lambda_m e_m e'_m$.

Define $R^+$ to be the matrix $\sum \left\{ \frac{1}{\lambda_m} e_m e'_m : \lambda_m \neq 0 \right\}$. $R$ and $R^+$ are mutually generalised inverses, that is, $R R^+ R = R$, $R^+ R R^+ = R^+$. Using the generalised inverse, equation 7-16 becomes:

$$
(R - \lambda_k I)^+ (R - \lambda_k I) \frac{\partial e_k}{\partial r_{ij}} = \frac{\partial \lambda_k}{\partial r_{ij}} (R - \lambda_k I)^+ e_k - (R - \lambda_k I)^+ \frac{\partial R}{\partial r_{ij}} e_k
$$

(7-17)

To simplify the above expression, consider:

$$
R - \lambda_k I = \sum_{m \neq k} \lambda_m e_m e'_m
$$

(7-18)

where: $R - \lambda_k I$ has the eigenvalues $(\lambda_k = 0, \lambda_2, \lambda_3, \ldots, \lambda_p)$

Now,

$$
(R - \lambda_k I)^+ = \sum_{m \neq k} \frac{1}{\lambda_m} e_m e'_m
$$

(7-19)

and

$$
(R - \lambda_k I)^+ e_k = \left( \sum_{m \neq k} \frac{1}{\lambda_m} e_m e'_m \right) e_k
$$

(7-20)

As $e_k$ is orthogonal to all other eigenvectors, equation 7-20 becomes:

$$
(R - \lambda_k I)^+ e_k = 0
$$

(7-21)
Also, consider:

\[
(R - \lambda_k I)^* (R - \lambda_k I) \frac{\partial e_k}{\partial r_{ij}} = \left( \sum_{m=k}^{\lambda_m} e_m e'_m \left( \sum_{m=k}^{\lambda_m} \lambda_m e_m e'_m \right) \right) \frac{\partial e_k}{\partial r_{ij}}
\]

\[
= \left( \sum_{m=k}^{\lambda_m} e_m e'_m \right) \frac{\partial e_k}{\partial r_{ij}} \tag{7-22}
\]

As the \( e_m \) are eigenvectors of a symmetric matrix, they form an orthonormal basis. Thus, the matrix \( Q \) with these vectors in its columns satisfies \( QQ' = I \). Written in summation form, \( QQ' = I \) has the form:

\[
\sum e_m e'_m = I \tag{7-23}
\]

Hence, equation 7-22 becomes:

\[
(R - \lambda_k I)^* (R - \lambda_k I) \frac{\partial e_k}{\partial r_{ij}} = (I - e_k e'_k) \frac{\partial e_k}{\partial r_{ij}}
\]

\[
= \frac{\partial e_k}{\partial r_{ij}} - e_k e'_k \frac{\partial e_k}{\partial r_{ij}} \tag{7-24}
\]

Substituting equation 7-9 into the above expression, gives:

\[
(R - \lambda_k I)^* (R - \lambda_k I) \frac{\partial e_k}{\partial r_{ij}} = (I - e_k e'_k) \frac{\partial e_k}{\partial r_{ij}}
\]

\[
= \frac{\partial e_k}{\partial r_{ij}} \tag{7-25}
\]

Substituting equations 7-21 and 7-25 into equation 7-17, gives:

\[
\frac{\partial e_k}{\partial r_{ij}} = -(R - \lambda_k I)^* \frac{\partial R}{\partial r_{ij}} e_k \tag{7-26}
\]
Using equation 7-19 and the form of $\frac{\partial R}{\partial r_{ij}}$, the above equation becomes:

$$
\frac{\partial e_k}{\partial r_{ij}} = -\sum_{m \neq k} 1 \frac{1}{\lambda_m} e_m (e_m e_{ki} + e_m e_{kj})
$$  \hspace{1cm} (7-27)

Equations 7-13 and 7-26 have similar forms to those derived by Sibson (1979). Sibson used expansions in Taylor-series form to express the effects of perturbations on eigenvalues and eigenvectors of a symmetric matrix.

### 7-4 CHANGES IN FACTOR LOADINGS WITH RESPECT TO CHANGES IN THE CORRELATION MATRIX

Now that expressions have been formed for the changes in eigenvalues and eigenvectors, an expression can be formed for the changes in the factor loadings. The loadings are given by:

$$
L_k = \sqrt{\lambda_k} e_k
$$  \hspace{1cm} (7-28)

Differentiating both sides of the above equation gives:

$$
\frac{\partial L_k}{\partial r_{ij}} = \frac{\sqrt{\lambda_k} \partial e_k}{\partial r_{ij}} - \frac{1}{2\sqrt{\lambda_k}} \frac{\partial \lambda_k}{\partial r_{ij}} e_k
$$  \hspace{1cm} (7-29)

Substituting equations 7-15 and 7-27 into the above equation gives:

$$
\frac{\partial L_k}{\partial r_{ij}} = -\sqrt{\lambda_k} \left( \sum_{m \neq k} 1 \frac{1}{\lambda_m} e_m (e_m e_{ki} + e_m e_{kj}) \right) + \frac{1}{\sqrt{\lambda_k}} e_k (e_k e_{ki})
$$

$$
= \sqrt{\lambda_k} \left( \frac{1}{\lambda_k} e_k (e_k e_{ki}) - \sum_{m \neq k} 1 \frac{1}{\lambda_m} e_m (e_m e_{ki} + e_m e_{kj}) \right)
$$  \hspace{1cm} (7-30)
Simplifying the above expression by combining the terms in the brackets gives:

$$\frac{\partial \mathbf{L}_k}{\partial r_{ij}} = \sqrt{\lambda_k} \sum_m \left( \frac{3 \delta_{mk}}{2} - 1 \right) \left( \frac{e_m}{\lambda_m} \right) \left( e_{mj} e_{li} + e_{mi} e_{kj} \right)$$

(7-31)

where: $\delta_{mk}$ is the Kronecker delta. That is, $\delta_{mk} = \begin{cases} 1 & m = k \\ 0 & m \neq k \end{cases}$

Equation 7-31 reveals that the change in a factor loading vector is a complicated expression involving all of the eigenvalues and components of eigenvectors. It is very difficult to determine the effect on loadings or which loadings are most affected by changes in the correlation matrix. However, the above equation is used for a very simple patterned correlation matrix in the next section to reveal some interesting properties of this equation.

### 7-5 EXAMPLE

Consider the correlation matrix:

$$\mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \text{ where } r > 0$$

The eigenvalues and one choice for the eigenvectors follows:

$$\lambda_1 = 1 + r \quad \mathbf{e}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\
\frac{1}{\sqrt{2}} \end{bmatrix}$$

(7-32)

$$\lambda_2 = 1 - r \quad \mathbf{e}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\
\frac{-1}{\sqrt{2}} \end{bmatrix}$$

(7-33)
The factor loading matrix is given by:

\[
L = \begin{bmatrix}
\frac{1+r}{2} & \frac{1-r}{2} \\
\sqrt{\frac{1+r}{2}} & \sqrt{\frac{1-r}{2}} \\
\sqrt{\frac{1+r}{2}} & -\sqrt{\frac{1-r}{2}}
\end{bmatrix}
\]  

(7-34)

Using equation 7-31, the changes in the loading vectors are given by:

\[
\frac{\partial L_1}{\partial r} = \begin{bmatrix}
1 \\
\frac{1}{2\sqrt{2}+2r} \\
\frac{1}{2\sqrt{2}+2r}
\end{bmatrix}
\]  

(7-35)

\[
\frac{\partial L_2}{\partial r} = \begin{bmatrix}
-1 \\
\frac{-1}{2\sqrt{2}-2r} \\
\frac{1}{2\sqrt{2}-2r}
\end{bmatrix}
\]  

(7-36)

Simply differentiating each of the elements in \( L \) produces the same results as equations 7-35 and 7-36 which were determined using equation 7-31. The above expressions (equations 7-35 and 7-36) indicate that both loadings increase with \( r \) for the first factor and decrease for the second factor. The changes in the first factor are both larger than the changes in the second factor.
Chapter 8

Conclusions and Implications

8-1 INTRODUCTION

This chapter concludes the thesis by discussing the findings reached about the main research problem and the individual research questions that were posed in chapter 1. The limitations of the research conducted are also discussed. The implications of the results from this thesis are summarised for researchers that utilise factor analysis. Finally, some key areas for future research are identified and discussed.

8-2 CONCLUSIONS ABOUT RESEARCH QUESTIONS

This section answers each of the research questions proposed in chapter 1 before examining the main research problem.
8.2.1 How are Random Multivariate Data Sets Generated that are Typical of that Encountered when a Factor Analysis is Applied?

Generating random data that are representative of that used by factor analysis researchers was no trivial task. Several authors have used simulations to investigate the abilities of various techniques to handle missing values in multivariate analysis. Of these authors, Krzanowski (1988), Timm (1970) and Buck (1960), simply selected a few data sets from the literature to analyse. There were two main problems with collecting data from the literature to analyse. It would be difficult to collect the large amounts of factor analysis data needed for a large simulation study. The second problem was the difficulty in comparing the results from the various data sets. Little (1988), Beale & Little (1975) and Haitovsky (1968) constructed regression data with correlated independent variables specifying the means, variances and correlations. These procedures were not used as this thesis was concerned with missing values in factor analysis as opposed to missing values in regression studies.

Bello (1993) generated a random covariance matrix as described by Bryce & Maynes (1979) having a predetermined geometric eigenvalue structure as used by Bendel and Mickey (1978). The eigenvalue structure is given by:

\[ \lambda_i = bv^{i-1} + \delta, \quad (i = 1, \ldots, p) \]

where: \[ b = \begin{cases} \frac{p(1-\delta)(1-\nu)}{1-\nu^p}, & 0 < \nu < 1 \\ 1-\delta, & \nu = 1 \end{cases} \]

\( p = \) number of variables
The lower bound $\delta = 0.1$ was introduced to avoid singularity problems although these problems still occurred. The values of $v$ represent a continuum where the level of interdependence among variables increases as $v$ decreases from one to zero.

The procedure used by Bello (1993) generated random covariance matrices, whereas this thesis required correlation matrices. Unfortunately, the covariance matrix could not be standardised as the desired eigenvalue structure would not be preserved. A series of orthogonal transformations were applied to the covariance matrices to produce correlation matrices with the same eigenvalue structure. The method described by Bryce and Maynes (1979) was then used to generate data that would have this correlation matrix.

Bendel and Mickey's model involved the use of an arbitrary choice of $v$ to determine the level of interdependence among variables. The measure $v$ is hard to interpret other than some sort of index of interdependency. As it is not possible to determine the level of $v$ for a real life data set, it would be difficult for researchers to extrapolate any conclusions from a simulation study to what occurs in real life. The measure

$$\varphi = \sqrt{\frac{\|R\|^2 - p}{p(p-1)}}$$

introduced by Gleason and Staelin (1975), produces a root-mean-square measure of the average level of correlation among the variables. The relationship between $\varphi$ and $v$ was derived and is shown below:

$$\varphi = \sqrt{\frac{(1-\delta)^2}{p-1} \left[ \frac{p(1-v)(1+v^p)}{(1-v^p)(1+v)} - 1 \right]}$$

(8-1)
The model used by Bendel and Mickey (1978) was modified using the above relationship. It was now possible to specify the average intercorrelation among the variables, convert it to a value of \( v \) and then use the model as described by Bendel and Mickey.

The main steps to generate random data \( X \) with a specified intercorrelation \( \varphi \) follow:

1. Specify the desired level of \( \varphi \).
2. Determine approximation for \( v \) using equation 8-1.
3. Calculate eigenvalues according to model used by Bendel and Mickey (1978).
4. Generate random orthogonal matrices \( Z \) and \( Q \) using the Gram-Schmidt orthogonalisation.
5. Perform orthogonal transformations on \( QDQ' \) to obtain the correlation matrix \( R \).
6. Calculate the upper-triangular matrix \( T \) by performing the Cholesky decomposition on \((n-1)R\).
7. Data matrix is formed by \( X = ZT \).

The main reason for the relative absence in the literature of a large factor analysis simulation study of missing values probably lies in the size and number of matrices that are necessary to be generated. Only two authors have designed studies to investigate missing values in a factor analysis. Finkbeiner (1979) worked with 50 replicates of sample size 64. Two factors, 6 variables and 2 ‘missingness’ patterns were used. Hendricks Brown (1983) examined a single factor model and used asymptotic comparison methods as opposed to Monte Carlo methods.
To be representative of typical factor analyses the following levels were chosen for the number of cases \( n \), the number of variables \( p \), the average intercorrelation among variables \( \varphi \) and the proportion of missing values \( k \):

\[
\begin{align*}
n &= 50, 100, 200 \\
p &= 15, 25, 50 \\
\varphi &= 0.3, 0.5, 0.7 \\
k &= 1\%, 5\%, 10\%
\end{align*}
\]

**8.2.2 Do all of the Missing Value Techniques Work for all Combinations when Data are Randomly Deleted?**

The techniques of complete cases only and all available cases might not work for all situations when data are randomly deleted (deletion patterns). Obviously, if any other technique employs complete cases only or all available cases as an initial estimator of parameters or missing values, then that particular technique might not work for some deletion patterns. Complete cases only will not work if there are no cases or only one case remaining after data are randomly deleted. An expression for the probability of complete cases only failing was derived and is shown below:

\[
\left[1 - (1 - k)^p\right]^{n-1} \left[1 + (n - 1)(1 - k)^p\right] 
\]

(8-2)

All available cases will not work if, for any pair of variables, there are no cases or only one case where data are present for both variables. Again, an expression for the probability of this technique failing was derived and is shown below:

\[
\left[\binom{p}{2} \left[1 - (1 - k)^2\right]^{n-1} \left[1 + (n - 1)(1 - k)^2\right]\right]^{p/2} 
\]

(8-3)
8-2.3 How is the Difference between Two Correlation Matrices Compared and Tested?

The correlation matrices obtained from a complete data set were compared to the
correlation matrices obtained from the various techniques applied to a data set with
missing values. Gleason and Staelin's (1975) root-mean-square measure

\[ E_a = \sqrt{\frac{\hat{R}_\alpha - R}{p(p-1)}} \]

was used as a single measure of the average error in the off-
diagonal terms of the correlation matrices. By using a single measure, it was possible
to graph the root-mean-square error of the different techniques over various
conditions and perform regression analyses of E versus n, p, \( \varphi \) and k.

The test developed by Jennrich (1970) that tested the difference between two
correlation matrices was used for the original and estimated correlation matrices.
Tables 5-9 to 5-14 were presented to show the number of significantly different
matrices at various significance levels. These were meant to give another indication
as to when each technique performed well or inadequately.

It was shown in section 5-3.5 that Jennrich's (1970) test of two correlation matrices
produced more significant results when there were NPD and ill-conditioned matrices
present. More NPD matrices were observed for imputing means as \( \varphi \) increased. All
available cases produced more NPD matrices at \( \varphi = 0.5 \) and 0.7. For both of these
techniques, Jennrich's test produced more significant differences at these same levels.
These findings were explained by assuming that the error produced from a technique
produced the same sized perturbation $\varepsilon$ to all off-diagonal elements of the correlation matrix $R$. It was shown for these assumptions that an NPD results when:

$$
\sum_{i \neq j} e_i e_j > \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon > 0 \quad \text{or} \quad \sum_{i \neq j} e_i e_j < \frac{\lambda}{2\varepsilon} \quad \text{for } \varepsilon < 0
$$

(8-4)

where: $e_i$ is the $i^{th}$ component of an eigenvector $e$ of $R$

The above equation shows that more NPD matrices result for smaller values of $\lambda$ and larger values of $\varepsilon$. This equation assisted in explaining the observed pattern in NPD’s and the unusual number of significant results in Jennrich’s test for all available cases and imputing means. Jennrich’s test was abandoned as it would be difficult to get meaningful results from this correlation difference test in the presence of NPD’s and ill-conditioned matrices.

### 8.2.4 Which Technique is the Best Estimator of the Correlation Matrix of a Data Set used for a Factor Analysis?

The iterative principal component method and the EM algorithm require initial estimates of the correlation matrix. The techniques complete cases only (cc), imputing means (im) and all available cases (ac) were examined for their ability to estimate the correlation matrix as described in section 8-2.2. Bello (1993) considered the abilities of these techniques to estimate the covariance matrix. Gleason and Staelin (1975) also examined these techniques to estimate the correlation matrix as well as the missing values themselves. Timm (1970) compared these techniques in terms of both the covariance and correlation matrices. However, none of these papers examined these techniques for the large number of variables $p$ usually present in a
factor analysis. Because of their design, Gleason and Staelin (1975) and Timm (1970) were not able to examine high levels of intercorrelation among the variables that could occur in a factor analysis.

Previous research from Timm (1970) indicated that complete cases only performed worse than the other techniques when \( p > 5 \). This was certainly confirmed in the current thesis. Complete cases only performed inadequately for 15 variables compared to the other techniques (refer to figures 5-1 to 5-4). This technique was removed from the remainder of the study. The graphs of the root-mean-square errors produced by imputing means and all available cases (refer to figures 5-5 to 5-8) revealed that all available cases performs better than imputing means for higher levels of intercorrelations \( \phi \) among the variables. The lowest level of \( \phi \) where all available cases continually outperforms imputing means lies between 0.1 and 0.3 for all levels of the parameters studied.

The root-mean-square error \( E \) for all available cases and imputing means were regressed against the variables \( n, p, \phi \) and \( k \). Multiplicative regression models were the most appropriate to explain \( E \), which agreed with the work done by Gleason and Staelin (1975). The final models are shown below:

\[
E_{im} = 1.855n^{-0.513}k^{0.539} (1-\phi)^{0.238} \quad (8-5)
\]
\[
E_{ac} = 2.396n^{-0.531}k^{0.537} (1-\phi)^{0.845} \quad (8-6)
\]

Both of these equations do not include \( p \) as a significant predictor. From the above equations, an expression (in terms of \( \phi \) and \( n \)) was derived to show when one
technique outperformed the other. Assuming the coefficient of $k$ is the same for both techniques, all available cases outperformed imputing means in predicting $R$ when:

$$\varphi > 1 - 0.656n^{0.030} \tag{8-7}$$

8-2.5 How are Two Factor Analyses Compared?

Comparisons have to be made between a factor analysis based on a complete data set and a factor analysis estimated from a technique that handles missing values. Section 6-3 discussed various methods that have been proposed in the literature to compare two factor analyses that might be appropriate in this thesis. To obtain meaningful results that are easy to compare over a large number of simulations and repetitions, only two measures were used.

Root-mean-square measures (equations 6-2 and 6-3 respectively) were used for both the loadings $FL$ and specific variances $SV$. The original unrotated factor loadings were considered to enable comparisons between these results and the theoretical nature of the changes in these unrotated loadings (refer to sections 8-2.6 and 8-2.7). The specific variances, as used by Finkbeiner (1979), were chosen because they are unaltered by rotations and for this reason are probably of more interest to researchers.

8-2.6 How do the Various Techniques Compare in Terms of their Costs in Computing Time?

As well as comparing the techniques abilities to reproduce factor loadings and specific variances, the computing time required for each technique to produce the correlation matrix that will be entered into the factor analysis was also examined.
Multiplicative regression models were used to examine the influence of various parameters on the computing time. The equations showing the significant parameters follow:

\[
T_{pc} = 0.008n^{0.767}p^{1.807}(1-\phi)^{0.206}k^{0.038} \\
T_{im} = 0.001n^{0.896}p^{2.004} \\
T_{ac} = 0.002n^{0.763}p^{2.046}k^{-0.023} \\
T_{cc} = 0.001n^{0.542}p^{1.265}(1-\phi)^{-0.023}k^{-0.257} \\
T_{em} = 0.107p^{3.332}(1-\phi)^{0.603}k^{0.594}
\]

\(R^2 = 0.996\) \hspace{1cm} (8-8)
\(R^2 = 0.999\) \hspace{1cm} (8-9)
\(R^2 = 0.998\) \hspace{1cm} (8-10)
\(R^2 = 0.896\) \hspace{1cm} (8-11)
\(R^2 = 0.912\) \hspace{1cm} (8-112)

Overall, the non-iterative techniques of complete cases only (cc), imputing means (im) and all available cases (ac) take the least time in that order. The iterative principal component method pc is the next quickest. The EM algorithm (em) takes much more time than the other techniques, particularly for large values of n, p and k.

### 8.2.7 What Effect does a Change in the Correlation Matrix have on the Factor Loadings?

This question was examined by investigating the partial derivative of the loadings with respect to a change in two symmetric elements of the correlation matrix, that is

\[
\frac{\partial L_k}{\partial r_{ij}}
\]

As \(L_k = f(\lambda_k, e_k)\), expressions had to be derived for \(\frac{\partial \lambda_k}{\partial r_{ij}}\) and \(\frac{\partial e_k}{\partial r_{ij}}\). The subscript k refers to the particular loadings, eigenvector and eigenvalue that are being considered. All three expressions are shown below:

\[
\frac{\partial \lambda_k}{\partial r_{ij}} = 2e_{ki}e_{kj}
\]

\(8-13\)
\[
\frac{\partial e_k}{\partial r_{ij}} = -\sum_{m \neq k} \frac{1}{\lambda_m} e_m (e_{mj} e_{ki} + e_{mi} e_{kj}) 
\] (8-14)

\[
\frac{\partial L_k}{\partial r_{ij}} = \sqrt{\lambda_k} \sum_m \left( \frac{3 \delta_{mk}}{2} - 1 \right) \left( \frac{e_m}{\lambda_m} \right) (e_{mj} e_{ki} + e_{mi} e_{kj}) 
\] (8-15)

The general forms of the first two expressions were shown to be equivalent to the work done by Sibson (1979) using perturbation theory. Sibson examined what happened to \( \lambda \) and \( e \) if there is a slight perturbation to a symmetric matrix. Equations 8-13 and 8-14 result when the changes are specifically to the elements of a correlation matrix \( (r_{ij} \text{ and } r_{ji}) \). The final expression shows the effect on the loadings due to changes in the correlation matrix. It is a complicated function that involves eigenvalues and elements of eigenvectors.

### 8-3 CONCLUSIONS ABOUT THE RESEARCH PROBLEM

The main objective of this thesis was to investigate which technique best handles missing values in a factor analysis. The conditions under which techniques perform well or very poorly, would also be of considerable interest to researchers.

Overall, there was no one technique that was best for all of the conditions studied in this thesis. Plots of FL and SV (figures 6-1 to 6-6) revealed that the EM algorithm outperformed all of the other techniques except when \( n = 50, p = 25; n = 50, p = 50; \) and \( n = 100, p = 50 \). In fact, the EM algorithm performed extremely poorly compared to the other techniques when \( n = 50 \) and \( p = 50 \). Ill-conditioned matrices were more likely to occur at these levels because of the presence of NPD matrices. The plots
showed that the EM algorithm, as programmed in this thesis, was not robust to matrices that were ill-conditioned.

When $n = 50$ and $p = 50$, the iterative principal component method and all available cases were the better techniques to estimate the factor loadings and all available cases was the best technique at estimating specific variances. For $n = 50$, $p = 25$ and $n = 100$, $p = 50$, the EM algorithm performed better when there were 1% of data missing. For more deletions, all available cases was the best technique to estimate both the factor loadings and specific variances.

For all of the situations investigated, complete cases only performed between 5 and 50 times worse than the other techniques. For factor loadings, imputing means was generally the next poorest technique except when there were 10% of the data missing. This is when the iterative principal component method started performing poorly. Apart from complete cases only, imputing means was usually the poorest performing technique when there were 1% of the data missing for specific variances. The iterative principal component method deteriorated for higher proportions of deletions compared to the other techniques.

Multiplicative regression models were used to investigate the effects of $n$, $p$, $\varphi$ and $k$ on FL and SV. These regression equations are shown on the following page.
\[
\begin{align*}
\text{FL}_{\text{pc}} &= 1.530n^{-0.399} p^{-0.729} (1-\varphi)^{0.521} k^{0.622} \\
\text{SV}_{\text{pc}} &= 0.433n^{-0.235} p^{-0.243} (1-\varphi)^{0.542} k^{0.710} \\
\text{FL}_{\text{im}} &= 1.806n^{-0.530} p^{-0.697} (1-\varphi)^{0.255} k^{0.541} \\
\text{SV}_{\text{im}} &= 0.819n^{-0.477} p^{-0.424} (1-\varphi)^{0.062} k^{0.534} \\
\text{FL}_{\text{ac}} &= 2.199n^{-0.535} p^{-0.726} (1-\varphi)^{0.613} k^{0.542} \\
\text{SV}_{\text{ac}} &= 1.309n^{-0.503} p^{-0.483} (1-\varphi)^{0.800} k^{0.532} \\
\text{FL}_{\text{cc}} &= 1.551n^{-0.492} p^{0.231} (1-\varphi)^{0.846} k^{0.830} \\
\text{SV}_{\text{cc}} &= 2.071n^{-0.503} p^{0.052} (1-\varphi)^{0.975} k^{0.794} \\
\text{FL}_{\text{em}} &= 3.924n^{-0.809} p^{-0.518} (1-\varphi)^{0.835} k^{0.580} \\
\text{SV}_{\text{em}} &= 3.121n^{-0.576} p^{-0.543} (1-\varphi)^{0.961} k^{0.712}
\end{align*}
\]

\(R^2 = 0.915\) \hfill (8-16)

\(R^2 = 0.922\) \hfill (8-17)

\(R^2 = 0.929\) \hfill (8-18)

\(R^2 = 0.859\) \hfill (8-19)

\(R^2 = 0.892\) \hfill (8-20)

\(R^2 = 0.886\) \hfill (8-21)

\(R^2 = 0.910\) \hfill (8-22)

\(R^2 = 0.891\) \hfill (8-23)

\(R^2 = 0.903\) \hfill (8-24)

\(R^2 = 0.885\) \hfill (8-25)

All of the variables investigated were significant in the above regression models. In terms of the factor loadings, all techniques improve as n increases with the EM algorithm being the most responsive. Complete cases only is the only technique to be adversely affected by increases in p. Complete cases only and the EM algorithm are more responsive to changes in \(\varphi\) than the other techniques. Complete cases only is the most adversely affected technique to increases in k, closely followed by the iterative principal component method.

For specific variances, the iterative principal component method is least influenced by n. Again, complete cases only is unduly influenced by increases in p. The coefficients of imputing means and all available cases are very similar with the exception of \(\varphi\). Imputing means is only slightly affected by \(\varphi\). The iterative principal component method, complete cases only and the EM algorithm are most affected by increases in k.
Generally, the EM algorithm outperformed the other techniques in estimating the original factor loadings and specific variances except for ill-conditioned matrices. In terms of factor loadings, the EM algorithm is outperformed by all other techniques (except complete cases only) when $n = 50$ and $p = 50$. The iterative principal component method and all available cases outperform the EM algorithm when $n = 50$ and $p = 25$. All available cases outperforms the EM algorithm when $n = 100$ and $p = 50$ when there were 10% of the data missing.

Examining specific variances, the EM algorithm is again outperformed by all other techniques (except complete cases only) when $n = 50$ and $p = 50$. At 10% missing values for $n = 100$ and $p = 50$, all available cases outperforms the EM algorithm. Overall, for most situations the EM algorithm outperformed all other techniques; however, it is important to bear in mind the relative cost in computing time of the technique (discussed in section 8-2.5).

8-4 LIMITATIONS

The obvious limitation in the factor analysis simulations concerns the actual levels chosen for the various parameters. Care should be taken when attempting to extrapolate the results presented to levels outside those that were investigated. For example, it would not be expected that complete cases only would continue to perform as poorly for smaller proportions of missing values than those investigated in this thesis. It is quite feasible that complete cases only would perform very well if only one case had some of its variables missing.
The results from this thesis were formed using the principal component method to estimate factor loadings. It is unlikely that the simulation results would be exactly reflected in a maximum likelihood factor analysis. Again, care should be taken in extrapolating these results to a factor analysis that uses maximum likelihood to estimate factor loadings.

There are numerous results that can be obtained from a factor analysis that may be of interest to researchers. The results from this thesis relate only to the abilities of various techniques to estimate specific variances and original factor loadings. Researchers that are interested in the rotated factor loadings, preserving variables in factors, preserving the order of factors or factor scores would need to make significant assumptions if attempting to extrapolate the results presented here.

Only five data sets were generated for each combination of n, p, φ and k. Different patterns of missing values associated with the level of k were used to produce more replicates. The constraints of time and computing power were the main reasons for this. The computing architecture that is available today would considerably shorten the time to perform the slower techniques like the EM algorithm compared to the quicker non-iterative techniques. It should be noted that Jamshidian and Jennrich (1997) and Meng and Van Dyk (1997) summarise and compare various methods that have been proposed to accelerate the EM algorithm.
8-5 IMPLICATIONS FOR RESEARCHERS

Taking into account the limitations discussed above, there are some guidelines that can be stated to assist the researcher interested in deciding which technique to use to handle missing values in a given situation. There was no single technique that was the best for all combinations of the parameters studied, but some general statements can still be made. The reader should note that the following are general recommendations. If there is not much difference between two techniques, then the technique that applies to most situations is suggested. This thesis used a simulation study that only had a few values for each of the parameters investigated. The following recommendations for researchers are made assuming that the results for these values can be interpolated.

If the computing time to perform a particular technique is not an important concern for a researcher, then the EM algorithm should be used for most situations. If the data produce a correlation matrix that is ill-conditioned (eigenvalues that are close to zero), then the researcher should be wary of the results using the EM algorithm. The following recommendations are suggested when there is some concern about ill-conditioned correlation matrices. Unfortunately, there was no information obtained about the eigenvalues of the correlation matrices when the EM algorithm performs inadequately compared to the other techniques. There were two different situations for ill-conditioned matrices where the EM algorithm did not outperform all other techniques. If the correlation matrix was very ill-conditioned (represented for situations in this thesis where \( n = 50 \)
and $p = 50$) then the EM algorithm should not be used. All available cases is the best technique to estimate specific variances for all situations. All available cases is also best at estimating factor loadings when there are more than 5% missing values. The iterative principal component method should be used to estimate factor loadings when there are no more than 5% missing values.

If the correlation matrix is moderately ill-conditioned (represented by $n = 50$, $p = 25$ and $n = 100$, $p = 50$ in this thesis) then the following recommendations are suggested for researchers. The EM algorithm is the best technique to estimate factor loadings when there are less than 5% missing values. All available cases should be used when there are at least 5% missing values. In terms of specific variances, the EM algorithm should be used when there are at most 5% missing values. When there are more than 5% missing values then all available cases should be used.

If computer time is an important concern to a researcher, then the EM algorithm should be avoided, particularly when there are more than 25 variables. In terms of estimating factor loadings, the iterative principal component method should be used when there are no more than 5% missing values. All available cases should be used when there are at least 5% missing values. Generally, all available cases is the best technique to estimate specific variances. When the number of cases is less than 100 and there are less than 5% missing values, the iterative principle component method is the best technique.

The above procedures are summarised in the flow charts depicted in figures 8-1 and 8-2. Other than using flow charts, the researcher could place particular values for $n$,
p, φ and k into the regression equations for SV and FL (equations 8-16 to 8-25). Care must be taken here, as the equations for the EM algorithm did not include n = 50 and p = 50, where it was not as robust to ill-conditioned matrices.

Figure 8-1 Flow Chart for Factor Loadings
Figure 8-2  Flow Chart for Specific Variances
8-6 IMPLICATIONS FOR FURTHER RESEARCH

Most researchers would be interested in a more precise criterion for deciding the situations when the EM algorithm performs inadequately compared to the other techniques. Very and moderate ill-conditioned matrices were terms used in section 8-5 to distinguish between the situations when the EM algorithm performed very poorly or moderately. Criteria need to be established to distinguish between these two situations. Examining the eigenvalues of the estimated correlation matrices might reveal some insight into determining these criteria. The use of variance inflation factors or the condition number \( \phi = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \) may prove useful in providing the desired criteria.

The singular value decomposition did not consistently converge for large matrices with a large amount of data deleted. The reasons for this should be explored further. Krzanowski (1988) showed for one small data set that this technique outperformed imputing means in estimating the missing values themselves. In comparing covariance matrices, Bello (1993) showed that the singular value decomposition method performed favourably with the EM algorithm and an iterative principal component method. This singular value decomposition method should be tested against the others techniques for the parameters used in a factor analysis.

The criterion for deciding between all available cases and imputing means to estimate the initial correlation matrix was \( \phi > 1 - 0.656n^{0.030} \). For the matrices used in this thesis, the values of \( \phi \) that decided which particular technique to use varied between
0.23 and 0.26. Matrices were generated with \( \phi = 0.1 \) and 0.3 for this investigation. More matrices could be generated with more values of \( \phi \) between 0.2 and 0.3 to provide a sharper selection criterion.

There are other aspects of a factor analysis that could be studied. For example, it would be of interest to investigate the ability of the various techniques to simulate the rotated factor loadings and factor scores using a similar simulation study to the one presented in this thesis. The techniques and methods used in this thesis could easily be extended to investigate missing values in a factor analysis with loadings estimated using maximum likelihood instead of principal components. This investigation might reveal some interesting results. For example, it could indicate whether the principal component method performs better for principal factoring compared to maximum likelihood factoring.

This thesis compared specific variances and unrotated factor loadings in terms of root-mean-square measures. The spaces defined by the original and estimated unrotated factor loadings could also be compared. Krzanowski (1979) compared the spaces defined by principal components. This method could easily be extended as another comparative measure of factor loadings.

Estimates of correlation matrices from data matrices that contain missing values using imputing means (adjusting for degrees of freedom) and all available cases can produce NPD matrices. The method proposed by Schwertman and Allen (1979) to smooth NPD matrices can be used with different weighting patterns. Gabriel and Zamir (1979) also examined several techniques that fit missing values based on
weighted least squares. A Monte Carlo study could be performed to investigate the best weighting pattern to use to smooth an NPD matrix obtained when there are missing values present.

Dong (1985) used a small ridge constant to the diagonal of $\mathbf{R}$ to force $\mathbf{R}$ to be positive definite. Dong’s procedure could be used when NPD matrices are present to determine if this procedure improves on the estimates obtained using the smoothing procedure discussed in this thesis.

The equation to determine the average intercorrelation $\varphi$ (equation 4-18) actually involves squared correlations. It is not clear whether there could be an effect due to the signs of individual correlations. Using the procedures shown in chapter 4, it is possible to generate random data that have a correlation matrix with a specified average intercorrelation. By simply changing some of the signs in the correlation matrix, it is then possible to generate data that have the same individual correlation magnitudes but different signs. Unfortunately, this simple sign changing would not preserve the eigenvalue structure. Despite this limitation, it could still be useful to investigate any effects due to the change in signs.

Rubin (1987) developed the technique of multiple imputation to estimate missing values and standard errors that reflect missing data uncertainty. In multiple imputation, each missing value is replaced by several plausible values from a predictive distribution. Multiple imputation has become popular in the last few years due to the increase in computing power. Schafer (1997) has adapted and implemented Markov chain Monte Carlo methods and written general purpose multiple imputation
software for incomplete multivariate data. The effects of using multiple imputation on factor loadings and specific variances could also be examined and compared with the results from this thesis.


List of References


Appendices

List of Appendices

Appendix A  Computer Programs of Techniques to Handle Missing Values
Appendix B  Computer Program to Generate Random Multivariate Data
Appendix C  Computer Program to Compare Correlation Matrices
Appendix D  Computer Program to Compare FL and SV
Appendix E  Derivations
Appendix F  Regression Analyses
MATHEMATICA CODE FOR COMPLETE CASES ONLY

(*
This procedure deletes an entire case from an input data matrix if
any of the variables are missing. The correlation matrix that
results from this procedure is returned.

data - data matrix with missing values
*)

CompCases[data_] :=
  Block[{alphadata},
    (* replace missing values with the letter 'a' *)
    alphadata = ReplacePart[data, a, Position[data, 999]];
    (* drop any cases that contain 'a' and determine the correlation
       matrix *)
    CorrelationMatrix[DropNonNumeric[alphadata]]]
MATHEMATICA CODE FOR IMPUTING MEANS

(*
This procedure replaces each missing value from an input data matrix
with the mean of the corresponding variable. The degrees of freedom
are adjusted using equation 3-10. If an NPD matrix results, then it
is smoothed. The correlation matrix of this procedure is returned.

data - data matrix with missing values
n - number of cases in the data matrix
p - number of variables in the data matrix
*)

ImpMeans[data, n, p] :=
  Block[{trdata, delmvdata, meandata, impmeandata, num1, num2, num12,
    impcorr, corr, eigval1, eigvec1, diagelems1, el, vl, cl, f1},
    trdata = Transpose[data];
    delmvdata = Table[DeleteCases[trdata[[i]], 999], {i, 1, p}];
    (* replace each missing value with mean of the variable *)
    meandata = Table[ReplacePart[trdata[[i]], Mean[delmvdata[[i]]],
      Position[trdata[[i]], 999]], {i, 1, p}];
    impmeandata = Transpose[meandata];
    (* initialise the correlation matrix of the data with missing
    values replaced by means *)
    impcorr = CorrelationMatrix[impmeandata];
    corr = IdentityMatrix[p];
    (* count how many values are not missing for each pair of
    variables and the number where both are present *)
    Do[
      num12 = 0;
      num1 = Count[trdata[[j]], x_/; x < 990];
      num2 = Count[trdata[[k]], x_/; x < 990];
      Do[
        If[
          trdata[[j, i]] != 999 && trdata[[k, i]] != 999,
          num12++,
          {i, 1, n}];
        (* adjust the degrees of freedom of the correlation matrix *)
        corr[[j, k]] = impcorr[[j, k]] Sqrt[(num1 - 1)

181
(num2 - 1)/(num12 - 1);
corr[[k, j]] = corr[[j, k]],
{j, 1, p - 1}, {k, j + 1, p}];
(* check if correlation matrix is NPD *)
eigvall = Eigenvalues[corr];
If[MemberQ[eigvall, ?Negative],
(* smooth NPD matrix *)
eigvec1 = Eigenvectors[corr];
diagelems1 = Delete[eigvall, Position[eigvall, ?Negative]]; 
e1 = DiagonalMatrix[diagelems1];
v1 = Delete[eigvec1, Position[eigvec1, ?Negative]]; 
c1 = Transpose[v1] . e1 . v1;
f1 = IdentityMatrix[Length[eigvall]]; 
Do[
    f1[[i, i]] = Sqrt[c1[[i, i]]], {i, 1, Length[eigvall]}];
Return[Inverse[f1] . c1 . Inverse[f1],
Return[corr//N]];]

A-3

MATHEMATICA CODE FOR ALL AVAILABLE CASES

(*
This procedure performs the technique of all available cases on an
input data matrix with missing values. If an NPD matrix results,
then it is smoothed. The correlation matrix of this procedure is
returned.

data - data matrix with missing values
n - number of cases in the data matrix
p - number of variables in the data matrix
*)

AvCas[data, n, p] :=
    Block[{trdata, corr, allpairdata, pairdata, ptcov, eigval, eigvec,
    diagelems, e, v, f, c},
    trdata = Transpose[data];
corr = IdentityMatrix[p];
Do[
    allpairdata = {};
(* set up a matrix that contains data where both variables are present *)
Do{
    If[trdata[[j, i]] != 999 && trdata[[k, i]] != 999,
        allpairdata = {allpairdata, trdata[[j, i]],
        trdata[[k, i]]},],
    {i, 1, n}];
(* determine the correlation matrix of the data where only both
variables are present *)
allpairdata = Flatten[allpairdata];
pairdata = Partition[allpairdata, 2];
ptcov = CovarianceMatrix[pairdata];
corr[[j, k]] = ptcov[[1, 2]] / Sqrt[ptcov[[1, 1]] ptcov[[2, 2]]];
corr[[k, j]] = corr[[j, k]],
    {j, 1, p - 1}, {k, j + 1, p}];
(* check if correlation matrix is NPD *)
eigval = Eigenvalues[corr];
If[MemberQ[eigval1, ?Negative],
    (* smooth NPD matrix *)
eigvec = Eigenvectors[corr];
diagelems = Delete[eigval, Position[eigval, ?Negative]]; e = DiagonalMatrix[diagelems];
v = Delete[eigvec, Position[eigval, ?Negative]]; c = Transpose[v] . e . v;
f = IdentityMatrix[Length[eigval]];
Do{
    f[[i, i]] = Sqrt[c[[i, i]]], {i, 1, Length[eigval]}];
Return[Inverse[f] . c . Inverse[f]],
Return[corr // N];}
MATHEMATICA CODE FOR THE ITERATIVE PRINCIPAL COMPONENT METHOD

(*
This procedure performs an iterated version of Dear's principal component method from an input data matrix containing missing values. The correlation matrix of this procedure is returned.

data - data matrix containing missing values
n - number of cases
p - number of variables
*)

DearPC[data_, n_, p_] :=
  Block[{nummiss, firstpc, trdata, stddata, deldata, corr, criteria,
    iterations, oldmat, newmat, eigsys, eig, eigvec, maxeigvec, m},
    (* initialise variables *)
    orig = data;
    (* count the number of missing values *)
    nummiss = Count[Flatten[orig], x_ /; x == 999];
    firstpc = Array[0&, {n}];
    (* initialise indicator matrix M with all elements equal to 1 *)
    m = Array[1&, {n, p}];
    trdata = Transpose[orig];
    stddata = trdata;
    deldata = Table[DeleteCases[trdata[[i]], 999], {i, 1, p}];
    (* standardise the data matrix *)
    Do[
      stddata[[i]] = (stddata[[i]] - Mean[deldata[[i]]]) / 
        StandardDeviation[deldata[[i]]],
      {i, 1, p}];
    (* missing values are assigned the value of 0 *)
    stddata = Table[ReplacePart[stddata[[i]], 0,
      Position[trdata[[i]], 999], {i, 1, p}];
    stddata = Transpose[stddata];
    (* initial estimate of correlation matrix using all available cases *)
    corr = AvCas[stddata, n, p];
    (* change indicator matrix to include zeros for missing values *)

184
Do[
    If[orig[[i, j]] == 999,
        m[[i, j]] = 0,
        {i, 1, n}, {j, 1, p}];
    (* initialise iterations and convergence criteria *)
    iterations = 0;
    criteria = 888;
    newmat = stddata;
While[criteria > 0.00001,
    (* update iterations *)
    iterations++;
    (* define oldmat as data matrix before the next iteration *)
    eigsyst = Eigensystem[corr];
    {eig, eigvec} = eigsyst;
    (* determine largest eigenvector of correlation matrix *)
    maxeigvec = Flatten[eigvec[[Flatten[Position[eig,
        Max[Chop[eig]]]]]]];
    (* calculate equation 3-17 *)
    Do[
        firstpc[[z]] = Sum[maxeigvec[[t]] newmat[[z, t]] m[[z, t]],
            {t, 1, p}],
            {z, 1, n}];
    (* calculate equation 3-18 *)
    Do[
        If[orig[[i, j]] == 999,
            newmat[[i, j]] = maxeigvec[[j]] firstpc[[i]],
            {i, 1, n}, {j, 1, p}];
    (* update convergence criteria *)
    criteria = Apply[Plus, Flatten[Abs[(oldmat - newmat) /
        nummiss]]];
    corr = CorrelationMatrix[newmat];
    ];
    newmat = Transpose[newmat];
    (* de-standardise *)
    Do[
        newmat[[i]] = newmat[[i]] StandardDeviation[deldata[[i]]] +
            Mean[deldata[[i]]],
            {i, 1, p}];
    Print["iterations from DearPC = ", iterations];
    (* determine the correlation matrix from this procedure *)
    CorrelationMatrix[Transpose[newmat]]]
MATHEMATICA CODE FOR THE SINGULAR VALUE DECOMPOSITION METHOD

(* This procedure performs the technique of singular value decomposition on a data matrix with missing values. The correlation matrix of this procedure is returned.

data - data matrix containing missing values
n - number of cases
p - number of variables
*)

SVD[data_, n_, p_] :=
   Block[{
      orig, nummiss, trdata, stddata, deldata, criteria,
      iterations, oldmat, newmat, uorig, mdorig, vorig, norowdata,
      nocoldata, svd1, u1, md1, v1, data2, svd2, u2, md2, v2, posy,
      foundmiss, a, olduorig, oldvorig, countmatch1, countmatch2},
      (* initialise variables *)
      orig = data;
      (* count the number of missing values *)
      nummiss = Count[Flatten[orig], x_ /; x == 999];
      trdata = Transpose[orig];
      stddata = trdata;
      deldata = Table[DeleteCases[trdata[[i]], 999], {i, 1, p}];
      (* standardise the data matrix *)
      Do[
         stddata[[i]] = (stddata[[i]] - Mean[deldata[[i]]]) / 
         StandardDeviation[deldata[[i]]],
         {i, 1, p}];
      (* missing values are assigned the value of 0 *)
      stddata = Table[ReplacePart[stddata[[i]], 0, 
         Position[trdata[[i]], 999]], {i, 1, p}];
      stddata = Transpose[stddata];
      (* initialise iterations and convergence criteria *)
      iterations = 0;
      criteria = 888;
      newmat = stddata;
      While[criteria > 0.00001,
(* update iterations *)
iterations ++;
oldmat = newmat;
olduorig = uorig;
oldvorig = vorig;

(* determine u, d, and v of the data matrix at the beginning of each iteration *)
{uorig, mdorig, vorig} = SingularValues[newmat];
Do[
  If[uorig[[i, j]] == 999,
    (* remove row with missing value *)
    norowdata = Drop[newmat, {i}];
    (* perform SVD on data matrix without row *)
    {ul, md1, vl} = svdl1;
    (* remove column with missing value *)
    data2 = Drop[Transpose[newmat], {j}];
    nocoldata = Transpose[data2];
    (* perform SVD on data matrix without column *)
    svd2 = SingularValues[nocoldata];
    {u2, md2, v2} = svd2;
    (* check for arbitrary sign changes *)
    Do[
      countmatch1 = 0;
      countmatch2 = 0;
      Do[
        If[Sign[uorig[[k, l]]] != Sign[olduorig[[k, l]]],
          countmatch1++;
          ],
        {l, 1, n}];
      If[countmatch1 >= 0.8 n,
        uorig[[k]] = uorig[[k]](-1);
        ];
      Do[
        If[Sign[vorig[[k, l]]] != Sign[oldvorig[[k, l]]],
          countmatch2++;
          ],
        {l, 1, p}];
      If[countmatch2 >= 0.8 p,
        vorig[[k]] = vorig[[k]](-1);
        ],
      {k, 1, p}]];
]
(* determine which singular value has to be removed from
the original matrix *)
(* create a list of numbers from 1 to p *)
posy = Table[b, {b, 1, p}];
(* establish flag when extra singular value is found *)
foundmiss = 0;
a = 2;
While[foundmiss == 0 && a < p,
  (* determine the absolute difference between the singular
values of original and column reduced matrices *)
If[Abs[md2[[a]]] - mdorig[[a]] < Abs[md2[[a]] -
   mdorig[[a + 1]]],
a++,
foundmiss = 1;
(* posy contains the p-1 singular values to be included
*)
posy = Delete[posy, a];
]
]
(* check if the last singular value was the one removed *)
If[a == p, posy = Delete[posy, p]];
(* calculate equation ??? *)
newmat[[i, j]] = Sum[(Abs[u2[[t, i]]] Sqrt[md2[[t]]]
   Sign[uorig[[t, i]]]) (Abs[v1[[posy[[t]], j]]]
   Sqrt[md1[[posy[[t]]]]] Sign[vorig[[posy[[t]], j]]]),
  {t, 1, p - 1}];
]
If[i == n && j == p,
  (* update convergence criteria *)
criteria = Apply[Plus, Flatten[Abs[(oldmat - newmat) /
   nummiss]]];
],
{i, 1, n}, {j, 1, p}];
(* de-standardise *)
Do[
  newmat[[i]] = newmat[[i]] StandardDeviation[deldata[[i]]] +
    Mean[deldata[[i]]],
  {i, 1, p}];
Print["iterations from SVD = ", iterations];
CorrelationMatrix[Transpose[newmat]]
MATHEMATICA CODE FOR THE EM ALGORITHM

(*
This procedure performs the EM algorithm on a data matrix with missing values. The correlation matrix of this procedure is returned.

data - data matrix containing missing values
n - number of cases
p - number of variables
*)

EM[data_, n_, p_] :=
  Block[{
    trdata, deldata, newdata, trnewdata, u, s, data2, partdata,
    ptcov, eigval, eigvec, diagelems, e, v, criterial, criteria2,
    iterations, textra, tr, tl, pos1, pos2, u1, u2, s11, s22, s12,
    flats11, flats22, flats12, mat, mtx, t2a, maty, t2, newu, news},
    trdata = Transpose[data];
    deldata = Table[DeleteCases[trdata[[i]], 999], {i, 1, p}];
    (* initial estimate of mean matrix μ *)
    newdata = Transpose[Table[ReplacePart[trdata[[i]],
      Mean[deldata[[i]]], Position[trdata[[i]], 999, {i, 1, p}]]];
    trnewdata = Transpose[newdata];
    u = Table[Mean[trnewdata[[i]]], {i, 1, p}];
    (* initial estimate of covariance matrix Σ using available cases *)
    s = Array[0&, {p, p}];
    Do[
      data2 = {};
      (* set up a matrix that contains data where both variables are present *)
      Do[
        If[trdata[[j, i]] != 999 && trdata[[k, i]] != 999,
          data2 = {data2, trdata[[j, i]], trdata[[k, i]]},
        ],
        {i, 1, n}];
      (* determine the covariance matrix of all variables where only both variables are present *)
      data2 = Flatten[data2];
      partdata = Partition[data2, 2];
ptcov = CovarianceMatrix[partdata];
s[[j, k]] = N[ptcov[[1, 2]]];
s[[k, j]] = s[[j, k]],
{j, 1, p-1}, {k, j+1, p}];
(* determine variance of each variable using non-missing values *)
Do[
s[[i, i]] = N[Variance[deldata[[i]]]],
{i, 1, p}];
(* check if correlation matrix is NPD *)
eigval = Eigenvalues[s];
If[MemberQ[eigval, _?Negative],
(* smooth NPD matrix *)
eigvec = Eigenvectors[s];
diagelems = Delete[eigval, Position[eigval, _?Negative]]; 
e = DiagonalMatrix[diagelems];
v = Delete[eigvec, Position[eigval, _?Negative]]; 
s = Transpose[v].e.v];
(* initialise convergence criteria for μ and Σ *)
criterial = 888;
criteria2 = 888;
iterations = 0;
While[(criterial > 0.0001 || criteria2 > 0.01) && iterations < 100,
iterations ++;
(* textra contains change in T_2 *)
textra = Array[0&, {p, p}];
(* tr contains the transpose of data with imputed means for 
missing values *)
tr = Transpose[newdata];
(* initially define T_1 as the addition of all observations for 
each variable *)
t1 = Table[Plus@@tr[[i]], {i, 1, p}];
Do[
(* pos1 and pos2 contain the positions of the missing and 
observed values respectively for each variable *)
pos1 = Position[data[[z]], _/; x == 999];
pos2 = Position[data[[z]], _/; x != 999];
If[Length[pos1] != 0,
(* partition μ into μ1 and μ2 (missing and observed 
respectively) *)
u1 = Delete[u, pos2];
u2 = Delete[u, pos1];
}
(* initialise the partition of the covariance matrix *)

s11 = Array[0& , {Length[pos1], Length[pos1]}];
s22 = Array[0& , {Length[pos2], Length[pos2]}];
s12 = Array[0& , {Length[pos1], Length[pos2]}];

(* place components into their appropriate positions *)

Do[
   s11[[i, j]] = s[[pos1[[i]], pos1[[j]]]],
   {i, 1, Length[pos1]}, {j, 1, Length[pos1]}];
flats11 = Flatten[s11];
s11 = Partition[flats11, {Length[pos1]}];

Do[
   s22[[i, j]] = s[[pos2[[i]], pos2[[j]]]],
   {i, 1, Length[pos2]}, {j, 1, Length[pos2]}];
flats22 = Flatten[s22];
s22 = Partition[flats22, {Length[pos2]}];

Do[
   s12[[i, j]] = s[[pos1[[i]], pos2[[j]]]],
   {i, 1, Length[pos1]}, {j, 1, Length[pos2]}];
flats12 = Flatten[s12];
s12 = Partition[flats12, {Length[pos2]}];

(* calculate equation 3-32 *)

mat = Delete[newdata[[z]], pos1] - u2;
matx = u1 + s12.Inverse[s22].Transpose[{mat}];

(* update T with the changes in values to the estimated missing values *)

Do[
   t1[[pos1[[i]]]] = t1[[pos1[[i]]]] + matx[[i]] -
   newdata[[z, pos1[[i]]]],
   {i, 1, Length[pos1]}];

(* update data matrix with new estimated missing values *)

Do[
   newdata[[z, pos1[[i]]]] = matx[[i]],
   {i, 1, Length[pos1]}];

(* calculate T *)

t2a = Transpose[newdata].newdata;
maty = s11 - s12.Inverse[s22].Transpose[s12];

Do[
   textra[[pos1[[i]], pos1[[j]]]] = textra[[pos1[[i]],
   pos1[[j]]]] + maty[[i, j]],
   {i, 1, Length[pos1]}, {j, 1, Length[pos1]}];
\[ t_2 = t_2 + t_{extra}; \]

\}

If[\[z == n,\]

(* update \(\mu\) and \(\Sigma\) estimates from sufficient statistics *)

newu = t_{1}/n;

news = t_{2}/n - Transpose[{newu}].{newu};

criterial = Apply[Plus, Flatten[Abs[(u - newu)/p]]];

criteria2 = Apply[Plus, Flatten[Abs[(s - news)/p p]]];

u = newu;

s = news;

},

\{z, 1, n\}];

Print["iterations from EM = ", iterations];

CorrelationMatrix[newdata]
Appendix B
Computer Program to Generate Random Multivariate Data

B-1

MATHEMATICA CODE TO GENERATE RANDOM DATA BY SPECIFYING THE AVERAGE INTERCORRELATION OF THE CORRELATION MATRIX

The following two procedures generate a random $n \times p$ data matrix that has a correlation matrix with an average intercorrelation $\varphi$ that is specified by the user.

(*
This procedure creates a diagonal matrix with containing the eigenvalues associated with average intercorrelation $\varphi$.

$p$ - number of variables in the data matrix
$\varphi$ - average intercorrelation of the data matrix
*)

Diag[p, $\varphi$] :=
Block[{$\delta$, v, vl, vlist, b, eigs},
  (* lower limit of eigenvalues *)
  $\delta$ = 0.1;
  (* determine v for input $\varphi$ using equation 4-26 *)
  $f[v_] = \text{Sqrt}[p (1 - v) (1 + v^p) (1 - \delta)^2 - (1 - \delta)^2 (1 - v^p)\)
\[
(1 + v) / ((1 - v^p) (1 + v) (p - 1))
\]

\(v1 = \text{FindRoot}[f[v] == \varphi, \{v, 1 - \varphi\}];\)

(* uses \(1 - \varphi\) as initial estimate *)

(* determine list of eigenvalues *)

\(vlist = v / . v1;\)

\(\text{If}[vlist == 1,\)

\( \quad b = 1 - \delta,\)

\( \quad b = p ((1 - \delta) (1 - vlist)) / (1 - vlist^p);\)

\(\text{eigs} = \text{Table}[b \text{vlist}^{(i - 1) + \delta}, \{i, 1, p\}];\)

\(\text{DiagonalMatrix}[\text{eigs}]\)

(*
This procedure generates a random data matrix with a specified average intercorrelation \(\varphi\).

\(n\) - required number of cases

\(p\) - required number of variables

\(\varphi\) - required average intercorrelation

*)

\(\text{GenData}[n, p, \varphi]:=\)

\(\text{Block}[\{\text{ranmat}, x1, \text{tranx1}, z1, \text{ranmat2}, \text{tranx2}, q, \text{precismat},\)

\(\quad \text{symmat}, \text{diagmat}, \text{mi}, \text{ma}, \text{porth}, \text{rotmat}, \text{anglerot}, \text{sols}, \text{realsols},\)

\(\quad \text{td}, \vartheta, \text{check1}\},\)

\(\text{check1} = 0;\)

\(\text{While}[\text{check1} == 0,\)

\(\quad (*\text{ generate random } n \times p \text{ matrix } *)\)

\(\text{ranmat} = \text{Table}[\text{Random[Real]}, \{i, 1, n\}, \{j, 1, p\}];\)

\(x1 = \text{ZeroMean[ranmat];}\)

\(\text{tranx1} = \text{Transpose[x1];}\)

\(\quad (*\text{ define a random orthonormal matrix } z1 *)\)

\(z1 = \text{Transpose[GramSchmidt[tranx1]]};\)

\(\text{diagmat} = \text{Diag[p, j];}\)

\(\text{ranmat2} = \text{Table[Real], \{i, 1, n\}, \{j, 1, p\}];\)

\(\text{tranx2} = \text{Transpose[ranmat2];}\)

\(\quad (*\text{ define another random orthonormal matrix } z1 *)\)

\(q = \text{Transpose[GramSchmidt[ranmat2]]};\)

\(\quad (*\text{ define the symmetric matrix } T'T *)\)

\(\text{symmat} = q \cdot \text{diagmat}. \text{Transpose[q]};\)

194
(* find the positions of the minimum and maximum diagonal elements of T'T *)

Do[
    mi = 1;
    ma = 1;
    Do[
        If[symmat[[i, i]] > symmat[[ma, ma]], ma = i];
        If[symmat[[i, i]] < symmat[[mi, mi]], mi = i],
        {i, 1, p}];
(* define the orthogonal transformation matrix P *)

If[symmat[[ma, ma]] != 1,
    porth = IdentityMatrix[p];
    porth[[ma, ma]] = Cos[θ];
    porth[[ma, mi]] = Sin[θ];
    porth[[mi, mi]] = Cos[θ];
    porth[[mi, ma]] = -Sin[θ];
    rotmat = porth . symmat . Transpose[porth];
    Off[Solve::"ifun" ]; Off[InverseFunction::"ifun" ];
    anglerot = Solve[rotmat[[ma, ma]] == 1, θ];
    On[Solve::"ifun" ]; On[InverseFunction::"ifun" ];
    sols = θ /. anglerot;
    realsols = Cases[sols, Real];
    (* the symmetric matrix takes the form of a correlation matrix with the eigenvalues preserved *)
    symmat = rotmat /. θ -> realsols[[1]];
    
    [{j, 1, p - 1}];
    precismat = SetPrecision[(n - 1) symmat, 10];
(* define the upper triangular matrix T *)

    td = CholeskyDecomposition[precismat];
    z1 . td]
Appendix C
Computer Program to Compare Correlation Matrices

C-1

MATHEMATICA CODE TO COMPARE ORIGINAL AND ESTIMATED CORRELATION MATRICES

The following program uses the procedures from Appendices A and B. Five data matrices are generated with n, p, φ, and k. The original correlation matrix is determined for each matrix. Random data is deleted 20 times (according to k) for each data matrix and the techniques of all available cases, imputing means, and complete cases only are used to estimate the correlation matrix. The original and estimated correlation matrices are tested for equality and the root-mean-square difference between the two matrices is also measured. A procedure to check whether complete cases only works for a particular deletion pattern is also included.

(*
This procedure produces a p-value for testing the equality of two correlation matrices using equation 5-7.

n - number of cases
p - number of variables
r1 and r2 - two correlation matrices to be compared
*)
CorrCheck[n_, p_, r1_, r2_] :=
   Block[{r, rinv, b, z, f, g, trace1, chisq},
      (* Define $\bar{R}$ *)
      r = (n r1 + n r2)/(2 n);
      (* Define $\bar{R}^{-1}$ *)
      rinv = Inverse[r];
      (* Define the elements of the inverse matrix *)
      b = Array[0&, {p, p}];
      Do[
         Do[
            If[i==j,
                b[[i, j]] = 1 + rinv[[i, j]],
                b[[i, j]] = r[[i, j]] rinv[[i, j]]],
            {j, 1, p}],
         {i, 1, p}];
      (* Define $Z$ *)
      z = Sqrt[n^2/(2 n)] (rinv.(r1 - r2));
      (* Determine $\text{tr}(Z^2)$ *)
      f = Array[0&, {1, p}];
      Do[
         f[[1, i]] = z[[i, i]],
         {i, 1, p}];
      g = z . z;
      trace1 = 0;
      Do[
         trace1 = trace1 + g[[i, i]],
         {i, 1, p}];
      (* Determine the test statistic *)
      chisq = 0.5 (trace1 - f.Inverse[b].Transpose[f]);
      (* Calculate the p-value *)
      1 - CDF[ChiSquareDistribution[p(p - 1)/2], chisq]
This procedure checks whether complete cases only will work on the deleted data.

data - data matrix with missing values
*)

TestOk[data_] :=
Block[{alphadata},
  (* replace any missing value with 'a' *)
  alphadata = ReplacePart[data, a, Position[data, 999]]; (* check if there are none or only one case remaining *)
  If[Length[DropNonNumeric[alphadata]] <= 1,
    Return[0],
    Return[1];]
]

This procedure randomly deletes data from a matrix and returns a matrix with the numbers 999 to indicate the deleted data.

data - complete data matrix
n - number of cases in the data matrix
p - number of variables in the data matrix
k - percentage of data to be deleted
*)

DelData[data_, n_, p_, k_] :=
Block[{rownum, colnum, newdata, c, numdel},
  C = 0;
  newdata = data;
  numdel = Round[k n p];
  While[C < numdel,
    (* determine the row and column of the data to delete *)
    rownum = Random[Integer, {1, n}];
    colnum = Random[Integer, {1, p}];
    (* check if position has already been deleted *)
    If[newdata[[rownum, colnum]] == 999,
      newdata[[rownum, colnum]] = 999;
      c++];
  ];
  newdata]
(*
This program generates an n x p data matrix with a specified average
intercorrelation \( \varphi \). Data is randomly deleted and the techniques:
complete cases only, all available cases, and imputing means are used
to estimate the correlation matrix. The error for each of these
techniques is determined and stored in files.
*)

storeddata = Array[0&, {1, 5, 50, 15}];
pvals = Array[0&, {3, 100}]
errormat = Array[0&, {3, 100}];
(* define the parameters *)
n1 = 50; pl = 15; \( \varphi l \) = 0.1, \( k l \) = 0.01;
Do[
   loopnum1 = 1;
edatamat = GenData[n1, pl, \( \varphi l \)];
origcorr = CorrelationMatrix[datamat];
(* store each generated matrix *)
storeddata[[1, loopnum1]] = datamat;
While[loopnum1 <= 20,
   delmat = DelData[datamat, n1, pl, \( \varphi l \)];
   checkdel = TestOk[delmat];
   (* proceed if all techniques works for the deleted pattern *)
   If[checkdel = 1,
      ccmat = CompCases[delmat];
      immat = ImpMeans[delmat, n1, pl];
      acmat = AvCas[delmat, n1, pl];
      (* determine the error from complete cases only *)
      errorcc = Sqrt[Apply[Plus, Flatten[(ccmat - origcorr)^2]] / (pl(pl - 1))]/N;
      (* determine the error from imputing means *)
      errorim = Sqrt[Apply[Plus, Flatten[(immat - origcorr)^2]] / (pl(pl - 1))]/N;
      (* determine the error from all available cases *)
      errorac = Sqrt[Apply[Plus, Flatten[(acmat - origcorr)^2]] / (pl(pl - 1))]/N;
      (* store all errors *)
      errmat[[1, 20(loopnum - 1) + loopnum1]] = errorcc;
      errmat[[2, 20(loopnum - 1) + loopnum1]] = errorim;
      errmat[[3, 20(loopnum - 1) + loopnum1]] = errorac;
]
(* store all p-values for correlation matrices equality test *)
pvals[[1, 20(loopnum - 1) + loopnuml]] = CorrCheck[n1, pl,
origcorr, cmat];
pvals[[2, 20(loopnum - 1) + loopnuml]] = CorrCheck[n1, pl,
origcorr, immat];
pvals[[3, 20(loopnum - 1) + loopnuml]] = CorrCheck[n1, pl,
origcorr, acmat];
loopnuml++;
];
{loopnum, 1, 5}];
storeddata >> dat1;
errformat >> err1;
pvals >> pvall;
Appendix D
Computer Program to Compare FL and SV

D-1
MATHEMATICA CODE TO COMPARE FACTOR LOADINGS AND SPECIFIC VARIANCES

The following procedures determine the original factor loadings and specific variances of an input complete data matrix. Data is randomly deleted and the techniques: all available cases, imputing means, complete cases only, EM algorithm, iterative principal component method, and singular value decomposition are utilised on the missing values. The time to complete each technique is calculated and the root mean square FL and SV are determined. All comparative measures are stored.

(*
This procedure calculates the factor loadings matrix of an input correlation matrix.

data - correlation matrix
fact - number of factors corresponding to eigenvalues > 1
p - number of variables
*)

Load[data_, fact_, p_] :=
  Block[{loadmat, eigval, eigvec},
    (*initialise factor loading matrix *)
    loadmat = Array[0&, {fact, p}];
    {eigval, eigvec} = Eigensystem[data];

(* sort eigenvalues into descending order *)
eigval = Sort[Chop[eigval], Greater];

(* remove all unwanted eigenvalues *)
Do[
loadmat[[i]] = Chop[Sqrt[eigval[[i]]]]
   Normalize[eigvec[[i]]/\[N],
   {i, 1, fact}];
Transpose[loadmat]]

(*
This procedure calculates the specific variances from an input factor
loading matrix.

data - factor loading matrix
p - number of variables
*)

SpVar[data_, p_] :=
Block[\{h\},
   (* initialise communality *)
   h = Array[0 &, \{p\}];
   Do[
      (* calculate communality *)
      h[[k]] = Apply[Plus, Flatten[data[[k]]^2]],
      \{k, 1, p\};
      1-h]

(*
This program loads an n x p data matrix. Data is randomly deleted
and the techniques: iterative principal component method, EM
algorithm, all available cases, imputing means, and complete cases
only are used to estimate the correlation matrices. The time to
complete this is also determined. The loadings and specific
variances are estimated for each technique. FL and SV are calculated
and stored in files.
*)

(* initialise variables *)
n1 = 50; p1 = 15; k1 = 0.01;
(* initialise arrays to contain the time it takes to perform each
technique *)
timdear = Array[0&, {50}];
timem = timdear;
timac = timdear;
timim = timdear;
timcc = timdear;

(* load previously generated data matrix *)
filedata = <<dat01;
(* calculate number of factors using eigenvalues > 1 *)
corr = CorrelationMatrix[filedata[[1, 1]]];
numfact = Count[Flatten[Eigenvalues[corr1]], x_ /; x >= 1.0];
(* initialise arrays *)
loadarray = Array[0&, {5, 5, 10, pl, numfact}];
loaddata = Array[0&, {5, pl, numfact}];
spvardata = Array[0&, {5, pl}];
rmsload = Array[0&, {5, numfact}];
spvararray = Array[0&, {5, 50}];
rmsspvar = Array[0&, {5, 50}];
meanrmsload = Array[0&, {5}];
meanrmsspvar = Array[0&, {5}];

(* calculate factor loading matrix and specific variances of original data matrix *)

Do[

data = filedata[[1, j]];  
corr = CorrelationMatrix[filedata[[1, j]]];
loaddata[[j]] = Load[corr, numfact, pl];
spvardata[[j]] = SpVar[loaddata[[j]], pl];

Do[

(* delete data to represent missing values *)
deldat = DelData[data, n1, pl, k1];
(* perform iterative principal component method *)
timdearl, dear = Timing[DearPC[deldat, n1, pl]];
(* calculate loading matrix *)
loadarray[[1, j, i]] = Load[dearl, numfact, pl];
(* perform EM algorithm *)
timeml, em = Timing[EM[deldat, n1, pl]];
loadarray[[2, j, i]] = Load[em, numfact, pl];
(* perform imputing means *)
timiml, im = Timing[ImpMeans[deldat, n1, pl]];
loadarray[[3, j, i]] = Load[im, numfact, pl];
(* perform all available cases *)
timac1, ac = Timing[AvCas[deldat, n1, pl]];
]
loadarray[[4, j, i]] = Load[ac, numfact, pl1];
(* perform complete cases only *)
{timcc1, cc} = Timing[CompCases[deldat]];
loadarray[[5, j, i]] = Load[cc, numfact, pl1];
(* remove seconds from times so they can be averaged *)
timdearl[[10 (j-1) + i]] = timdearl /. Second -> 1;
timem[[10 (j-1) + i]] = timeml /. Second -> 1;
timim[[10 (j-1) + i]] = timiml /. Second -> 1;
timac[[10 (j-1) + i]] = timacl /. Second -> 1;
timcc[[10 (j-1) + i]] = timcc1 /. Second -> 1;
{i, 1, 10}],
{j, 1, 5}];
Do[
(* calculate root mean square loadings *)
rmsload[[k, 10 (j-1) + i]] = Sqrt[Apply[Plus,
   Flatten[(loaddata[[j]] - loadarray[[k, j, i]])^2]] / (pl1^2
   numfact^2)]/N;
(* calculate root mean square specific variances *)
rmsspvar[[k, 10 (j-1) + i]] = Sqrt[Apply[Plus,
   Flatten[(spvardata[[j]] - spvararray[[k, j, i]])^2]] / (pl1^2
   numfact^2)]/N,
{k, 1, 5}, {j, 1, 5}, {i, 1, 10}];
Do[
   meanrmsload[[i]] = Mean[rmsload[[i]]],
   meanrmsspvar[[i]] = Mean[rmsspvar[[i]]],
   {i, 1, 5}];
Print["average time for dpc = ", Mean[timdearl]];
Print["average time for em = ", Mean[timem]];
Print["average time for im = ", Mean[timim]];
Print["average time for ac = ", Mean[timac]];
Print["average time for cc = ", Mean[timcc]];
Print["mean rms load for each technique = ", meanrmsload];
Print["mean rms sp. var. for each technique = ", meanrmsspvar];
(* save files *)
loadarray >> loadarr1;
loaddata >> loaddat1;
spvardata >> svdata1;
spvararray >> svarr1;
rmsload >> rmsload1;
rmsspvar >> rmssv1;
Appendix E
Derivations

DERIVATION OF SOLUTIONS TO $\alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta = 1$

Consider:

$$\alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta = 1 \quad (9-1)$$

Squaring both sides of equation 9-1:

$$(\alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta)^2 = 1$$

$$\alpha^2 \cos^4 \theta + \gamma^2 \sin^4 \theta + 4\beta^2 \cos^2 \theta \sin^2 \theta \cos \theta + 2\alpha \gamma \cos^2 \theta \sin \theta \cos \theta + 4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta = 1$$

$$\alpha^2 \cos^4 \theta + \gamma^2 \sin^4 \theta + \frac{4\beta^2 + 2\alpha \gamma}{1} \cos^2 \theta \sin^2 \theta + \frac{4\alpha \beta}{2} \cos^3 \theta \sin \theta + \frac{4\beta \gamma}{3} \cos \theta \sin^3 \theta = 1 \quad (9-2)$$

Consider 1 from equation 9-2:

$$\gamma^2 \sin^4 \theta = (\gamma \sin^2 \theta)^2$$
$$\gamma^2 \sin^4 \theta = [\gamma (1 - \cos^2 \theta)]^2$$
$$\gamma^2 \sin^4 \theta = (\gamma - \gamma \cos^2 \theta)^2$$
$$\gamma^2 \sin^4 \theta = \gamma^2 + \gamma^2 \cos^4 \theta - 2\gamma^2 \cos^2 \theta \quad (9-3)$$

Consider 2 from equation 9-2:

$$\cos^2 \theta \sin^2 \theta = (4\beta^2 + 2\alpha \gamma) \cos^2 \theta \sin^2 \theta - (4\beta^2 + 2\alpha \gamma) \cos^4 \theta$$

$$\cos^2 \theta \sin^2 \theta = (4\beta^2 + 2\alpha \gamma) \cos^2 \theta \sin^2 \theta - (4\beta^2 + 2\alpha \gamma) \cos^4 \theta \quad (9-4)$$
Consider 3 from equation 9-2:

\[ 4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta = 4\beta \sin \theta \cos \theta (\alpha \cos^2 \theta + \gamma \sin^2 \theta) \]
\[ 4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta = 4\beta \sin \theta \cos \theta (\alpha \cos^2 \theta + \gamma - \gamma \cos^2 \theta) \]
\[ 4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta = 4\beta \sin \theta \cos \theta [\gamma - \gamma \cos^2 \theta + \gamma] \quad (9-5) \]

From equation 9-1:

\[ 4\beta \sin \theta \cos \theta = 2(1 - \alpha \cos^2 \theta - \gamma \sin^2 \theta) \quad (9-6) \]

Substituting equation 9-6 into equation 9-5:

\[ 4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta = 2(1 - \alpha \cos^2 \theta - \gamma \sin^2 \theta)[(\alpha - \gamma)\cos^2 \theta + \gamma] \]
\[ = 2[(1 - \alpha \cos^2 \theta - \gamma + \gamma \cos^2 \theta)](\alpha - \gamma)\cos^2 \theta + \gamma \]
\[ = 2[(1 - \gamma)(\alpha - \gamma)\cos^2 \theta + (\gamma - \alpha)(\alpha - \gamma)\cos^4 \theta + \gamma(1 - \gamma) + \gamma(\gamma - \alpha)\cos^2 \theta] \]
\[ = 2[(4\gamma - 2)(\gamma - \alpha)\cos^2 \theta + 2(\gamma - \alpha)(\alpha - \gamma)\cos^4 \theta + 2\gamma(1 - \gamma)] \quad (9-7) \]

Substituting equations 9-3, 9-4, and 9-7 back into equation 9-2:

\[ \alpha^2 \cos^4 \theta + \gamma^2 \sin^4 \theta + \frac{(4\beta^2 + 2\alpha \gamma)\cos^2 \theta \sin^2 \theta}{2} + \frac{4\alpha \beta \cos^3 \theta \sin \theta + 4\beta \gamma \cos \theta \sin^3 \theta}{3} = 1 \]

\[ \alpha^2 \cos^4 \theta + \gamma^2 \cos^4 \theta - 2\gamma^2 \cos^2 \theta + (4\beta^2 + 2\alpha \gamma)\cos^2 \theta - (4\beta^2 + 2\alpha \gamma)\cos^4 \theta + (4\gamma - 2)(\gamma - \alpha)\cos^2 \theta + 2(\gamma - \alpha)(\alpha - \gamma)\cos^4 \theta + 2\gamma(1 - \gamma) - 1 = 0 \]

\[ [\alpha^2 + \gamma^2 - (4\beta^2 + 2\alpha \gamma) + 2(\gamma - \alpha)(\alpha - \gamma)]\cos^4 \theta + [-2\gamma^2 + (4\beta^2 + 2\alpha \gamma) + (4\gamma - 2)(\gamma - \alpha)]\cos^2 \theta + [\gamma^2 + 2\gamma(1 - \gamma) - 1] = 0 \quad (9-8) \]

Let \( \cos^2 \theta = x \) and simplify:

\[ -(\alpha^2 + \gamma^2 - 2\alpha \gamma + 4\beta^2)x^2 + (2\gamma^2 - 2\alpha \gamma - 2\gamma + 2\alpha + 4\beta^2)x + \gamma^2 - 1 + 2\gamma(1 - \gamma) = 0 \]
\[ [(\alpha - \gamma)^2 + 4\beta^2]x^2 + [2(\gamma - 1)(\alpha - \gamma) - 4\beta^2]x + (\gamma - 1)^2 = 0 \quad (9-9) \]
Equation 9-9 can be solved using the quadratic equation:

$$\cos^2 \theta = \frac{4\beta^2 - 2(\gamma - 1)(\alpha - \gamma) \pm \sqrt{[2(\gamma - 1)(\alpha - \gamma) - 4\beta^2]^2 - 4[(\alpha - \gamma)^2 + 4\beta^2](\gamma - 1)^2}}{2[(\alpha - \gamma)^2 + 4\beta^2]}$$

(9-10)

The expression under the square root sign will produce a real solution if:

- $[2(\gamma - 1)(\alpha - \gamma) - 4\beta^2]^2 - 4[(\alpha - \gamma)^2 + 4\beta^2](\gamma - 1)^2 \geq 0$
- $4(\gamma - 1)^2(\alpha - \gamma)^2 + 16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - \gamma) - 4(\gamma - 1)^2[(\alpha - \gamma)^2 + 4\beta^2] \geq 0$
- $16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - \gamma) - 16\beta^2(\gamma - 1)^2 \geq 0$
- $\beta^2 - (\gamma - 1)(\alpha - \gamma) - (\gamma - 1)^2 \geq 0$
- $\beta^2 - (\gamma - 1)[(\alpha - \gamma) + (\gamma - 1)] \geq 0$
- $\beta^2 - (\gamma - 1)(\alpha - 1) \geq 0$
- $\beta^2 \geq (\gamma - 1)(\alpha - 1)$

Since $\alpha > 1$ and $\gamma < 1$, the above expression is true. Hence, the expression under the square root sign in equation 9-10 will produce a real solution.

Simplifying equation 9-10:

$$\cos^2 \theta = \frac{4\beta^2 - 2(\gamma - 1)(\alpha - \gamma) \pm \sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)}}{2[(\alpha - \gamma)^2 + 4\beta^2]}$$

(9-11)

A quick examination of this equation will reveal that the right-hand side is positive when considering the plus sign from the $\pm$ sign. Hence, a real solution will exist for this equation if the right-hand side is less than one. That is,
\[
\frac{4\beta^2 - 2(\gamma - 1)(\alpha - \gamma) + \sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)}}{2[(\alpha - \gamma)^2 + 4\beta^2]} \leq 1
\]
\[
4\beta^2 - 2(\gamma - 1)(\alpha - \gamma) + \sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)} \leq 2[(\alpha - \gamma)^2 + 4\beta^2]
\]
\[
\sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)} \leq 2[(\alpha - \gamma)^2 + 4\beta^2 - 2\beta^2 + (\gamma - 1)(\alpha - \gamma)]
\]
\[
\sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)} \leq 2[\alpha^2 - \alpha\gamma - \alpha + \gamma + 2\beta^2]
\]
\[
\sqrt{16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1)} \leq 2(\alpha - 1)(\alpha - \gamma) + 4\beta^2
\]
\[
16\beta^4 - 16\beta^2(\gamma - 1)(\alpha - 1) \leq 4(\alpha - 1)^2(\alpha - \gamma)^2 + 16\beta^4 + 16\beta^2(\alpha - 1)(\alpha - \gamma)
\]
\[
0 \leq 4(\alpha - 1)^2(\alpha - \gamma)^2 + 16\beta^2(\alpha - 1)[((\alpha - \gamma) + (\gamma - 1))]
\]
\[
0 \leq 4(\alpha - 1)^2(\alpha - \gamma)^2 + 16\beta^2(\alpha - 1)
\]
\[
0 \leq 4(\alpha - 1)^2[(\alpha - \gamma)^2 + 4\beta^2]
\]
\[
0 \leq (\alpha - \gamma)^2 + 4\beta^2
\]

This equation must be true as the right hand side is always positive. Therefore, it has been shown that there must be at least one real solution to the equation
\[
\alpha \cos^2 \theta + \gamma \sin^2 \theta + 2\beta \sin \theta \cos \theta = 1.
\]
Appendix F
Regression Analyses

F-1
REGRESSION MODELS FOR \( E_{\text{in}} \)

This section includes various regression models attempted for \( E_{\text{in}} \). SPSS output is shown for the stepwise procedures as well as residual plots for each model.

F-1A
LINEAR REGRESSION OF \( E_{\text{in}} \)

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.748*</td>
<td>.560</td>
<td>.560</td>
<td>.000E-02</td>
</tr>
<tr>
<td>2</td>
<td>.874*</td>
<td>.764</td>
<td>.764</td>
<td>.000E-03</td>
</tr>
<tr>
<td>3</td>
<td>.910*</td>
<td>.838</td>
<td>.838</td>
<td>.637E-03</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), K
b. Predictors: (Constant), K, N
c. Predictors: (Constant), K, N, J
Appendix F - Regression Analyses

F.1B
REGRESSION OF $E_{\text{am}}$ USING NATURAL LOG TRANSFORMATION ON DEPENDENT VARIABLE

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.755</td>
<td>.570</td>
<td>.570</td>
<td>.3691</td>
</tr>
<tr>
<td>2</td>
<td>.690</td>
<td>.790</td>
<td>.792</td>
<td>.2735</td>
</tr>
<tr>
<td>3</td>
<td>.937</td>
<td>.878</td>
<td>.878</td>
<td>.2085</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), K
b. Predictors: (Constant), K, N
c. Predictors: (Constant), K, N, J

d. ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>2978.710</td>
<td>19255.651</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1</td>
<td>14553</td>
<td>.155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2</td>
<td>2248.541</td>
<td>2068.482</td>
<td>.0000</td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>2</td>
<td>4136.963</td>
<td>27058.029</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>2</td>
<td>14554</td>
<td>7.479E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4</td>
<td>5255.251</td>
<td>4.388E-02</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>3</td>
<td>1529.925</td>
<td>34844.562</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>3</td>
<td>14551</td>
<td>4.388E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>6</td>
<td>5255.251</td>
<td>4.388E-02</td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), K
b. Predictors: (Constant), K, N
c. Predictors: (Constant), K, N, J
d. Dependent Variable: LN$E_{\text{am}}$
## Appendix F - Regression Analyses

### Coefficientsa

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Sig.</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>Tolerance</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>4.296</td>
<td>.005</td>
<td>.795</td>
</tr>
<tr>
<td>K</td>
<td>12.921</td>
<td>.003</td>
<td>.755</td>
<td>136.801</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>3.589</td>
<td>.006</td>
<td>.536</td>
</tr>
<tr>
<td>K</td>
<td>12.946</td>
<td>.006</td>
<td>.756</td>
<td>159.889</td>
</tr>
<tr>
<td>N</td>
<td>-4.586E-03</td>
<td>.000</td>
<td>.471</td>
<td>-124.556</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>3.589</td>
<td>.006</td>
<td>.536</td>
</tr>
<tr>
<td>K</td>
<td>12.972</td>
<td>.006</td>
<td>.758</td>
<td>261.492</td>
</tr>
<tr>
<td>N</td>
<td>-4.586E-03</td>
<td>.000</td>
<td>.470</td>
<td>-162.172</td>
</tr>
<tr>
<td>J</td>
<td>-529</td>
<td>.008</td>
<td>.205</td>
<td>-101.249</td>
</tr>
</tbody>
</table>

*a. Dependent Variable: LNIM

### Charts

- **Histogram**
  - Dependent Variable: LNIM

- **Normal P-P Plot of Regression Standardized Residuals**
  - Dependent Variable: LNIM

- **Scatterplot**
  - Dependent Variable: LNIM

- **Residual Plot for r**

### F.1.3

**Regression of Emax Using Natural Log Transformation on Dependent and Independent Variables**

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.762</td>
<td>.616</td>
<td>.233</td>
<td>.1423</td>
</tr>
<tr>
<td>2</td>
<td>.591</td>
<td>.514</td>
<td>.233</td>
<td>.1423</td>
</tr>
<tr>
<td>3</td>
<td>.971</td>
<td>.944</td>
<td>.233</td>
<td>.1423</td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), LNK

*b. Predictors: (Constant), LNK, LNN

*c. Predictors: (Constant), LNK, LNN, LNJ

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>32192.219</td>
<td>1</td>
<td>32192.219</td>
<td>23354.201</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>2006.033</td>
<td>14553</td>
<td></td>
<td>.138</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>34198.251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>4401.913</td>
<td>2</td>
<td>2205.956</td>
<td>40648.730</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>763.338</td>
<td>14552</td>
<td></td>
<td>5.42E-02</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5165.251</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>4935.738</td>
<td>3</td>
<td>1645.779</td>
<td>91204.279</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>294.513</td>
<td>14551</td>
<td></td>
<td>2.08E-05</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5230.251</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), LNK

*b. Predictors: (Constant), LNK, LNN

*c. Predictors: (Constant), LNK, LNN, LNJ

*d. Dependent Variable: LNIM
### Appendix F - Regression Analyses

#### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficient</th>
<th>Standardized Coefficient</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>1.963</td>
<td>.013</td>
<td>.785</td>
<td>-156.853</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.536</td>
<td>.004</td>
<td>-152.825</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>.528</td>
<td>.018</td>
<td>.785</td>
<td>22.546</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.538</td>
<td>.002</td>
<td>243.711</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>LNN</td>
<td>-5.515</td>
<td>.003</td>
<td>-148.145</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>3 (Constant)</td>
<td>.518</td>
<td>.011</td>
<td>.785</td>
<td>50.859</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.539</td>
<td>.001</td>
<td>400.795</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>LNN</td>
<td>-5.515</td>
<td>.002</td>
<td>-244.170</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>LNNJ</td>
<td>2.58</td>
<td>.002</td>
<td>156.899</td>
<td>.000</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNIM

### Charts

- Histogram
- Normal P-P Plot of Regression Standardized Residual
- Scatterplot
- Residual Plot for ln(y)

#### F-2

**REGRESSION MODELS FOR $E_{ac}$**

This section includes various regression models attempted for $E_{ac}$. SPSS output is shown for the stepwise procedures as well as residual plots for each model.

#### F-2A

**LINEAR REGRESSION OF $E_{ac}$**

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.984</td>
<td>.966</td>
<td>.964</td>
<td>1.398E-06</td>
</tr>
<tr>
<td>2</td>
<td>.912</td>
<td>.836</td>
<td>.824</td>
<td>1.399E-06</td>
</tr>
<tr>
<td>3</td>
<td>.888</td>
<td>.804</td>
<td>.796</td>
<td>7.395E-06</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), JK
b. Predictors: (Constant), J, K
c. Predictors: (Constant), J, K, N
**Appendix F - Regression Analyses**

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>1.289</td>
<td>1.289</td>
<td>7532.688</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>14553</td>
<td>14554</td>
<td>1.711E-04</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3.779</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>2</td>
<td>1.248</td>
<td>1.248</td>
<td>14155.523</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>14552</td>
<td>14554</td>
<td>8.18E-05</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3.779</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>3</td>
<td>.995</td>
<td>.995</td>
<td>18221.928</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>14551</td>
<td>14554</td>
<td>5.46E-05</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>3.779</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* a. Predictors: (Constant), K
* b. Predictors: (Constant), K, J
* c. Predictors: (Constant), K, J, N
* d. Dependent Variable: ACERROR

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>6.71E-03</td>
<td>.000</td>
<td>49.908</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>.269</td>
<td>.003</td>
<td>.564</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>2.47E-02</td>
<td>.000</td>
<td>133.066</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>.270</td>
<td>.002</td>
<td>.567</td>
</tr>
<tr>
<td></td>
<td>J</td>
<td>-3.21E-02</td>
<td>.000</td>
<td>-.568</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>3.540E-02</td>
<td>.000</td>
<td>191.718</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>.271</td>
<td>.002</td>
<td>.588</td>
</tr>
<tr>
<td></td>
<td>J</td>
<td>-3.21E-02</td>
<td>.000</td>
<td>-.564</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>-9.43E-05</td>
<td>.000</td>
<td>-.360</td>
</tr>
</tbody>
</table>

* a. Dependent Variable: ACERROR

### Charts

- **Histogram**
  - Dependent Variable: ACERROR

- **Normal P-P Plot of Regression Standardized Residual**
  - Dependent Variable: ACERROR

- **Scatterplot**
  - Dependent Variable: ACERROR

- **Residual Plot for k**

- **Residual Plot for n**

- **Residual Plot of Regression Standardized Predicted Value**
Appendix F- Regression Analyses

REGRESSION OF E<sub>60</sub> USING NATURAL LOG TRANSFORMATION ON DEPENDENT VARIABLE

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.678&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.459</td>
<td>.469</td>
<td>.4652</td>
</tr>
<tr>
<td>2</td>
<td>.851&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.725</td>
<td>.725</td>
<td>.4660</td>
</tr>
<tr>
<td>3</td>
<td>.914&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.835</td>
<td>.835</td>
<td>.3564</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), J
b. Predictors: (Constant), J, K
c. Predictors: (Constant), J, K, N

c.<sup>a</sup>

ANOVAS<sup>a</sup>

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>5140.877</td>
<td>1</td>
<td>5140.877</td>
<td>12350.784</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>6057.555</td>
<td>1455</td>
<td>.416</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>11198.432</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>8118.120</td>
<td>2</td>
<td>4059.560</td>
<td>19184.583</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>3078.531</td>
<td>14552</td>
<td>.212</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>11196.651</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>9349.629</td>
<td>3</td>
<td>3116.548</td>
<td>24529.315</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1848.762</td>
<td>14551</td>
<td>.127</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>11198.401</td>
<td>14554</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), J
b. Predictors: (Constant), J, K
c. Predictors: (Constant), J, K, N
d. Dependent Variable: LNAC

c.<sup>a</sup>

Coefficients<sup>a</sup>

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant) -3.153</td>
<td>.011</td>
<td>-291.386</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>J -2.097</td>
<td>.019</td>
<td>-111.134</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>(Constant) -3.729</td>
<td>.009</td>
<td>-609.104</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>J -2.106</td>
<td>.013</td>
<td>-184.490</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>K 12.926</td>
<td>.100</td>
<td>118.632</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>(Constant) -3.195</td>
<td>.009</td>
<td>-358.671</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>J -2.101</td>
<td>.010</td>
<td>-201.530</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>K 12.900</td>
<td>.064</td>
<td>153.403</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>N -4.72E-03</td>
<td>.000</td>
<td>-48.412</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNAC

c.<sup>a</sup>

Charts

- Histogram: Dependent Variable: LNAC
- Normal P-P Plot of Regression Standardized Residual: Dependent Variable: LNAC
- Scatterplot: Dependent Variable: LNAC
- Residual Plot for n: Standardized Residual
F-2C

REGRESSION OF E_{op} USING NATURAL LOG TRANSFORMATION ON
DEPENDENT AND INDEPENDENT VARIABLES

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.747</td>
<td>.559</td>
<td>.559</td>
<td>.3828</td>
</tr>
<tr>
<td>2</td>
<td>.919²</td>
<td>.845</td>
<td>.845</td>
<td>.3450</td>
</tr>
<tr>
<td>3</td>
<td>.966²</td>
<td>.961</td>
<td>.961</td>
<td>.1740</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LN1J
b. Predictors: (Constant), LN1J, LNK
c. Predictors: (Constant), LN1J, LNK, LNN

d. Dependent Variable: LNAC

ANOVA²

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6255.971</td>
<td>1</td>
<td>6255.971</td>
<td>16420.725</td>
<td>.000²</td>
</tr>
<tr>
<td></td>
<td>4942.430</td>
<td></td>
<td>.340</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11198.401</td>
<td></td>
<td>11198.401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9468.539</td>
<td>2</td>
<td>4733.265</td>
<td>38771.131</td>
<td>.000²</td>
</tr>
<tr>
<td></td>
<td>1731.871</td>
<td></td>
<td>.119</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11198.401</td>
<td></td>
<td>11198.401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10757.852</td>
<td>3</td>
<td>3585.951</td>
<td>119441.1</td>
<td>.000²</td>
</tr>
<tr>
<td></td>
<td>440.500</td>
<td></td>
<td>3.0285-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11198.401</td>
<td></td>
<td>11198.401</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LN1J
b. Predictors: (Constant), LN1J, LNK
c. Predictors: (Constant), LN1J, LNK, LNN
d. Dependent Variable: LNAC

c. Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant) -3.415</td>
<td>.943</td>
<td>-453.846</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LN1J .843</td>
<td>.006</td>
<td>.747</td>
<td>135.723</td>
</tr>
<tr>
<td>2</td>
<td>LN1J -.1561</td>
<td>.012</td>
<td>-128.939</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK .846</td>
<td>.004</td>
<td>.750</td>
<td>20.813</td>
</tr>
<tr>
<td></td>
<td>LN1J .535</td>
<td>.003</td>
<td>.535</td>
<td>104.848</td>
</tr>
<tr>
<td>3</td>
<td>LN1J .845</td>
<td>.002</td>
<td>.749</td>
<td>65.804</td>
</tr>
<tr>
<td></td>
<td>LNK .537</td>
<td>.002</td>
<td>.537</td>
<td>32.621</td>
</tr>
<tr>
<td></td>
<td>LNN -.531</td>
<td>.003</td>
<td>-.530</td>
<td>206.522</td>
</tr>
</tbody>
</table>

b. Dependent Variable: LNAC

Charts

Histogram

Dependent Variable: LNAC

Normal P-P Plot of Regression Standardized Residual

Dependent Variable: LNAC

Scatterplot

Dependent Variable: LNAC

Residual Plot for ln(11)
F.3
REGRESSION MODELS FOR FL

This section includes only the multiplicative regression models of FL for the various techniques.

F.3A
MULTIPLICATIVE REGRESSION MODEL FOR FLpc

<table>
<thead>
<tr>
<th>Model Summary</th>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.750a</td>
<td>.563</td>
<td>.563</td>
<td>.5310</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.786b</td>
<td>.788</td>
<td>.788</td>
<td>.3685</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.922c</td>
<td>.889</td>
<td>.889</td>
<td>.2910</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.992</td>
<td>.915</td>
<td>.914</td>
<td>.2243</td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNP
c. Predictors: (Constant), LNK, LNP, LNN
d. Predictors: (Constant), LNK, LNP, LNN, LNIJ

Appendix F- Regression Analyses

<table>
<thead>
<tr>
<th>ANOVA*</th>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression Residual Total</td>
<td>1292.479</td>
<td>1</td>
<td>1292.479</td>
<td>4583.878</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1003.221</td>
<td>3558</td>
<td>3559</td>
<td>.262</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2295.700</td>
<td>3559</td>
<td>.262</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression Residual Total</td>
<td>1810.115</td>
<td>2</td>
<td>955.058</td>
<td>6629.721</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>485.584</td>
<td>3557</td>
<td>3559</td>
<td>.137</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2295.700</td>
<td>3559</td>
<td>.137</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression Residual Total</td>
<td>1984.516</td>
<td>3</td>
<td>664.839</td>
<td>7849.585</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>301.184</td>
<td>3556</td>
<td>3559</td>
<td>8.470E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2295.700</td>
<td>3559</td>
<td>8.470E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression Residual Total</td>
<td>2099.581</td>
<td>4</td>
<td>524.895</td>
<td>9514.665</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>196.119</td>
<td>3555</td>
<td>3559</td>
<td>5.517E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2295.700</td>
<td>3559</td>
<td>5.517E-02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNP
c. Predictors: (Constant), LNK, LNP, LNN
d. Predictors: (Constant), LNK, LNP, LNN, LNIJ
e. Dependent Variable: LNDFPC

<table>
<thead>
<tr>
<th>Coefficients*</th>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-4.152</td>
<td>.032</td>
<td>-130.259</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.625</td>
<td>.009</td>
<td>.750</td>
<td>67.704</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-1.557</td>
<td>.048</td>
<td>-32.700</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.625</td>
<td>.006</td>
<td>.750</td>
<td>97.271</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.910</td>
<td>.013</td>
<td>-475</td>
<td>-61.577</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>.282</td>
<td>.054</td>
<td>5.187</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.625</td>
<td>.005</td>
<td>.750</td>
<td>123.492</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.910</td>
<td>.010</td>
<td>-475</td>
<td>-78.177</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.282</td>
<td>.009</td>
<td>-283</td>
<td>-46.990</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>.425</td>
<td>.044</td>
<td>9.546</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.625</td>
<td>.004</td>
<td>.746</td>
<td>152.232</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.729</td>
<td>.009</td>
<td>-427</td>
<td>-85.043</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.399</td>
<td>.007</td>
<td>-283</td>
<td>-57.915</td>
</tr>
<tr>
<td></td>
<td>LNIJ</td>
<td>.521</td>
<td>.012</td>
<td>.219</td>
<td>43.640</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNDFPC
Appendix F- Regression Analyses

F-3b
MULTIPLICATIVE REGRESSION MODEL FOR F1 dm

Model Summary*

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.716</td>
<td>.513</td>
<td>.512</td>
<td>.5100</td>
</tr>
<tr>
<td>2</td>
<td>.868</td>
<td>.738</td>
<td>.736</td>
<td>.5742</td>
</tr>
<tr>
<td>3</td>
<td>.870</td>
<td>.739</td>
<td>.739</td>
<td>.2207</td>
</tr>
<tr>
<td>4</td>
<td>.905</td>
<td>.909</td>
<td>.909</td>
<td>.2042</td>
</tr>
</tbody>
</table>

a. Predictors: [Constant], LNK
b. Predictors: [Constant], LNK, LNP
c. Predictors: [Constant], LNK, LNP, LNN
d. Predictors: [Constant], LNK, LNP, LNN, LN1J
e. Dependent Variable: LNIM
Appendix F- Regression Analyses

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>973.412</td>
<td>973.412</td>
<td>3742.368</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td></td>
<td></td>
<td>.200</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>974.612</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>2</td>
<td>700.414</td>
<td>.140</td>
<td>5002.351</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td></td>
<td></td>
<td>458.040</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>1158.454</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>3</td>
<td>575.201</td>
<td>4.873E-02</td>
<td>11804.980</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td></td>
<td></td>
<td>437.672</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>1013.873</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>4</td>
<td>437.672</td>
<td>4.168E-02</td>
<td>10500.328</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td></td>
<td></td>
<td>437.672</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
<td>875.344</td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNp
c. Predictors: (Constant), LNK, LNp, LNN
d. Predictors: (Constant), LNK, LNp, LNN, LN1J
e. Dependent Variable: LNM

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficient</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-4.278</td>
<td>.051</td>
<td>-139.736</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.542</td>
<td>.009</td>
<td>.716</td>
</tr>
<tr>
<td></td>
<td>LNp</td>
<td>-.736</td>
<td>.013</td>
<td>-.474</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-1.320</td>
<td>.048</td>
<td>-32.814</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.542</td>
<td>.007</td>
<td>.716</td>
</tr>
<tr>
<td></td>
<td>LNp</td>
<td>-.736</td>
<td>.013</td>
<td>-.474</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.530</td>
<td>.008</td>
<td>-.414</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>.521</td>
<td>.041</td>
<td>12.625</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.542</td>
<td>.004</td>
<td>.716</td>
</tr>
<tr>
<td></td>
<td>LNp</td>
<td>-.736</td>
<td>.008</td>
<td>-.474</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.530</td>
<td>.008</td>
<td>-.414</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>.591</td>
<td>.038</td>
<td>15.432</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.541</td>
<td>.004</td>
<td>.714</td>
</tr>
<tr>
<td></td>
<td>LNp</td>
<td>-.697</td>
<td>.007</td>
<td>-.449</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.530</td>
<td>.006</td>
<td>-.414</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>.555</td>
<td>.010</td>
<td>.118</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNM
Appendix F - Regression Analyses

F-3C
MULTIPLICATIVE REGRESSION MODEL FOR FLac

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adj R Square</th>
<th>Std. Error of Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.677</td>
<td>.459</td>
<td>.459</td>
<td>.5713</td>
</tr>
<tr>
<td>2</td>
<td>.844</td>
<td>.707</td>
<td>.707</td>
<td>.4005</td>
</tr>
<tr>
<td>3</td>
<td>.922</td>
<td>.861</td>
<td>.861</td>
<td>.2898</td>
</tr>
<tr>
<td>4</td>
<td>.964</td>
<td>.928</td>
<td>.928</td>
<td>.2077</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNP
c. Predictors: (Constant), LNK, LNP, LNN
d. Predictors: (Constant), LNK, LNP, LNN, LNJ
e. Dependent Variable: LNAC

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>985.279</td>
<td>1</td>
<td>985.279</td>
<td>326</td>
<td>.002</td>
</tr>
<tr>
<td>Residual Total</td>
<td>1161.321</td>
<td>3558</td>
<td>2146.600</td>
<td>3559</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1517.458</td>
<td>2</td>
<td>758.729</td>
<td>4289.647</td>
<td>.002</td>
</tr>
<tr>
<td>Residual Total</td>
<td>629.142</td>
<td>3557</td>
<td>2146.600</td>
<td>3559</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1847.458</td>
<td>3</td>
<td>615.982</td>
<td>7354.700</td>
<td>.002</td>
</tr>
<tr>
<td>Residual Total</td>
<td>298.642</td>
<td>3556</td>
<td>2146.600</td>
<td>3559</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1993.219</td>
<td>4</td>
<td>498.305</td>
<td>11549.547</td>
<td>.002</td>
</tr>
<tr>
<td>Residual Total</td>
<td>153.360</td>
<td>3555</td>
<td>4.314-02</td>
<td>3559</td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNP
c. Predictors: (Constant), LNK, LNP, LNN
d. Predictors: (Constant), LNK, LNP, LNN, LNJ

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coef</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sg</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-4.473</td>
<td>.504</td>
<td>109.427</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.545</td>
<td>.100</td>
<td>.877</td>
<td>54.942</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-8.226</td>
<td>.15</td>
<td>-.498</td>
<td>-54.852</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-1.942</td>
<td>.504</td>
<td>-33.982</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.546</td>
<td>.007</td>
<td>.877</td>
<td>74.608</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-8.226</td>
<td>.010</td>
<td>-.498</td>
<td>-76.604</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>.621</td>
<td>.054</td>
<td>106.254</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.546</td>
<td>.005</td>
<td>.877</td>
<td>106.254</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-8.226</td>
<td>.010</td>
<td>-.498</td>
<td>-76.604</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-5.355</td>
<td>.009</td>
<td>-.392</td>
<td>-62.732</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>.788</td>
<td>.039</td>
<td>20.620</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.542</td>
<td>.004</td>
<td>.873</td>
<td>150.016</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-7.226</td>
<td>.008</td>
<td>-.440</td>
<td>-95.726</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-5.355</td>
<td>.006</td>
<td>-.392</td>
<td>-87.523</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>.613</td>
<td>.011</td>
<td>.267</td>
<td>58.024</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNAC
Appendix F - Regression Analyses

Charts

Histogram
Dependent Variable: LNAC

Normal P-P Plot of Regression Standardized Residual
Dependent Variable: LNAC

Scatterplot
Dependent Variable: LNAC

Residual Plot for In(p)

Residual Plot for ln(1-\(\hat{e}\))

Residual Plot for ln(|\(\hat{e}\)|)

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.831</td>
<td>.660</td>
<td>.660</td>
<td>.552</td>
</tr>
<tr>
<td>2</td>
<td>.880</td>
<td>.774</td>
<td>.774</td>
<td>.460</td>
</tr>
<tr>
<td>3</td>
<td>.920</td>
<td>.847</td>
<td>.847</td>
<td>.379</td>
</tr>
<tr>
<td>4</td>
<td>.977</td>
<td>.959</td>
<td>.959</td>
<td>.363</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LN1J
d. Predictors: (Constant), LNK, LNN, LN1J, LNP
e. Dependent Variable: LNCC
### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Regression</td>
<td>2306.446</td>
<td>1</td>
<td>2306.446</td>
<td>7930.873</td>
<td>.000²</td>
</tr>
<tr>
<td>Residual</td>
<td>1024.733</td>
<td>3558</td>
<td>.291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3341.179</td>
<td>3559</td>
<td>.291</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Regression</td>
<td>2585.889</td>
<td>2</td>
<td>1292.945</td>
<td>6088.058</td>
<td>.000²</td>
</tr>
<tr>
<td>Residual</td>
<td>755.290</td>
<td>3557</td>
<td>.212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3341.179</td>
<td>3559</td>
<td>.212</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Regression</td>
<td>2830.477</td>
<td>3</td>
<td>943.492</td>
<td>6569.507</td>
<td>.000²</td>
</tr>
<tr>
<td>Residual</td>
<td>510.702</td>
<td>3556</td>
<td>.144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3341.179</td>
<td>3559</td>
<td>.144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Regression</td>
<td>2870.554</td>
<td>4</td>
<td>717.638</td>
<td>5420.383</td>
<td>.000²</td>
</tr>
<tr>
<td>Residual</td>
<td>470.625</td>
<td>3555</td>
<td>.132</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3341.179</td>
<td>3559</td>
<td>.132</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- a. Predictors: (Constant), LNK
- b. Predictors: (Constant), LNK, LNN
- c. Predictors: (Constant), LNK, LNN, LNJ
- d. Predictors: (Constant), LNK, LNN, LNJ, LNP
- e. Dependent Variable: LNCC

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>-1.740</td>
<td>.050</td>
<td>-53.764</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.835</td>
<td>.009</td>
<td>.831</td>
<td>89.015</td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>.924</td>
<td>.068</td>
<td>.926</td>
<td>104.221</td>
</tr>
<tr>
<td>LNK</td>
<td>.839</td>
<td>.008</td>
<td>.839</td>
<td>104.221</td>
</tr>
<tr>
<td>LNN</td>
<td>-.492</td>
<td>-.014</td>
<td>-.299</td>
<td>-66.277</td>
</tr>
<tr>
<td>3 (Constant)</td>
<td>1.124</td>
<td>.058</td>
<td>19.387</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.830</td>
<td>.007</td>
<td>.826</td>
<td>125.964</td>
</tr>
<tr>
<td>LNN</td>
<td>-.492</td>
<td>.011</td>
<td>-.299</td>
<td>-44.111</td>
</tr>
<tr>
<td>LNJ</td>
<td>.776</td>
<td>.019</td>
<td>.271</td>
<td>41.268</td>
</tr>
<tr>
<td>4 (Constant)</td>
<td>.492</td>
<td>.068</td>
<td>6.435</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.830</td>
<td>.006</td>
<td>.826</td>
<td>131.171</td>
</tr>
<tr>
<td>LNN</td>
<td>-.492</td>
<td>.011</td>
<td>-.299</td>
<td>-45.944</td>
</tr>
<tr>
<td>LNJ</td>
<td>.546</td>
<td>.018</td>
<td>.295</td>
<td>45.740</td>
</tr>
<tr>
<td>LNP</td>
<td>.231</td>
<td>.013</td>
<td>.112</td>
<td>17.399</td>
</tr>
</tbody>
</table>

- a. Dependent Variable: LNCC

### Charts

- Histogram
  - Dependent Variable: LNCC
- Normal P-P Plot of Regression Standardized Residual
  - Dependent Variable: LNCC
- Scatterplot
  - Dependent Variable: LNCC
- Residual Plot for In(y)
- Residual Plot for In(1-y)
Appendix F- Regression Analyses

F-3E
MULTIPLICATIVE REGRESSION MODEL FOR FL_{\text{est}}

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.629a</td>
<td>.395</td>
<td>.395</td>
<td>.7544</td>
</tr>
<tr>
<td>2</td>
<td>.848b</td>
<td>.716</td>
<td>.716</td>
<td>.5174</td>
</tr>
<tr>
<td>3</td>
<td>.889c</td>
<td>.791</td>
<td>.791</td>
<td>.4438</td>
</tr>
<tr>
<td>4</td>
<td>.869d</td>
<td>.792</td>
<td>.792</td>
<td>.4429</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LN1J
d. Predictors: (Constant), LNK, LNN, LN1J, LNP

c. Dependent Variable: LNEM

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>914.688</td>
<td>1</td>
<td>914.688</td>
<td>1567.244</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1398.286</td>
<td>2457</td>
<td>.569</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2312.974</td>
<td>2458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>1955.558</td>
<td>2</td>
<td>927.778</td>
<td>3092.433</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>657.418</td>
<td>2456</td>
<td>.268</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2312.974</td>
<td>2458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>1825.406</td>
<td>3</td>
<td>609.802</td>
<td>3095.870</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>463.568</td>
<td>2455</td>
<td>.197</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2312.974</td>
<td>2458</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>1832.516</td>
<td>4</td>
<td>458.129</td>
<td>2339.951</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>480.458</td>
<td>2454</td>
<td>.196</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2312.974</td>
<td>2458</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LN1J
d. Predictors: (Constant), LNK, LNN, LN1J, LNP
e. Dependent Variable: LNEM

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td></td>
<td>-4.306</td>
<td>.055</td>
<td></td>
<td>-79.170</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td></td>
<td>.015</td>
<td>.015</td>
<td>.020</td>
<td>40.090</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td></td>
<td>8.114e-02</td>
<td>.003</td>
<td></td>
<td></td>
<td>.381</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td></td>
<td>.017</td>
<td>.011</td>
<td>.012</td>
<td>58.608</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td></td>
<td>.964</td>
<td>.018</td>
<td>.964</td>
<td>-52.609</td>
<td>.000</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td></td>
<td>.649</td>
<td>.082</td>
<td></td>
<td>7.942</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td></td>
<td>.612</td>
<td>.008</td>
<td>.620</td>
<td>67.779</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td></td>
<td>.962</td>
<td>.016</td>
<td>.962</td>
<td>-61.237</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td></td>
<td>.772</td>
<td>.039</td>
<td>.772</td>
<td>27.090</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td></td>
<td>.903</td>
<td>.103</td>
<td></td>
<td>8.750</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td></td>
<td>.508</td>
<td>.009</td>
<td>.521</td>
<td>67.039</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td></td>
<td>.981</td>
<td>.016</td>
<td>.981</td>
<td>-61.371</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td></td>
<td>.760</td>
<td>.025</td>
<td>.760</td>
<td>29.159</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td></td>
<td>.936e-02</td>
<td>.024</td>
<td></td>
<td>-4.386</td>
<td>.000</td>
</tr>
</tbody>
</table>

b. Dependent Variable: LNEM
Appendix F: Regression Analyses

Charts

Histogram
Dependent Variable: LNEM

Normal P-P Plot of Regression Standardized Residual
Dependent Variable: LNEM

Scatterplot
Dependent Variable: LNEM

Residual Plot for ln(p)

Residual Plot for ln(s)

Residual Plot for ln(LNK)

Residual Plot for ln(LNP)

F-3F
MULTIPLICATIVE REGRESSION MODEL FOR FL-om WITH N=50 AND P=50 REMOVED

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.670a</td>
<td>.449</td>
<td>.448</td>
<td>.6792</td>
</tr>
<tr>
<td>2</td>
<td>.668a</td>
<td>.440</td>
<td>.450</td>
<td>.6589</td>
</tr>
<tr>
<td>3</td>
<td>.524a</td>
<td>.393</td>
<td>.393</td>
<td>.5246</td>
</tr>
<tr>
<td>4</td>
<td>.946a</td>
<td>.927</td>
<td>.892</td>
<td>.2369</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LNJ
d. Predictors: (Constant), LNK, LNN, LNJ, LNP
e. Dependent Variable: LNEM
### Appendix F - Regression Analyses

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>867.133</td>
<td>1</td>
<td>867.133</td>
<td>1930.656</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>1066.256</td>
<td>2374</td>
<td>.449</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1933.388</td>
<td>2375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>1431.428</td>
<td>2</td>
<td>715.714</td>
<td>3383.517</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>501.960</td>
<td>2373</td>
<td>.212</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1933.388</td>
<td>2375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>1649.283</td>
<td>3</td>
<td>549.761</td>
<td>4589.966</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>294.105</td>
<td>2372</td>
<td>.120</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1933.388</td>
<td>2375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>1724.414</td>
<td>4</td>
<td>431.103</td>
<td>4891.246</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>206.875</td>
<td>2371</td>
<td>8.814E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1933.388</td>
<td>2375</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a. Predictors: (Constant), LNK  
b. Predictors: (Constant), LNK, LNN  
c. Predictors: (Constant), LNK, LNN, LN1J  
d. Predictors: (Constant), LNK, LNN, LN1J, LNP  
e. Dependent Variable: LNEM

#### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>Std. Error</th>
<th>Beta</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-4.459</td>
<td>-867.133</td>
<td>-867.133</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.609</td>
<td>.609</td>
<td>.609</td>
<td>230.299</td>
<td>.000</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-4.445</td>
<td>-5.236</td>
<td>-5.236</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.611</td>
<td>.611</td>
<td>.611</td>
<td>230.299</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.864</td>
<td>-.540</td>
<td>-.540</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>-1.42</td>
<td>.666</td>
<td>.666</td>
<td>2.171</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.605</td>
<td>.605</td>
<td>.605</td>
<td>2.171</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.851</td>
<td>-.532</td>
<td>-.532</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>.877</td>
<td>.338</td>
<td>.338</td>
<td>42.648</td>
<td>.000</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>1.387</td>
<td>93.591</td>
<td>93.591</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.590</td>
<td>.638</td>
<td>.638</td>
<td>9.520</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.809</td>
<td>-.506</td>
<td>-.506</td>
<td>7.498</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>.835</td>
<td>.320</td>
<td>.320</td>
<td>47.169</td>
<td>.000</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-.518</td>
<td>-.299</td>
<td>-.299</td>
<td>29.196</td>
<td>.000</td>
</tr>
</tbody>
</table>

*a. Dependent Variable: LNEM

### Charts

- **Histogram**  
  - Dependent Variable: LNEM
  - Observed Cum Prob
  - Normal P-P Plot of Regression Standardized Residual

- **Scatterplot**  
  - Dependent Variable: LNEM
  - Observed Cum Prob
  - Regression Standardized Residual

- **Residual Plot for In(n)**
  - Observed Cum Prob
  - Residual Plot for ln(n)

- **Residual Plot for ln(1-δ)**
  - Observed Cum Prob
  - Standardized Residual
F-4
MULTIPLICATIVE REGRESSION MODEL FOR SV

This section includes only the multiplicative regression models for SV for the various techniques.

F-4A
MULTIPLICATIVE REGRESSION MODEL FOR SVpc

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.959b</td>
<td>.876</td>
<td>.876</td>
<td>.4782</td>
</tr>
<tr>
<td>2</td>
<td>.918b</td>
<td>.843</td>
<td>.843</td>
<td>.3329</td>
</tr>
<tr>
<td>3</td>
<td>.932c</td>
<td>.889</td>
<td>.888</td>
<td>.3045</td>
</tr>
<tr>
<td>4</td>
<td>.941d</td>
<td>.888</td>
<td>.886</td>
<td>.2833</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNJ
c. Predictors: (Constant), LNK, LNJ, LNN
d. Predictors: (Constant), LNK, LNJ, LNN, LNP
e. Dependent Variable: LNDPC

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1694.146</td>
<td>1</td>
<td>1694.146</td>
<td>7409.473</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>813.294</td>
<td>3557</td>
<td>.229</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2507.440</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2113.350</td>
<td>2</td>
<td>1056.675</td>
<td>9534.710</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>394.090</td>
<td>3556</td>
<td>.111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2507.440</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2177.838</td>
<td>3</td>
<td>725.946</td>
<td>7929.998</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>329.092</td>
<td>3555</td>
<td>9.27E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2507.440</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2222.115</td>
<td>4</td>
<td>555.529</td>
<td>6919.649</td>
<td>.0002</td>
</tr>
<tr>
<td></td>
<td>285.325</td>
<td>3554</td>
<td>8.00E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2507.440</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
<td></td>
<td>-3.432</td>
<td>.009</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.716</td>
<td>.008</td>
<td>.622</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
<td></td>
<td>-2.645</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.710</td>
<td>.006</td>
<td>.615</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>1.016</td>
<td>.017</td>
<td>.409</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
<td></td>
<td>-1.558</td>
<td>.047</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.708</td>
<td>.005</td>
<td>.815</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>1.016</td>
<td>.015</td>
<td>.409</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-2.36</td>
<td>.009</td>
<td>-1.60</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
<td></td>
<td>-8.377</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.710</td>
<td>.006</td>
<td>.015</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>.942</td>
<td>.014</td>
<td>.379</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-2.36</td>
<td>.008</td>
<td>-1.60</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-2.43</td>
<td>.010</td>
<td>-1.36</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNDPC
Appendix F - Regression Analyses

F-4B
MULTIPlicative REGRESSION MODEL FOR SV_mn

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.787</td>
<td>.620</td>
<td>.620</td>
<td>.4053</td>
</tr>
<tr>
<td>2</td>
<td>.890</td>
<td>.792</td>
<td>.792</td>
<td>.2982</td>
</tr>
<tr>
<td>3</td>
<td>.945</td>
<td>.890</td>
<td>.890</td>
<td>.2175</td>
</tr>
<tr>
<td>4</td>
<td>.944</td>
<td>.891</td>
<td>.891</td>
<td>.2165</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LNP
d. Predictors: (Constant), LNK, LNN, LNP, LNJ
e. Dependent Variable: LNIM
### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>944.111</td>
<td>1</td>
<td>944.111</td>
<td>599.122</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>578.626</td>
<td>356</td>
<td>.163</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1522.737</td>
<td>357</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>1206.621</td>
<td>2</td>
<td>603.310</td>
<td>8784.739</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>316.117</td>
<td>355</td>
<td>8.882E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1522.737</td>
<td>357</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>1354.617</td>
<td>3</td>
<td>451.539</td>
<td>9645.372</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>168.120</td>
<td>354</td>
<td>4.720E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1522.737</td>
<td>357</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>1356.125</td>
<td>4</td>
<td>339.031</td>
<td>7229.840</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>168.612</td>
<td>353</td>
<td>4.665E-02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1522.737</td>
<td>357</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LNP
d. Predictors: (Constant), LNK, LNN, LNP, LN1J
e. Dependent Variable: LNM

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Std. Error</th>
<th>Beta</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-3.00</td>
<td>-156.916</td>
<td>0.007</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>0.534</td>
<td>0.787</td>
<td>76.172</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-1.904</td>
<td>-38.300</td>
<td>0.044</td>
<td>0.787</td>
<td>103.040</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>0.534</td>
<td>0.005</td>
<td>78.767</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-0.477</td>
<td>-0.415</td>
<td>54.354</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>-0.217</td>
<td>-141.250</td>
<td>0.041</td>
<td>0.787</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>0.534</td>
<td>0.004</td>
<td>141.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-0.477</td>
<td>-0.415</td>
<td>74.743</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>0.433</td>
<td>-0.312</td>
<td>55.934</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>-0.200</td>
<td>-4.920</td>
<td>0.041</td>
<td>0.787</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>0.534</td>
<td>0.004</td>
<td>141.745</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-0.477</td>
<td>-0.415</td>
<td>74.743</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>0.424</td>
<td>-0.305</td>
<td>53.597</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>6.245E-02</td>
<td>0.011</td>
<td>0.008</td>
<td>5.671</td>
<td>0.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNM
**F-4C**

MULTIPLICATIVE REGRESSION MODEL FOR SV

---

### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.668a</td>
<td>.457</td>
<td>.457</td>
<td>.5321</td>
</tr>
<tr>
<td>2</td>
<td>.829b</td>
<td>.772</td>
<td>.772</td>
<td>.4253</td>
</tr>
<tr>
<td>3</td>
<td>.906c</td>
<td>.921</td>
<td>.921</td>
<td>.3143</td>
</tr>
<tr>
<td>4</td>
<td>.854d</td>
<td>.910</td>
<td>.910</td>
<td>.2228</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LN1J
c. Predictors: (Constant), LNK, LN1J, LNN
d. Predictors: (Constant), LNK, LN1J, LNN, LNP
e. Dependent Variable: LNAC

---

### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>954.102</td>
<td>1</td>
<td>954.102</td>
<td>3372.829</td>
<td>.0002</td>
</tr>
<tr>
<td>Residual</td>
<td>1007.044</td>
<td>3557</td>
<td>0.283</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1961.946</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1318.736</td>
<td>2</td>
<td>659.368</td>
<td>3645.331</td>
<td>.0002</td>
</tr>
<tr>
<td>Residual</td>
<td>643.210</td>
<td>3556</td>
<td>0.181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1961.946</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1610.23</td>
<td>3</td>
<td>536.941</td>
<td>5438.342</td>
<td>.0002</td>
</tr>
<tr>
<td>Residual</td>
<td>351.123</td>
<td>3555</td>
<td>9.837E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1961.946</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1785.587</td>
<td>4</td>
<td>448.397</td>
<td>8995.835</td>
<td>.0002</td>
</tr>
<tr>
<td>Residual</td>
<td>176.359</td>
<td>3554</td>
<td>4.962E-02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1961.946</td>
<td>3558</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>B</th>
<th>Std. Error</th>
<th>Beta</th>
<th>1</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td>-4.210</td>
<td></td>
<td></td>
<td>.092</td>
<td></td>
<td>-131.810</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.537</td>
<td></td>
<td></td>
<td>.099</td>
<td></td>
<td>58.076</td>
<td>.000</td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>-5.477</td>
<td></td>
<td></td>
<td>.030</td>
<td></td>
<td>-114.743</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.532</td>
<td></td>
<td></td>
<td>.007</td>
<td></td>
<td>71.870</td>
<td>.000</td>
</tr>
<tr>
<td>LN1J</td>
<td>.947</td>
<td></td>
<td></td>
<td>.021</td>
<td></td>
<td>44.849</td>
<td>.000</td>
</tr>
<tr>
<td>3 (Constant)</td>
<td>-1.162</td>
<td></td>
<td></td>
<td>.048</td>
<td></td>
<td>-24.164</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.532</td>
<td></td>
<td></td>
<td>.005</td>
<td></td>
<td>97.257</td>
<td>.000</td>
</tr>
<tr>
<td>LN1J</td>
<td>.946</td>
<td></td>
<td></td>
<td>.016</td>
<td></td>
<td>60.688</td>
<td>.000</td>
</tr>
<tr>
<td>LNN</td>
<td>-5.03</td>
<td></td>
<td></td>
<td>.009</td>
<td></td>
<td>-54.381</td>
<td>.000</td>
</tr>
<tr>
<td>4 (Constant)</td>
<td>.269</td>
<td></td>
<td></td>
<td>.042</td>
<td></td>
<td>6.438</td>
<td>.000</td>
</tr>
<tr>
<td>LNK</td>
<td>.532</td>
<td></td>
<td></td>
<td>.004</td>
<td></td>
<td>137.414</td>
<td>.000</td>
</tr>
<tr>
<td>LN1J</td>
<td>.800</td>
<td></td>
<td></td>
<td>.011</td>
<td></td>
<td>70.651</td>
<td>.000</td>
</tr>
<tr>
<td>LNN</td>
<td>-5.03</td>
<td></td>
<td></td>
<td>.007</td>
<td></td>
<td>-76.701</td>
<td>.000</td>
</tr>
<tr>
<td>LNP</td>
<td>.493</td>
<td></td>
<td></td>
<td>.008</td>
<td></td>
<td>-95.345</td>
<td>.000</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNAC

---

### Residual Plot of ln(1-y)

The Residual Plot of ln(1-y) shows the distribution of residuals, indicating that the model residuals are randomly distributed around zero, suggesting a good fit of the model to the data.
F.4D  
MULTIPlicative REGRESSION MODEL FOR $S_{ve}$

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.629*</td>
<td>.866</td>
<td>.688</td>
<td>.123</td>
</tr>
<tr>
<td>2</td>
<td>.899*</td>
<td>.809</td>
<td>.808</td>
<td>.093</td>
</tr>
<tr>
<td>3</td>
<td>.959*</td>
<td>.905</td>
<td>.903</td>
<td>.095</td>
</tr>
<tr>
<td>4</td>
<td>.969*</td>
<td>.903</td>
<td>.903</td>
<td>.263</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNJ

c. Predictors: (Constant), LNK, LNJ, LNN
d. Predictors: (Constant), LNK, LNJ, LNN, LNP

e. Dependent Variable: LNCC
Appendix F - Regression Analyses

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1</td>
<td>2117.056</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Residual</td>
<td>1</td>
<td>3559</td>
<td>.272</td>
<td>7790.529</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3</td>
<td>3559</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>2</td>
<td>1245.547</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Residual</td>
<td>2</td>
<td>3559</td>
<td>.167</td>
<td>7473.187</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>4</td>
<td>3559</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>3</td>
<td>927.925</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Residual</td>
<td>3</td>
<td>3556</td>
<td>8.441E-02</td>
<td>10993.241</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>5</td>
<td>3556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>4</td>
<td>696.452</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Residual</td>
<td>4</td>
<td>3555</td>
<td>8.388E-02</td>
<td>8304.855</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>6</td>
<td>3555</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LN1J
c. Predictors: (Constant), LNK, LN1J, LNN
d. Predictors: (Constant), LNK, LN1J, LNN, LNP

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-2.178</td>
<td>.051</td>
<td>-69.586</td>
</tr>
<tr>
<td>LNK</td>
<td>.800</td>
<td>.009</td>
<td>.829</td>
<td>88.284</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-1.435</td>
<td>.029</td>
<td>-49.334</td>
</tr>
<tr>
<td>LNK</td>
<td>.794</td>
<td>.007</td>
<td>.823</td>
<td>111.867</td>
</tr>
<tr>
<td>LN1J</td>
<td>.960</td>
<td>.020</td>
<td>.823</td>
<td>47.373</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>.882</td>
<td>.044</td>
<td>19.809</td>
</tr>
<tr>
<td>LNK</td>
<td>.704</td>
<td>.005</td>
<td>.823</td>
<td>157.194</td>
</tr>
<tr>
<td>LN1J</td>
<td>.960</td>
<td>.014</td>
<td>.823</td>
<td>66.568</td>
</tr>
<tr>
<td>LNN</td>
<td>-5.023</td>
<td>.009</td>
<td>-.308</td>
<td>-58.885</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>.728</td>
<td>.054</td>
<td>15.405</td>
</tr>
<tr>
<td>LNK</td>
<td>.794</td>
<td>.005</td>
<td>.822</td>
<td>157.690</td>
</tr>
<tr>
<td>LN1J</td>
<td>.975</td>
<td>.015</td>
<td>.823</td>
<td>68.255</td>
</tr>
<tr>
<td>LNN</td>
<td>-5.033</td>
<td>.009</td>
<td>-.308</td>
<td>-59.077</td>
</tr>
<tr>
<td>LNP</td>
<td>6.294E-02</td>
<td>.011</td>
<td>.026</td>
<td>4.952</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNCC
F-4E
MULTIPLICATIVE REGRESSION MODEL FOR $S_{V_m}^n$

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.878$^a$</td>
<td>.459</td>
<td>.459</td>
<td>.7977</td>
</tr>
<tr>
<td>2</td>
<td>.796$^b$</td>
<td>.637</td>
<td>.637</td>
<td>.5533</td>
</tr>
<tr>
<td>3</td>
<td>.844$^c$</td>
<td>.712</td>
<td>.711</td>
<td>.5828</td>
</tr>
<tr>
<td>4</td>
<td>.844$^d$</td>
<td>.712</td>
<td>.712</td>
<td>.5820</td>
</tr>
</tbody>
</table>

- $a$: Predictors: (Constant), LNK
- $b$: Predictors: (Constant), LNK, LNN
- $c$: Predictors: (Constant), LNK, LNN, LNJ
- $d$: Predictors: (Constant), LNK, LNN, LNJ, LNP
- $e$: Dependent Variable: LNEM

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1526.154</td>
<td>1</td>
<td>1326.154</td>
<td>2083.886</td>
<td>.000$^a$</td>
</tr>
<tr>
<td>Residual</td>
<td>1562.326</td>
<td>2445</td>
<td>.636</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2988.480</td>
<td>2446</td>
<td>.636</td>
<td>2083.886</td>
<td>.000$^a$</td>
</tr>
<tr>
<td>Regression</td>
<td>1841.004</td>
<td>2</td>
<td>920.502</td>
<td>2156.528</td>
<td>.000$^b$</td>
</tr>
<tr>
<td>Residual</td>
<td>1047.476</td>
<td>2445</td>
<td>.457</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2888.480</td>
<td>2446</td>
<td>.457</td>
<td>2156.528</td>
<td>.000$^b$</td>
</tr>
<tr>
<td>Regression</td>
<td>2055.416</td>
<td>3</td>
<td>685.139</td>
<td>2017.425</td>
<td>.000$^c$</td>
</tr>
<tr>
<td>Residual</td>
<td>833.065</td>
<td>2453</td>
<td>.340</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2888.480</td>
<td>2456</td>
<td>.340</td>
<td>2017.425</td>
<td>.000$^c$</td>
</tr>
<tr>
<td>Regression</td>
<td>2057.988</td>
<td>4</td>
<td>514.497</td>
<td>1519.034</td>
<td>.000$^d$</td>
</tr>
<tr>
<td>Residual</td>
<td>830.493</td>
<td>2452</td>
<td>.339</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2888.480</td>
<td>2456</td>
<td>.339</td>
<td>1519.034</td>
<td>.000$^d$</td>
</tr>
</tbody>
</table>

- $a$: Predictors: (Constant), LNK
- $b$: Predictors: (Constant), LNK, LNN
- $c$: Predictors: (Constant), LNK, LNN, LNJ
- $d$: Predictors: (Constant), LNK, LNN, LNJ, LNP
- $e$: Dependent Variable: LNEM

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>.128</td>
<td>.128</td>
<td>.128</td>
<td>.128</td>
</tr>
<tr>
<td>LNK</td>
<td>.215</td>
<td>.215</td>
<td>.215</td>
<td>.215</td>
</tr>
<tr>
<td>LNN</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
<td>.625</td>
</tr>
<tr>
<td>LNJ</td>
<td>.215</td>
<td>.215</td>
<td>.215</td>
<td>.215</td>
</tr>
<tr>
<td>LNP</td>
<td>.583</td>
<td>.583</td>
<td>.583</td>
<td>.583</td>
</tr>
</tbody>
</table>

- $a$: Dependent Variable: LNEM
F.4F

MULTIPLICATIVE REGRESSION MODEL FOR SV_{em} WITH N = 50 AND P = 50 REMOVED

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.782*</td>
<td>.581</td>
<td>.581</td>
<td>.063</td>
</tr>
<tr>
<td>2</td>
<td>.948b</td>
<td>.719</td>
<td>.718</td>
<td>.035</td>
</tr>
<tr>
<td>3</td>
<td>.921c</td>
<td>.847</td>
<td>.847</td>
<td>.037</td>
</tr>
<tr>
<td>4</td>
<td>.941d</td>
<td>.865</td>
<td>.864</td>
<td>.033</td>
</tr>
</tbody>
</table>

- Predictors: (Constant), LNK
- Predictors: (Constant), LNK, LNN
- Predictors: (Constant), LNK, LNN, LN1N
- Predictors: (Constant), LNK, LNN, LN1N, LN1P
- Dependent Variable: LNEM
### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>1291.108</td>
<td>1</td>
<td>1291.108</td>
<td>.392</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>930.712</td>
<td>2373</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2221.819</td>
<td>2374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>1596.401</td>
<td>2</td>
<td>798.201</td>
<td>.264</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>625.418</td>
<td>2372</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2221.819</td>
<td>2374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>1882.843</td>
<td>3</td>
<td>627.514</td>
<td>.143</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>338.976</td>
<td>2371</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2221.819</td>
<td>2374</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>1965.545</td>
<td>4</td>
<td>491.386</td>
<td>.108</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>256.274</td>
<td>2370</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2221.819</td>
<td>2374</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNK
b. Predictors: (Constant), LNK, LNN
c. Predictors: (Constant), LNK, LNN, LN1J
d. Predictors: (Constant), LNK, LNN, LN1J, LNP
e. Dependent Variable: LNEM

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-.777</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.749</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-.920</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.744</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.636</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>-.146</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.738</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.651</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>1.006</td>
</tr>
<tr>
<td>4</td>
<td>(Constant)</td>
<td>1.138</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.712</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>-.576</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>.961</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>-.543</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNEM
F-5
MULTIPLICATIVE REGRESSION MODEL FOR T

This section includes only the multiplicative regression models of T for the various techniques.

F-5A
MULTIPLICATIVE REGRESSION MODEL FOR T

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.895</td>
<td>.801</td>
<td>.799</td>
<td>.409</td>
</tr>
<tr>
<td>2</td>
<td>.968</td>
<td>.991</td>
<td>.991</td>
<td>.524E-02</td>
</tr>
<tr>
<td>3</td>
<td>.988</td>
<td>.996</td>
<td>.996</td>
<td>7.122E-02</td>
</tr>
<tr>
<td>4</td>
<td>.988</td>
<td>.996</td>
<td>.996</td>
<td>6.243E-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>64.409</td>
<td>1</td>
<td>.64.409</td>
<td>318.595</td>
<td>.000*</td>
</tr>
<tr>
<td>Residual</td>
<td>15.571</td>
<td>79</td>
<td>.202</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80.380</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>79.672</td>
<td>2</td>
<td>39.835</td>
<td>4392.072</td>
<td>.000**</td>
</tr>
<tr>
<td>Residual</td>
<td>.707</td>
<td>76</td>
<td>9.070E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80.380</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>79.999</td>
<td>3</td>
<td>26.663</td>
<td>5255.976</td>
<td>.000**</td>
</tr>
<tr>
<td>Residual</td>
<td>.391</td>
<td>77</td>
<td>5.073E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80.380</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>80.083</td>
<td>4</td>
<td>20.021</td>
<td>5137.484</td>
<td>.000**</td>
</tr>
<tr>
<td>Residual</td>
<td>.296</td>
<td>76</td>
<td>3.897E-03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>80.380</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Predictors: (Constant), LNP, LNN, LNJ

d. Predictors: (Constant), LNP, LNN, LNJ, LNK

Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Ustandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>[Constant]</td>
<td>-1.478</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>1.807</td>
</tr>
<tr>
<td>2</td>
<td>[Constant]</td>
<td>-5.011</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>1.807</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.767</td>
</tr>
<tr>
<td>3</td>
<td>[Constant]</td>
<td>-4.876</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>1.807</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.767</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>.180</td>
</tr>
<tr>
<td>4</td>
<td>[Constant]</td>
<td>-4.771</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>1.807</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.767</td>
</tr>
<tr>
<td></td>
<td>LNJ</td>
<td>.206</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>2.798E-02</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNDPCSTM
Appendix F - Regression Analyses

F-5B
MULTIPLICATIVE REGRESSION MODEL FOR $T_{in}$

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>$R$</th>
<th>$R^2$</th>
<th>Adjusted $R^2$</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.899</td>
<td>.791</td>
<td>.789</td>
<td>.2145</td>
</tr>
<tr>
<td>2</td>
<td>1.000</td>
<td>.999</td>
<td>.999</td>
<td>3.174E-02</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Dependent Variable: LNIMTM

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>$F$</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>79.221</td>
<td>1</td>
<td>79.221</td>
<td>299.193</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>20.918</td>
<td>79</td>
<td>.265</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100.139</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>100.001</td>
<td>2</td>
<td>50.000</td>
<td>49647.713</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>7.860E-02</td>
<td>76</td>
<td>1.000E-03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>100.139</td>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Dependent Variable: LNIMTM
Appendix F - Regression Analyses

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>-3.092</td>
<td>.384</td>
<td></td>
<td>-8.044</td>
</tr>
<tr>
<td>LNP</td>
<td>2.004</td>
<td>.116</td>
<td>.889</td>
<td>17.297</td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>-7.219</td>
<td>.037</td>
<td>.889</td>
<td>-193.813</td>
</tr>
<tr>
<td>LNP</td>
<td>2.004</td>
<td>.007</td>
<td>.889</td>
<td>280.386</td>
</tr>
<tr>
<td>LNN</td>
<td>.896</td>
<td>.006</td>
<td>.456</td>
<td>143.905</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNIMTM

### Charts

- **Histogram**: Dependent Variable: LNIMTM
- **Normal P-P Plot of Regression Standardized Residual**: Dependent Variable: LNIMTM
- **Scatterplot**: Dependent Variable: LNIMTM
- **Residual Plot for In(y)**

### F-5C

**MULTIPLICATIVE REGRESSION MODEL FOR T_{nc}**

#### Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.918²</td>
<td>.943</td>
<td>.941</td>
<td>.0412</td>
</tr>
<tr>
<td>2</td>
<td>.999²</td>
<td>.999</td>
<td>.997</td>
<td>.0000E-02</td>
</tr>
<tr>
<td>3</td>
<td>.999²</td>
<td>.998</td>
<td>.998</td>
<td>.0000E-02</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Predictors: (Constant), LNP, LNN, LNK
d. Dependent Variable: LNACTM

#### ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>82.536</td>
<td>1</td>
<td>82.536</td>
<td>424.067</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>15.376</td>
<td>79</td>
<td>.195</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>97.911</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>97.648</td>
<td>2</td>
<td>48.824</td>
<td>14485.474</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>.263</td>
<td>78</td>
<td>3.371E-03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>97.911</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>97.714</td>
<td>3</td>
<td>32.571</td>
<td>12723.605</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>.197</td>
<td>77</td>
<td>2.560E-03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>97.911</td>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Predictors: (Constant), LNP, LNN, LNK
d. Dependent Variable: LNACTM
Appendix F- Regression Analyses

F.5.D
MULTIPLICATIVE REGRESSION MODEL FOR T^D

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.778</td>
<td>.602</td>
<td>.597</td>
<td>.5141</td>
</tr>
<tr>
<td>2</td>
<td>.904</td>
<td>.747</td>
<td>.741</td>
<td>.4120</td>
</tr>
<tr>
<td>3</td>
<td>.944</td>
<td>.891</td>
<td>.867</td>
<td>.2720</td>
</tr>
<tr>
<td>4</td>
<td>.949</td>
<td>.901</td>
<td>.866</td>
<td>.2614</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNN
c. Predictors: (Constant), LNP, LNN, LNK
d. Predictors: (Constant), LNP, LNN, LNK, LN1J
e. Dependent Variable: LNACTM

Charts

- Residual Plot for ln(\theta)
- Residual Plot for ln(\eta)
- Normal P-P Plot of Regression Standardized Residuals
- Histogram
- Scatterplot

Coefficients*

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficient</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>-2.948</td>
<td>.528</td>
<td>.916</td>
<td>-5.642</td>
</tr>
<tr>
<td>LNP</td>
<td>2.046</td>
<td>.099</td>
<td>.916</td>
<td>20.590</td>
</tr>
<tr>
<td>2 (Constant)</td>
<td>-3.303</td>
<td>.068</td>
<td>.916</td>
<td>-39.453</td>
</tr>
<tr>
<td>LNP</td>
<td>2.046</td>
<td>.013</td>
<td>.916</td>
<td>155.484</td>
</tr>
<tr>
<td>LNN</td>
<td>.782</td>
<td>.011</td>
<td>.933</td>
<td>68.981</td>
</tr>
<tr>
<td>3 (Constant)</td>
<td>-4.311</td>
<td>.031</td>
<td>.933</td>
<td>-105.889</td>
</tr>
<tr>
<td>LNP</td>
<td>2.046</td>
<td>.011</td>
<td>.918</td>
<td>179.559</td>
</tr>
<tr>
<td>LNN</td>
<td>.763</td>
<td>.010</td>
<td>.933</td>
<td>76.826</td>
</tr>
<tr>
<td>LNK</td>
<td>2.23E-02</td>
<td>.004</td>
<td>.904</td>
<td>-5.070</td>
</tr>
</tbody>
</table>

* Dependent Variable: LNACTM
### ANOVA*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>31.531</td>
<td>1</td>
<td>.0044</td>
<td>119.319</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>20.877</td>
<td>79</td>
<td>.2605</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>52.408</td>
<td>80</td>
<td>.6546</td>
<td>.0000</td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>39.169</td>
<td>2</td>
<td>.1953</td>
<td>115.353</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>13.342</td>
<td>78</td>
<td>.1705</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>52.408</td>
<td>80</td>
<td>.6546</td>
<td>.0000</td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>46.712</td>
<td>3</td>
<td>15.571</td>
<td>210.502</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>5.896</td>
<td>77</td>
<td>7.397E-02</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>52.408</td>
<td>80</td>
<td>.6546</td>
<td>.0000</td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>47.213</td>
<td>4</td>
<td>11.803</td>
<td>173.675</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>5.195</td>
<td>76</td>
<td>6.835E-02</td>
<td>.0000</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>52.408</td>
<td>80</td>
<td>.6546</td>
<td>.0000</td>
</tr>
</tbody>
</table>

---

**a. Predictors:** (Constant), LNP

**b. Predictors:** (Constant), LNP, LNN

**c. Predictors:** (Constant), LNP, LNN, LNK

**d. Predictors:** (Constant), LNP, LNN, LNK, LN1J

**e. Dependent Variable:** LNCCTM

### Coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-3.526</td>
<td>.1265</td>
<td>.776</td>
<td>-9.185</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.1265</td>
<td>.116</td>
<td>10.023</td>
</tr>
<tr>
<td>2</td>
<td>-6.024</td>
<td>.493</td>
<td>.776</td>
<td>-12.488</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.542</td>
<td>.093</td>
<td>13.628</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.542</td>
<td>.382</td>
<td>8.706</td>
</tr>
<tr>
<td>3</td>
<td>-6.757</td>
<td>.327</td>
<td>.776</td>
<td>-20.558</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.542</td>
<td>.093</td>
<td>20.646</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.542</td>
<td>.382</td>
<td>10.160</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>-2.413</td>
<td>.337</td>
<td>-10.100</td>
</tr>
<tr>
<td>4</td>
<td>-6.984</td>
<td>.559</td>
<td>.776</td>
<td>-21.461</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>.542</td>
<td>.093</td>
<td>21.476</td>
</tr>
<tr>
<td></td>
<td>LNN</td>
<td>.542</td>
<td>.382</td>
<td>10.569</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>-2.257</td>
<td>.093</td>
<td>-10.849</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>-2.234</td>
<td>.093</td>
<td>-2.706</td>
</tr>
</tbody>
</table>

---

**a. Dependent Variable:** LNCCTM

### Charts

- **Histogram**: Dependent Variable: LNCCTM
- **Normal P-P Plot**: of Regression Standardized Residual
- **Scatterplot**: Dependent Variable: LNCCTM
- **Residual Plot for ln(p)**
- **Residual Plot for ln(1-p)**
Appendix F - Regression Analyses

F-5E
MULTIPLICATIVE REGRESSION MODEL FOR T_{on}

Model Summary

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.873a</td>
<td>.762</td>
<td>.759</td>
<td>.0035</td>
</tr>
<tr>
<td>2</td>
<td>.946b</td>
<td>.900</td>
<td>.898</td>
<td>.0032</td>
</tr>
<tr>
<td>3</td>
<td>.955c</td>
<td>.912</td>
<td>.908</td>
<td>.0036</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNK
c. Predictors: (Constant), LNP, LNK, LN1J
d. Dependent Variable: LNEMTM

ANOVA

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>218.932</td>
<td>1</td>
<td>218.932</td>
<td>253.383</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>66.259</td>
<td>79</td>
<td>.864</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>287.190</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Regression</td>
<td>258.531</td>
<td>2</td>
<td>129.266</td>
<td>355.815</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>19.659</td>
<td>76</td>
<td>.367</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>278.190</td>
<td>80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td>281.859</td>
<td>3</td>
<td>97.926</td>
<td>285.322</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>25.332</td>
<td>77</td>
<td>.329</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>287.190</td>
<td>80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), LNP
b. Predictors: (Constant), LNP, LNK
c. Predictors: (Constant), LNP, LNK, LN1J
d. Dependent Variable: LNEMTM

coefficients

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-4.498</td>
<td>-.664</td>
<td>-6.749</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>3.332</td>
<td>.209</td>
<td>15.918</td>
</tr>
<tr>
<td>2</td>
<td>(Constant)</td>
<td>-2.220</td>
<td>.451</td>
<td>-5.867</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>3.332</td>
<td>.157</td>
<td>24.410</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.501</td>
<td>.053</td>
<td>10.382</td>
</tr>
<tr>
<td>3</td>
<td>(Constant)</td>
<td>-2.235</td>
<td>.451</td>
<td>-4.556</td>
</tr>
<tr>
<td></td>
<td>LNP</td>
<td>3.332</td>
<td>.129</td>
<td>25.797</td>
</tr>
<tr>
<td></td>
<td>LNK</td>
<td>.594</td>
<td>.052</td>
<td>11.421</td>
</tr>
<tr>
<td></td>
<td>LN1J</td>
<td>.603</td>
<td>.190</td>
<td>3.180</td>
</tr>
</tbody>
</table>

a. Dependent Variable: LNEMTM
Techniques to Handle Missing Values in a Factor Analysis

by

Christopher Turville

A thesis presented to the University of Western Sydney Macarthur in partial fulfilment of the requirements for the degree of Doctor of Philosophy

March, 2000

© C Turville March 2000
Dedication

This thesis is dedicated to the memory of Dr Bobbie Vaile. Underneath her quiet demeanour and wonderful sense of humour was a fierce determination and drive. Her first and second year Physics lectures were inspirational. The splattering of eggs and flying squash balls in lectures are a few of the highlights etched in my memory forever. All that knew Bobbie were touched in different ways. Her words of encouragement and attitudes to life have provided me with a role model for my career. Bobbie’s life was cut short by a brain tumour, but she showed how to accomplish much in a short period of time.
Acknowledgements

I would like to acknowledge Assoc/Prof. Robert Mellor for his supervision, support, and contributions during the writing of this thesis. Thanks must also go to Dr Kevin Donegan for his attention to detail and grammar at the final stages of the writing process. I will never again be ambiguous with the word ‘this’. I would also like to acknowledge the contributions of Dr Alec Lee for all of the discussions that we had about various subjects throughout this thesis. His good sense of humour, gentle demeanor and open-door policy were invaluable as he listened to my rational and irrational arguments.

My eternal thanks must go to my parents Ken and Audrey Turville. Their encouragement and financial support have allowed me to achieve my goals. I must give a special thanks to Jacqueline Bond for her complete understanding and support throughout my PhD. The support and encouragement of my entire family and those mentioned above have helped me to achieve far more than I could have ever imagined.
Statement of Authentication

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged by the text. I hereby declare that I have not submitted this material, either in whole or in part, for a degree at this or any other institution.

Signed,

Christopher Turville
# Table of Contents

List of Tables vi
List of Figures vii
Abbreviations and Symbols x
Abstract xiii

<table>
<thead>
<tr>
<th>Chapter 1</th>
<th>Introduction</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Background to the Research</td>
<td>1</td>
</tr>
<tr>
<td>1-2</td>
<td>Research Problem</td>
<td>3</td>
</tr>
<tr>
<td>1-3</td>
<td>Justification for the Research</td>
<td>4</td>
</tr>
<tr>
<td>1-4</td>
<td>Methodology</td>
<td>5</td>
</tr>
<tr>
<td>1-5</td>
<td>Outline of this Thesis</td>
<td>6</td>
</tr>
<tr>
<td>1-6</td>
<td>Major Contributions</td>
<td>7</td>
</tr>
<tr>
<td>1-7</td>
<td>Conclusion</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 2</th>
<th>Basic Concepts</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>2-2</td>
<td>Multivariate Analysis</td>
<td>9</td>
</tr>
<tr>
<td>2-2.1</td>
<td>Multivariate Concepts</td>
<td>11</td>
</tr>
<tr>
<td>2-3</td>
<td>Factor Analysis</td>
<td>14</td>
</tr>
<tr>
<td>2-3.1</td>
<td>Model and Assumptions</td>
<td>15</td>
</tr>
<tr>
<td>2-3.2</td>
<td>Factor Rotation</td>
<td>18</td>
</tr>
<tr>
<td>2-3.3</td>
<td>Methods of Estimating Loadings</td>
<td>19</td>
</tr>
<tr>
<td>2-3.4</td>
<td>Methods that Estimate the Number of Factors</td>
<td>21</td>
</tr>
<tr>
<td>2-4</td>
<td>Missing Values</td>
<td>23</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>Techniques Investigated</td>
<td>Contents</td>
</tr>
<tr>
<td>------------</td>
<td>-------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>3-1</td>
<td>Introduction</td>
<td>26</td>
</tr>
<tr>
<td>3-2</td>
<td>Complete Cases Only</td>
<td>26</td>
</tr>
<tr>
<td>3-3</td>
<td>Imputing Means</td>
<td>31</td>
</tr>
<tr>
<td>3-4</td>
<td>All Available Cases</td>
<td>35</td>
</tr>
<tr>
<td>3-4.1</td>
<td>Covariance Mean from All Available Cases; Variance Mean from All Available Cases</td>
<td>35</td>
</tr>
<tr>
<td>3-4.2</td>
<td>Covariance Mean from Paired Cases, Variance Mean from All Available Cases</td>
<td>36</td>
</tr>
<tr>
<td>3-4.3</td>
<td>Covariance Mean from Paired Cases; Variance Mean from Paired Cases</td>
<td>36</td>
</tr>
<tr>
<td>3-5</td>
<td>Iterative Principal Component Method</td>
<td>38</td>
</tr>
<tr>
<td>3-6</td>
<td>Singular Value Decomposition Method</td>
<td>40</td>
</tr>
<tr>
<td>3-7</td>
<td>Regression Method</td>
<td>44</td>
</tr>
<tr>
<td>3-8</td>
<td>EM Algorithm</td>
<td>47</td>
</tr>
<tr>
<td>3-9</td>
<td>Other Computing Considerations</td>
<td>49</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Data Simulation</th>
<th>Contents</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-1</td>
<td>Introduction</td>
<td>53</td>
</tr>
<tr>
<td>4-2</td>
<td>Previous Attempts to Generate Random Multivariate Data</td>
<td>53</td>
</tr>
<tr>
<td>4-3</td>
<td>Outline of Method</td>
<td>55</td>
</tr>
<tr>
<td>4-4</td>
<td>Generating Data by Specifying all of the Eigenvalues of the Covariance Matrix</td>
<td>56</td>
</tr>
<tr>
<td>4-5</td>
<td>Generating Data with a Correlation Matrix of Specified Eigenvalue Structure</td>
<td>58</td>
</tr>
<tr>
<td>4-6</td>
<td>Generating Eigenvalues by Specifying the Number of Variables and an Index of Degree of Interdependence</td>
<td>63</td>
</tr>
<tr>
<td>4-7</td>
<td>Generating Eigenvalues by Specifying the Number of Variables and the Average Intercorrelation</td>
<td>65</td>
</tr>
<tr>
<td>4-7.1</td>
<td>Determining the Relationship Between the Index of Interdependence and the Average Intercorrelation</td>
<td>66</td>
</tr>
<tr>
<td>4-8</td>
<td>An Example to Produce Random Data by Specifying the Average Intercorrelation of the Correlation Matrix</td>
<td>71</td>
</tr>
</tbody>
</table>
Chapter 5  Analysis of Techniques to Estimate the Correlation Matrix  74

5-1 Introduction  74
5-2 Monte Carlo Design  75
  5-2.1 Study Design  75
  5-2.2 Measurements to Compare the Original and Estimated Correlation Matrices  77
  5-2.3 Tests for the Equality of Correlation Matrices  78
5-3 Results  81
  5-3.1 Comparison of Simulation Patterns  81
  5-3.2 Graphical Comparison of Three Techniques to Estimate the Correlation Matrix  85
  5-3.3 Regression Analysis of Errors  94
  5-3.4 Comparison with Other Proposed Methods  97
  5-3.5 Testing the Equality of the Original and Estimated Correlation Matrices  99
5-4 Summary of Study and Results  110

Chapter 6  Analysis of Techniques to Estimate Loadings and Specific Variances  113

6-1 Introduction  113
6-2 Study Design  114
6-3 Comparing Two Factor Analyses  116
6-4 Results  120
  6-4.1 Graphical Comparison of Factor Loadings and Specific Variances  120
  6-4.2 Regression Analysis of Factor Loadings  129
  6-4.3 Regression Analysis of Specific Variances  131
  6-4.4 Analysis of Computing Time  132
6-5 Summary of Results  135
## Chapter 7  Derivatives of Factor Loadings

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-1</td>
<td>Introduction</td>
<td>139</td>
</tr>
<tr>
<td>7-2</td>
<td>Changes in Eigenvalues with Respect to Changes in the Correlation Matrix</td>
<td>140</td>
</tr>
<tr>
<td>7-3</td>
<td>Changes in Eigenvectors with Respect to Changes in the Correlation Matrix</td>
<td>143</td>
</tr>
<tr>
<td>7-4</td>
<td>Changes in Factor Loadings with Respect to Changes in the Correlation Matrix</td>
<td>146</td>
</tr>
<tr>
<td>7-5</td>
<td>Example</td>
<td>147</td>
</tr>
</tbody>
</table>

## Chapter 8  Conclusions and Implications

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-1</td>
<td>Introduction</td>
<td>149</td>
</tr>
<tr>
<td>8-2</td>
<td>Conclusions about Research Questions</td>
<td>149</td>
</tr>
<tr>
<td>8-2.1</td>
<td>How are Random Multivariate Data Sets Generated that are Typical of that Encountered when a Factor Analysis is Applied?</td>
<td>150</td>
</tr>
<tr>
<td>8-2.2</td>
<td>Do all of the Missing Values Techniques Work for all Combinations when Data is Randomly Deleted?</td>
<td>153</td>
</tr>
<tr>
<td>8-2.3</td>
<td>How is the Difference Between Two Correlation Matrices Compared and Tested?</td>
<td>154</td>
</tr>
<tr>
<td>8-2.4</td>
<td>Which Technique is the Best Estimator of the Correlation Matrix of a Data Set Used for a Factor Analysis?</td>
<td>155</td>
</tr>
<tr>
<td>8-2.5</td>
<td>How are Two Factor Analyses Compared?</td>
<td>157</td>
</tr>
<tr>
<td>8-2.6</td>
<td>How do the Various Techniques Compare in Terms of Their Computing Time Costs?</td>
<td>157</td>
</tr>
<tr>
<td>8-2.7</td>
<td>What Effect does a Change in the Correlation Matrix have on the Factor Loadings?</td>
<td>158</td>
</tr>
<tr>
<td>8-3</td>
<td>Conclusions about the Research Problem</td>
<td>159</td>
</tr>
<tr>
<td>8-4</td>
<td>Limitations</td>
<td>162</td>
</tr>
<tr>
<td>8-5</td>
<td>Implications for Researchers</td>
<td>164</td>
</tr>
<tr>
<td>8-6</td>
<td>Implications for Further Research</td>
<td>168</td>
</tr>
<tr>
<td>Contents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>List of References</td>
<td>172</td>
<td></td>
</tr>
<tr>
<td>Appendices</td>
<td>179</td>
<td></td>
</tr>
<tr>
<td>Appendix A</td>
<td>Computer Programs of Techniques to Handle Missing Values</td>
<td>180</td>
</tr>
<tr>
<td>Appendix B</td>
<td>Computer Program to Generate Random Data</td>
<td>193</td>
</tr>
<tr>
<td>Appendix C</td>
<td>Computer Program to Compare Correlation Matrices</td>
<td>196</td>
</tr>
<tr>
<td>Appendix D</td>
<td>Computer Program to Compare FL and SV</td>
<td>201</td>
</tr>
<tr>
<td>Appendix E</td>
<td>Derivations</td>
<td>205</td>
</tr>
<tr>
<td>Appendix F</td>
<td>Regression Analyses</td>
<td>209</td>
</tr>
</tbody>
</table>
List of Tables

Table 3-1  Percentage of Cases Remaining Using Complete Cases Only
Table 3-2  Percentage of Situations when Complete Cases Only Fails
Table 5-1  Computing Time Required to Generate Random Data Matrices
Table 5-2  Comparing Average Errors for Both Simulation Patterns for \( n = 100, p = 15, k = 1\% \)
Table 5-3  Comparing Standard Deviations for Both Simulation Patterns of Errors for \( n = 100, p = 15, k = 1\% \)
Table 5-4  Comparing Average Errors for Both Simulation Patterns for \( n = 100, p = 15, k = 10\% \)
Table 5-5  Comparing Standard Deviations for Both Simulation Patterns of Errors for \( n = 100, p = 15, k = 10\% \)
Table 5-6  Summary of Regression Transformations for Imputing Means
Table 5-7  Summary of Regression Transformations for All Available Cases
Table 5-8  Summary of Gleason & Staelin’s Study Design
Table 5-9  Number of \( \hat{R} \) Significantly Different from \( R \) for Imputing Means (\( n = 50, p = 15 \))
Table 5-10 Number of \( \hat{R} \) Significantly Different from \( R \) for Imputing Means (\( n = 100, p = 15 \))
Table 5-11 Number of \( \hat{R} \) Significantly Different from \( R \) for Imputing Means (\( n = 200, p = 15 \))
Table 5-12 Number of \( \hat{R} \) Significantly Different from \( R \) for All Available Cases (\( n = 50, p = 15 \))
Table 5-13 Number of \( \hat{R} \) Significantly Different from \( R \) for All Available Cases (\( n = 100, p = 15 \))
Table 5-14 Number of \( \hat{R} \) Significantly Different from \( R \) for All Available Cases (\( n = 200, p = 15 \))
List of Tables

Table 5-15  Number of NPD Matrices for Imputing Means (n = 50, p = 15)
Table 5-16  Number of NPD Matrices for Imputing Means (n = 100, p = 15)
Table 5-17  Number of NPD Matrices for Imputing Means (n = 200, p = 15)
Table 5-18  Number of NPD Matrices for All Available Cases (n = 50, p = 15)
Table 5-19  Number of NPD Matrices for All Available Cases (n = 100, p = 15)
Table 5-20  Number of NPD Matrices for All Available Cases (n = 200, p = 15)
Table 5-21  Last Ten Eigenvalues for the Model Developed in Section 4-7
Table 6-1  Number of Deletion Patterns used for the EM Algorithm
Table 6-2  Computing Time Required for Iterative Principal Component Method
Table 6-3  Computing Time Required for Imputing Means
Table 6-4  Computing Time Required for All Available Cases
Table 6-5  Computing Time Required for Complete Cases Only
Table 6-6  Computing Time Required for EM Algorithm
List of Figures

Figure 4-1  Plot of $1.5 \cos^2 \theta + 0.5 \sin^2 \theta + 2 \sin \theta \cos \theta$
Figure 4-2  Three Dimensional Plot Showing the Relationship Between $\phi$ and $\nu$
Figure 4-3  Rotated Three Dimensional Plot of Figure 4-2
Figure 4-4  Rotated Three Dimensional Plot of Figure 4-2
Figure 5-1  Comparison of $E$ for Imputing Means, All Available Cases and Complete Cases Only ($p = 15$, $k = 1\%$)
Figure 5-2  Comparison of $E$ for Imputing Means, All Available Cases and Complete Cases Only ($p = 15$, $k = 2\%$)
Figure 5-3  Comparison of $E$ for Imputing Means, All Available Cases and Complete Cases Only ($p = 25$, $k = 1\%$)
Figure 5-4  Comparison of $E$ for Imputing Means, All Available Cases and Complete Cases Only ($p = 25$, $k = 2\%$)
Figure 5-5  Graphical Comparison of $E$ for All Available Cases and Imputing Means ($k = 1\%$)
Figure 5-6  Graphical Comparison of $E$ for All Available Cases and Imputing Means ($k = 2\%$)
Figure 5-7  Graphical Comparison of $E$ for All Available Cases and Imputing Means ($k = 5\%$)
Figure 5-8  Graphical Comparison of $E$ for All Available Cases and Imputing Means ($k = 10\%$)
Figure 6-1  Comparison of Four Techniques to Estimate Factor Loadings ($k = 1\%$)
Figure 6-2  Comparison of Four Techniques to Estimate Factor Loadings ($k = 5\%$)
Figure 6-3  Comparison of Four Techniques to Estimate Factor Loadings ($k = 10\%$)
Figure 6-4  Comparison of Four Techniques to Estimate Specific Variances ($k = 1\%$)
Figure 6-5  Comparison of Four Techniques to Estimate Specific Variances
            (k = 5%)
Figure 6-6  Comparison of Four Techniques to Estimate Specific Variances
            (k = 10%)
Figure 8-1  Flow Chart for Factor Loadings
Figure 8-2  Flow Chart for Specific Variances
Abbreviations and Symbols

ac  technique of all available cases
b   constant in geometric progression
C   positive semi-definite symmetric matrix
cc  technique of complete cases only
D   diagonal matrix
e   eigenvector
E   diagonal matrix of positive eigenvalues
$E_\alpha$ root-mean-square measure of the difference between two correlation matrices
em  EM algorithm
F   common factor vector
FL  root-mean-square measure of the difference between two factor loading vectors
$h_i^2$ communality
im  technique of imputing means
k   proportion of missing values
L   factor loading matrix
$\hat{L}$ estimated factor loading matrix
$\ell_{ij}$ loading of $i^{th}$ variable on $j^{th}$ factor
m   number of factors
M   indicator matrix
MAR missing at random
MCAR missing completely at random
n  number of cases
NPD  non-positive definite
p  number of variables
P  orthogonal matrix
pc  iterative principal component method
Q  random orthonormal matrix
R  sample correlation matrix
\hat{R}  estimated correlation matrix
S  sample covariance matrix
\hat{S}  estimated covariance matrix
s_{ii}  sample variance for i^{th} variable
sd  technique of singular value decomposition
SV  root-mean-square measure of the difference between two specific variance vectors
SVD  singular value decomposition
T  upper triangular matrix obtained from Cholesky decomposition
T_1, T_2  sufficient statistics
U  matrix whose columns contain eigenvectors of \( XX' \)
v  index of degree of interdependence among variables
V  matrix whose columns contain eigenvectors of \( X'X \)
W  matrix of weights
X  data matrix
\bar{x}  sample mean vector
X^*  random matrix of numbers
X_i  data matrix with i^{th} row deleted
X_j  data matrix with j^{th} column deleted
Z  standardised data matrix
\epsilon  vector of specific factors
\phi  condition number
\gamma_i  first principal component for i^{th} case
\phi  average intercorrelation among variables
\lambda  eigenvalue
\mu  vector of population means
Abbreviations and Symbols

\[ \rho \quad \text{population correlation matrix} \]
\[ \Sigma \quad \text{population covariance matrix} \]
\[ \Psi \quad \text{vector of specific variances} \]
\[ \psi_i \quad \text{specific variance} \]
A factor analysis typically involves a large collection of data on many variables. For various reasons, it is quite common for some of the data to be unrecorded and researchers must decide what to do about these missing values. There have been various techniques proposed in the literature to deal with the problem of missing values in multivariate analyses. Several comparative studies have been conducted that examine the ability of these techniques to estimate the missing values themselves, the covariance matrix or the correlation matrix of the data. This thesis specifically investigates the ability of these techniques to handle missing values in a factor analysis. The techniques investigated in this thesis are: complete cases only, all available cases, imputing means, an iterative principal component method, singular value decomposition and the EM algorithm.

Initially, a data set that is representative of that used for a factor analysis is simulated. Some of this data set are then randomly removed to represent missing values. The number of variables, the number of cases, the average intercorrelation among variables and the proportion of missing values are varied to investigate the performance of the various techniques over a wide range of conditions.

Some of the iterative techniques that handle missing values require an initial estimate of the correlation matrix. Graphical comparisons and regression analyses are utilised
to examine the abilities of complete cases only, imputing means and all available cases to estimate the correlation matrix. The conditions under which one technique performs better than the others are determined.

Several criteria are used to investigate the abilities of the techniques to handle missing values in a factor analysis. Graphical and regression analyses are utilised to examine the various techniques’ abilities to estimate the original factor loadings and specific variances. The computing time required to produce the correlation matrix that is input into the factor analysis is also used as a comparative measure. Overall, there is no one technique that performs best for all of the conditions studied. If the computing time is not an important concern for a researcher, then the EM algorithm is generally the most effective technique except when there are ill-conditioned matrices present. When computing time is of concern or ill-conditioned matrices are present, then the conditions when a particular technique outperforms the others are identified.

Finally, some theoretical considerations are introduced regarding the effects that changes in the correlation matrix will have on the loadings of a factor analysis. A complicated expression is derived that shows that the change in factor loadings as a result of a change in the elements of a correlation matrix involves components of eigenvectors and eigenvalues.