D.C.e

MISSING VALUES IN MULTIVARIATE STATISTICAL ANALYSIS

by

Roderick J. A. Little.

Thesis submitted for the degree of Doctor of Philosophy in the University of London, and the Diploma of Imperial College of Science and Technology.

October 1974.

ABSTRACT

The problem of carrying out standard statistical analyses when the data are incomplete is considered from two standpoints. A "maximum likelihood approach" postulates a distribution for the incomplete data and estimates the parameters by maximum likelihood. A "fitting approach" finds suitable fitted values for the missing data, and carries out a modified standard analysis on the completed data. If maximum likelihood estimates are found by Orchard and Woodbury's Missing Information Principle, the resulting methods also make sense from the "fitting approach" point of view, and are robust against departures from the underlying model. This is illustrated by the problem of estimating the means and covariances of p variables from N incomplete observations; multiple regression and autoregression analyses are also considered. The idea of "randomly missing" data is formalized, and some non-random deletion patterns, for example in censored data, are analysed.

ACKNOWLEDGEMENTS.

I am greatly indebted to Professor E.M.L. Beale for introducing me to this subject, and for his advice and many helpful suggestions at all stages of this work. I wish to thank my supervisor, Professor D.R. Cox, for his encouragement and his illuminating comments on missing value problems. My thanks are due also to Mrs. M. White for typing the thesis so expertly.

I am grateful to the Science Research Council for financial support.

TABLE OF CONTENTS.

Table of Contents ctd.

1 INTRODUCTION

1.1. The Problem and the Literature.

Consider the following situation. We have N experimental units, and for each unit we wish to observe the values of p variables $X_1, X_2, \ldots X_p$, some of which are stochastic. A complete data matrix $\mathbf{X}$ consists of the (Nxp) matrix of values

$$(x_{ij}) \qquad , (i = 1,2,\ldots,N \; ; \; j = 1,2,\ldots,p).$$

Given such a matrix, we can proceed to some kind of statistical analysis, such as multiple regression analysis, factor analysis, or principal component analysis.

Now suppose some of the x_{ij}'s are missing. For example, they have not been measured, or measured and then lost, or are known to have been misrecorded. The question is how to modify the standard analysis, given such a set of incomplete data.

We describe $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ as the ith observation, and this observation is complete if all the x_{ij}'s are observed, (j=1,2,...,p). A simple way of avoiding the problem of incomplete data is to reject any individual with incomplete observations, and to analyse the remaining complete data. This is feasible and widely practised when a large number of complete observations exist, but in many cases such a procedure would be impossible, or highly inefficient. For example in a linear regression, with p large, an observation may be rejected when the value of one covariate is missing, and this variable may be insignificant in the final regression equation. Then intuitively speaking this observation contains as much information as a complete observation as regards estimating the regression. Further it is quite conceivable to have no complete observations, particularly when p is large. In such cases we are compelled to find some way of exploiting the information in the incomplete data.

A number of authors have tackledthis problem. For reviews of the literature, see Afifi and Elashoff (1966), or more recently Orchard and Woodbury (1971). Two distinct approaches to the subject can be recognised, which are simply described as the "fitting" approach, and the "maximum likelihood" approach.

In a fitting approach we seek to fill the gaps in the data by estimates of the missing variables, which are in some sense close to the true values. We then proceed with the standard analysis of the completed data, perhaps with some modification to allow for any bias introduced in the fitting. It seems intuitively reasonable that a method constructed according to the following <u>Principles of Fitting</u> will be sensible:

(P1): Find good estimates of the missing values, according to some criterion such as unbiasedness and small mean square error.

(P2): Use a standard method on the completed data.

(P3): If necessary adjust the standard method to correct for systematic bias caused by the fitted values.

In the chapters that follow we consider more precisely what these principles involve.

In a maximum likelihood approach we postulate a distribution for some or all of the variables, and estimate unknown parameters of this distribution by maximum likelihood. A common model is to assume

$$\boldsymbol{x}_i \overset{i.i.d}{\sim} MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) ,$$

the multivariate normal distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and to estimate $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for incomplete data where the missing values are in a sense "randomly missing". Even for this, the simplest of distributional assumptions, the resulting maximum likelihood equations for a general pattern of missing values are complicated, and an iterative solution is needed. Anderson (1957) deserves special mention here, for introducing a simplifying factorization of the likelihood for the special case of <u>monotone</u> data: that is, when the variables $(X_1, X_2, \ldots, X_p)$ can be renumbered so that x_{ij} is always observed if $x_{i,j+1}$ is observed. Eventually, perhaps, the increased power of computers provided the necessary spur for the iterative calculation of the maximum likelihood estimates (m.l.e's) of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for a general pattern of missing values. See

Hartley and Hocking (1972).

We shall return to this solution later, but it is interesting to compare it with the fitting approach to this problem. In its simplest form this involves fitting values for the missing x_{ij}'s, and then forming the sample means

$$\hat{\bar{x}}_j = \frac{1}{N}\sum_{i=1}^{N} \hat{x}_{ij} \;, \qquad (j = 1,2,\ldots,p) \,,$$

and $\hat{S}$, the sample sum of squares and cross products (S.S.C.P.) matrix about the means, with (j,k)th element

$$\hat{s}_{jk} = \sum_{i=1}^{N} (\hat{x}_{ij} - \hat{\bar{x}}_j)(\hat{x}_{ik} - \hat{\bar{x}}_k) \;,$$

where $\hat{x}_{ij}$ refers to the observed or fitted value of x_{ij} . A crude estimate of a missing value is the mean of that variable over all the observations for which it is present. Another method is to estimate $\bar{x}_j$ and s_{jk} separately from all the observations where X_j is observed, and all the observations where X_j and X_k are observed. Here an adjustment may be necessary to ensure that $\hat{S}$ is positive definite. Such methods can produce considerable bias in the usual estimate, $\frac{1}{N-1}\hat{S}$, of the covariance matrix, and simulation studies have shown that the estimates can be misleading. See for example Haitovsky (1968).

Let us construct more efficient estimators by using the principles of fitting. Suppose that x_{ij} is missing, and that we fit $\hat{x}_{ij}$, some function of the data. If we assume a distribution for the data and assume x_{ij} has the same mean whether it is present or missing, we may seek $\hat{x}_{ij}$ so that

$$E(x_{ij}-\hat{x}_{ij}) = 0 \quad , \tag{1.1.1.}$$

for repeated sampling with the same pattern of missing values. Clearly the fitted value found by averaging X_j over the observations for which it is observed satisfies (1.1.1.). However a closer fit to x_{ij} is obtained if we ask that

$$E(x_{ij}-\hat{x}_{ij} \mid P_i) = 0 \quad , \tag{1.1.2.}$$

where P_i stands for the set of observed variables in the .ith observation, and the expectation is taken with these variables fixed. If $\boldsymbol{x}_i$ is multivariate normally distributed, $E(x_{ij} \mid P_i)$ is a linear combination of the known variables P_i, with regression coefficients which are known functions of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We are led to the method of fitting proposed by Buck (1960). An initial estimate of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is found from the complete observations, and these estimates are then used to fit for the missing values linear combinations of the known variables in each observation.

Sample means and covariances are then formed from the completed data. We can improve these estimates by considering the principle P3, for although the sample means are unbiased estimates of $\boldsymbol{\mu}$, $\frac{1}{N-1}\hat{\boldsymbol{S}}$ is a biased estimate of $\boldsymbol{\Sigma}$. To correct for this, we must add certain adjustments to the S.S.C.P. matrix of the completed data. Buck gave the correct adjustment for one missing variable, but failed to consider the necessary modification to $\hat{S}_{jk}$, when x_{ij} and x_{ik} are both missing in some observation i. We consider these adjustments in some detail in Chapter 3.

A final modification to Buck's method is to make the process iterative. The new estimates of (μ,Σ) replace the original estimates from the complete observations, and the procedure is repeated. Then iterate until there is no significant change in the estimates. We call the resulting method Iterated Buck. Figure 1 gives a diagrammatical summary of the procedure.

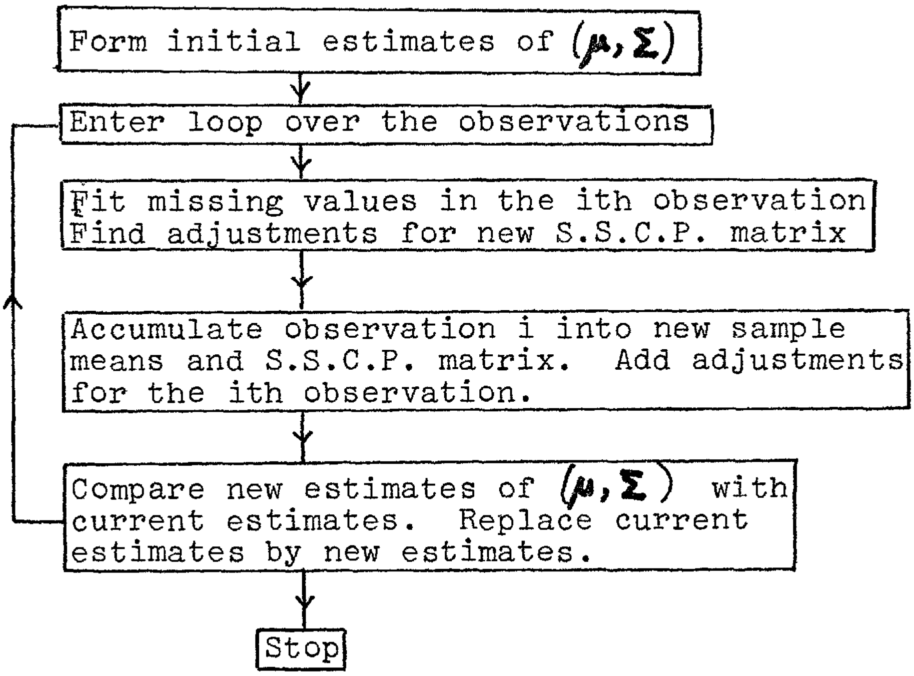

Form initial estimates of (μ,Σ)
Enter loop over the observations
Fit missing values in the ith observation
Find adjustments for new S.S.C.P. matrix
Accumulate observation i into new sample means and S.S.C.P. matrix. Add adjustments for the ith observation.
Compare new estimates of (μ,Σ) with current estimates. Replace current estimates by new estimates.
Stop

Figure 1.Flow Diagram for the estimation of (μ,Σ) by Iterated Buck

We consider the asymptotic unbiasedness of the estimates obtained by this method in Chapter 3, for a general distribution of x_i . A less detailed analysis is given by Beale and Little (1973).

We now return to the maximum likelihood approach, and the method of Iterated Buck forms a bridge. Orchard and Woodbury (1972) produced a set of iterative equations for obtaining m.l.e's of (μ, Σ) when x_i is multivariate normally distributed. These equations are much simpler to solve than those given by Hartley and Hocking, referred to above; in fact they are nearly identical to the equations of Iterated Buck, the difference being that the maximum likelihood method (called here M.L.M.) makes no correction for the degree of freedom in estimating the mean. Of interest here is Woodbury's contribution to the discussion of Hartley and Hocking's paper.

Orchard and Woodbury derived these equations by an application of a general principle for finding m.l.e's from incomplete data, their Missing Information Principle (M.I.P.). This simple and powerful idea plays an important role in this thesis, and in Chapter 2 we give a derivation of the principle, also given in Beale and Little (1973). The principle indicates the connection between the fitting approach and the maximum likelihood approach for a large class of problems, and the intuitive appeal of the equations found by applying

M.I.P. suggests that the resulting estimators will be robust against normality assumptions in the underlying model.

This survey of the literature has been very selective and biased, for example no mention has been made of many useful papers on univariate missing data problems. A more detailed treatment of past work may be found in the review papers mentioned above.

Chapters 2 and 3 of the thesis concern the problem of estimating the means and covariances of p variables; Chapter 2 considers the maximum likelihood approach and Chapter 3 the fitting approach. Chapter 4 concerns the precision of the estimates. Chapter 5 considers the linear regression of one variable on the other p-1 variables, and Chapter 6 considers the precision of the resulting estimates of the regression coefficients. Chapter 7 gives the result of simulation work to compare some of the methods proposed in Chapter 5. Finally Chapter 8 is a simple application of the Missing Information Principle to time series data, as an illustration of the potential value of the Principle in this field.

2. MAXIMUM LIKELIHOOD ESTIMATION OF THE MEANS AND COVARIANCE MATRIX.

2.1. Introduction.

A complete set of data consists of N independent observations $\mathbf{x}_i$ on a set of variables $X_1, X_2, \ldots, X_p$, which we suppose are multivariate normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We shall write $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T \overset{i.i.d.}{\sim} MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We consider the problem of estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by maximum likelihood, when some of the x_{ij}'s are missing. First we exclude cases where a value is missing because it is in some sense unusual (for example, too high to measure). The natural assumption is that P_i, the set of variables present in the ith observation, have the joint distribution found by integrating the missing variables, M_i, out of $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This property, assumed for each observation i, corresponds to a random pattern of deletions.

For this data and model, it is possible to write down the maximum likelihood equations for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and to solve them iteratively. (See, for example, Hartley and Hocking (1972)). However these equations are very involved, and a much simpler set of equations, which also give maximum likelihood estimates (m.l.e's) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, were found by Orchard and Woodbury (1971).

These were derived by an application of their Missing Information Principle (M.I.P.), which they explained in the same paper.

In §2.2 we derive the M.I.P. The argument follows that of Orchard and Woodbury, but we include a formal definition of the Principle, and emphasize that the effect of the principle is to replace a maximization problem by a fixed point problem. This section is the result of joint work with E.M.L. Beale, and is a revised form of the derivation of M.I.P. given at a R.S.S. Multivariate Study Group conference at Hull, as part of a joint paper (Beale and Little, 1973). This paper has been submitted for publication to the Journal of the Royal Statistical Society, Series 'B'.

We follow Orchard and Woodbury in showing that the principle leads to a simple iterative algorithmfor finding m.l.e's of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for our problem.

We noted that for the resulting estimates to be maximum likelihood, certain assumptions are involved about the circumstances which cause the data to be incomplete. These assumptions are formalized in §2.3, and some examples are given to indicate how to proceed when they do not hold ,that is the pattern of missing values is in a sense non-random. Included in this category are censored data.

2.2. Orchard and Woodbury's Missing Information Principle.

The Missing Information Principle is concerned with the situation in which there are random variables that can be grouped into two sets P and M, with a joint distribution depending on the vector θ of parameters, and where P have been observed but M have not been observed. In our application of the principle θ represents the set of means and the covariance matrix for the multivariate normal distribution, P represents the complete observations and the known variables in the incomplete observations, and M represents the missing values in the incomplete observations.

We wish to find $\hat{\theta}$, the estimate of θ that maximises the log-likelihood $\ell(P;\theta)$ of P given θ. But it may not be easy to compute this directly. On the other hand it may be much easier to find the value of θ that maximizes the log-likelihood $\ell(P,M;\theta)$ of P and M given θ, for any complete set of data (P,M). Furthermore we may be able to find the value of θ that maximizes the expected value of $\ell(P,M;\theta)$, if M is treated as a random variable with some known distribution. The appropriate formulae can sometimes be derived by imagining that the sample is replicated an arbitrary number of times, with P taking the same values in all replications, but with M having its known distribution. This procedure is central to the M.I.P., which is now described.

Let $f(M|P;\theta)$ denote the probability density function for the conditional distribution of M given P and θ, and let $\ell(M|P;\theta)$ denote $\ln f(M|P;\theta)$. Then

$$\ell(M,P;\theta) = \ell(P;\theta) + \ell(M|P;\theta) \,. \qquad (2.2.1)$$

Now take any assumed value θ_A for θ. This, together with the observed value of P, defines a conditional distribution for M, given the data. Take expectations of both sides of (2.2.1) over this distribution. We obtain

$$E\{\ell(M,P;\theta)|P;\theta_A\} = \ell(P;\theta) + E\{\ell(M|P;\theta)|P;\theta_A\} \qquad (2.2.2)$$

The left hand side of equation (2.2.2) is a function of P, θ and θ_A. We find θ_M, the value of θ that maximizes this function. This may depend on θ_A, so we write

$$\theta_M = \phi(\theta_A) \quad . \qquad (2.2.3)$$

Equation (2.2.3) represents a transformation from the vector θ_A to the vector θ_M. We now define

The Missing Information Principle

Estimate θ by a fixed point of the transformation ϕ, i.e. a value of θ such that

$$\theta = \phi(\theta) \quad . \qquad (2.2.4)$$

The set of equations (2.2.4) are called the <u>fixed point equations</u>, and they are analogues of the likelihood equations. This approach is justified by the following two theorems, which show that the m.l.e. of θ satisfies (2.2.4), and conversely, that every solution of the fixed point equations is a stationary value of the likelihood function. (Orchard and Woodbury implicitly define ϕ by differentiating the left hand side of (2.2.2) w.r.t. θ, and setting the result equal to zero. Defining ϕ as a maximization reduces the possibility of finding turning values of the likelihood other than local maxima). We assume regularity conditions which allow us to differentiate with respect to the parameters inside the expectation sign.

<u>Theorem 2.2.1</u>.

The maximum likelihood estimator $\hat{\theta}$ satisfies equation (2.2.4).

<u>Theorem 2.2.2</u>.

If $\ell(M|P;\theta)$ is a differentiable function of θ, then any other value of θ satisfying (2.2.4) is a maximum or stationary value of $\ell(P;\theta)$.

To prove the theorems, we observe that if the distribution of M has a probability density element $f(M|P;\theta_A)dM$, then

$$E[\ell(M|P;\theta)|P;\theta_A] = \int \ell(M|P;\theta)f(M|P;\theta_A)\,dM,$$

and regarded as a function of θ, this is maximized at $\theta = \theta_A$. This is simply Jensen's inequality. The proof is elementary: see for example Kendall and Stuart (1967), pp.39-40. Thus setting $\theta_A = \hat{\theta}$ in (2.2.2), the value $\theta = \hat{\theta}$ maximizes both terms on the right hand side of (2.2.2). It therefore maximizes their sum. This proves Theorem 1. To prove Theorem 2, we note that $\theta = \theta_A$ maximizes the second term on the right hand side (2.2.2), and by hypothesis this is differentiable. It cannot then be a maximum of the left hand side of (2.2.2) unless it is either a maximum or a stationary value of the first term on the right hand side.

The proofs carry over to the situation with discrete random variables in the set M; integrals becomes sums in the usual way. Also the theory can be rephrased to allow for partial information about the distribution of M. Then $\ell(P;\theta)$ represents the log likelihood of the data including this information, and $f(M|P;\theta)$ is the density for the missing data, given the data and the partial information. This is illustrated in §2.3.

We now apply this theory to our problem. Denote by X the complete $(N \times p)$ matrix of variables, by P_i the set of variables observed in observation i, and by P the total set of variables observed. Then in the above notation,

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

The log likelihood for the multivariate normal distribution is

$$\ell(X;\mu,\Sigma) = -\tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{p} \sum_{k=1}^{p} (x_{ij} - \mu_j)\sigma^{jk}(x_{ik} - \mu_k) - \tfrac{1}{2} N \ln(\det \Sigma),$$

where σ^{jk} denotes the(j,k)th element of Σ^{-1}. Taking expectations with $\theta = \theta_A$ and P fixed, we have that

$$E\{\ell(X;\mu,\Sigma)\,|\,P;\mu_A,\Sigma_A\} = -\tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{p} \sum_{k=1}^{p} [(\hat{x}_{ijA} - \mu_j)(\hat{x}_{ikA} - \mu_k) + \sigma_{jkA\cdot P_i}]\sigma^{jk} - \tfrac{1}{2} N \ln(\det \Sigma)\ ,$$

where

$$\hat{x}_{ijA} = E\{x_{ij}\,|\,P_i;\mu_A,\Sigma_A\}$$

and

$$\sigma_{jkA\cdot P_i} = Cov\{(x_{ij}, x_{ik})\,|\,P_i;\mu_A,\Sigma_A\}\ .$$

Maximization with respect to μ and Σ gives the analogue of (2.2.3):

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

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

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

$$\hat{x}_{ij} = E\{ x_{ij} \mid P_i ; \mu , \Sigma \} , \tag{2.2.5}$$

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ij} , \tag{2.2.6}$$

$$\sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} \{ (\hat{x}_{ij} - \mu_j)(\hat{x}_{ik} - \mu_k) + \sigma_{jk \cdot P_i} \} , \tag{2.2.7}$$

$$\sigma_{jk \cdot P_i} = Cov \{ (x_{ij} , x_{ik}) \mid P_i ; \mu , \Sigma \} . \tag{2.2.8}$$

These are the equationsfound by Orchard and Woodbury. To find m.l.e's we obtain initial estimates of (μ , Σ) and cycle through (2.2.5) - (2.2.8) until we find no significant changes in the estimates between successive iterations. Note that

$$\hat{x}_{ij} = \begin{cases} x_{ij} , \text{ if } x_{ij} \text{ is observed ;} \\ \text{a linear combination of the variable in } P_i \\ \qquad , \text{ if } \quad x_{ij} \quad \text{ is missing.} \end{cases}$$

At each iteration the data are completed by equation (2.2.5), and the means, and a sum of squares and products (S.S.C.P.) matrix found for the variables. This matrix is adjusted by adding $\sigma_{jk \cdot P_i}$ to the (j,k)th element for each observation i. This adjustment is zero unless both x_{ij} and x_{ik} are missing, and it depends only on the pattern of missing values in observation i, and not on the values themselves.

It is important that the computing involved in this procedure is very simple. For a missing x_{ij}, we fit

$$\hat{x}_{ijA} = \mu_{jA} + \sum_{\ell \in P_i} b_{j\ell A \cdot P_i}(x_{i\ell} - \mu_{\ell A}) , \qquad (2.2.9)$$

where $b_{j\ell A \cdot P_i}$ is the partial regression coefficient between X_j and X_ℓ when X_j is regressed on the variables in P_i, calculated from the current estimate Σ_A of the covariance matrix. These coefficients, and the adjustments $\sigma_{jkA \cdot P_i}$, are found at the same time by pivoting on the elements of Σ_A corresponding to the variables in P_i. In symbols,

if $$\mathrm{PIV}(P_i)\,\Sigma_A = V_{(p \times p)} ,$$

then $$v_{jk} = b_{jkA \cdot P_i} , \quad \text{for } x_{ij} \notin P_i , x_{ik} \in P_i ;$$

$$v_{jk} = \sigma_{jkA \cdot P_i} , \quad \text{for } x_{ij} \notin P_i , x_{ik} \notin P_i .$$

The PIV operator is defined in Appendix I, and some basic properties given. A discussion of this important computational point is given by Jowett (1963), or an expository account with more emphasis on computational aspects, by Beale (1970).

We describe the method of estimating (μ, Σ) in this section as M.L.N., to indicate maximum likelihood for a normal distribution. The method assumes a random pattern of deletions, and we consider this assumption in more detail in the next section.

2.3. Random Deletion Patterns.

In the previous section, we analysed a given partition (P,M) of the data, under the assumption that the observed variables had their marginal distribution after the missing variables had been integrated out of the joint distribution of all the variables. We now consider the partitioning process in more detail, by embedding the model of §2.2. in a larger model, where the partitioning process is stochastic. We consider when maximum likelihood for this larger model corresponds to maximum likelihood for the conditional model of §2.2.

We suppose data to be generated from the following model, which has two components:

(a) A joint distribution for a complete set of data X, indexed by a set of parameters θ, which we wish to estimate.

(b) A deletion mechanism, which causes the partition of X into two sets P and M, where P is the set of variables present, and M is the set of missing variables.

The data consists of P, and a deletion pattern D which is a set of binary variables, one corresponding to each variable in X, and taking value 0 if that variable is present, 1 if that variable is missing. The deletion pattern is considered stochastic, and has a distribution from a class defined by the deletion mechanism, indexed in general by θ and a set of additional parameters, ϕ.

In the following examples X represents an Nxp data matrix. We consider some possible deletion mechanisms.

DM1: Each x_{ij} has a known or unknown probability p_{ij} of deletion, which is functionally independent of the parameter θ. The parameter ϕ is the set of distinct p_{ij}'s, and we may specify $p_{ij}=p_i$, or more commonly $p_{ij}=p_j$, or some such relation. Anyway we assume the factorization of the underlying parameter space, $\Omega_{(\theta,\phi)} = \Omega_\theta \times \Omega_\phi$.

DM2: (Censored Data). We assume x_{ij} is observed if and only if $x_{ij} \in R_j$, where R_j is some range of values of the variable, usually an interval. This interval may be known, or the cut-off points may be additional parameters to be estimated.

DM3: Now suppose x_{ij} is observed if $x_{ik} \in R_k$, where R_k is some range of values of the variable X_k, $k \neq j$. Again R_k may be known or unknown. For example X_j may be the result of a medical test, which is not carried out if some other aspect of a patient's condition, as measured by X_k, renders the test dangerous.

DM4: The value x_{ij} is observed if and only if it is within k standard deviations of its mean. For such a standardized type of censoring the a priori probability of obtaining the deletion pattern does not depend on the parameters θ.

DM5: Suppose x_{ij} has probability $p_j(\mu_j)$ of deletion, which depends on the population mean of X_j.

These mechanisms can be combined or made more elaborate. We see how the deletion pattern may be determined by part of the data, or by missing data (e.g. a mixture of DM1 and DM3), or it may depend on the parameter θ (e.g. DM2, DM5).

Now let $\ell(P, D; \theta, \phi)$ be the log likelihood of the data, under the full model. We have

$$\ell(P,D;\theta,\phi) = \ell_1'(P|D;\theta,\phi) + \ell_2(D;\theta,\phi). \quad (2.3.1)$$

Also component (a) of the model defines the likelihood $L(P,M;\theta)$ of a complete set of data. The log-likelihood of the marginal distribution of P is thus, to within a constant,

$$\ell_1(P;\theta) = \ln\left[\int L(P,M;\theta)\,dM\right] . \qquad (2.3.2)$$

We define a random deletion mechanism to be one such that for all possible sets of data (P,D), the value of θ which maximizes $\ell_1(P;\theta)$ also maximizes $\ell(P,D;\theta,\phi)$ By the partition (2.3.1), we see that sufficient conditions for this are

S1: $$\ell_1'(P|D;\theta,\phi) = \ell_1(P;\theta) . \qquad (2.3.3)$$

S2: $$\ell_2(D;\theta,\phi) = \ell_2(D;\phi) \text{, and } \Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi . \qquad (2.3.4)$$

Condition S1 is self-explanatory, and S2 implies that the deletion pattern D is ancillary as regards the estimation of θ.

Turning to the examples, we see that DM1 is random, DM2 fails S1 and S2, DM3 fails S1 and S2, DM4 fails S1 and DM5 fails S2.

Less stringent conditions are sufficient if the deletion mechanism is such that a subset P_c of the variables X have probability 1 of observation. Then for any deletion pattern these variables have their marginal distribution from component (a) of the model. If $\ell_c(P_c;\theta)$ is the corresponding log-likelihood, then

$$\ell(P,D;\theta,\phi) = \ell_c(P_c;\theta) + \ell_3'(P|P_c,D;\theta,\phi) + \ell_4(D|P_c;\theta,\phi),$$

and $\ell_1(P;\theta) = \ell_c(P_c;\theta) + \ell_3(P|P_c;\theta)$.

Hence the deletion mechanism is random if

C1 $\ell_3'(P|P_c,D;\theta,\phi) = \ell_3(P|P_c;\theta)$. (2.3.5)

C2 $\ell_4(D|P_c;\theta,\phi) = \ell_4(D|P_c;\phi)$, and $\Omega_{\theta,\phi} = \Omega_\theta \times \Omega_\phi$. (2.3.6)

For if C1 and C2 are satisfied,

$$\begin{aligned}\ell(P,D;\theta,\phi) &= \ell_c(P_c;\theta) + \ell_3(P|P_c;\theta) + \ell_4(D|P_c;\phi)\\ &= \ell_1(P;\theta) + \ell_4(D|P_c;\phi),\end{aligned}$$

and so maximizing $\ell_1(P;\theta)$ with respect to θ is equivalent to maximizing $\ell(P,D;\theta,\phi)$ with respect to θ, and the deletion mechanism is random.

The conditions C1 and C2 parallel S1 and S2, but we can allow the deletion probabilities to depend on the values of variables which are always observed. Considering the examples, we see that DM2 fails C1 and C2, DM4 fails C1, DM5 fails C2, DM3 fails C1 and C2 except if X_k is always present, since then conditioning on that variable shows thatDM3 <u>is</u> random. The latter example is important, for it shows that it is possible to have a random deletion pattern, although the sample of observed values of a variable which is sometimes missing does not have the marginal distribution of the variable when it is always observed.

Before working out some examples, we indicate when we are justified in conditioning on the deletion pattern obtained, and in using conditional maximum likelihood to estimate the parameters. The conditions are C2 and

<u>C1</u>$'$ $\quad \ell_3'(P \mid P_c, D; \theta, \phi) = \ell_3'(P \mid P_c; \theta)$,

which is weaker than C1. If C1$'$ and C2 are satisfied D is statistically ancillary to the estimation of θ. Two examples are worth a brief mention.

(a) <u>DM2</u>: For censored data, C1$'$ and C2 are satisfied if the cut-off points are known. However if the cut-off points are unknown, we must use the deletion pattern to estimate the additional parameters.

(b) <u>DM5:</u> Here the log likelihood $\ell_1'(P|D;\theta)$ is the marginal likelihood $\ell_1(P;\theta)$ from component (a) of the model, but the deletion pattern contains additional information about the means.

Thus strictly speaking we are losing information by maximizing $\ell_1(P;\theta)$, although in practice this loss may be small. Example (2.3.3) is a simple illustration of this.

We now apply the M.I.P. to some further examples.

<u>Example (2.3.1).</u> To illustrate DM1 and DM3, mixed. Suppose $(X_1,X_2)^T \sim MN_2(\mu,\Sigma)$, and X_2 is not observed if X_1 is greater than a known constant c. After measurement, some proportion of the measurements on X_1 are lost. These losses do not depend on the values of the variables.

<u>Analysis</u>. According to M.I.P., from current estimates (μ_A, Σ_A) of the parameters, we find the fitted value

$$x_{i2}^* = E(x_{i2} \,|\, x_{i1}, x_{i1} > c) = E(x_{i2} \,|\, x_{i1})$$

for an observation with X_1 present and X_2 missing. Here the information in the deletion pattern, that $x_{i1} > c$, does not affect the conditional distribution of x_{i2} given x_{i1}, which is normal. Thus the fitted value x_{i2}^* and the adjustment for the variance is the same as for M.L.N. However for an observation with X_1 missing but X_2 present, the fitted value is

$$x_{i1}^{**} = E\left(x_{i1} \mid x_{i2}, x_{i1} < c\right) = \frac{\int_{-\infty}^{c} x_1 \, \phi\left(\frac{x_1 - x_{i1}^{*}}{v}\right) dx_1}{\int_{-\infty}^{c} \phi\left(\frac{x_1 - x_{i1}^{*}}{v}\right) dx_1},$$

where ϕ represents the standard normal density, and (x_{i1}^{*}, v) are the mean and variance of the conditional distribution of $(x_{i1} \mid x_{i2})$, which is normal. We see that the deletion pattern tells us that $x_{i1} < c$, so x_{i1} has a truncated normal distribution. The fitted value from M.L.N. is x_{i1}^{*}, which is different from x_{i1}^{**}. Hence M.L.N. is maximum likelihood only if X_1 is always observed. This is indicated by the previous theory.

Example 2.3.2. Censored Data.

Suppose $\mathbf{x}_i \overset{i.i.d.}{\sim} MN_p(\mu, \Sigma)$, but that the variable X_j $(1 \leq j \leq p)$ is only observed in a known range R_j. We can find m.l.e's of (μ, Σ) using the M.I.P. If x_{ij} is missing, we substitute

$$x_{ijA}^{*} = E\left(x_{ij} \mid P_i, x_{ij} \in \bar{R}_j; \mu_A, \Sigma_A\right) = \frac{\int_{\bar{R}_j} x f(x)\, dx}{\int_{\bar{R}_j} f(x)\, dx},$$

where $\bar{R}_j = (-\infty, \infty) \setminus R_j$, and $f(x)$ is the normal density for the conditional distribution of x_{ij} given the variables present in observation i. The corresponding adjustment to the estimate of variance is

$$Var\,(x_{ij}\,|\,P_i\,,x_{ij}\in\bar{R}_j\,;\mu_A,\Sigma_A) \;=\; \frac{\int_{\bar{R}_j} x^2 f(x)\,dx}{\int_{\bar{R}_j} f(x)\,dx} - x^{*2}_{ijA}.$$

Hence we can solve the fixed point equations, calculating these integrals at each iteration. We can also incorporate missing values in the other variables, which are considered randomly deleted. Suppose x_{ik} is such a missing value, and x_{ij} is also missing. Then we fit

$$\begin{aligned} x^*_{ikA} &= E\,(x_{ik}\,|\,P_i\,,\,x_{ij}\in\bar{R}_j\,;\mu_A,\Sigma_A) \\ &= \frac{\int_{\bar{R}_j} E\,(x_{ik}\,|\,P_i\,,x_j\,;\mu_A,\Sigma_A)\,f(x_j)\,dx_j}{\int_{\bar{R}_j} f(x_j)\,dx_j}\,, \end{aligned}$$

and the adjustments are

$$\begin{aligned} &Cov\,(x_{ij},x_{ik}\,|\,P_i\,,x_{ij}\in\bar{R}_j\,;\mu_A,\Sigma_A) \\ &\qquad= \frac{\int_{\bar{R}_j} x_j\;E\,(x_{ik}\,|\,P_i\,,x_j\,;\mu_A,\Sigma_A)\,f(x_j)\,dx_j}{\int_{\bar{R}_j} f(x_j)\,dx_j} - x^*_{ijA}\,x^*_{ikA}\,, \end{aligned}$$

$$Var\,(x_{ik}\,|\,P_i\,,x_{ij}\in\bar{R}_j\,;\mu_A,\Sigma_A) = \frac{\int_{\bar{R}_j}[E(x_{ik}\,|\,P_i,x_j\,;\mu_A,\Sigma_A)]^2 f(x_j)\,dx_j}{\int_{\bar{R}_j} f(x_j)\,dx_j} - x^{*2}_{ikA}.$$

More complicated models for deleting multivariate normal data can be handled in a similar way. The quadratic nature of the log likelihood function for complete data ensures that first and second moments of the truncated distribution are all that are required to solve the fixed point equations.

Example 2.3.3. (An example of DM5).

Suppose again $\mathbf{x}_i \overset{i.i.d.}{\sim} MN_p(\mu,\Sigma)$, but X_j has a probability p_j of deletion in any observation, where

$$p_j = \frac{e^{\phi\mu_j}}{1+e^{\phi\mu_j}} ,$$

where ϕ is an unknown parameter, independent of (μ,Σ).

Then if $\theta = (\mu,\Sigma)$,

$$\ell(P,D|\theta,\phi) = \ell_1(P|D;\theta) + \ell_2(D|\mu,\phi).$$

M.L.N. maximizes the log likelihood $\ell_1(P|D;\theta)$, but this is not the true m.l.e. of θ, since the deletion pattern itself contains information about μ.

Applying the M.I.P. the transformation $\phi(\theta_A)$ is the value θ which maximizes

$$E\{\ell(M,P,D|\theta,\phi)|P,D;\theta_A,\phi_A\}$$

instead of

$$E\{\ell(M,P|D;\theta,\phi)|P,D;\theta_A,\phi_A\}.$$

If n_j are the number of observations for which X_j is observed, the difference between these terms is the factor

$$\ell(D|\mu_j,\phi) = n_j \log p_j(\mu_j) - (N-n_j)\log[1-p_j(\mu_j)],$$

which modifies the estimates found by M.L.N.

3. Robustness of Estimators of the Means and Covariances.

3.1. Introduction.

In the previous Chapter we considered the maximum likelihood approach to the problem of estimating the parameters of a multivariate normal distribution, from an incomplete random sample. We now examine the methods of Buck and Iterated Buck, derived in Chapter 1 by adopting the fitting approach. As indicated there, virtually the same method as M.L.N. can be arrived at by considering "good" fitted values for the missing variables and adjusting the usual estimates of the means and covariances, formed from the completed data, to elimate bias.

In §3.2 the goodness of fit of the linear estimators of the missing values is considered, using the criterion of mean square error. In §3.3 the bias corrections are considered in some detail for Buck's (1960) method, from an unbiasedness viewpoint. In the following section the iterative version of the method is considered. The estimates are found to be consistent as N tends to infinity in such a way that the number of observations with an observed deletion pattern also tends to infinity, provided every pair of variables is observed together in at least one of the observed patterns.

The "robustness" in the title of this chapter is justified by the fact that no normality assumptions are made about the joint distribution of $X_1, X_2, \ldots, X_p$. For sections 3.2 - 3.4 we assume that

(i) $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are independent, with mean $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$.

(ii) The deletion pattern is random, in the sense that the means and covariances of the set P_i of variables observed in observation i are the same as those given by (i).

(iii) The distribution of $\mathbf{x}_i$ has finite fourth moments (i = 1,2,...,N).

The condition (ii) is the analogue of S1, equation (2.3.3), but here we are concerned only with the first two moments of the underlying population, rather than a specific distribution. It allows us to take expectations for repetitions of the sample with the same pattern of missing values.

Finally we consider a finite sample argument, where moments are taken with respect to a population consisting of the hypothetical complete observations. If "random deletion" is considered as meaning that every observation has an equal chance of coinciding with any of the N deletion patterns $(P_1, P_2, \ldots, P_N)$ in the observed sample, then good estimates of the sample means and sample covariances of the undeleted sample are obtained by Iterated Buck.

3.2. The fitted values.

We have an incomplete data matrix $\mathbf{X}$, which satisfies the conditions of §3.1. A certain value x_{ij} is missing, and we fit $\hat{x}_{ij}$, a function of the data. If we knew completely the distribution of $(x_{i1}, x_{i2}, \ldots x_{ip})$, then the best fitted value in the sense of minimizing the mean square error

$$E\left\{ (x_{ij} - \hat{x}_{ij})^2 \mid P_i \right\} \qquad (3.2.1)$$

is evidently $E(x_{ij} \mid P_i)$. In this expression we consider replications of the sample with the values of P_i fixed. For multivariate normal data this value is

$$x_{ij}^* = \mu_j + \sum_{\ell \in P_i} \beta_{j\ell \cdot P_i} (x_{i\ell} - \mu_\ell), \qquad (3.2.2)$$

where $\beta_{j\ell \cdot P_i}$ are partial regression coefficients for the regression of X_j on P_i, calculated from $\mathbf{\Sigma}$. This value cannot be fitted in practice since we do not know $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$. However in the M.L.N. method this value is estimated for assumed values $(\boldsymbol{\mu}_A, \mathbf{\Sigma}_A)$ of the unknown parameters. See Equation (2.2.9).

In general $E(x_{ij} \mid P_i)$ will not be a linear combination of the variables P_i, and for a particular distribution of $\mathbf{x}_i$, better fitted values can be found. However for any underlying population, x_{ij}^* is the best linear combination of the variables in P_i, in the sense of minimizing

$$E\,(x_{ij} - \hat{x}_{ij})^2 \;, \qquad (3.2.3)$$

where we average over the values of P_i. For any linear fitted value $\hat{x}_{ij}$,

$$\begin{aligned} E\,(x_{ij} - \hat{x}_{ij})^2 &= E\,(x_{ij} - x_{ij}^* + x_{ij}^* - \hat{x}_{ij})^2 \\ &= E\,(x_{ij} - x_{ij}^*)^2 + E\,(x_{ij}^* - \hat{x}_{ij})^2, \end{aligned}$$

since the residual $(x_{ij} - x_{ij}^*)$ is uncorrelated with the variables in P_i, and hence uncorrelated with $(x_{ij}^* - \hat{x}_{ij})$. Thus (3.2.3) is minimized at $\hat{x}_{ij} = x_{ij}^*$.

Hence with a mean square error criterion, M.L.N. fits estimates of the best fitted values when the data are multivariate normal, and estimates of the best linear fitted values when the data are sampled from a general underlying population. Improved non-linear fitted values for certain non-normal populations are considered briefly in Chapter 5.

Buck's method estimates the means and partial regression coefficients of (3.2.2) by their sample analogues, calculated from the set of complete observations. Hence for the observed or fitted value,

$$\hat{x}_{ij} = \tilde{x}_j + \sum_{\ell \in P_i} b_{j\ell \cdot P_i} (x_{i\ell} - \tilde{x}_\ell) , \qquad (3.2.4)$$

where $\tilde{x}_j$ is the sample mean of X_j from the complete observations, and

$$b_{j\ell \cdot P_i} = \begin{cases} \text{sample partial regression coefficient, found from the complete observations,} & \text{if } x_{ij} \in M_i ; \\ \delta_{j\ell} , & \text{if } x_{ij} \in P_i , \end{cases} \qquad (3.2.5)$$

where M_i are the missing variables in the ith observation, and $\delta_{j\ell}$ is the Kronecker Delta. Then form

$$\hat{\hat{x}}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ij} , \qquad (j=1,2\ldots,p) ,$$

$$a_{jk} = \sum_{i=1}^{N} (\hat{x}_{ij} - \hat{\hat{x}}_j)(\hat{x}_{ik} - \hat{\hat{x}}_k) , \qquad (j,k=1,2,\ldots,p),$$

and estimate μ_j by $\hat{\hat{x}}_j$, and σ_{jk} by

$$\hat{\sigma}'_{jk} = \frac{1}{N-1} \sum_{i=1}^{N} \left[(\hat{x}_{ij} - \hat{\hat{x}}_j)(\hat{x}_{ik} - \hat{\hat{x}}_k) + c_{ijk} \right] , \qquad (3.2.6)$$

where c_{ijk} is a correction term to eliminate bias introduced by the fitted values. We now consider what the correction terms should be to produce an unbiased estimate of Σ.

3.3 Bias Correction for Buck's Method.

In the rest of this chapter we shall be concerned with asymptotic bias in the estimates, as the number of observations tends to infinity in some way. We shall require the following result: if y_N is a statistic based on N observations, and $g(y)$ is a function of y, independent of N, then subject to mild regularity conditions on y_N,

$$\left.\begin{aligned} & E(y_N) = \theta + O(\tfrac{1}{N}) \quad , \text{ as } N \to \infty , \\ \text{and} \quad & Cov(y_N) = O(\tfrac{1}{N}) \quad , \text{ as } N \to \infty \\ \text{together imply} \quad & E(g(y_N)) = g(\theta) + O(\tfrac{1}{N}) , \text{as } N \to \infty \end{aligned}\right\} \quad (3.3.1)$$

This follows by considering the first three terms of the Taylor expansion of $g(y_N)$ about θ.

Now let us consider the bias in the estimate of Σ formed by Buck's Method. The estimate σ'_{jk} of equation (3.2.6) is unbiased if

$$E\left\{\frac{1}{N-1}\left(a_{jk}+\sum_{i=1}^{N}c_{ijk}\right)\right\} = \sigma_{jk},$$

where the expectation is for repetitions of the sample with the same deletion pattern. Hence c_{ijk} must satisfy

$$E\left(\sum_{i=1}^{N}c_{ijk}\right) = (N-1)\sigma_{jk} - E(a_{jk}) . \qquad (3.3.2)$$

We expand $E(a_{jk})$. Write

$$\hat{x}_{ij} = x^{*}_{ij} + \varepsilon_{ij},$$
$$\varepsilon_{ij} = \nu^{(i)}_{j} + \gamma_{ij},$$

where x^{*}_{ij} is the limiting value of $\hat{x}_{ij}$ as the number of complete observations, n_c, tends to infinity, and

$$\nu^{(i)}_{j} = \tilde{x}_j - \sum_{\ell\in P_i} b_{j\ell\cdot P_i}\tilde{x}_\ell \text{ , } (=0 \text{ if } x_{ij}\in P_i), \qquad (3.3.3)$$

$$\gamma_{ij} = \sum_{\ell\in P_i}(b_{j\ell\cdot P_i}-\beta_{j\ell\cdot P_i})x_{i\ell}, (=0 \text{ if } x_{ij}\in P_i) . \qquad (3.3.4)$$

In terms of these variables,

$$\begin{aligned} a_{jk} &= \frac{N-1}{N}\sum_{i=1}^{N}\hat{x}_{ij}\,\hat{x}_{ik} - \frac{1}{N}\sum_{\substack{i_1=1\\ i_1\neq i_2}}^{N}\sum_{i_2=1}^{N}\hat{x}_{i_1,j}\,\hat{x}_{i_2,k} \\ &= \frac{N-1}{N}\sum_{i=1}^{N}x^*_{ij}\,x^*_{ik} - \frac{1}{N}\sum_{\substack{i_1=1\\ i_1\neq i_2}}^{N}\sum_{i_2=1}^{N}x^*_{i_1,j}\,x^*_{i_2,k} \\ &\quad + \frac{N-1}{N}\sum_{i=1}^{N}\left(\zeta_{ij}\,\zeta_{ik} + \zeta_{ij}\,x^*_{ik} + \zeta_{ik}\,x^*_{ij}\right) \\ &\quad - \frac{1}{N}\sum_{\substack{i_1=1\\ i_1\neq i_2}}^{N}\sum_{i_2=1}^{N}\left(\zeta_{i_1,j}\,\zeta_{i_2,k} + \zeta_{i_1,j}\,x^*_{i_2,k} + \zeta_{i_2,k}\,x^*_{i_1,j}\right). \end{aligned}$$

We assume without loss of generality that the means of $X_1, X_2, \dots, X_p$ are zero. Then the "partial covariance" of $x_{i_1,j}$ and $x_{i_2,k}$,

$$E\left(x^*_{i_1,j}\,x^*_{i_2,k}\right) = \begin{cases} \sigma_{jk} - \sigma_{jk\cdot P_i}, & \text{if } i_1 = i_2 = i \\ 0, & \text{if } i_1 \neq i_2 \end{cases}, \qquad (3.3.5)$$

where $\sigma_{jk\cdot P_i}$ is the residual covariance of X_j and X_k after fitting linear regressions on the variables in P_i. This partial covariance is the conditional covariance $Cov(x_{i_1,j}, x_{i_2,k} | P_i)$ in the multivariate normal case.

Next consider (3.3.3) and (3.3.4). The terms $\tilde{x}_j$, $b_{j\ell\cdot P_i}$, and hence also $\nu_j^{(i)}$, are functions of the complete data, and hence independent of all functions of the incomplete observations with constant coefficients, and in particular the x^*_{ik}'s for an incomplete observations i. Using this, and the fact that the means of the variables are zero, we have

$$E(\zeta_{i_1,j}\, x^*_{i_2,k}) = \begin{cases} E(\nu_j^{(i_1)} x_{i_2,k}) & \text{, if } x_{i_1,j} \in M_{i_1}, \text{ observation } i_2 \text{ complete} \\ 0 & \text{, otherwise} \end{cases} \qquad (3.3.6)$$

$$E(\zeta_{i_1,j}\, \zeta_{i_2,k}) = \begin{cases} E(\nu_j^{(i_1)} \nu_k^{(i_2)}) & \text{, if } i_1 \neq i_2 \text{ ;} \qquad (3.3.7) \\ E(\nu_j^{(i)} \nu_k^{(i)}) + E(\gamma_{ij}\gamma_{ik}) & \text{, if } i_1 = i_2 = i \text{ .} \end{cases}$$

Now suppose the first n_c observations are complete. Taking the expected value of a_{jk} and using (3.3.5) - (3.3.7), we have

$$E(a_{jk}) = \frac{N-1}{N} \sum_{i=1}^{N} (\sigma_{jk} - \sigma_{jk\cdot P_i}) + T_1 + T_2 + T_3 \text{ ,} \qquad (3.3.8)$$

where

$$T_1 = \frac{N-1}{N} \sum_{i=n_c+1}^{N} \left[E(\nu_j^{(i)} \nu_k^{(i)}) + E(\gamma_{ij}\gamma_{ik}) \right] \text{ ,}$$

$$T_2 = -\frac{1}{N} \sum_{\substack{i_1=n_c+1 \\ i_1 \neq i_2}}^{N} \sum_{i_2=n_c+1}^{N} E(\nu_j^{(i_1)} \nu_k^{(i_2)}) \text{ ,}$$

$$T_3 = -\frac{1}{N} \sum_{i_1=1}^{N_c} \sum_{i_2=n_c+1}^{N} E(\nu_j^{(i_2)} x_{i_1,k} + \nu_k^{(i_2)} x_{i_1,j}) .$$

First consider T_1. By the independence of the observations,

$$E(\gamma_{ij}\gamma_{ik}) = \sum_{l \in P_i} \sum_{m \in P_i} \left[(b_{jl\cdot P_i} - \beta_{jl\cdot P_i})(b_{km\cdot P_i} - \beta_{km\cdot P_i}) \sigma_{lm} \right] . \qquad (3.3.9)$$

Let $A_{(i)}$ denote the $(r_i \times r_i)$ S.S.C.P. matrix of the complete observations, for the r_i variables in P_i, $(1 \leqslant r_i \leqslant p)$. Let $\Sigma_{(i)}$ denote the submatrix of Σ corresponding to these variables, and $\hat{\Sigma}_{(i)} = \frac{1}{n_c - 1} A_{(i)}$. Then

$$E(\hat{\Sigma}_{(i)}) = \Sigma_{(i)} \quad , \; Cov(\hat{\Sigma}_{(i)}) = O(\tfrac{1}{n_c}) \; , \text{ as } n_c \rightarrow \infty \, ,$$

since the distribution of $(X_1, X_2, \ldots, X_p)$ has finite fourth moments. Hence by (3.3.1), if $g(\Sigma_{(i)})$ is a known function of $\Sigma_{(i)}$, independent of n_c,

$$E(g(\hat{\Sigma}_{(i)})) = g(\Sigma_{(i)}) + O(\tfrac{1}{n_c}) . \qquad (3.3.10)$$

Setting $g(\hat{\Sigma}_{(i)}) = \gamma_{ij}\gamma_{ik}$, we see that $E(\gamma_{ij}\gamma_{ik})$ is $O(\frac{1}{n_c})$ as $n_c \rightarrow \infty$. We now work out the $O(\frac{1}{n_c})$ term under the multivariate normal assumption. Then conditionally on P_i,

$$E(b_{jl \cdot P_i} - \beta_{jl \cdot P_i}) = 0 \quad ; \; Cov(b_{jl \cdot P_i}, b_{km \cdot P_i} | P_i) = a^{lm}_{(i)} \, \sigma_{jk \cdot P_i} \, ,$$

where $a^{lm}_{(i)}$ is the (l,m)th element of $A^{-1}_{(i)}$. Also

$$E(\hat{\Sigma}^{-1}_{(i)}) = \Sigma^{-1}_{(i)} + O(\tfrac{1}{n_c}) ,$$

by another application of (3.3.10). Hence

$$E(\gamma_{ij}\gamma_{ik}) = \sum_{\ell\in P_i}\sum_{m\in P_i}\left(\frac{1}{n_c-1}\right)\sigma_{(i)}^{\ell m}\,\sigma_{jk\cdot P_i}\,\sigma_{\ell m} + O\left(\frac{1}{n_c^2}\right),$$

i.e. $$E(\gamma_{ij}\gamma_{ik}) = \frac{r_i}{n_c-1}\,\sigma_{jk\cdot P_i} + O\left(\frac{1}{n_c^2}\right). \tag{3.3.11}$$

Now $$E(\nu_j^{(i)}\nu_k^{(i)}) = E\left[\left(\tilde{x}_j - \sum_{\ell\in P_i} b_{j\ell\cdot P_i}\tilde{x}_\ell\right)\left(\tilde{x}_k - \sum_{m\in P_i} b_{km\cdot P_i}\tilde{x}_m\right)\right],$$

and by a similar argument to that given above this term is $O(\frac{1}{n_c})$. Again we find the $O(\frac{1}{n_c})$ term when the variables are multivariate normal. Then the means $\tilde{x}_j$ are uncorrelated with the partial regression coefficients, so

$$\begin{aligned} E(\nu_j^{(i)}\nu_k^{(i)}) &= E\left[\left(\tilde{x}_j - \sum_{\ell\in P_i}\beta_{j\ell\cdot P_i}\tilde{x}_\ell\right)\left(\tilde{x}_k - \sum_{m\in P_i}\beta_{km\cdot P_i}\tilde{x}_m\right)\right] \\ &\quad + E\left[\sum_{\ell\in P_i}(b_{j\ell\cdot P_i}-\beta_{j\ell\cdot P_i})\tilde{x}_\ell \sum_{m\in P_i}(b_{km\cdot P_i}-\beta_{km\cdot P_i})\tilde{x}_m\right] \\ &= \frac{1}{n_c}\sigma_{jk\cdot P_i} + \frac{1}{n_c}E(\gamma_{ij}\gamma_{ik}) = \frac{1}{n_c}\sigma_{jk\cdot P_i} + O\left(\frac{1}{n_c^2}\right). \end{aligned}$$

Adding and summing over i, we have

$$T_1 = \sum_{i=n_c+1}^{N}\left[\left(\frac{r_i+1}{n_c}\right)\sigma_{jk\cdot P_i} + O\left(\frac{1}{n_c^2}\right)\right] \tag{3.3.12}$$

for the normal case, and some other $O(\frac{1}{n_c})$ term in other cases.

To calculate T_2, we need $E(\nu_j^{(i_1)}\nu_k^{(i_2)})$ for $i_1 \neq i_2$. As for the case $i_1 = i_2$, this is $O(\frac{1}{n_c})$, and for the normal case

$$E(\nu_j^{(i_1)}\nu_k^{(i_2)}) = E\Big[\frac{1}{n_c}\Big\{(X_j - \sum_{\ell \in P_{i_1}} \beta_{j\ell\cdot P_{i_1}} X_\ell)(X_k - \sum_{m \in P_{i_2}} \beta_{km\cdot P_{i_2}} X_m)\Big\}\Big]$$
$$+ O(\tfrac{1}{n_c^2})$$
$$= \frac{1}{n_c}\sigma_{jk\cdot P_{i_2}} \quad \text{if} \quad P_{i_1} \subseteq P_{i_2},$$

but the right hand side has no simple form unless $P_{i_1} \subseteq P_{i_2}$ or $P_{i_2} \subseteq P_{i_1}$. Therefore we consider the $O(\frac{1}{n_c})$ term only when the data is monotone, i.e. we can arrange the observations and variables so that

$$P_{i_1} \subseteq P_{i_2} \qquad \text{if} \qquad i_1 \geqslant i_2 . \tag{3.3.13}$$

Then

$$T_2 = -\frac{2}{Nn_c}\sum_{i_2=n_c+1}^{N}\sum_{i_1>i_2}^{N} \sigma_{jk\cdot P_{i_2}} + O(\tfrac{1}{n_c^2}),$$

i.e. $$T_2 = -\frac{2}{Nn_c}\sum_{i=n_c+1}^{N}(N-i)\,\sigma_{jk\cdot P_i} + O(\tfrac{1}{n_c^2}). \tag{3.3.14}$$

Finally $T_3 = -\frac{n_c}{N}\sum_{i=n_c+1}^{N}\big[E(\nu_j^{(i)}\tilde{x}_k) + E(\nu_k^{(i)}\tilde{x}_j)$; this term is found in the same way as $E(\nu_j^{(i)}\nu_k^{(i)})$. Again it is $O(\frac{1}{n_c})$, and in the normal case,

$$T_3 = -\frac{1}{N}\sum_{i=n_c+1}^{N}\left(\sigma_{jk\cdot p_i}+\sigma_{jk\cdot p_i}\right) = -\frac{2}{N}\sum_{i=n_c+1}^{N}\sigma_{jk\cdot p_i} . \tag{3.3.15}$$

Substituting (3.3.12), (3.3.14) and (3.3.15) into (3.3.8),

$$E(a_{jk}) = \frac{N-1}{N}\sum_{i=1}^{N}\left(\sigma_{jk}-\sigma_{jk\cdot p_i}\right) + O\left(\frac{1}{n_c}\right),$$

and if the data are from $MN_p(\mu,\Sigma)$ and are monotone,

$$E(a_{jk}) = \frac{N-1}{N}\sum_{i=1}^{N}\left(\sigma_{jk}-\sigma_{jk\cdot p_i}\right) + \sum_{i=1}^{N}\sigma_{jk\cdot p_i}\left[\frac{r_i+1}{n_c} - \frac{2(N-i)}{Nn_c} - \frac{2}{N}\right] + O\left(\frac{1}{n_c^2}\right) .$$

Substituting this in (3.3.2), for unbiasedness we require

$$E(c_{ijk}) = \sigma_{jk\cdot p_i} + O\left(\frac{1}{n_c}\right) , \tag{3.3.16}$$

and for monotone data from $MN_p(\mu,\Sigma)$,

$$E(c_{ijk}) = \sigma_{jk\cdot p_i}\left[\frac{N+1}{N} - \frac{r_i+1}{n_c} + \frac{2(N-i)}{Nn_c}\right] + O\left(\frac{1}{n_c^2}\right).$$

If we replace (N-i) by its average value for $i=n_c+1,n_c+2,\ldots,N$, we obtain

$$E(c_{ijk}) \doteq \sigma_{jk\cdot P_i}\left[\frac{N+1}{N} - \frac{r_i+1}{n_c} + \frac{N-n_c-1}{Nn_c}\right] + O\left(\frac{1}{n_c^2}\right),$$

i.e. $E(c_{ijk}) \doteq \left(\frac{n_c-r_i-1}{n_c-1}\right)\sigma_{jk\cdot P_i} + O\left(\frac{1}{n_c^2}\right)$. (3.3.17)

Now if $\tilde{\Sigma}$ is the estimate of Σ from the complete observations, and $v_{jk\cdot P_i}$ is the (j,k)th element of

$$V_{(i)} = PIV(P_i)\,\tilde{\Sigma} \quad ,$$

then $E(v_{jk\cdot P_i}) = \frac{n_c-r_i-1}{n_c-1}\,\sigma_{jk\cdot P_i}$; (3.3.18)

the numerator n_c-r_i-1 results from losing r_i degrees of freedom by the pivoting process.

Comparing (3.3.17) and (3.3.18), we see that $v_{jk\cdot P_i}$ is the correct adjustment, ignoring terms of $O\left(\frac{1}{n_c^2}\right)$, and subject to the approximation given above. This is exact when all the incomplete observations have the same deletion pattern. (The simpler case of one incomplete observation was considered by Beale and Little (1973), and achieved the same result by a somewhat different methdd). For non-normal data, or when the pattern of missing values is not monotone, the correction $v_{jk\cdot P_i}$ is correct to $O\left(\frac{1}{n_c}\right)$. For these situations no simple formula can improve on $v_{jk\cdot P_i}$ from this unbiasedness viewpoint, and the final estimates of (μ,Σ) are still correct to $O\left(\frac{1}{n_c}\right)$. In the next section we consider the adjustments for Iterated Buck. The iterative nature of this method appears to preclude a detailed analysis based on unbiasedness. However a consideration of the consistency of the estimates leads to an adjustment similar to $V_{jk \cdot P_i}$.

3.4. The Bias Correction for Iterated Buck.

We write down the set of iterative equations explicitly, and consider the limiting equations as N tends to infinity. First we need to define limiting properties for the deletion pattern. Let $\mathcal{P}$ be the set of distinct patterns among $(P_1, P_2, \ldots, P_N)$, and let P be a typical pattern. (The usage of the letter P in this section differs from elsewhere in the thesis, where it means the set of data). Let n_P be the number of observations in S_P, the set of observations with pattern P. Then

$N \overset{*}{\rightarrow} \infty$ denotes $N \rightarrow \infty$ such that $\frac{n_P}{N} = \lambda_P$, a constant $> 0, \forall P \in \mathcal{P}$.

In this section we do not assume $n_c > 0$, i.e. that complete observations exist. If there are no complete observations the iterative process can be started by some other estimate of (μ, Σ), for example

$$\tilde{\mu}_j = \frac{1}{m_j} \sum_{i, s.t. x_{ij} \in P_i} x_{ij} \quad , \qquad (3.4.1)$$

$$\tilde{\sigma}_{jk} = \frac{1}{m_{jk}-1} \sum_{\substack{i, s.t. \\ x_{ij} \in P_i, x_{ik} \in P_i}} (x_{ij} - \tilde{\mu}_j)(x_{ik} - \tilde{\mu}_k) \quad , \qquad (3.4.2)$$

where m_j and m_{jk} are the number of elements in the corresponding summations. The resulting estimate of Σ may require a correction to make it positive semi-definite. This is done by pivoting on the matrix, and setting "negative variance" terms, and the corresponding covariance terms, equal to zero.

The equations of Iterated Buck are

$$\begin{aligned}
\hat{x}_{ij} &= \hat{\mu}_j + \sum_{\ell \in P_i} b_{j\ell \cdot P_i} (x_{i\ell} - \hat{\mu}_\ell) \quad , \\
\hat{\mu}_j &= \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ij} \quad , \\
\hat{\sigma}_{jk} &= \frac{1}{N-1} \sum_{i=1}^{N} \left[(\hat{x}_{ij} - \hat{\mu}_j)(\hat{x}_{ik} - \hat{\mu}_k) + c_{ijk} \right] ,
\end{aligned}$$

where $b_{j\ell \cdot P_i}$ are defined by (3.2.5), except that they are estimated from $\hat{\Sigma}$, rather than the complete observations. Hence

$$\hat{\mu}_j = \frac{1}{N} \sum_{P \in \mathcal{P}} \sum_{i \in S_P} \left[\hat{\mu}_j + \sum_{\ell \in P} b_{j\ell \cdot P} (x_{i\ell} - \mu_\ell) \right] , \qquad (3.4.3)$$

$$\hat{\sigma}_{jk} = \frac{1}{N-1} \sum_{P \in \mathcal{P}} \sum_{i \in S_p} \left[\sum_{\ell \in P} \sum_{m \in P} b_{j\ell \cdot p} \, b_{km \cdot p} (x_{i\ell} - \hat{\mu}_\ell)(x_{im} - \hat{\mu}_m) + c_{ijk} \right] . \qquad (3.4.4)$$

Now for $P \in \mathcal{P}$, the statistics

$$\hat{\mu}_{p\ell} = \frac{1}{n_p} \sum_{i\in S_p} x_{i\ell} \quad , \tag{3.4.5}$$

$$\hat{\sigma}_{p\ell m} = \frac{1}{n_p} \sum_{i\in S_p} (x_{i\ell} - \hat{\mu}_{p\ell})(x_{im} - \hat{\mu}_{pm}) \tag{3.4.6}$$

form a condensed set of data which is fixed over iterations. In terms of these quantities, (3.4.3) and (3.4.4) become

$$\sum_{p\in\mathcal{P}} \frac{n_p}{N} \sum_{\ell\in p} b_{j\ell\cdot p} (\hat{\mu}_{p\ell} - \hat{\mu}_\ell) = 0 \quad , \tag{3.4.7}$$

$$\hat{\sigma}_{jk} = \sum_{p\in\mathcal{P}} \frac{n_p}{N-1} \Big[\sum_{\ell\in p} \sum_{m\in p} b_{j\ell\cdot p}\, b_{km\cdot p} \{ \hat{\sigma}_{p\ell m} + (\hat{\mu}_{p\ell} - \hat{\mu}_\ell)(\hat{\mu}_{pm} - \hat{\mu}_m) \} + c_{pjk} \Big] , \tag{3.4.8}$$

where $c_{pjk} = \frac{1}{n_p} \sum_{i\in S_p} c_{ijk}$.

Now let $N \overset{*}{\rightarrow} \infty$, and consider the limiting equations. As $N \overset{*}{\rightarrow} \infty$, $\hat{\mu}_{p\ell} = \mu_\ell + O(\frac{1}{\sqrt{N}})$, $\hat{\sigma}_{p\ell m} = \sigma_{\ell m} + O(\frac{1}{\sqrt{N}})$, $c_{pjk} \rightarrow c^*_{pjk}$ (say), and (3.4.7), (3.4.8) tend to

$$\sum_{p\in\mathcal{P}} \lambda_p \sum_{\ell\in p} b_{j\ell\cdot p} (\mu_\ell - \hat{\mu}_\ell) = O(\tfrac{1}{\sqrt{N}}) \quad , \tag{3.4.9}$$

$$\hat{\sigma}_{jk} = \sum_{p\in\mathcal{P}} \lambda_p \Big[\sum_{\ell\in p} \sum_{m\in p} b_{j\ell\cdot p}\, b_{km\cdot p} \{ \sigma_{\ell m} + (\mu_\ell - \hat{\mu}_\ell)(\mu_m - \hat{\mu}_m) \} + c^*_{pjk} \Big] + O(\tfrac{1}{\sqrt{N}}). \tag{3.4.10}$$

The limiting equations are (3.4.9) and (3.4.10) with the $O(\frac{1}{\sqrt{N}})$ terms ignored. For consistent estimates,

$$(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

must be a solution of the limiting equations. Substituting this solution, with the consequent population partial regression coefficients, we have from (3.4.10)

$$\sigma_{jk} = \sum_{P\in\mathcal{P}} \lambda_P \Big[\sum_{\ell\in P} \sum_{m\in P} \beta_{j\ell\cdot P}\beta_{km\cdot P}\,\sigma_{\ell m} + C^*_{Pjk} \Big]$$

$$= \sum_{P\in\mathcal{P}} \lambda_P \Big[\sigma_{jk} - \sigma_{jk\cdot P} + C^*_{Pjk} \Big] \quad .$$

Hence $C^*_{Pjk} = \sigma_{jk\cdot P}$.

Therefore for consistent estimates, we choose c_{ijk} so that

$$\lim_{N\to\infty} c_{ijk} = \sigma_{jk\cdot P_i} \quad . \qquad (3.4.11)$$

Replacing $\hat{\sigma}_{jk\cdot P_i}$ for c_{ijk} in the equations of Iterated Buck, we obtain the same equations as M.L.N., (2.2.5) - (2.2.8), except that the factor $\frac{1}{N}$ in (2.2.7) is replaced by $\frac{1}{N-1}$, the standard correction when the data are complete. Thus Iterated Buck is "corrected maximum likelihood" when the data are multivariate normal, and although in the Chapters that follow the method M.L.N. will sometimes be quoted for theoretical reasons, in practice the corrected method can always be used. Of course the practical difference is small.

We now ask the question: under what conditions do the limiting equations have a unique solution? If X_j and X_k are never observed together, then the set of solutions

$$\left.\begin{array}{lll} \hat{\mu}_l = \mu_l & , (l = 1,2,\ldots,p), \\ \hat{\sigma}_{lm} = \sigma_{lm} & , (l,m) \neq (j,k), \\ \hat{\sigma}_{jk\cdot\bar{p}} = \rho\sqrt{\hat{\sigma}_{jj\cdot\bar{p}}\,\sigma_{kk\cdot\bar{p}}} & , (-1 < \rho < 1). \end{array}\right\} , \qquad (3.4.12)$$

where $\bar{P}$ is the set of variables $(X_1, X_2, \ldots, X_p)$ with X_j and X_k removed, satisfy the limiting equations, since σ_{jk} does not appear in the set of equations (3.4.10), and thus can be chosen arbitrarily in the sense of (3.4.12). Consequently the estimate of σ_{jk} from Iterated Buck is not consistent, as one might expect since we have no information about the partial correlation ρ of (3.4.12). In practice, convergence is speeded up by setting $\hat{\sigma}_{jk\cdot\bar{p}} = 0$ at the start of each iteration. However, if the deletion patterns are such that every pair of variables are observed, the limiting equations have a unique solution, and Iterated Buck is consistent.

Asymptotic Unbiasedness of the Estimates.

Evalhation of $O(\frac{1}{N})$ bias in the estimates of (μ, Σ) is not feasible, but we can show that if the initial estimates of (μ, Σ) are unbiased to $O(\frac{1}{N})$ as $N \overset{*}{\to} \infty$, then the final estimates $(\hat{\mu}, \hat{\Sigma})$ are also unbiased to $O(\frac{1}{N})$ as $N \overset{*}{\to} \infty$. We sketch a proof.

Suppose current estimates (μ_A, Σ_A) are unbiased to $O(\frac{1}{N})$, and (μ_M, Σ_M) are new estimates, found by one iteration of the method. It is sufficient to show (μ_M, Σ_M) are unbiased to $O(\frac{1}{N})$. By an application of (3.3.1) with $y_N = (\mu_A, \Sigma_A)$, $\theta = (\mu, \Sigma)$,

$$E(\hat{\sigma}_{jkA\cdot P_i}) = \sigma_{jk\cdot P_i} + O(\tfrac{1}{N}) \quad \text{as} \quad N \overset{*}{\rightarrow} \infty \, , \qquad (3.4.13)$$

$$E(b_{jkA\cdot P_i}) = \beta_{jk\cdot P_i} + O(\tfrac{1}{N}) \quad \text{as} \quad N \overset{*}{\rightarrow} \infty \, , \qquad (3.4.14)$$

with the notation of §2.2. Using (3.4.13), it is sufficient to show

$$E(\hat{x}_{ijA}) = E(x_{ij}^*) + O(\tfrac{1}{N}) \, , \qquad (3.4.15)$$

$$E(\hat{x}_{i_1,jA}\,\hat{x}_{i_2,kA}) = E(x_{i_1,j}^*\, x_{i_2,k}^*) + O(\tfrac{1}{N}), \quad 1 \leq i_1, i_2 \leq N \, , \qquad (3.4.16)$$

where $\quad \hat{x}_{ijA} = \hat{\mu}_{jA} + \sum_{\ell \in P_i} b_{j\ell A\cdot P_i}(x_{i\ell} - \hat{\mu}_{\ell A})$,

and $\quad x_{ij}^* = \mu_j + \sum_{\ell \in P_i} \beta_{j\ell\cdot P_i}(x_{i\ell} - \mu_\ell)$.

Equation (3.4.15) and (3.4.16) follow from Lemma (3.3), provided also

$$E(b_{j\ell A\cdot P_i}\, x_{i\ell}) = E(\beta_{j\ell\cdot P_i}\, x_{i\ell}) + O(\tfrac{1}{N}) \, , \qquad (3.4.17)$$

and $$E(b_{j\ell A\cdot P_{i_1}}\, x_{i_1,\ell}\, b_{kmA\cdot P_{i_2}}\, x_{i_2,m}) = E(\beta_{j\ell\cdot P_{i_1}}\, x_{i_1,\ell}\, \beta_{km\cdot P_{i_2}}\, x_{i_2,m}) + O(\tfrac{1}{N}). \quad (3.418)$$

These would follow by an application of (3.3.1) if $(b_{j\ell A\cdot P_{i_1}}, b_{kmA\cdot P_{i_2}})$ was independent of $(x_{i_1\ell}\ x_{i_2,m})$ and $b_{j\ell A\cdot P_i}$ was independent of $x_{i\ell}$. This difficulty is overcome by considering $b'_{j\ell A\cdot P_i}$, equal to $b_{j\ell A\cdot P_i}$ but with a new independently sampled observation replacing the ith observation, with the same deletion pattern. Clearly $b'_{j\ell A\cdot P_i}$ is independent of $x_{i\ell}$, and

$$b'_{j\ell A\cdot P_i} = b_{j\ell A\cdot P_i} + O(\tfrac{1}{N}) \quad \text{as} \quad N \overset{*}{\rightarrow} \infty .$$

This is sufficient to prove (3.4.17). Similar tactics prove (3.4.18), and hence the result.

3.5. Unbiasedness under permutation of the observations.

So far we have considered the data as part of a random sample from a joint distribution with means and covariances (μ, Σ), and finite fourth moments. We now consider a finite sample approach to the problem. We make no distributional assumptions about the variables, and the problem is to estimate the sample means $\bar{x}_j$, and the sample S.S.C.P. matrix S of the undeleted sample, given the incomplete data.

For a given pattern of deletions, Iterated Buck produces estimates of $\bar{x}_j$ and S. Now consider a hypothetical permutation of the underlying undeleted observations. Every permutation π produces a different set of data, and different estimates $\bar{x}_{j\pi}$ and $S_{jk\pi}$ of $\bar{x}_j$ and S_{jk}. Now suppose every permutation is equally probable, and consider the average values of the estimates over all the possible permutations. If E_N denotes this averaging process, and if initial estimates $\bar{x}_{jo}, s_{jko}$ are such that

$$E_N(\bar{x}_{jo}) = \bar{x}_j + O(\tfrac{1}{N}) \quad , \tag{3.5.1}$$

$$E_N(s_{jko}) = s_{jk} + O(1) \quad , \tag{3.5.2}$$

then the final estimates $\hat{\bar{x}}_j, \hat{s}_{jk}$ from applying Iterated Buck satisfy

$$E_N(\hat{\bar{x}}_j) = \bar{x}_j + O(\tfrac{1}{N}) \quad , \tag{3.5.3}$$

$$E_N(\hat{s}_{jk}) = s_{jk} + O(1) \quad . \tag{3.5.4}$$

To prove this, we require the finite sample analogue of (3.3.1), together with the following theorem. Consider a hypothetical method where fitted values and adjustments for one iteration of Iterated Buck are found from the complete undeleted sample, i.e. are functions of $\{\bar{x}_j, s_{jk}\}$. These fitted values are finite sample analogues of the x_{ij}^{*}'s for an infinite population. Let $\bar{x}_{j\pi}, s_{jk\pi}$ be the estimates of $\bar{x}_j, s_{jk}$ from this method, for permutation π of the observations.

<u>Theorem 3.5.1.</u> The estimates $\bar{x}_{j\pi}, s_{jk\pi}$ satisfy

$$E_N(\bar{x}_{j\pi}) = \bar{x}_j \quad , \tag{3.5.5}$$

$$E_N(s_{jk\pi}) = s_{jk} + \gamma_{jk} \quad , \tag{3.5.6}$$

where γ_{jk} is $O(\frac{1}{N})$ as $N \overset{*}{\to} \infty$.

Proof. We treat the deletion patterns $(P_1, P_2, \ldots, P_N)$ as distinct, and let $\hat{x}_{ij(P_r)}$ be the observed or fitted value of x_{ij} when pattern P_r coincides with observation i. For the hypothetical fitting procedure we are considering this value is invariant under permutation of the other (N-1) observations. Hence

$$E_N(\bar{x}_{j\pi}) = \frac{1}{N!}\left[\frac{1}{N}\sum_{i=1}^{N} \hat{x}_{ij\pi}\right] = \frac{1}{N^2}\sum_{r=1}^{N}\sum_{i=1}^{N} \hat{x}_{ij(P_r)} \quad , \tag{3.5.7}$$

with the obvious notation for $\hat{x}_{ij\pi}$. Similarly

$$E_N\left(\sum_{i=1}^{N} \hat{x}_{ij\pi}\hat{x}_{ik\pi}\right) = \frac{1}{N}\sum_{r=1}^{N}\sum_{i=1}^{N} \hat{x}_{ij(P_r)}\,\hat{x}_{ik(P_r)} \quad . \tag{3.5.8}$$

Now fix P_r , and suppose $X_j \in M_r, X_k \in M_r$. Then $\hat{x}_{ij(P_r)}$, $\hat{x}_{ik(P_r)}$ are the fitted values from the regression of X_j and X_k on the variables in P_r, calculated from the undeleted sample. Thus by the geometry of least squares,

$$\frac{1}{N}\sum_{i=1}^{N} \hat{x}_{ij(P_r)} = \frac{1}{N}\sum_{i=1}^{N} x_{ij} = \bar{x}_j \quad , \tag{3.5.9}$$

since the residuals sum to zero, and denoting the left hand side of (3.5.9) by $\hat{\bar{x}}_{j(P_r)}$,

$$\sum_{i=1}^{N} \left(\hat{x}_{ij(P_r)} - \hat{\bar{x}}_{j(P_r)}\right)\left(\hat{x}_{ik(P_r)} - \hat{\bar{x}}_{k(P_r)}\right) = s_{jk} - s_{jk\cdot P_r} \quad , \tag{3.5.10}$$

where $S_{jk\cdot P_r}$ is the jk'th element of

$$PIV(P_r)\,S\,,$$

and is the residual sum of cross products between X_j and X_k. Equation (3.5.10) follows by noting

$$S_{jk\cdot P_r} = \sum_{i=1}^{N} \left(x_{ij} - \hat{x}_{ij(P_r)}\right)\left(x_{ik} - \hat{x}_{ik(P_r)}\right)$$

$$= \sum_{i=1}^{N} \left(x_{ij}\,x_{ik} - \hat{x}_{ij(P_r)}\,\hat{x}_{ik(P_r)}\right) .$$

If $X_j \in P_r$, $\hat{x}_{ij(P_r)} = x_{ij}$ (i=1,2,...,N), and (3.5.9) and (3.5.10) remain valid with $S_{jk\cdot P_r}$ defined as zero. Using these expressions, (3.5.7) and (3.5.8) become

$$E_N\left(\bar{x}_{j\pi}\right) = \bar{x}_j \quad ,$$

as required, and

$$E_N\left[\sum_{i=1}^{N} \hat{x}_{ij\pi}\,\hat{x}_{ik\pi} - N\,\bar{x}_{j\pi}\,\bar{x}_{k\pi}\right] = \frac{1}{N}\sum_{i=1}^{N}\left(S_{jk} - S_{jk\cdot P_i}\right) + \gamma_{jk}\,, \tag{3.5.11}$$

where
$$\gamma_{jk} = \sum_{r=1}^{N} \hat{\bar{x}}_{j(P_r)}\,\hat{\bar{x}}_{k(P_r)} - N\,E_N\left(\bar{x}_{j\pi}\,\bar{x}_{k\pi}\right) . \tag{3.5.12}$$

Equation (3.5.11) is equivalent to (3.5.8), since the estimate of S_{jk} is

$$S_{jk\pi} = \sum_{i=1}^{N}\left[\left(\hat{x}_{ij\pi} - \bar{x}_{j\pi}\right)\left(\hat{x}_{ik\pi} - \bar{x}_{k\pi}\right) + \frac{1}{N}\,S_{jk\cdot P_i}\right] .$$

Thus it remains to show that γ_{jk} is $O(1)$ as $N \overset{*}{\rightarrow} \infty$. But

$$E_N\left(\bar{x}_{j\pi}\right) = \bar{x}_j = \hat{\bar{x}}_{j(P_r)}\,,$$

So
$$\gamma_{jk} = -N\,Cov_N\left[\bar{x}_{j\pi}\,,\,\bar{x}_{k\pi}\right] = O(1),$$

by analogy with infinite populations, or by explicit calculation, assuming $\bar{x}_j = \bar{x}_k = 0$; Cov_N here denotes covariance for the finite sample.

The argument from (3.5.1) and (3.5.2) to (3.5.3) and (3.5.4) now follows the corresponding argument of the previous section. The details are omitted. Note that initial estimates obtained by either (i) forming estimates from the complete observations, or (ii) forming $(\tilde{\mu}_j, \tilde{\sigma}_{jk})$ as in (3.4.1) and (3.4.2), satisfy equations (3.5.1) and (3.5.2).

Two properties of this finite sample argument illustrate the appeal of the approach. Firstly, no independence assumptions are made about the observations. Secondly, Theorem (3.5.1) illuminates the geometrical aspects of the method of Iterated Buck, by appealing to the geometry of least squares.

4. ASYMPTOTIC COVARIANCE MATRIX OF THE MAXIMUM LIKELIHOOD ESTIMATES.

4.1. Introduction.

In Chapter 2 we found m.l.e's of the means and covariances, for an incomplete sample from the $MN_p(\mu,\Sigma)$ distribution. We now consider the precision of these estimates, and hence derive confidence intervals and classical tests of significance for the means.

The standard method of obtaining an asymptotic covariance matrix for the m.l.e. of θ is to find the expected information matrix $J_p(\theta,\theta)$, by differentiating the log-likelihood $\ell(P;\theta)$ of the incomplete data, and then inverting this matrix. However the M.I.P. provides a simpler way of finding $J_p(\theta,\theta)$, which we describe in the next section. We apply the method to our problem in §4.3, and discuss the resulting matrix, and its potential use in choosing an experimental design. In §4.4 $J_p(\theta,\theta)$ is inverted for the simple case of two variables, and we propose approximate t-tests for hypotheses of the form

$$H_0 : \mathbf{c}^T\mu = \mathbf{c}^T\mu_0 ,$$

for a constant vector $\mathbf{c}$. We also touch on the extension to more general linear hypotheses where $\mathbf{c}$ is replaced by an (r x p) matrix C.

4.2. The Expected Information Matrix for Incomplete Data .

This section is an extension of §2.2, and is given more concisely by Orchard and Woodbury (1971). We adopt the notation of §2.2. Recall equation (2.2.2):

$$E\{\ell(P,M;\theta)\,|\,P;\theta_A\} = \ell_1(P;\theta) + E\{\ell_2(M|P;\theta)|P;\theta_A\}.$$

We differentiate both sides with respect to θ_j , and assume the regularity conditions which allow us to commute the derivative with the expectation sign. Then

$$E\left\{\frac{\partial\ell(P,M;\theta)}{\partial\theta_j}\,\Big|\,P;\theta_A\right\} = \frac{\partial\ell_1(P;\theta)}{\partial\theta_j} + E\left\{\frac{\partial\ell_2(M|P;\theta)}{\partial\theta_j}\,\Big|\,P;\theta_A\right\}.$$

The expression $E\{\ell_2(M|P;\theta)|P;\theta_A\}$ is a maximum at $\theta=\theta_A$, so setting $\theta=\theta_A$,

$$E\left\{\frac{\partial\ell(P,M;\theta)}{\partial\theta_j}\,\Big|\,P;\theta\right\} = \frac{\partial\ell_1(P;\theta)}{\partial\theta_j} \quad . \tag{4.2.1}$$

Also writing ℓ for $\ell(P,M;\theta)$,

$$\begin{aligned} \mathrm{Cov}\left\{\left(\frac{\partial\ell}{\partial\theta_j},\frac{\partial\ell}{\partial\theta_k}\right);\theta\right\} &= E\left\{\mathrm{Cov}\left[\left(\frac{\partial\ell}{\partial\theta_j},\frac{\partial\ell}{\partial\theta_k}\right)\Big|P;\theta\right];\theta\right\} \\ &\quad + \mathrm{Cov}\left\{\left[E\left(\frac{\partial\ell}{\partial\theta_j}\Big|P;\theta\right),E\left(\frac{\partial\ell}{\partial\theta_k}\Big|P;\theta\right)\right];\theta\right\}. \end{aligned} \tag{4.2.2}$$

The left hand side of (4.2.2) is the (j,k)th element of the expected information matrix for a complete set of data, say $J_{P,M}(\theta,\theta)$. By equation (4.2.1), the second term on the right hand side of (4.2.2) is the (j,k)th element of the expected information matrix for the incomplete data, $J_P(\boldsymbol{\theta},\boldsymbol{\theta})$. The first term on the right hand side of (4.2.2) represents the "lost information" in the missing data. Thus

$$J_{P,M}(\theta_j,\theta_k) = J_P(\theta_j,\theta_k) + J_{MIP}(\theta_j,\theta_k) \quad , \tag{4.2.3}$$

where $$J_{MIP}(\theta_j,\theta_k) = E\left\{ Cov\left[\left(\frac{\partial\ell}{\partial\theta_j},\frac{\partial\ell}{\partial\theta_k}\right)\Big| P;\boldsymbol{\theta}\right];\boldsymbol{\theta}\right\} . \tag{4.2.4}$$

The lost information, calculated from (4.2.4), may be simple to calculate for any set of data P. Then $J_P(\boldsymbol{\theta},\boldsymbol{\theta})$ is found using (4.2.3). We apply this procedure to the multivariate normal example in the next section.

4.3. The Multivariate Normal Case.

Now write $\boldsymbol{\theta} = (\boldsymbol{\mu},\boldsymbol{\Sigma})$, the means and covariances of the p-variate normal distribution. We carry out the analysis of the previous section.

The log-likelihood of a complete set of data $\mathbf{X}$ is

$$\ell(\mathbf{X};\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{p}\sum_{k=1}^{p}(x_{ij}-\mu_j)(x_{ik}-\mu_k)\sigma^{jk} - \frac{1}{2}N\ln(\det\boldsymbol{\Sigma}) .$$

Differentiating with respect to $\boldsymbol{\theta}$, we find

$$\frac{\partial\ell}{\partial\mu_j} = \sum_{i=1}^{N}\sum_{r=1}^{p}(x_{ir}-\mu_r)\sigma^{jr} \quad , \tag{4.3.1}$$

$$\frac{\partial \ell}{\partial \sigma_{\ell m}} = -\tfrac{1}{2}N(2-\delta_{\ell m})\Big[\sigma^{\ell m} - \tfrac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{p}\sum_{s=1}^{p}\sigma^{\ell r}\sigma^{ms}(x_{ir}-\mu_r)(x_{is}-\mu_s)\Big] , \qquad (4.3.2)$$

where $\delta_{\ell m} = \begin{cases} 1 & ,\ell = m \\ 0 & ,\ell \neq m \end{cases}$, the Kronecker Delta.

The factor $(2-\delta_{\ell m})$ results from the symmetry of Σ . We now find the expected information for a complete set of data, by finding the variances and covariances of the "scores" of (4.3.1), (4.3.2). We require the following simple properties of the moments of $MN_p(\mu,\Sigma)$: for $1 \leqslant r,s,t,u \leqslant p$,

$$\left.\begin{aligned} &\mathrm{Cov}\big[(x_{ir}-\mu_r),(x_{is}-\mu_s)(x_{it}-\mu_t)\big] = 0 \quad , \\ &\mathrm{Cov}\big[(x_{ir}-\mu_r)(x_{is}-\mu_s),(x_{it}-\mu_t)(x_{iu}-\mu_u)\big] = \sigma_{rt}\sigma_{su}+\sigma_{ru}\sigma_{st} \end{aligned}\right\} . \qquad (4.3.3)$$

We find

$$J_{P,M}(\theta,\theta) = \begin{array}{c} \\ [\mu_k] \\ [\sigma_{rs}] \end{array}\!\!\begin{array}{c} \begin{array}{cc} \qquad [\mu_j] \qquad & \qquad\qquad [\sigma_{\ell m}] \qquad \end{array} \\ \begin{bmatrix} N\sigma^{jk} & 0 \\ 0 & \tfrac{1}{4}N(2-\delta_{\ell m})(2-\delta_{rs})(\sigma^{\ell r}\sigma^{ms}+\sigma^{\ell s}\sigma^{mr}) \end{bmatrix} \end{array} , \qquad (4.3.4)$$

where the elements of (4.3.4) correspond to submatrices as

$$1 \leqslant j,k,\ell,m,r,s \leqslant p \quad ; \quad r \leqslant s \ , \ \ell \leqslant m .$$

We now find the Lost Information, given by (4.2.4). The following generalization of (4.3.3) simplifies the calculations:-

Lemma 4.3. Suppose $(V_1, V_2, \dots, V_p) \sim MN_p(\mathbf{0}, \boldsymbol{\Sigma})$. Let A be any subset of the variables. Then for $1 \leq r,s,t,u \leq p$,

$$E\{Cov[(V_r, V_sV_t)|A;\Sigma];\Sigma\} = 0 \quad , \tag{4.3.5}$$

$$\begin{aligned} E\{Cov[(V_rV_s, V_tV_u)|A;\Sigma];\Sigma\} &= \sigma_{rt}\sigma_{su} - (\sigma_{rt}-\sigma_{rt\cdot A})(\sigma_{su}-\sigma_{su\cdot A}) \\ &+ \sigma_{ru}\sigma_{st} - (\sigma_{ru}-\sigma_{ru\cdot A})(\sigma_{st}-\sigma_{st\cdot A}) \ , \end{aligned} \tag{4.3.6}$$

where on the left hand side the inner covariances are taken with the variables in A fixed, and the outer expectations are over the distribution of the variables in A. In (4.3.6), $\sigma_{rt\cdot A} = Cov[(V_r, V_t)|A;\Sigma]$, etc.

The proof of the Lemma is straightforward. Writing $\mu_{j\cdot A} = E(V_j|A;\Sigma)$, apply (4.3.3) to the conditional distribution of the variables, with A fixed.

$$\left.\begin{array}{r} Cov[(V_r-\mu_{r\cdot A}), (V_s-\mu_{s\cdot A})(V_t-\mu_{t\cdot A})] = 0 \quad , \\ Cov[(V_r-\mu_{r\cdot A})(V_s-\mu_{s\cdot A}), (V_t-\mu_{t\cdot A})(V_u-\mu_{u\cdot A})] = \sigma_{rt\cdot A}\sigma_{su\cdot A} + \sigma_{ru\cdot A}\sigma_{st\cdot A} \end{array}\right\} . \tag{4.3.7}$$

Now expand (4.3.7) and take expectations, noting that

$$E(\mu_{j\cdot A};\Sigma) = 0 \quad ; \quad E(\mu_{j\cdot A}\mu_{k\cdot A};\Sigma) = \sigma_{jk} - \sigma_{jk\cdot A} \ .$$

Hence obtain (4.3.5) and (4.3.6).

Now we can find the Lost Information matrix, by applying the Lemma with $V_j = x_{ij} - \mu_j$ and $A = P_i$, the set of variables present in the ith observation. From (4.3.1),

$$J_{MIP}(\mu_j, \mu_k) = E\left\{ \mathrm{Cov}\left[\left(\frac{\partial \ell}{\delta \mu_j}, \frac{\partial \ell}{\partial \mu_k} \right) \middle| P; \Sigma\right]; \Sigma \right\}$$

$$= \sum_{i=1}^{N} \sum_{\ell=1}^{p} \sum_{m=1}^{p} \sigma^{j\ell} \sigma^{km} \sigma_{\ell m \cdot P_i} .$$

From equation (4.3.5), we obtain

$$J_{MIP}(\mu_j, \sigma_{\ell m}) = 0 .$$

From equation (4.3.6), we obtain

$$J_{MIP}(\sigma_{\ell m}, \sigma_{rs}) = \frac{1}{4}(2-\delta_{\ell m})(2-\delta_{rs}) \sum_{i=1}^{N} \sum_{t=1}^{p} \sum_{u=1}^{p} \sum_{v=1}^{p} \sum_{w=1}^{p} \sigma^{\ell t} \sigma^{mu} \sigma^{rv} \sigma^{sw} \gamma_{ituvw} ,$$

where $\gamma_{ituvw} = \sigma_{tv}\sigma_{uw} - (\sigma_{tv} - \sigma_{tv \cdot P_i})(\sigma_{uw} - \sigma_{uw \cdot P_i})$
$\qquad + \sigma_{tw}\sigma_{uv} - (\sigma_{tw} - \sigma_{tw \cdot P_i})(\sigma_{uv} - \sigma_{uv \cdot P_i})$.

Subtracting $J_{MIP}(\theta, \theta)$ from $J_{P,M}(\theta,\theta)$, we find the expected information matrix for the incomplete data, P :

$$J_p(\theta, \theta) = \begin{array}{c|cc} & [\mu_j] & [\sigma_{\ell m}] \\ \hline [\mu_k] & \sum_{i=1}^{N} \psi_{jk \cdot P_i} & 0 \\ [\sigma_{rs}] & 0 & \frac{1}{4}(2-\delta_{\ell m})(2-\delta_{rs}) \sum_{i=1}^{N} (\psi_{\ell r \cdot P_i}\psi_{ms \cdot P_i} + \psi_{\ell s \cdot P_i}\psi_{mr \cdot P_i}) \end{array} , \qquad (4.3.8)$$

where $\psi_{jk \cdot P_i} = \sum_{r=1}^{p} \sigma^{jr} \sigma^{ks} (\sigma_{rs} - \sigma_{rs \cdot P_i})$. (4.3.9)

Properties of $J_P(\theta, \theta)$

First we notice the formal resemblance of $J_P(\theta,\theta)$ and $J_{P,M}(\theta,\theta)$. The elements σ^{jk} in $J_{P,M}(\theta,\theta)$ are replaced by $\psi_{jk \cdot P_i}$ in $J_P(\theta,\theta)$. Two extreme patterns of missing values are :

(i) all the variables in observation i observed: $\psi_{jk \cdot P_i} = \sigma^{jk} \; \forall j,k.$

(ii) no variables in observation i observed : $\psi_{jk \cdot P_i} = 0 \; \forall j,k.$

Furthermore we can show

$$\psi_{jk \cdot P_i} = 0 \text{ if } x_{ij} \text{ or } x_{ik} \text{ is missing .} \tag{4.310}$$

This can be proved from (4.3.9), but we defer the proof until Chapter 6, when the result follows by a reparametrization of $\boldsymbol{\theta}$. From (4.3.10), we have

$J_p(\mu_j, \mu_k) = 0 \; \forall k$, if x_{ij} is missing in all the observations.

$J_p(\sigma_{jk}, \sigma_{\ell m}) = 0 \; \forall \ell, m$, if x_{ij} and x_{ik} are never observed together.

In both these situations we have a lack of information about the parameters, and the expected information matrix is singular. This parallels the results of §3.3, concerning the non-uniqueness of the solutions of the limiting equations of Iterated Buck when two variables are never observed together.

The expression $J_p(\boldsymbol{\theta}, \boldsymbol{\theta})$ has potential value in certain design problems. For example, we may wish to estimate the means of p correlated variables, but the nature of our experimental units limits us to measuring any r of the p variables for each unit, where $1 < r \leq p$. How do we allocate variables to units, given some a priori knowledge about the covariances between the variables? For the pattern $(P_1, P_2, \ldots, P_N)$, we have

$$J_p(\mu,\mu) = \Sigma^{-1} V \Sigma^{-1}, \text{ where } \quad V_{rs} = \sum_{i=1}^{N} (\sigma_{rs} - \sigma_{rs\cdot P_i}) \, ;$$
$$\det J_p(\mu,\mu) = (\det \Sigma)^{-2} \det V .$$

Thus if $\det J_p(\mu,\mu)$ is considered a good overall measure of information about the means, a design should be chosen to maximize $\det V$.

Another application of the matrix $J_p(\theta,\theta)$ lies in the calculation of confidence intervals and tests of significance of hypotheses concerning the means. This involves the inversion of $J_p(\theta,\theta)$. We now do this analytically for the case of two variables, and derive some results from the resulting asymptotic covariance matrix.

4.4. The Two Variable Problem .

Suppose we have the following paired data, with extra observations in one or both of the variables:

$$\left.\begin{array}{lll} n_c \text{ observations} & (x_{i1}, x_{i2}), \; (i=1,2,\ldots,n_c) \, ; & \lambda_c = \frac{n_c}{N} \, ; \\ n_1 \text{ observations} & (x_{i1}, -), \; (i=n_c+1, n_c+2,\ldots,n_c+n_1) ; & \lambda_1 = \frac{n_1}{N} \, ; \\ n_2 \text{ observations} & (-, x_{i2}), \; (i=n_c+n_1+1, n_c+n_1+2,\ldots, N) ; & \lambda_2 = \frac{n_2}{N} \, . \end{array}\right\} , \quad (4.4.1)$$

so that $\lambda_c + \lambda_1 + \lambda_2 = 1$.

First we find $J_p(\theta,\theta)$. From (4.3.10),

$$\psi_{12\cdot 1} = \psi_{12\cdot 2} = \psi_{11\cdot 2} = \psi_{22\cdot 1} = 0 .$$

Also $\psi_{11\cdot 1} = \sigma^{11} - \sum_{r=1}^{2}\sum_{s=1}^{2} \sigma^{1r}\sigma^{1s}\sigma_{rs\cdot 1} = \frac{1}{\sigma_{11}}$; $\psi_{22\cdot 2} = \frac{1}{\sigma_{22}}$.

So

$$J_p(\boldsymbol{\theta},\boldsymbol{\theta}) = \begin{bmatrix} J_p(\boldsymbol{\mu},\boldsymbol{\mu}) & \mathbf{0} \\ \mathbf{0} & J_p(\boldsymbol{\Sigma},\boldsymbol{\Sigma}) \end{bmatrix} ,$$

where $J_p(\boldsymbol{\mu},\boldsymbol{\mu}) = n_c\begin{bmatrix} \sigma^{11} & \sigma^{12} \\ \sigma^{12} & \sigma^{11} \end{bmatrix} + n_1\begin{bmatrix} \frac{1}{\sigma_{11}} & 0 \\ 0 & 0 \end{bmatrix} + n_2\begin{bmatrix} 0 & 0 \\ 0 & \frac{1}{\sigma_{22}} \end{bmatrix}$

$$= \frac{N}{1-\rho^2}\begin{bmatrix} [\lambda_c+\lambda_1(1-\rho^2)]\sigma_{11}^{-1} & -\lambda_c\rho^2\sigma_{12}^{-1} \\ -\lambda_c\rho^2\sigma_{12}^{-1} & [\lambda_c+\lambda_2(1-\rho^2)]\sigma_{22}^{-1} \end{bmatrix} ; \rho = \mathrm{Corr}(X_1,X_2).$$

Similarly

$$J_p(\boldsymbol{\Sigma},\boldsymbol{\Sigma}) = \frac{N}{2(1-\rho^2)^2}\begin{bmatrix} [\lambda_c+\lambda_1(1-\rho^2)^2]\sigma_{11}^{-2} & -2\lambda_c\rho^2\sigma_{11}^{-1}\sigma_{12}^{-1} & \lambda_c\rho^2\sigma_{11}^{-1}\sigma_{22}^{-1} \\ -2\lambda_c\rho^2\sigma_{11}^{-1}\sigma_{12}^{-1} & 2\lambda_c(1+\rho^2)\sigma_{11}^{-1}\sigma_{22}^{-1} & -2\lambda_c\rho^2\sigma_{22}^{-1}\sigma_{12}^{-1} \\ \lambda_c\rho^2\sigma_{11}^{-1}\sigma_{22}^{-1} & -2\lambda_c\rho^2\sigma_{22}^{-1}\sigma_{12}^{-1} & [\lambda_c+\lambda_2(1-\rho^2)^2]\sigma_{22}^{-2} \end{bmatrix} .$$

Inversion gives the asymptotic covariance matrix

$$J_p^{-1}(\boldsymbol{\theta},\boldsymbol{\theta}) = \begin{bmatrix} J_p^{-1}(\boldsymbol{\mu},\boldsymbol{\mu}) & \mathbf{0} \\ \mathbf{0} & J_p^{-1}(\boldsymbol{\Sigma},\boldsymbol{\Sigma}) \end{bmatrix} , \qquad (4.4.2)$$

where $J_p^{-1}(\boldsymbol{\mu},\boldsymbol{\mu}) = \frac{1}{NK_1}\begin{bmatrix} [\lambda_c+\lambda_2(1-\rho^2)]\sigma_{11} & \lambda_c\sigma_{12} \\ \lambda_c\sigma_{12} & [\lambda_c+\lambda_1(1-\rho^2)]\sigma_{22} \end{bmatrix}$,

(4.4.3)

$J_p^{-1}(\boldsymbol{\Sigma},\boldsymbol{\Sigma}) =$

$$\frac{1}{NK_2}\begin{bmatrix} [\lambda_c+\lambda_2(1-\rho^4)]\sigma_{11}^2 & [\lambda_c+\lambda_2(1-\rho^2)]\sigma_{11}\sigma_{12} & \lambda_c\sigma_{12}^2 \\ [\lambda_c+\lambda_2(1-\rho^2)]\sigma_{11}\sigma_{12} & \frac{1}{2\lambda_c}[\lambda_c^2(1+\rho^2)+(1-\rho^2)\{\lambda_c\lambda_1+\lambda_c\lambda_2+\lambda_1\lambda_2(1-\rho^2)^2\}]\sigma_{11}\sigma_{22} & [\lambda_c+\lambda_1(1-\rho^2)]\sigma_{22}\sigma_{12} \\ \lambda_c\sigma_{12}^2 & [\lambda_c+\lambda_1(1-\rho^2)]\sigma_{22}\sigma_{12} & [\lambda_c+\lambda_1(1-\rho^4)]\sigma_{22}^2 \end{bmatrix} ,$$

(4.4.4)

where $K_1 = (\lambda_c+\lambda_1)(\lambda_c+\lambda_2)-\lambda_1\lambda_2\rho^2$ and $K_2 = (\lambda_c+\lambda_1)(\lambda_c+\lambda_2)-\lambda_1\lambda_2\rho^4$.

We now concentrate on (4.4.3). Rearrangement gives

$$\mathrm{Var}\,\hat{\mu}_1 \overset{\text{asympt.}}{=} \frac{\sigma_{11}}{n_c+n_1}\left[1-\frac{\lambda_c\lambda_2\rho^2}{(\lambda_c+\lambda_1)(\lambda_c+\lambda_2)-\lambda_1\lambda_2\rho^2}\right] ,$$

(4.4.5)

and the second term of the right hand side of (4.4.5) represents asymptotically the gain in precision in using the extra observations on X_2 to estimate μ_1; for the naive estimate, the sample mean of X_1 over the first n_c+n_1 observations, has variance $\frac{\sigma_{11}}{n_c+n_1}$. This increase in precision increases with ρ^2, and as $\rho^2 \to 1$, $\mathrm{Var}\,\hat{\mu}_1 \to \frac{\sigma_{11}}{N}$. This is as it should be.

Now suppose we wish to test the hypothesis:

$$H_0 : \mu_1 = \mu_{10}$$

against ageneral alternative. A simple procedure is to construct a test based on the m.l.e. of μ_1. Asymptotically,

$$\hat{\mu}_1 \sim N\left(\mu_{10}, \frac{[\lambda_c+\lambda_2(1-\rho^2)]\sigma_{11}}{NK_1}\right) ,$$

Hence $$Z_1 = \frac{\hat{\mu}_1-\mu_{10}}{\sqrt{\frac{1}{NK_1}[\lambda_c+\lambda_2(1-\hat{\rho}^2)]\hat{\sigma}_{11}}} \overset{\text{asympt}}{\sim} N(0,1), \text{ under } H_0. \quad (4.4.6)$$

We substitute m.l.e's for ρ, σ_{11} and σ_{22} in the denominator of Z_1. We can use Z_1 with standard normal tables to obtain confidence intervals for μ_1, and to test H_0.

For small samples this statistic suffers the usual drawbacks of a normal approximation, and hence we seek to approximate to a t-distribution. From (4.4.4), after a little manipulation, we have

$$\mathrm{Var}\,\hat{\sigma}_{11} \overset{\text{asympt.}}{=} \frac{2\sigma_{11}^2}{n_c+n_1}\left[1-\frac{\rho^4\lambda_c\lambda_2}{K_2}\right] .$$

(4.4.7)

This suggests that approximately

$$Z_1 \sim t_\nu \tag{4.4.8}$$

under the null hypothesis, where t_ν is the standard t distribution with ν degrees of freedom, and

$$\nu = \frac{n_c+n_1-1}{\left[1-\frac{\rho^4\lambda_c\lambda_2}{K_2}\right]} \;. \tag{4.4.9}$$

In (4.4.9) we replace n_c+n_1 by n_c+n_1-1, so that when $n_2=0$, $\hat{\mu}_1 = \frac{1}{n_c+n_1}\sum_{i=1}^{n_c+n_1} x_{i_1}$, $Z_1 = \frac{\hat{\mu}_1-\mu_{10}}{\sqrt{\frac{1}{N}\hat{\sigma}_{11}}}$, and then $\nu = n_c+n_1-1$ is such that (4.4.8) is exact. In postulating ν we ignore the fact that the factor of $\hat{\sigma}_{11}$ in the denominator of Z_1 has to be estimated. Nevertheless (4.4.8) should be an improvement on the normal approximation.

It would be possible to construct the generalized likelihood ratio (g.l.r.) test of H_0. The resulting statistic, like Z_1, does not have a closed form, and it requires more computing than Z_1, since two iterative processes are involved for estimation under the null and alternative hypotheses. Both tests are asymptotically efficient, and for small samples an approximate distribution such as (4.4.8) appears to be less easy to construct. We do not consider the g.l.r. statistic here, although a comparison with Z_1 would be interesting.

The same remarks apply to the problem of estimating the difference of two means, $\delta = \mu_1-\mu_2$. We propose estimating δ by its m.l.e.

$$\hat{\delta} = \hat{\bar{x}}_1 - \hat{\bar{x}}_2 ,$$

where $\hat{\bar{x}}_j$ is the mean of X_j over the final observed or fitted values found by Iterated Buck. The asymptotic variance of $\hat{\delta}$ is found from (4.4.3):

$$\left.\begin{aligned} \text{Var}\,\hat{\delta} &= \sigma_\delta^2 = c_{11}\sigma_{11} + c_{22}\sigma_{22} - 2c_{12}\sigma_{12} , \\ c_{11} &= \frac{\lambda_c + \lambda_2(1-\rho^2)}{NK_1} , \quad c_{22} = \frac{\lambda_c + \lambda_1(1-\rho^2)}{NK_1} , \\ c_{12} &= \frac{\lambda_c}{NK_1} . \end{aligned}\right\} . \qquad (4.4.10)$$

Thus to test the null hypothesis $H_0' : \delta = \delta_0$, or to find confidence intervals for δ, we propose the statistic

$$Z = \frac{\hat{\delta} - \delta_0}{\hat{\sigma}_\delta} , \qquad (4.4.11)$$

where $\hat{\sigma}_\delta^2$ is the m.l.e. of σ_δ^2, found by substituting the m.l.e. of Σ in (4.4.10). As before under H_0', Z is asymptotically normal, but we find an approximate t-distribution from the asymptotic covariance matrix of $\hat{\Sigma}$. The variance of

$$\hat{\sigma}_\delta^2 = \hat{c}_{11}\hat{\sigma}_{11} + \hat{c}_{22}\hat{\sigma}_{22} - 2\hat{c}_{12}\hat{\sigma}_{12}$$

is approximately

$$\hat{c}_{11}^2 \text{Var}\,\hat{\sigma}_{11} + 2\hat{c}_{11}\hat{c}_{22}\,\text{Cov}(\hat{\sigma}_{11}, \hat{\sigma}_{22}) + \hat{c}_{22}^2\,\text{Var}(\hat{\sigma}_{22})$$
$$+ 4\hat{c}_{12}^2\,\text{Var}(\hat{\sigma}_{12}) - 4\hat{c}_{11}\hat{c}_{12}\,\text{Cov}(\hat{\sigma}_{11}, \hat{\sigma}_{12}) - 4\hat{c}_{22}\hat{c}_{12}\,\text{Cov}(\hat{\sigma}_{22}, \hat{\sigma}_{12}).$$

Substituting the corresponding elements of $J_\rho^{-1}(\Sigma, \Sigma)$

for the variance and covariance terms, we obtain an expression for $Var\,\hat{\sigma}_\delta^2$. Writing this in the form

$$Var\ \hat{\sigma}_\delta^2 = \frac{2\sigma_\delta^4}{K_\delta} , \tag{4.4.12}$$

we assign K_δ degrees of freedom to the approximate t-distribution for Z . Again we ignore the approximation in estimating the c_{jk}'s and the variance terms in $\hat{\delta}$. We find K_δ for two special cases:-

(a) $\lambda_c = 1, \ \lambda_1 = \lambda_2 = 0$.

For complete data $\sigma_\delta^2 = \frac{1}{N}(\sigma_{11} + \sigma_{22} - 2\sigma_{12})$, and

$$Var(\hat{\sigma}_\delta^2) = \frac{2}{N^3}\left[\sigma_{11}^2 + 2\sigma_{12}^2 + \sigma_{22}^2 + 2\sigma_{11}\sigma_{22}(1+\rho^2) - 4\sigma_{11}\sigma_{12} - 4\sigma_{22}\sigma_{12}\right]$$

$$= \frac{2}{N}\sigma_\delta^4 .$$

Hence $K_\delta = N$. In fact Z is simply the paired t statistic, which has $N-1$ degrees of freedom. To bring the approximate method in line with the exact distribution for this special case, there is something to be said for assigning $K_\delta - 1$ rather than K_δ degrees of freedom to the t-statistic, in the general case.

(b) $\lambda_c = 0$.

With no complete observations, we are left with two independent samples of size n_1 and n_2 , i.e. the Fisher-Behrens problem. As pointed out in the previous chapter, we have no information for estimating σ_{12} . Nevertheless Iterated Buck converges slowly to a solution, and convergence is speeded up by setting $\hat{\sigma}_{12} = 0$ after each iteration. The resulting estimates of μ_1 and μ_2 are the sample means, and the variances σ_{11} and σ_{22} are estimated by the sample variances. Also

$$C_{11} = \frac{1}{n_1} \quad , \; C_{22} = \frac{1}{n_2} \quad , \; C_{12} = 0 \, ,$$

and

$$K_\delta = \frac{\left(\frac{\sigma_{11}}{n_1} + \frac{\sigma_{22}}{n_2}\right)^2}{\left(\frac{\sigma_{11}^2}{n_1^3} + \frac{\sigma_{22}^2}{n_2^3}\right)} \, , \tag{4.4.13}$$

which is similar to the approximate degrees of freedom for the Fisher-Behrens problem suggested by Welch (1947), equation 26, p.32:-

$$f = \frac{\left(\frac{\sigma_{11}}{n_1} + \frac{\sigma_{22}}{n_2}\right)^2}{\left(\frac{\sigma_{11}^2}{n_1^2(n_1+1)} + \frac{\sigma_{22}^2}{n_2^2(n_2+1)}\right)} \; ; \tag{4.4.14}$$

K_δ and f differ in that f takes into account the estimation of μ_1 and μ_2 .

<u>Alternative assumptions about $\boldsymbol{\Sigma}$</u>.

The statistic Z belongs to a class $\mathcal{C}$ of statistics of the form

$$Z_A = \frac{\hat{\delta}_A - \delta_0}{\hat{\sigma}_{\delta A}} \, , \tag{4.4.15}$$

where $\hat{\delta}_A$, $\hat{\sigma}_{\delta A}^2$ are m.l.e's of $\delta_A, \sigma_\delta^2$ under some assumption A about the covariance matrix Σ . A general approach to the problem of estimating and testing is to find a statistic in $\mathcal{C}$ for the relevant assumption A, and then find an approximate t-distribution, that is an appropriate number of degrees of freedom, ν_A .

With a reasonable proportion of complete observations, a lower bound for ν_A is found by considering the degrees of freedom when incomplete observations are rejected. The statistics in $\mathcal{C}$ formed from the n_c complete observations are described in Table 1.

	Assumption about Σ (A)	Statistic	Degrees of Freedom
A1 :	Σ arbitrary.	paired t	n_c-1
A2 :	$\frac{\sigma_{11}}{\sigma_{22}}$ known; ρ unknown.	paired t	n_c-1
A3 :	$\frac{\sigma_{11}}{\sigma_{22}}$ unknown; $\rho=0$.	"unpaired t"	$(n_c-1) < \nu_{A_3} < 2(n_c-1)$
A4 :	$\frac{\sigma_{11}}{\sigma_{22}}$ known; $\rho=0$.	"unpaired t"	$2(n_c-1)$
A5 :	Σ known	normal	∞

Table 1. Statistics for testing $H_0': \delta=\delta_0$, based on the complete observations.

Also A3* and A4* are analogues of A3 and A4, when ρ is non-zero and known, with the same degrees of freedom as when $\rho=0$. Assumption A3 is the Fisher-Behrens problem with equal sample sizes, so ν_{A_3} is formed by setting $n_c=n_1=n_2$ in (4.4.14). In all cases $\hat{\delta}$ is the difference in sample means, and "paired" and "unpaired" refers to the estimate of variance.

Now consider statistics in $\mathcal{C}$ which use all the data, and compare them with other approaches in the literature. Lin and Stivers (1974) finds $\hat{\delta}_{A_S}$, the m.l.e. of δ when Σ is known. This is normal, with a variance given exactly by (4.4.10). For unknown Σ, they propose estimating δ and σ_δ^2 by substituting the estimate of Σ found from the complete observations. The resulting statistic lies in $\mathcal{C}$ if n_1 or n_2 is zero. They propose n_c-1 degrees of freedom for the approximate t-statistic. With extra observations in both variables, this statistic differs from Z in the estimate of Σ, which does not use all the available information. An iterative calculation is avoided, but for small numbers of complete observations one would expect Z to be more powerful.

Morrison (1972) tests H_0' for extra observations in one variable only (say $n_2=0$), and an assumption about Σ similar to A4*:

$$\Sigma = \sigma^2\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} , \rho \text{ known, } \sigma^2 \text{ unknown.} \qquad (4.4.16)$$

He calculates the g.l.r. statistic, and slightly modifies the estimate of σ^2 to obtain a statistic

$$Z_M = \frac{\hat{\delta}_M - \delta_0}{\hat{\sigma}_{\delta_M}} ,$$

which is distributed exactly as a t_{N+n_c-3} distribution under $H_0': \delta=\delta_0$. In fact Z_M is nearly in $\mathcal{C}$, since $\hat{\delta}_M$ can be shown to be the m.l.e. of δ under (4.416), and is nearly maximum likelihood. For unknown ρ, Morrison (1973) replaces ρ by its m.l.e. when the variances are assumed equal, viz.

$$\hat{\rho} = \frac{2a_{12}}{a_{11}+a_{22}} ,$$

where $\mathbf{A}$ is the S.S.C.P. matrix based on the complete observations. Thus the estimate of δ is still maximum likelihood. Morrison proposes n_c-1 degrees of freedom for this statistic. These results may be compared with A2 and A4 of the table.

The general approach of this section can of course be applied to a hypothesis about any linear contrast of the means, for two or more variables. Furthermore to test the more general linear hypothesis

$$H_0'' : \mathbf{C}\boldsymbol{\mu} = \mathbf{C}\boldsymbol{\mu}_0$$

for a constant (rxp) matrix $\mathbf{C}$, we have

$$\mathbf{C}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0) \overset{asympt}{\sim} MN_r(\mathbf{0}, \mathbf{C}J_p^{-1}(\boldsymbol{\mu},\boldsymbol{\mu})\mathbf{C}^T) ,$$

from which we can construct in the usual way a test statistic which is asymptotically χ^2 with r degrees of freedom, under H_0 .

5. MULTIPLE LINEAR REGRESSION.

5.1. Introduction.

So far we have treated the variables $X_1, X_2, \dots, X_p$ symmetrically; now we write $Y \equiv X_p$ and consider the regression of Y on $X_1, X_2, \dots, X_{p-1}$. First consider the following model:-

$$\left.\begin{aligned} y_i &= \sum_{j=0}^{p-1} \beta_j x_{ij} + \epsilon_i \\ E(\epsilon_i) &= 0 \\ E(\epsilon_{i_1}\epsilon_{i_2}) &= \delta_{i_1, i_2}\, \sigma_Y^2 \end{aligned}\right\} , \qquad (5.1.1)$$

where $1 \leqslant i, i_1, i_2 \leqslant N$, and x_{i0} is identically equal to 1 for all i, so β_0 is a constant term. (The analysis which follows is easily adapted to a regression through the origin.) We write

$$\boldsymbol{\theta}_1 = (\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma_Y^2)^T ,$$

the parameters of interest, and compare estimators of $\boldsymbol{\theta}_1$ when values are missing in the dependent and independent data.

We can estimate $\boldsymbol{\theta}_1$ by maximum likelihood. For an incomplete sample from $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, as in Chapter 2 writing $\boldsymbol{\theta} = (\mu_1, \mu_2, \dots \mu_p, \sigma_{11}, \sigma_{12}, \sigma_{22}, \dots, \sigma_{pp})$, $\boldsymbol{\theta}_1$ is a (1-1) vector function of $\boldsymbol{\theta}$, and hence the m.l.e. of $\boldsymbol{\theta}_1$ is

$$\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1(\hat{\boldsymbol{\theta}}) , \qquad (5.1.2)$$

where $\hat{\theta}$ is the m.l.e. of θ, found by M.L.N. In practice $\hat{\theta}_1$ is found by pivoting on the matrix $\hat{\Sigma}$.

However the multivariate normality assumption is often unrealistic in the regression situation. Indeed some of the independent variables may be points of a design matrix, and therefore fixed. Thus it is desirable to weaken the assumptions under which $\hat{\theta}_1$ is the m.l.e. of θ_1. We can certainly do this if the data are complete. Then $\hat{\theta}_1$ also maximizes the conditional likelihood of the data with the independent variables fixed. No distributional assumptions are needed for $X_1, X_2, \cdots, X_{p-1}$, and we are left with the standard model (5.1.1) with an i.i.d. normal structure of error.

With missing independent data this model is not enough, since we require a distribution for missing independent variables in order to use the information in the incomplete observations. However if a set of the independent variables, say $X_1, X_2, \ldots, X_r$ $(r \leq p-1)$, are present in every observation, then $\hat{\theta}_1$ maximizes the conditional likelihood of the data with $X_1, X_2, \ldots, X_r$ fixed, so no distributional assumptions are needed for these variables. This is stated more generally and proved in §5.2.

For $\hat{\theta}_1$ given by (5.1.2) to be the m.l.e. of θ_1 we require a multivariate normal distribution for Y and the independent variables which are sometimes missing. In §5.3 we consider a general distribution for the missing independent variables, with a normal structure of error, and we see how this affects the estimate of θ_1 found by applying the M.I.P. In general M.L.N. no longer finds the m.l.e. of θ_1, but since it fits the best linear approximations to the missing values, it remains a valid method. The m.l.e. found by solving the fixed point equations is generally much harder to compute, and this is illustrated by some examples.

Whereas in §5.3 we generalize the distribution of the independent variables, in §5.4 we generalize the distribution of the errors. Nelder and Wedderburn (1972) give a concise formulation of how to construct and solve the maximum likelihood equations, for regression with a non-normal structure of error, when the error variance is proportional to a known function of the mean. The equations are solved by Iterative Weighted Least Squares. In §5.4 the method is modified to deal with missing values in the independent variables.

Both for the "maximum likelihood" and the "fitting" approaches to the problems of this chapter, we cannot proceed without a distribution for the missing variables. So far we have estimated this distribution from the data, by maximum likelihood. But it is desirable from a theoretical point of view to provide a framework for incorporating other information about a missing variable into the analysis. Consider the following 5 observations on 3 variables, with Y the response:

$$\begin{bmatrix} X_1 \\ X_2 \\ Y \end{bmatrix} = \begin{bmatrix} 1 \\ 2.2 \\ 1.3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3.1 \\ 1.1 \end{bmatrix}, \begin{bmatrix} 3 \\ 0.6 \\ -2.0 \end{bmatrix}, \begin{bmatrix} ? \\ 2.8 \\ 2.9 \end{bmatrix}, \begin{bmatrix} 5 \\ 4.2 \\ -0.5 \end{bmatrix} .$$

A regression of Y on X_1 and X_2 using the 4 complete observations, gives $E(Y) \hat{=} X_2 - X_1$. Using this information, the fitted value for x_{41} is about zero. But suppose we know that X_1 is a controlled variable, and we have external information that indicates a value of 4. We may wish to incorporate this by some prior distribution for x_{41}, with high probability at $x_{41} = 4$. This will evidently reduce the goodness of fit of the regression equation. The hazards of such a procedure are obvious, and the example is illustrative rather than practical in nature.

A theoretical basis for such an approach is outlined in §5.5. We construct a prior distribution for the missing independent variables, which may be regarded as subjective, or as having a frequency interpretation based on past data. We also allow the prior to depend on the independent data, but not on the values of the response variable Y. In this way the maximum likelihood methods we have been considering are included in the general framework. The prior distribution is converted into a posterior distribution by the dependent data, via the model (5.1.1) with normal errors. Hence data on the dependent variable have a different logical status to the independent data, which we consider as "prior knowledge". The resulting structure is flexible, but its usefulness will depend on the feasibility of specifying the prior to a practical situation. Some applications are discussed briefly.

Finally we include a brief note about the uses of M.L.N. or Iterated Buck for estimating the parameters of a Multivariate Linear Model. The ideas are a simple generalization from the univariate response variable to a multivariate response variable.

In this chapter we classify each variable x_{ij} in the data matrix as "present" (P) or "missing" (M), and "independent" (x) or "dependent" (y) . We adopt the following notation, applied to (a) all the data and (b) the ith observation. The marginal sets areunions over classifications, so $P = P_x \cup P_y$, etc.

(a) All the data:

$$\begin{array}{l|c:c|l} & \text{present} & \text{missing} & \\ \hline \text{independent} & P_x & M_x & X' \\ \hdashline \text{dependent} & P_y & M_y & y \\ \hline & P & M & \end{array}$$

(b) observation i:

$$\begin{array}{l|c:c|l} & \text{present} & \text{missing} & \\ \hline \text{independent} & P_{x_i} & M_{x_i} & x_i' \\ \hdashline \text{dependent} & P_{y_i} & M_{y_i} & y_i \\ \hline & P_i & M_i & x_i \end{array}$$

5.2. Maximum Likelihood Estimates of θ_1

We divide the independent variables $X_1, X_2, \ldots, X_{p-1}$ into two groups: X_c are the variables present in all N observations, and X_I are the variables which are sometimes missing. Let P_c be the set of data on X_c, and P_I the set of data on X_I, so $P_X = P_c \cup P_I$. Let $\ell(P;\theta)$ be the log-likelihood of the data, for some distribution for P indexed by θ. If the factorization

$$\ell(P;\theta) = \ell_1(P|P_c;\phi_1) + \ell_2(P_c;\phi_2) , \qquad (5.2.1)$$

where ϕ_1 and ϕ_2 are functions of θ, is such that ϕ_1 and ϕ_2 are disjoint sets of parameters, and $\Omega_{\phi_1,\phi_2} = \Omega_{\phi_1} \times \Omega_{\phi_2}$, then the estimate of ϕ_1 found by maximizing $\ell_1(P|P_c;\phi_1)$ is the same as that found by maximizing $\ell(P;\theta)$ with respect to θ and setting $\hat{\phi}_1 = \phi(\hat{\theta})$.

Now suppose the distribution of Y given $X_1, X_2, \ldots, X_{p-1}$ is indexed by θ_1. We do not specify θ_1 in this general formulation, but in the context of this chapter θ_1 represents the coefficients of a linear regression, and parameters associated with the residual error. Now θ_1 is a function of ϕ_1, and so the m.l.e. of θ_1 is

$$\hat{\theta}_1 = \theta_1(\hat{\phi}_1) = \theta_1[\phi_1(\hat{\theta})] ,$$

where $\hat{\phi}_1$ can be found by maximizing $\ell(P;\theta)$ or $\ell_1(P|P_c;\phi_1)$.

Now suppose $\boldsymbol{\theta}$ are the means and covariances of the multivariate normal distribution, and $\hat{\boldsymbol{\theta}}$ is the m.l.e. of $\boldsymbol{\theta}$, found by M.L.N. Then for the distribution of Y given $X_1, X_2, \ldots, X_{p-1}$,

$$\left.\begin{aligned} \boldsymbol{\theta}_1 &= (\beta_0, \beta_1, \ldots, \beta_{p-1}, \sigma_Y^2)^T \\ y_i &= \sum_{j=0}^{p-1} \beta_j x_{ij} + \epsilon_i, \; (i=1,2,\ldots,N) \\ \epsilon_i &\overset{i.i.d}{\sim} N(0, \sigma_Y^2) \end{aligned}\right\} \tag{5.2.2}$$

The above theory implies that the m.l.e. $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1(\hat{\boldsymbol{\theta}})$ found by pivoting on the estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$ is also the m.l.e. of $\boldsymbol{\theta}_1$ under a model which fixes P_c, and results in a log-likelihood $\ell_1(P|P_c; \phi_1)$ which satisfies (5.2.1). This model is given by (5.2.2), with the additional assumptions: (a) $(X_I | X_c)$ is multivariate normally distributed, with constant covariance matrix, and a mean which is a linear combination of the variables in X_c; (b) the observations are independent, in the sense that the distribution of $(P_I | P_c)$ factorizes into N terms corresponding to the N observations; (c) the deletion pattern is random, and in particular $(P_I | P_c)$ has its marginal distribution after the missing variables have been integrated out of the distribution of $(X | P_c)$, given by (a).

Two extreme patterns of missing values are of some interest. First if P_c is empty, we are led to the full multivariate normal model of Chapter 2, as one might expect. If $P_c = P_x$, that is all the independent variables are present, then the independent data are fixed, and we have the model (5.2.2). Observations with Y missing contribute no information to the estimation of θ_1, and Iterated Buck is equivalent to ordinary least squares on the complete observations. Even in this situation the method has some computational value. Orchard and Woodbury (1971) suggest that Iterated Buck may be quicker to compute than a least squares analysis, when extra design points can be added to make a balanced design. Such a procedure is equivalent to standard missing value techniques as used by Yates (1933), Tocher (1952), and others.

§5.3. Regression with Normal Errors.

We have seen that in order to estimate the linear regression of Y on $X_1, X_2, \ldots, X_{p-1}$ by maximum likelihood, given a random deletion pattern, our model must include the following characteristics.

(i) A distribution for $(Y | X_1, X_2, \ldots, X_{p-1})$, indexed by an unknown parameter θ_1;

(ii) A distribution for $(X_I | X_c)$, possibly indexed by an unknown parameter θ_2.

We apply the Missing Information Principle to a model of this type. Let $\ell(X', y \mid P_c; \theta_1, \theta_2)$ be the log-likelihood of a complete set of data. Then according to the principle, we seek a fixed point of the transformation ϕ, where $\phi(\theta_{1A}, \theta_{2A})$ is defined as the value of (θ_1, θ_2) which maximizes

$$E_A[\ell(X', y \mid P_c; \theta_1, \theta_2)] .$$

Here E_A refers to expectation over the conditional distribution of the missing data M, given the data P, at assumed values $(\theta_{1A}, \theta_{2A})$ of the parameters. Now

$$E_A[\ell(X', y \mid P_c; \theta_1, \theta_2)] = E_A[\ell_1(y \mid X', \theta_1)] + E_A[\ell_2(X' \mid P_c; \theta_2)] \tag{5.3.1}$$

where ℓ_1 and ℓ_2 are the log-likehoods of a complete set of data, corresponding to the components (i) and (ii) of the model given above. Hence if the parameter space factorizes, i.e.

$$\Omega_{(\theta_1, \theta_2)} = \Omega_{\theta_1} \times \Omega_{\theta_2} \quad , \tag{5.3.2}$$

then ϕ is equivalent to two separate maximizations:

(a) maximize $E_A[\ell_1(y \mid X'; \theta_1)]$ with respect to θ_1, leading to θ_{1M} ; (5.3.3)

(b) maximize $E_A[\ell_2(X' \mid P_c; \theta_2)]$ with respect to θ_2, leading to θ_{2M} . (5.3.4)

We consider (5.3.3) for the model with normal errors (5.2.2), without specifying the distribution of the missing independent variables. Then

$$\ell_1(y \mid X'; \theta_1) = -\frac{1}{2\sigma_y^2} \sum_{i=1}^{N} [y_i - \sum_{j=0}^{p-1} \beta_j x_{ij}]^2 - \tfrac{1}{2} N \ln \sigma_y^2 + \text{const},$$

so (5.3.3) is equivalent to solving

$$E_A\left(\frac{\partial \ell_1}{\partial \beta_j}\right) = \frac{1}{\sigma_y^2} E_A\left[\sum_{i=1}^{N} x_{ij}\left(y_i - \sum_{k=0}^{p-1} \beta_k x_{ik}\right)\right] = 0 \ , (j=0,1,\ldots,p-1),$$

$$E_A\left(\frac{\partial \ell_1}{\partial \sigma_y^2}\right) = E_A\left[\frac{1}{2\sigma_y^4}\sum_{i=1}^{N}\left(y_i - \sum_{k=0}^{p-1} \beta_k x_{ik}\right)^2 - \frac{N}{2\sigma_y^2}\right] = 0 \ .$$

Hence for the assumed distribution of $(M_i \mid P_i ; \theta_{1A}, \theta_{2A})$ we fit

$$\hat{x}_{ijA} = E(x_{ij} \mid P_i ; \theta_{1A}, \theta_{2A}) \tag{5.3.5}$$

if x_{ij} is missing, and then form the S.S.C.P. matrix of the completed data. Then for each observation i, we add to the (j,k)th element of this matrix the adjustment

$$Cov(x_{ij}, x_{ik} \mid P_i ; \theta_{1A}, \theta_{2A}) \qquad (1 \leq j,k \leq p) \ . \tag{5.3.6}$$

This adjustment is non-zero only if both x_{ij} and x_{ik} are missing. We then pivot on this adjusted matrix in the usual way, to obtain a new estimate θ_{1M} of θ .

Thus the influence of the distribution of the independent variables in the estimation of θ_1 appears solely in the resulting fitted values and adjustments of (5.3.5) and (5.3.6). We have seen that for the multivariate normal case the fitted values are linear combinations of the data, and the set of adjustments are the same for different observations with the same pattern of deletions. We consider this to be a good approximate procedure for many other problems, particularly when $\hat{x}_{ij}$ can be well approximated by a linear combination of the variables P_i. For non-normal distributions, the estimation of θ_2 and the fitted values and adjustments often involves a lot of computing, as can be seen from the examples which follow.

Example (5.3.1).

We wish to estimate the regression of $Y(\equiv X_3)$ on X_1 and X_2, and we suppose that the distribution of $(Y|X_1,X_2)$ is normal, with mean $\beta_0+\beta_1X_1+\beta_2X_2$, and variance σ_Y^2. The data consist of N independent observations, in which X_1 and Y are always observed, but X_2 is present for the first n_c observations, $(i=1,2,\dots,n_c)$, and missing for the remaining $N-n_c$ observations, $(i=n_c+1,n_c+2,\dots,N)$. We suppose X_2 is a binary variable, and

$$p_i = pr(x_{i2}=1|x_{i1}) = 1-pr(x_{i2}=0|x_{i1}),$$

where

$$p_i = \frac{exp(\lambda_0+\lambda_1x_{i1})}{1+exp(\lambda_0+\lambda_1x_{i1})} . \tag{5.3.7}$$

The parameters (λ_0,λ_1) are unknown, so in the general notation

$$\theta_1 = (\beta_0,\beta_1,\beta_2,\sigma_Y^2)^T \quad , \quad \theta_2 = (\lambda_0,\lambda_1)^T .$$

For such a model, the distribution of the missing data given the data is given by

$$pr\,(x_{i2}=1\,|\,x_{i1}, y_i) \;=\; \frac{p_i\, e^{-\phi_{i1}}}{p_i\, e^{-\phi_{i1}} + (1-p_i)\, e^{-\phi_{i0}}} \;, \tag{5.3.8}$$

where $\phi_{i1} = \frac{1}{2\sigma_Y^2}\left(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2\right)^2$,

$\phi_{i0} = \frac{1}{2\sigma_Y^2}\left(y_i - \beta_0 - \beta_1 x_{i1}\right)^2$.

Also $\ell_2(P_I|P_c\,;\boldsymbol{\theta}_2) = \sum_{i=1}^{N}\left[x_{i2}\ln p_i + (1-x_{i2})\ln(1-p_i)\right]$,

so $E_A\left[\ell_2(P_I|P_c\,;\boldsymbol{\theta}_2)\right] = \sum_{i=1}^{N}\left[\hat{x}_{i2A}\ln p_i + (1-\hat{x}_{i2A})\ln(1-p_i)\right]$, (5.3.9)

where $\hat{x}_{i2A} = E\left(x_{i2}\,|\,x_{i1}\,,\, y_i\,;\,\boldsymbol{\theta}_{1A}, \boldsymbol{\theta}_{2A}\right)$. (5.3.10)

Given assumed values of the parameters, we find the fitted values (5.3.10) and the adjustments from the distribution given by (5.3.8), and hence find a new estimate of $\boldsymbol{\theta}_1$. The new estimate of $\boldsymbol{\theta}_2$ is found by maximizing (5.3.9) with respect to λ_0 and λ_1 . This involves a logistic analysis of the model (5.3.8), for the completed set of data. The algorithm usually used is iterative weighted least squares, but probably one iteration will be sufficient within the overall iterative scheme for solving the fixed point equations.

Notice in this example the lack of distributional assumptions about X_1, which is always present. In the simple case with one binary variable missing, the specification of the model is straight forward and the analysis is not too involved. With several missing binary variables, which may be correlated, the logistic model tends to create a lot of nuisance parameters, and the analysis is complicated by the inconvenient forms of the marginal and conditional distributions required for different patterns of missing variables.

Example 5.3.2.

Now suppose we have a set of N observations $(\mathbf{z}_i, y_i)$, where

$$\left.\begin{aligned} \mathbf{z}_i &= (z_{i1}, z_{i2}, \ldots, z_{iq})^T \overset{i.i.d}{\sim} MN_q(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \\ (y_i \mid \mathbf{z}_i) &\sim N(\nu_i, \sigma_y^2) \quad , \\ \nu_i &= \sum_{j=0}^{p-1} \beta_j x_{ij} \end{aligned}\right\} . \qquad (5.3.11)$$

In (5.3.11), $x_{ij} = x_{ij}(\mathbf{z}_i)$ are known functions of the underlying $\mathbf{z}$-variables, for example polynomials in the in the components of $\mathbf{z}_i$. If $q = p-1$ and $x_{ij} = z_{ij}$, (j=1,2,...,p-1), we have the multivariate normal model of Chapter 2.

Once again the fitted values and adjustments required for a new estimate of θ_1 are given by (5.3.5) and (5.3.6). The distribution of the x_{ij}'s is in general not simple, and the calculations may involve numerical integration. The parameter θ_2 here represents the means and covariances of the z's, and the equations for a new estimate θ_{2M} of θ_2 are

$$\hat{\mu}_{zjM} = \frac{1}{N}\sum_{i=1}^{N} \hat{z}_{ijA} \ ,$$

$$\hat{\sigma}_{zjkM} = \frac{1}{N}\sum_{i=1}^{N}\left[(\hat{z}_{ijA}-\hat{\mu}_{zjA})(\hat{z}_{ikA}-\mu_{zkA}) + \hat{\sigma}_{zjkA\cdot P_i}\right] ,$$

where
$$\hat{z}_{ijA} = E\left[z_{ij} \mid P_i \,;\, \theta_{1A},\theta_{2A}\right] ,$$
$$\hat{\sigma}_{zjkA\cdot P_i} = Cov\left[z_{ij}, z_{ik} \mid P_i \,;\, \theta_{1A},\theta_{2A}\right] .$$

If y_i is observed, the distribution of the missing data M_i given P_i is in general non-normal, and the fitted values are not linear combinations of the known variables. We illustrate the problems of finding m.l.e's for such a model with a simple regression with a quadratic term.

<u>Example 5.3.3.</u>

This is a special case of the previous example. The model is

$$\left.\begin{aligned} & y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \\ & x_{i1} = z_{i1} \ , \ x_{i2} = z_{i2} \ , \ x_{i3} = z_{i1}^2 \\ & \begin{pmatrix} z_{i1} \\ z_{i2} \end{pmatrix} \overset{i.i.d}{\sim} MN_2\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} , \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}\right] \\ & \epsilon_i \overset{i.i.d}{\sim} N(0, \sigma_Y^2) \end{aligned}\right\} . \qquad (5.3.12)$$

We consider the fitted values and adjustments for different patterns of missing values, given assumed values of the parameters.

Case 1: Patterns with X_1 (and hence X_3) present. The distribution of $(X_2, Y | X_1)$ is normal, and so the fitted values and adjustments are the same as in Iterated Buck.

Case 2: Y and X_2 present, X_1 missing. For the fitted values and adjustments we need the mean and variance of X_1 and X_1^2, or equivalently the first four moments of the distribution of $(X_1 | Y, X_2)$

Now the density function

$$p(X_1 | Y, X_2 ; \theta_{1A}, \theta_{2A}) \propto p(Y | X_1, X_2 ; \theta_{1A}) \, p(X_1 | \theta_{2A}) ,$$

$$p(X_1 | Y, X_2 ; \theta_{1A}, \theta_{2A}) \propto \exp\left[-\frac{1}{2\sigma_{YA}^2} (Y - \beta_{0A} - \beta_{1A} X_1 - \beta_{2A} X_2 - \beta_{3A} X_1^2)^2 - \frac{(X_1 - \mu_{1A})^2}{2\sigma_{11A}} \right] .$$

Hence

$$E(X_1^r | Y, Y_2 ; \theta_{1A}, \theta_{2A}) = \frac{I_{rA}(Y, X_2)}{I_{0A}(Y, X_2)} ,$$ where

$$I_{rA}(Y, X_2) = \int_{-\infty}^{\infty} x_1^r \exp\left[-\frac{1}{2\sigma_{YA}^2} (Y - \beta_{0A} - \beta_{1A} x_1 - \beta_{2A} X_2 - \beta_{3A} x_1^2)^2 - \frac{(x_1 - \mu_{1A})^2}{2\sigma_{11A}} \right] dx_1 .$$

In practice the integrals I_{rA} are reparameterized to depend on **four** distinct parameters. Thus simplified the values of I_{0A}, I_{1A} and I_{2A} are found by numerical integration for each observation with this deletion pattern. The integrals I_{3A} and I_{4A} are found from I_{0A}, I_{1A} and I_{2A} by using the identities

$$\int_{-\infty}^{\infty} \psi'(x)\, e^{-\psi(x)}\, dx = 0 \ ,$$

$$\int_{-\infty}^{\infty} \frac{d}{dx}\left(x\, e^{-\psi(x)}\right) dx = \int_{-\infty}^{\infty} \left(1 - x\,\psi'(x)\right) e^{-\psi(x)} dx = 0,$$

where $\psi(x)$ is the exponent of the integrals I_{rA}. These expressions are also useful for checking the subroutines which carry out the numerical integrations.

Case 3: Y present, X_1 and X_2 missing.

For such a pattern we require moments of the distribution of $(X_1, X_2 | Y)$, which is not bivariate normal. Calculation of these moments involves double numerical integrations. The information recovered will only improve the estimate of the mean of Y, and in practice one might hope that this information is small, and that little is lost by discarding these observations.

Case 4: X_2 present , Y and X_1 missing.

Again double numerical integration is required. As in the previous case the information is heavily dependent on the multivariate normal assumption about the distribution of the X's. With the dependent variable missing, the value of these observations is even more questionable. We conclude from this example that

(1) Maximum Likelihood for higher order polynomial models requires in general much more computing time than Iterated Buck.

(2) The maximum likelihood equations should be treated with discretion when dealing with "sparse" observations, with several missing values; if these observations are used, the distributional assumptions should be tested by plotting or goodness-of-fit techniques.

Example (5.3.3) is perhaps the simplest practical example of a higher order polynomial model, for comparing the maximum likelihood estimates with those found by Iterated Buck. Observations with Y and X_2 present, X_1 missing, can carry considerable information and a pattern of deletions can be constructed with every pair of variables observed together, but with no complete observations. We report on a simulation study based on this example, in Chapter 7. However the practical importance of these techniques is greater for problems with more variables.

5.4. Regression with Non-Normal Errors - the Generalized Linear Model.

In the previous section we considered models with a normal distribution for $(Y|X_1,X_2,\ldots,X_{p-1})$. Nelder and

Wedderburn (1972) consider a more general class of regression problems, which arise from assuming the following Generalized Linear Model (G.L.M.). The density of $(y_i | x_{i1}, x_{i2}, \ldots, x_{ip})$ has the form

$$\pi(y_i\, ; \theta_i, \phi_0) = \exp\{\alpha(\phi_0)[y_i\theta_i - g(\theta_i) + h(y_i)] + \beta(\phi_0, y_i)\}\ , \tag{5.4.1}$$

where ϕ_0 are considered nuisance parameters, for example the variance of a normal distribution, and θ_i is a scalar, which depends on the independent variables through a systematic component

$$S_i = \sum_{j=0}^{p-1} \beta_j x_{ij}\ , \tag{5.4.2}$$

combined with a known link function

$$\theta_i = f(S_i)\ . \tag{5.4.3}$$

Finally the observations $y_1, y_2, \ldots, y_N$ are independent. The density (5.4.1) characterizes the unexplained variation in Y, and includes as special cases a Normal, Poisson, Binomial, or Gamma distribution of errors.

For a complete set of data, we can find the m.l.e. of $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$. If l_1 is the log-likelihood of the sample, then by differentiation we find

$$\frac{\partial l_1}{\partial \beta_j} = \alpha(\phi_0) \sum_{i=1}^{N} \omega_i (x_{ij} - \mu_i) \cdot \frac{1}{\left(\frac{d\mu_i}{dS_i}\right)} \tag{5.4.4}$$

$$-E\left(\frac{\partial^2 \ell_1}{\partial\beta_j\partial\beta_k}\right) = \alpha(\phi_0)\sum_{i=1}^{N} w_i x_{ij} x_{ik} \quad , \tag{5.4.5}$$

where the weight w_i is defined as

$$w_i = \frac{1}{V_i}\left(\frac{d\mu_i}{d s_i}\right)^2 \quad , \tag{5.4.6}$$

and μ_i, V_i are proportional to the mean and variance of y_i:

$$g'(\theta_i) = E(y_i) = \mu_i \quad , \tag{5.4.7}$$

$$g''(\theta_i) = \alpha(\phi_0)\mathrm{Var}(y_i) = V_i \, . \tag{5.4.8}$$

These equations are found directly from (5.4.1). The likelihood equations result from equating (5.4.4) to zero from j=0,1,...,p-1. One way of solving them is by Fisher's method of scoring. Given current estimates $\boldsymbol{\beta}_A$ of $\boldsymbol{\beta}$, calculate

$$\mathbf{M} = \left[-E\left(\frac{\partial^2 \ell_1}{\partial\beta_j\partial\beta_k}\right)\right] \; ; \; \mathbf{c} = \left(\frac{\partial \ell_1}{\partial\beta_0}, \frac{\partial \ell_1}{\partial\beta_1}, \ldots, \frac{\partial \ell_1}{\partial\beta_{p-1}}\right)^T \tag{5.4.9}$$

from (5.4.4.) and (5.4.5), and then form new estimates $\boldsymbol{\beta}_A + \delta\boldsymbol{\beta}$, where

$$\mathbf{M}\,\delta\boldsymbol{\beta} = \mathbf{c} \quad . \tag{5.4.10}$$

Then proceed iteratively. Alternatively, the estimates can be found by Iterative Weighted Least Squares (I.W.L.S.). Add $\mathbf{M}\boldsymbol{\beta}$ to both sides of (5.4.10), using

$$(\mathbf{M}\boldsymbol{\beta})_j = \alpha(\phi_0) \sum_{i=1}^{N} \sum_{k=0}^{p-1} w_i x_{ij} x_{ik} \beta_k = \alpha(\phi_0) \sum_{i=1}^{N} w_i x_{ij} S_i .$$

We obtain

$$\sum_{i=1}^{N} \sum_{k=0}^{p-1} w_i x_{ij} x_{ik} (\beta_k + \delta\beta_k) = \sum_{i=1}^{N} w_i x_{ij} \eta_i , \qquad (5.4.11)$$

where the weights w_i are calculated at each step from (5.4.6) and η_i is a modified dependent variable:

$$\eta_i = S_i + (y_i - \mu_i) \cdot \frac{1}{\left(\frac{d\mu_i}{dS_i}\right)} . \qquad (5.4.12)$$

Note that $\alpha(\phi_0)$ does not affect the estimation of $\boldsymbol{\beta}$, and ϕ_0 is not estimated in this process. Nelder and Wedderburn adopt a criterion of goodness-of-fit based on the likelihood ratio, called the deviance, and this is in general only proportional to the usual (asymptotic or exact) χ^2 statistic. However for Poisson or Binomial errors, ϕ_0 is not needed.

Now suppose we have an incomplete set of data, but all the y_i's are present. As in the previous section we assume the deletions are random, and the missing independent variables have some distribution indexed by $\boldsymbol{\theta}_2$. For the distribution of $(Y | X_1, X_2, \ldots, X_{p-1})$,

$$\theta_1 = (\beta_0, \beta_1, \ldots, \beta_{p-1}\,;\, \phi_0).$$

Assuming θ_1, θ_2 are disjoint sets, in the sense of (5.3.2), $\phi(\theta_A)$is given by the maximizations (5.3.3) and (5.3.4). Again we consider the maximization of $E_A[\ell_1(y|X';\theta_1)]$ with respect to β, for assumed values $(\theta_{1A}, \theta_{2A})$ of the parameters. Differentiating, we must solve

$$E_A\left(\frac{\partial \ell_1}{\partial \beta_j}\right) = E_A\left[\ \alpha(\phi_0) \sum_{i=1}^{N} w_i(x_{ij} - \mu_i)\frac{1}{\left(\frac{d\mu_i}{ds_i}\right)}\right]\ , \qquad (j=0,1,\ldots p-1)$$

this from (5.4.4). These equations can be solved by adapting the method of scoring, so that (5.4.10) becomes

$$E_A(M)\,\delta\beta = E_A(c)\,. \tag{5.4.13}$$

The equations for the increment in β are

$$\sum_{i=1}^{N}\sum_{k=0}^{p-1} E_A(w_i x_{ij} x_{ik})\,\delta\beta_k = \sum_{i=1}^{N} E_A\left[w_i x_{ij}(y_i - \mu_i)\frac{1}{\left(\frac{d\mu_i}{ds_i}\right)}\right]. \tag{5.4.14}$$

The evaluation of (5.4.14) is complicated by the fact that w_i, μ_i and $\frac{d\mu_i}{ds_i}$ are in general functions of the x_{ij}'s, and hence not constant with respect to E_A. We propose to approximate (5.4.14) by

$$\sum_{i=1}^{N-1}\sum_{k=0}^{p-1} w_i(\hat{x}_{ijA}\hat{x}_{ikA} + c_{ijkA})\,\delta\beta_k = \sum_{i=1}^{N} \frac{w_i}{\left(\frac{d\mu_i}{ds_i}\right)}\left[\hat{x}_{ijA}(y_i - \mu_i)\right], \tag{5.4.15}$$

where $$\hat{x}_{ijA} = E\left(x_{ij} \,|\, P_i\,;\, \theta_{1A}, \theta_{2A}\right)\ , \tag{5.4.16}$$

$$c_{ijkA} = \mathrm{Cov}\left[(x_{ij}, x_{ik}) \,|\, P_i\,;\, \theta_{1A}, \theta_{2A}\right]. \tag{5.4.17}$$

This assumes that approximately , ω_i and $\frac{\omega_i}{\left(\frac{d\mu_i}{dS_i}\right)}$ are fixed with respect to E_A . The weights ω_i are in general not too critical to the answer, so this should be a reasonable approximation, particularly for the important sufficient statistics case, when $\theta_i = S_i$. For then equations (5.4.6) and (5.4.8) imply that $\omega_i = \frac{d\mu_i}{dS_i}$.

For the normal errors model (5.2.2), $\omega_i = \frac{d\mu_i}{dS_i} = 1$, and (5.4.15) is exact and can be solved non-iteratively.

The modified I.W.L.S. equations (5.4.11) are found by adding $E_A(M\beta)_j$ to both sides of (5.4.15). Again the influence of the distribution of the missing variables in the estimation of β is solely in the fitted values and adjustments (5.4.16) and (5.4.17). We make the simplest assumption about this distribution, and propose the resulting method as an approximate procedure in the general case. The method is

Weighted Adjusted Iterative Least Squares (W.A.I.L.S.)

The method entails two distinct iterative cycles . First Iterated Buck is applied to the data, and the final fitted values $\hat{x}_{ij}$ and the adjustments $\hat{\sigma}_{jk\cdot P_i}$ for each pattern P_i of values present, are retained. Then (5.4.15), or the modified I.W.L.S. equations, are solved iteratively, with $\hat{x}_{ijA} = \hat{x}_{ij}$ and $c_{ijkA} = \hat{\sigma}_{jk\cdot P_i}$ for every iteration.

By using Iterated Buck to fill in the missing variables we are fitting linear approximations to the true values, and in this sense W.A.I.L.S. is the analogue of Iterated Buck for the Generalized Linear Model with missing data. With M.L.N. instead of Iterated Buck, W.A.I.L.S. would be maximum likelihood if the distribution of $(X_I | X_c, Y)$ was multivariate normal, with constant covariance matrix and a mean which depends linearly on X_c and Y. This condition, together with the required partition of the log-likelihood,

$$\ell(X_I, Y | X_c; \theta_1, \theta_2) = \ell_1(Y | X_1, X_2, \ldots, X_{p-1}; \theta_1) + \ell_2(X_I | X_c; \theta_2)$$

is only possible if $\mu_i = S_i$ and the errors are normal; then W.A.I.L.S. is equivalent to Iterated Buck.

For the complete maximum likelihood solution we must specify the distribution of $(X_I | X_c)$ and find the fitted values and adjustments from the first two moments of the distribution of $(X_I | X_c, Y)$. This has density

$$p(X_I | X_c, Y) \propto \pi(Y | \theta, \phi_0) p(X_I | X_c; \theta_2).$$

Even for a simple choice of $p(X_I | X_c; \theta_2)$, such as a multivariate normal density, $p(X_I | X_c, Y)$ will be a non-standard density if the error distribution π is non-normal, and except in special cases the calculations of the moments will require numerical integration. Any theoretical gain in finding the maximum likelihood estimate of θ_1 is counterbalanced by the large amount of computing involved in its calculation, as compared with W.A.I.L.S., and the feasibility of specifying the distribution of $(X_I | X_c)$. Thus W.A.I.L.S. would seem to be a useful generalization of Iterated Buck to a large class of generalized linear models.

5.5. Likelihood Methods with Missing Variables considered as Parameters.

So far we have treated independent variables which are missing in some observations as random, and have maximized

$$\ell(P;\theta) \quad \text{or} \quad \ell(P|P_c;\phi_1)$$

by maximum likelihood. In what follows we consider the known independent variables as fixed, and the unknown independent variables as parameters, and maximize the conditional log likelihood

$$\ell(P_y|P_x;M_x,\theta_1) \tag{5.5.1}$$

to obtain estimates of θ_1. We assume $P_y = \mathbf{y}$, that is all the y_i's are present; observations with Y missing may be regarded as containing extraneous information about M_x.

To maximize (5.5.1) jointly with respect to M_x and θ_1, we simply discard the incomplete observations for the estimation of θ_1. For example with normal errors, (5.5.1) becomes

$$-\frac{1}{2\sigma_Y^2}\sum_{i=1}^{N}\Big(y_i-\sum_{j=0}^{p-1}\beta_j x_{ij}\Big)^2-\frac{1}{2}N\ln\sigma_Y^2 ,$$

and now if x_{ik} is missing we can choose its estimate so that

$y_i - \sum_{j=0}^{p-1} \hat{\beta}_j x_{ij} = 0$. This analysis corresponds to total lack of knowledge about the missing values: we cannot predict them from the data, and have no other information. We also assume implicitly that the distribution of Y depends on the missing variables, i.e. in the example above $\beta_k \neq 0$.

We may consider M_x as incidental parameters, and θ_1 as structural parameters, fixed over observations. Suppose we can specify some prior distribution for the incidental parameters, say

$$\pi(M_x | P_x) \quad , \qquad (5.5.2)$$

which may depend not only on the independent data, but also on external information. We can then remove M_x from the likelihood or log-likelihood, and maximize the resulting function with respect to θ_1. We consider two ways of doing this, one suggested by the theory of likelihoods and one suggested by the procedures of §5.2.

The first method is to integrate M_x out of the likelihood function, with respect to the prior $\pi(M_x | P_x)$. The resulting <u>Integrated Likelihood</u>

$$IL(y | P_x ; \theta_1) = \int L(y | P_x ; M_x , \theta_1) \pi(M_x | P_x) dM_x \qquad (5.5.3)$$

is then maximized with respect to θ_1. We call this method Maximum Integrated Likelihood (M.I.L.). For a discussion of Integrated Likelihood and other techniques for removing incidental parameters, see Kalbfleisch and Sprott (1970).

This method has certain similarities with the maximum likelihood procedures of §5.2 - §5.4. The estimate of θ_1 found by (5.2.3) is found by maximizing with respect to θ_1

$$Il_A(y|P_x;\theta_1) = \int \ell(y|X';\theta_1)\, p(M_x|P_x,y;\theta_{1A},\theta_{2A})\,dM_x\,, \qquad (5.5.4)$$

where $p(M_x|P_x,y;\theta_{1A},\theta_{2A})$ is the distribution of the missing variables given the data, at $\theta_1=\theta_{1A}, \theta_2=\theta_{2A}$. We contrast the two procedures represented by (5.5.3) and (5.5.4):

(1) Considered as functions , $\ln L(y|P_x;M_x,\theta_1) = \ell(y|X';\theta_1)$.

(2) M.I.L. is non iterative, given the prior for whilst the maximization of Il_A is part of the iterative procedure for finding the fixed point of the transformation ϕ. Thus the parameters θ_2 do not appear in (5.5.3).

(3) The set M_x are considered logically as parameters with a prior distribution in M.I.L. They are random variables in the M.I.P. procedure.

The prior distribution $\pi(M_x|P_x)$, together with the model and the data y, induce a posterior distribution for the missing variables:

$$p(M_x|P_x, y; \theta_{1A}) \propto \pi(M_x|P_x)\, p(y|P_x; M_x, \theta_{1A}).$$

If we integrate M_x out of the log-likelihood $\ell(y|X'; \theta_1)$ with respect to this posterior distribution, we are left with a function of θ_1 which is formally the same as $I\ell_A(y|P_x;\theta_1)$ in (5.5.4). Then define the transformed value of θ_{1A} to be the value of θ_1 which maximizes this function, and find a fixed point of this transformation. The resulting method is a "generalization" of the M.I.P. procedure, with the distribution $p(M_x|P_x;\theta_2)$ replaced by the prior $\pi(M_x|P_x)$. We call this Integrated Pseudo-Maximum Likelihood (I.P.M.L.), and this is our second method of removing M_x.

Choosing the prior.

The flexibility of M.I.L. or I.P.M.L. comes from allowing extraneous information to affect the choice of prior, but first we consider a prior distribution formed from the data P_x alone. Suppose we assume a multivariate normal distribution for M_x, and estimate the means and covariances by Iterated Buck, applied to the independent data only. I.P.M.L. with this prior is easily seen to be the same as Iterated Buck on the whole data P, the fixed point equations being solved in a different way. Logically speaking the prior distribution for M_x then has a frequency interpretation, and I.P.M.L. becomes a true maximum likelihood method, under a suitable model.

M.I.L. with this prior produces the following method:

(a) Apply Iterated Buck to the independent data, and fit for M_x the final fitted values found by this procedure.

(b) Analyse the completed data, without adjustments, by Iterative Weighted Least Squares, with weights

$$w_i = \frac{\hat{\sigma}_Y^2}{\hat{\sigma}_{Y \cdot P_{X_i}}^2} , \qquad (5.5.5)$$

where $\hat{\sigma}_Y^2$ is the estimated residual variance when Y is fitted as a linear combination of $X_1, X_2, \ldots, X_{p-1}$, and $\hat{\sigma}_{Y \cdot P_{X_i}}^2$ is the estimated residual variance when Y is fitted as a linear combination of the independent variables present in the ith observation. These estimates are found iteratively by pivoting on the current weighted S.S.C.P. matrix. Notice $w_i = 1$ for complete observations, $w_i \leq 1$ for incomplete observations. Observations with Y missing are given weight zero, although they can be used in (a).

This is quite a sensible method, first suggested by E.M.L. Beale as a straight modification of ordinary least squares when data are incomplete. It suffers from not using the Y-variable when fitting missing values, but it should be robust against non-normality of the X's, and by excluding the dependent data during the fitting process, residuals can be calculated in the usual way to test the model. A comparison with Iterated Buck by simulation, and an improvement of the method, is given in Chapter 7.

Subjective Priors.

No practical work has been done with subjective priors in this thesis. We discuss briefly two examples.

Example 5.5.1. Suspect x-variables.

Suppose in a complete set of data we have a value ξ_{ik} of a variable which we have reason to suspect as inaccurate or misrecorded. We replace ξ_{ik} by a normal distribution, centred at ξ_{ik}, with a variance σ_ξ^2 chosen a priori. With this prior for x_{ik}, we can apply M.I.L. or I.P.M.L. As in the previous example, M.I.L. gives a weighted least squares analysis. The uncertainty about x_{ik} produces a weight,

$$w_i = \frac{\sigma_y^2}{\sigma_y^2 + \hat{\beta}_k^2 \sigma_\xi^2} \tag{5.5.6}$$

for the ith observation , compared with weights of 1 for the other observations. Since $\hat{\beta}_k$ is the estimated regression coefficient of X_k, the analysis is iterative. Note that w_i decreases as σ_ξ^2 or $\hat{\beta}_k^2$ increases. This method would seem to have possibilities as a robust regression technique, although the problem lies more in the detection of suspect variables than in the analysis.

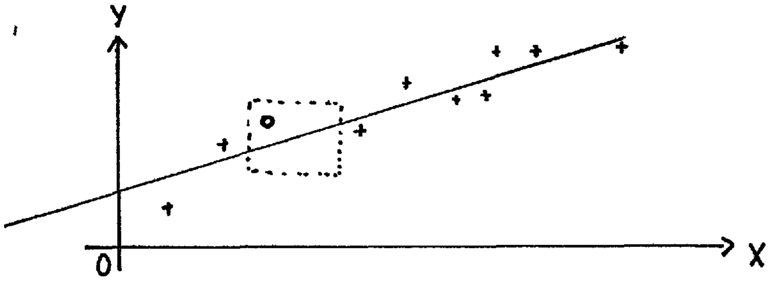
With the normal prior the method is in a sense a generalization of the model proposed by Berkson (1950) when the x's have "target values". Other priors may be appropriate, for example skew distributions to deal with punching errors.

In I.P.M.L. the fitted value for x_{ik} is changed from ξ_{ik} to

$$\hat{x}_{ikA} = E(x_{ik} \mid P_i ; \theta_{iA}) = \xi_{ik} + \frac{\beta_{kA}\,\sigma_\xi^2\,R_{iA}}{\sigma_{YA}^2 + \beta_{kA}^2\,\sigma_\xi^2} , \tag{5.5.7}$$

where $$R_{iA} = y_i - \sum_{\substack{j=0 \\ j\neq k}}^{p-1} \beta_{jA}\,x_{ij} - \beta_{kA}\,\xi_{ik} .$$

Also an adjustment $\dfrac{\sigma_{YA}^2\,\sigma_\xi^2}{\sigma_{YA}^2 + \beta_{kA}^2\,\sigma_\xi^2}$ is added to the (k,k)th element of the S.S.C.P. matrix of the data. Equation (5.5.7) represents a shift of the quantity ξ_{ik} towards the regression line. This is illustrated for two variables in Figures 2 and 3.



o = suspect reading
+ = other data.

Figure 2 . Plot of data on two variables with a suspect reading.

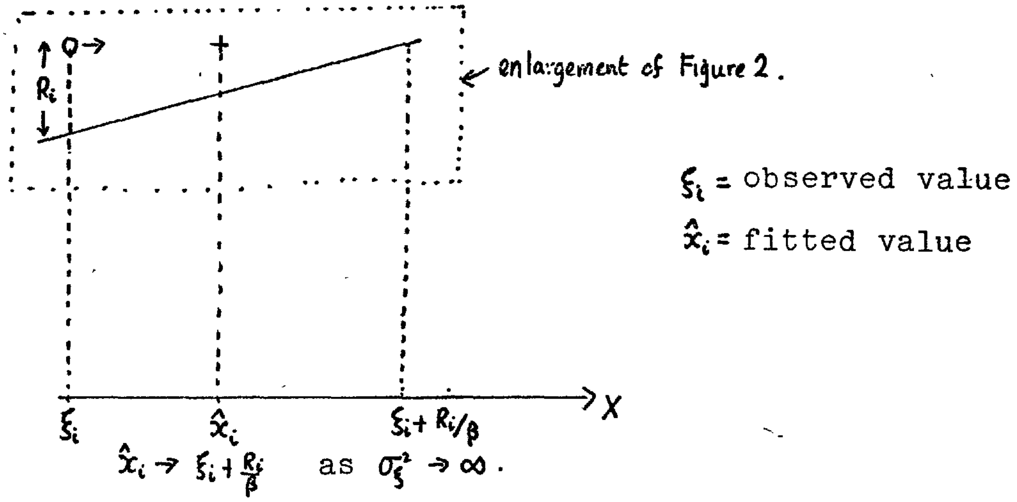

Q→
R_i
enlargement of Figure 2.
ξ_i = observed value
$\hat{x}_i$ = fitted value
X
ξ_i
$\hat{x}_i$
$\xi_i + R_i/\beta$
$\hat{x}_i \to \xi_i + \frac{R_i}{\beta}$ as $\sigma_\xi^2 \to \infty$.

Figure 3. Observed and fitted value of a suspect variable, for I.P.M.L.

A comparison between these methods would be interesting. The equations are easy to derive, and readily extend to problems with more than one suspect variable.

Example 5.5.2. A uniform prior on a missing value.

Suppose we attempt to express indifference about the value of a continuous variable which is missing by an (improper) uniform prior: $\pi(x_{ik} \mid P_x) \propto$ const.

For the normal errors model the posterior distribution of $(x_{ik} \mid P_x, y_i)$ is then normal. One might expect that the result of applying I.P.M.L. with this prior would be to reject the observation i, but this is not so. A simple calculation shows that the effect of the ith observation is to shrink the regression coefficient $\hat{\beta}_{kc}$, estimated from the complete data. This results from allowing for the possibility of x_{ik} taking extreme values in comparison with the rest of the data. Thus the uniform prior is not an expression of indifference about the value of x_{ik}.

If x_{ik} is binary, we might set a priori

$$pr(x_{ik}=1) = 1 - pr(x_{ik}=0) = \tfrac{1}{2} .$$

The analysis then follows Example 5.3.1.

5.6. The Multivariate Linear Model.

So far we have considered linear regression with one dependent variable. We consider briefly how to analyse a multivariate linear model, when r variables, say $X_1, X_2, \ldots, X_r$, are independent, and $p-r$ variables $X_{r+1}, X_{r+2}, \ldots, X_p$ are dependent, for some $r<p$. We adopt the obvious generalization of the notation of §5.1. Thus $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y)$, where $\boldsymbol{\beta}$ is the matrix of regression coefficients, and $\boldsymbol{\Sigma}_Y$ the residual covariance matrix. We adopt the usual assumption about the vector of errors, that of a zero-centred multivariate normal distribution. With a random pattern of deletions, $\boldsymbol{\theta}_1$ can be estimated by applying Iterated Buck, and then pivoting on the estimate of the covariance matrix:

$$\mathrm{PIV}(1,2,\ldots,r)\,\hat{\Sigma} = \begin{bmatrix} V_{(r)}^{-1} & \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\beta}} & \hat{\boldsymbol{\Sigma}}_Y \end{bmatrix} .$$

The conditions under which $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_Y)$ are (corrected) maximum likelihood estimates parallel those of §5.2. We simply fix the variables in the set $(X_1, X_2, \ldots, X_r)$ which are always observed, and then make the obvious assumptions of normality, independence, and random deletion for the remaining variables.

McDonald (1971) considers the extreme case where all the independent variables are observed, and all the dependent variables are missing from the incomplete observations. This generalizes the problem of missing design points to a multivariate response. By fitting missing values by treating dependent variables separately, and using standard missing value techniques, McDonald finds best linear unbiased estimates of $\boldsymbol{\beta}$, in the sense of minimising the trace of the residual S.S.C.P. matrix.

As in the univariate response case, Iterated Buck produces best linear unbiased estimates of $\boldsymbol{\beta}$ for this pattern of missing data. When the dependent variables in some observations are partly observed, the best linear unbiasedness criterion ceases to mean much. In this more general situation Iterated Buck is justified by Maximum Likelihood, or the asymptotic unbiasedness considerations of Chapter 3.

6. ASYMPTOTIC COVARIANCE MATRIX OF THE ESTIMATES FROM A LINEAR REGRESSION.

6.1. Introduction.

In Chapter 4 we considered the asymptotic covariance matrix $J_p^{-1}(\theta,\theta)$, where θ represents the means and covariances of the $MN_p(\mu,\Sigma)$ distribution. In Chapter 5 we estimated the linear regression of X_p on $X_1, X_2, \ldots, X_{p-1}$. This involved estimating the alternative set of parameters

$$\psi = (\mu_1, \mu_2, \ldots, \mu_p, \sigma_{11}, \sigma_{12}, \sigma_{22}, \ldots, \sigma_{p-1,p-1}, \beta_1, \beta_2, \ldots, \beta_{p-1}, \sigma_Y^2), \tag{6.1.2}$$

where the β_j's are regression coefficients and σ_Y^2 is the residual variance. We now find the expected information matrix $J_p(\psi,\psi)$ corresponding to this parametrization, in order to estimate the precision of our estimates of the regression coefficients.

One way of doing this is to transform the matrix $J_p(\theta,\theta)$ using the Jacobian of the transformation from θ to ψ. It is less arduous, however, to simply adopt the same approach as Chapter 4, that is to work out $J_{p,M}(\psi,\psi)$ and subtract the Lost Information. The final expression $J_p(\psi,\psi)$ is rewritten in terms of the weights which appear in the Weighted Least Squares procedure outlined in §5.5 (Equation(5.5.5)). This indicates a way of finding an approximate covariance matrix for estimates of the regression coefficients, by assigning a weight to each observation and forming a weighted S.S.C.P. matrix. This is simpler than inverting the whole matrix $J_p(\psi, \psi)$. The two methods are compared for the two variable problem, and in less detail in the general case.

6.2. Calculation of the Expected Information Matrix.

We express the log-likelihood of a complete set of data as

$$\ell(X;\psi) = -\tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{p-1}\sum_{k=1}^{p-1}(x_{ij}-\mu_j)\sigma_x^{jk}(x_{ik}-\mu_k) - \tfrac{1}{2}N \ln \det \Sigma_x - \sum_{i=1}^{N}\frac{\epsilon_i^2}{2\sigma_Y^2} - \tfrac{1}{2}N \ln \sigma_Y^2 \quad , \qquad (6.2.1)$$

where $\epsilon_i = y_i - \sum_{j=0}^{p-1}\beta_j x_{ij} = x_{ip} - \mu_p - \sum_{j=1}^{p-1}\beta_j(x_{ij}-\mu_j)$,

Σ_x is the covariance matrix of $(X_1, X_2, \dots, X_{p-1})$, and σ_x^{jk} is the (j,k)th element of the inverse of Σ_x. Notice that μ_p rather than β_0 is included in the parametrization (6.1.2). This allows us to calculate the Lost Information $J_{M|P}(\psi, \psi)$ by applying Lemma (4.3) to the set of variables

$$(X_1 - \mu_1, X_2 - \mu_2, \dots, X_{p-1} - \mu_{p-1}, \epsilon),$$

which have zero mean. The scores are

$$\left.\begin{aligned}
\frac{\partial \ell}{\partial \mu_j} &= \sum_{i=1}^{N}\Big[\sum_{r=1}^{p-1}(x_{ir}-\mu_r)\sigma_x^{jr} - \frac{\beta_j\epsilon_i}{\sigma_Y^2}\Big] \qquad (j=1,2,\dots,p-1)\\
\frac{\partial \ell}{\partial \mu_p} &= \sum_{i=1}^{N}\frac{\epsilon_i}{\sigma_Y^2}\\
\frac{\partial \ell}{\partial \sigma_{\ell m}} &= -\tfrac{1}{2}N(2-\delta_{\ell m})\Big[\sigma_x^{\ell m} - \frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{p-1}\sum_{s=1}^{p-1}\sigma_x^{\ell r}\sigma_x^{ms}(x_{ir}-\mu_r)(x_{is}-\mu_s)\Big]\\
\frac{\partial \ell}{\partial \beta_k} &= \sum_{i=1}^{N}\frac{\epsilon_i}{\sigma_Y^2}(x_{ik}-\mu_k) \qquad (k=1,2,\dots,p-1)\\
\frac{\partial \ell}{\partial \sigma_Y^2} &= -\tfrac{1}{2}N\Big[\frac{1}{\sigma_Y^2} - \frac{1}{N\sigma_Y^4}\sum_{i=1}^{N}\epsilon_i^2\Big]
\end{aligned}\right\} \qquad (6.2.2)$$

Forming covariances of the elements of (6.2.2), we find

$$J_{P,M}(\psi,\psi) = \begin{array}{c|ccccc|} & (\mu_j) & (\mu_p) & (\sigma_{lm}) & (\beta_j) & (\sigma_Y^2) \\ (\mu_k) & N(\sigma_x^{jk}+\frac{1}{\sigma_Y^2}\beta_j\beta_k) & -N\frac{\beta_k}{\sigma_Y^2} & 0 & 0 & 0 \\ (\mu_p) & -N\frac{\beta_j}{\sigma_Y^2} & \frac{N}{\sigma_Y^2} & 0 & 0 & 0 \\ (\sigma_{rs}) & 0 & 0 & \frac{N}{4}(2-\delta_{lm})(2-\delta_{rs})[\sigma_x^{lr}\sigma_x^{ms}+\sigma_x^{ls}\sigma_x^{mr}] & 0 & 0 \\ (\beta_k) & 0 & 0 & 0 & \frac{N}{\sigma_Y^2}\sigma_{jk} & 0 \\ (\sigma_Y^2) & 0 & 0 & 0 & 0 & \frac{N}{2\sigma_Y^4} \end{array}$$

(6.2.3)

The right hand side of (6.2.3) is a condensed form of the matrix, with a similar notation to (4.3.4). We write $\boldsymbol{\beta} = (\beta_1,\beta_2,\dots,\beta_p)^T$; then inversion of the submatrix corresponding to $\boldsymbol{\beta}$ gives

$$Cov(\hat{\boldsymbol{\beta}}) \overset{asympt}{=} \frac{1}{N}\boldsymbol{\Sigma}_x^{-1}\sigma_Y^2 ,$$

which is the analogue of the standard least squares estimate for complete data

$$Cov(\hat{\boldsymbol{\beta}}) = \mathbf{S}_x^{-1}\sigma_Y^2 ,$$

where $\mathbf{S}_x$ is the S.S.C.P. matrix of the independent variables, held fixed.

The Lost Information terms are found as in Chapter 4; the details are omitted. Suffices run from 1,2,...,p-1, except where stated:

$$J_{MIP}(\mu_j,\mu_k) = \sum_{i=1}^{N}\Big[\sum_{\ell=1}^{p-1}\sum_{m=1}^{p-1}\sigma_x^{j\ell}\sigma_x^{km}\sigma_{\ell m\cdot p_i} - \frac{\beta_j}{\sigma_Y^2}\sum_{m=1}^{p-1}\sigma_x^{jm}\sigma_{m\epsilon\cdot p_i} - \frac{\beta_k}{\sigma_Y^2}\sum_{\ell=1}^{p-1}\sigma_x^{k\ell}\sigma_{\ell\epsilon\cdot p_i} + \frac{\beta_j\beta_k}{\sigma_Y^4}\sigma_{\epsilon\epsilon\cdot p_i}\Big],$$

$$J_{MIP}(\mu_j,\mu_p) = \sum_{i=1}^{N}\Big[\frac{1}{\sigma_Y^2}\sum_{\ell=1}^{p-1}\sigma_x^{j\ell}\sigma_{\ell\epsilon\cdot p_i} - \frac{\beta_j}{\sigma_Y^4}\sigma_{\epsilon\epsilon\cdot p_i}\Big],$$

$$J_{MIP}(\mu_p,\mu_p) = \sum_{i=1}^{N}\frac{\sigma_{\epsilon\epsilon\cdot p_i}}{\sigma_Y^4},$$

$$J_{MIP}(\mu_j,\sigma_{\ell m}) = J_{MIP}(\mu_j,\beta_k) = J_{MIP}(\mu_j,\sigma_Y^2) = 0 \quad, (j=1,2,\ldots,p),$$

$$J_{MIP}(\sigma_{\ell m},\sigma_{rs}) = \frac{1}{4}(2-\delta_{\ell m})(2-\delta_{rs})\sum_{i=1}^{N}\sum_{t=1}^{p-1}\sum_{u=1}^{p-1}\sum_{v=1}^{p-1}\sum_{w=1}^{p-1}\sigma_x^{\ell t}\sigma_x^{mu}\sigma_x^{rv}\sigma_x^{sw}\gamma_{ituvw},$$

where γ_{ituvw} is the same as in Chapter 4;

$$J_{MIP}(\beta_k,\sigma_{\ell m}) = (2-\delta_{\ell m})\frac{1}{2\sigma_Y^2}\sum_{i=1}^{N}\Big\{\sum_{u=1}^{p-1}\sum_{v=1}^{p-1}\sigma_x^{\ell u}\sigma_x^{mv}\big[\sigma_{u\epsilon\cdot p_i}(\sigma_{kv}-\sigma_{kv\cdot p_i})+\sigma_{v\epsilon\cdot p_i}(\sigma_{ku}-\sigma_{ku\cdot p_i})\big]\Big\},$$

$$J_{MIP}(\sigma_Y^2,\sigma_{\ell m}) = (2-\delta_{\ell m})\frac{1}{2\sigma_Y^4}\sum_{i=1}^{N}\Big[\sum_{u=1}^{p-1}\sum_{v=1}^{p-1}\sigma_x^{\ell u}\sigma_x^{mv}\sigma_{u\epsilon\cdot p_i}\sigma_{v\epsilon\cdot p_i}\Big],$$

$$J_{MIP}(\beta_j,\beta_k) = \frac{1}{\sigma_Y^2}\sum_{i=1}^{N}\big[(\sigma_{jk}-\sigma_{jk\cdot p_i})\sigma_{\epsilon\epsilon\cdot p_i} + \sigma_Y^2\,\sigma_{jk\cdot p_i} - \sigma_{j\epsilon\cdot p_i}\sigma_{k\epsilon\cdot p_i}\big],$$

$$J_{MIP}(\beta_k,\sigma_Y^2) = \frac{1}{\sigma_Y^2}\sum_{i=1}^{N}\sigma_{k\epsilon\cdot p_i}(\sigma_Y^2-\sigma_{\epsilon\epsilon\cdot p_i}),$$

$$J_{MIP}(\sigma_Y^2,\sigma_Y^2) = \frac{1}{2\sigma_Y^8}\sum_{i=1}^{N}\sigma_{\epsilon\epsilon\cdot p_i}^2 .$$

In these expressions, $\sigma_{j\epsilon\cdot P_i} = Cov(x_{ij}, \epsilon_i \mid P_i)$ and $\sigma_{\epsilon\epsilon\cdot P_i} = Var(\epsilon_i \mid P_i)$.

Now if P_{x_i} is the set of independent variables present in observation i, we have

$$Cov(x_{ij}, \epsilon_i \mid P_{x_i}) = 0 \ , \quad (j=1,2,\ldots,p-1), \text{ and } \quad Var(\epsilon_i \mid P_{x_i}) = \sigma_Y^2 .$$

Hence

$$Var(\epsilon_i \mid P_i) = Var(\epsilon_i \mid P_{x_i}) - \frac{[Cov(x_{ip}, \epsilon_i \mid P_{x_i})]^2}{Var(x_{ip} \mid P_{x_i})} \ ,$$

i.e. $$\sigma_{\epsilon\epsilon\cdot P_i} = (1 - w_i)\sigma_Y^2 \ , \tag{6.2.4}$$

and $$Cov(x_{ij}, \epsilon_i \mid P_i) = Cov(x_{ij}, \epsilon_i \mid P_{x_i}) - \frac{Cov(x_{ij}, x_{ip} \mid P_{x_i})\, Cov(x_{ip}, \epsilon_i \mid P_{x_i})}{Var(x_{ip} \mid P_{x_i})} \ ,$$

i.e. $$\sigma_{j\epsilon\cdot P_i} = - w_i \, \sigma_{jp\cdot P_{x_i}} \ , \tag{6.2.5}$$

where $$w_i = \begin{cases} \dfrac{\sigma_Y^2}{Var(x_{ip} \mid P_{x_i})} & \text{, if } x_{ip} \text{ is present ;} \\ 0 & \text{, if } x_{ip} \text{ is missing .} \end{cases} \tag{6.2.6}$$

The definition of w_i is that given in §5.5, and may be considered intuitively as the weight of observation i. Subtracting the Lost Information matrix form (6.2.3), and substituting (6.2.4) - (6.2.6), we find

$$J_P(\psi, \psi) = \begin{bmatrix} \sum_{i=1}^{N} J_{P_i}(\mu, \mu) & 0 \\ 0 & \sum_{i=1}^{N} J_{P_i}(\sigma, \sigma) \end{bmatrix} \ , \tag{6.2.7}$$

where $\sigma = (\sigma_{11}, \sigma_{12}, \sigma_{22}, \ldots, \sigma_{p-1,p-1}, \beta_1, \beta_2, \ldots, \beta_{p-1}, \sigma_Y^2)^T$,

$$J_{P_i}(\boldsymbol{\mu},\boldsymbol{\mu}) = \begin{array}{c} \\ (\mu_k) \\ (\mu_p) \end{array} \begin{array}{cc} (\mu_j) & (\mu_p) \\ \left[\begin{array}{c} \phi_{jk\cdot P_i} + \frac{w_i}{\sigma_Y^2}\left(\beta_j\beta_k - \beta_j a_{kp\cdot P_{x_i}} - \beta_k a_{jp\cdot P_{x_i}}\right) \\ -\frac{w_i}{\sigma_Y^2}\left(\beta_j - a_{jp\cdot P_{x_i}}\right) \end{array} \right. & \left. \begin{array}{c} -\frac{w_i}{\sigma_Y^2}\left(\beta_k - a_{kp\cdot P_{x_i}}\right) \\ \frac{w_i}{\sigma_Y^2} \end{array} \right] \end{array} ,$$

(6.2.8)

$$J_{P_i}(\boldsymbol{\sigma},\boldsymbol{\sigma}) =$$

$$\begin{bmatrix} \frac{1}{4}(2-\delta_{\ell m})(2-\delta_{rs})(\phi_{\ell r\cdot P_i}\phi_{ms\cdot P_i} + \phi_{\ell s\cdot P_i}\phi_{mr\cdot P_i}) & (2-\delta_{rs})\frac{w_i}{2\sigma_Y^2}(a_{rp\cdot P_{x_i}}b_{sj\cdot P_i} + a_{sp\cdot P_{x_i}}b_{rj\cdot P_i}) & \frac{w_i^2}{2\sigma_Y^4}a_{rp\cdot P_{x_i}}a_{sp\cdot P_{x_i}} \\ (2-\delta_{\ell m})\frac{w_i}{2\sigma_Y^2}(a_{\ell p\cdot P_{x_i}}b_{mk\cdot P_i} + a_{mp\cdot P_{x_i}}b_{\ell k\cdot P_i}) & \frac{w_i}{\sigma_Y^4}(\sigma_{jk} + \sigma_{jk\cdot P_{x_i}} - 2\sigma_{jk\cdot P_i}) & \frac{w_i^2}{\sigma_Y^4}\sigma_{kp\cdot P_{x_i}} \\ \frac{w_i^2}{2\sigma_Y^4}\, a_{\ell p\cdot P_{x_i}}\, a_{mp\cdot P_{x_i}} & \frac{w_i^2}{\sigma_Y^4}\,\sigma_{jp\cdot P_{x_i}} & \frac{w_i^2}{2\sigma_Y^4} \end{bmatrix} .$$

(6.2.9)

In these matrices

$$\phi_{jk\cdot P_i} = \sum_{r=1}^{p-1}\sum_{s=1}^{p-1}\sigma_x^{jr}\sigma_x^{ks}(\sigma_{rs} - \sigma_{rs\cdot P_i}) ,$$ (6.2.10)

$$a_{jp\cdot P_{x_i}} = \sum_{\ell=1}^{p-1}\sigma_x^{j\ell}\sigma_{\ell p\cdot P_{x_i}} ,$$ (6.2.11)

$$b_{jk\cdot P_i} = \sum_{\ell=1}^{p-1}\sigma_x^{j\ell}(\sigma_{\ell k} - \sigma_{\ell k\cdot P_i}) ;$$ (6.2.12)

Properties of $J_p(\psi,\psi)$

First compare (6.2.7) with (4.3.8), the expression for $J_p(\theta,\theta)$. The matrix $J_p(\psi,\psi)$ decomposes into two submatrices $J_p(\mu,\mu)$ and $J_p(\sigma,\sigma)$, corresponding to the means and the other parameters. Equation (6.2.8), summed over the observations, is thus another expression for $J_p(\mu,\mu)$. Since $w_i=0$ when x_{ip} is missing, comparison of the two expressions shows that $\psi_{jp\cdot p_i}=0$ if x_{ip} is missing. This proves the assertion (4.3.10) of Chapter 4.

Now let us turn to $J_p(\sigma,\sigma)$, since in this Chapter we are more interested in the regression coefficients and residual variance. When x_{ip} is missing, $w_i=0$ and all the elements of $J_{p_i}(\sigma,\sigma)$ vanish except those corresponding to the covariance matrix of the independent variables. Such observations are useful only in that they improve the estimate of Σ_x, as one might expect.

An unfortunate aspect of (6.2.9) is that the cross terms $J_{p_i}(\beta_j,\sigma_{rs})$, $J_{p_i}(\sigma_Y^2,\sigma_{rs})$ and $J_{p_i}(\beta_j,\sigma_Y^2)$, which vanish when all the variables are measured, or the dependent variable is not measured, do not in general vanish for intermediate cases. Therefore we are not justified in simply inverting the submatrix $J_p(\beta,\beta)$ to obtain the asymptotic covariance matrix of $\hat{\beta}$. Instead we must invert the complete matrix $J_p(\sigma,\sigma)$. We now do this analytically for the simple case of one regressor variable, and consider an approximation for the general case.

Example 6.2.1.

Suppose we have two variables $X_1, X_2 = Y$, and the data of the example of §4.4. Then

$$J_p(\sigma,\sigma) =$$

$$n_c \begin{bmatrix} \frac{1}{2\sigma_{11}^2} & 0 & 0 \\ 0 & \frac{\sigma_{11}}{\sigma_Y^2} & 0 \\ 0 & 0 & \frac{1}{2\sigma_Y^4} \end{bmatrix} + n_1 \begin{bmatrix} \frac{1}{2\sigma_{11}^2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + n_2 \begin{bmatrix} \frac{\rho^4}{2\sigma_{11}^2} & \frac{\rho^2\beta_1}{\sigma_{22}} & \frac{\beta_1^2}{2\sigma_{22}^2} \\ \frac{\rho^2\beta_1}{\sigma_{22}} & \frac{2\rho^2\sigma_{11}}{\sigma_{22}} & \frac{\beta_1\sigma_{11}}{\sigma_{22}} \\ \frac{\beta_1^2}{2\sigma_{22}^2} & \frac{\beta_1\sigma_{11}}{\sigma_{22}} & \frac{1}{2\sigma_{22}^2} \end{bmatrix},$$

where $\rho = Corr(X_1, X_2)$, $\beta_1^2 = \rho^2 \frac{\sigma_{22}}{\sigma_{11}}$, $\sigma_Y^2 = (1-\rho^2)\sigma_{22}$.

Adding and Inverting,

$$J_p^{-1}(\sigma,\sigma) = \begin{bmatrix} 2\frac{\sigma_{11}^2}{NK_2}\left[\lambda_c + \lambda_2(1-\rho^4)\right] & -\frac{2\sigma_{11}\beta_1\lambda_2\rho^2(1-\rho^2)}{NK_2} & -\frac{2\sigma_Y^2\sigma_{11}\lambda_2\rho^2(1-\rho^2)}{NK_2} \\ -\frac{2\sigma_{11}\beta_1\lambda_2\rho^2(1-\rho^2)}{NK_2} & \frac{\sigma_Y^2}{N\lambda_c\sigma_{11}}\left[1 - \frac{2\lambda_2\rho^2(1-\rho^2)(\lambda_c+\lambda_1)}{K_2}\right] & -\frac{2\sigma_Y^2\beta_1\lambda_2(\lambda_c+\lambda_1)(1-\rho^2)^2}{N\lambda_c K_2} \\ -\frac{2\sigma_Y^2\sigma_{11}\rho^2(1-\rho^2)}{NK_2} & -\frac{2\sigma_Y^2\beta_1\lambda_2(\lambda_c+\lambda_1)(1-\rho^2)^2}{N\lambda_c K_2} & \frac{2\sigma_Y^4}{N\lambda_c}\left[1 - \frac{\lambda_2(\lambda_c+\lambda_1)(1-\rho^2)^2}{K_2}\right] \end{bmatrix},$$

(6.2.13)

where $K_2 = (\lambda_c+\lambda_1)(\lambda_c+\lambda_2) - \lambda_1\lambda_2\rho^4$, as before. Note the asymptotic variance of $\hat{\sigma}_{11}$ agrees with the expression in (4.4.4). The asymptotic variance of $\hat{\beta}_1$ is

$$Var\,\hat{\beta}_1 = \frac{\sigma_y^2}{N\lambda_c\sigma_{11}}\left[1 - \frac{2\lambda_2\rho^2(1-\rho^2)(\lambda_c+\lambda_1)}{K_2}\right] . \tag{6.2.14}$$

It is interesting to compare this with $\hat{\beta}_{1c}$, the least squares estimate of β_1 based on the complete observations. Asymptotically,

$$Var\,\hat{\beta}_{1c} = \frac{\sigma_y^2}{N\lambda_c\sigma_{11}} . \tag{6.2.15}$$

Thus incomplete observations are most useful for values of ρ^2 such that $\rho^2(1-\rho^2)$ is high, that is for values in the middle of its range. Also

$$Var\,\hat{\beta}_1 \to Var\,\hat{\beta}_{1c} \quad \text{as} \quad \rho \to 0 \quad \text{or} \quad \rho \to 1 .$$

If $n_2 = 0$, $Var\,\hat{\beta}_1 = Var\,\hat{\beta}_{1c}$, and then observations on X_1 alone do not contribute information to $\hat{\beta}_1$ in fact $\hat{\beta}_1 = \hat{\beta}_{1c}$. Observations on X_1 alone increase the value of observations on X_2 alone, by the factor $\frac{\lambda_c+\lambda_1}{(\lambda_c+\lambda_1)(\lambda_c+\lambda_2)-\lambda_1\lambda_2\rho^4}$, found by setting $\lambda_1 = 0$ in (6.2.14). This improvement is greatest when ρ^2 is high, and results from the increased precision of the estimate of σ_{11} .

If σ_{11} is known, the asymptotic variance of $\hat{\beta}_1$ is

$$Var\ \hat{\beta}_1' = \frac{\sigma_y^2}{N\lambda_c\sigma_{11}}\left[1 - \frac{2\lambda_2\rho^2(1-\rho^2)}{\lambda_c+\lambda_2(1-\rho^4)}\right] . \tag{6.2.16}$$

This is found by inverting the (2 x 2) submatrix of $J_p(\boldsymbol{\sigma},\boldsymbol{\sigma})$ corresponding to β_1 and σ_y^2. Naturally this expression does not involve λ_1, since observations on X_1 have no value.

6.3. A Simple Approximate Solution.

For more than two variables, explicit inversion of $J_p(\boldsymbol{\sigma},\boldsymbol{\sigma})$ for a general pattern of missing values is impractical. For a given set of data we can find $J_p(\boldsymbol{\sigma},\boldsymbol{\sigma})$ and then invert, but for large numbers of variables this may involve a lot of computing. Beale and Little (1973) suggest an approximate method of estimating $Cov\,\hat{\boldsymbol{\beta}}$ which involves less calculation. We now describe this method.

For a complete set of data, we have by least squares theory

$$Cov\,\hat{\boldsymbol{\beta}} = \mathbf{S}_x^{-1}\sigma_y^2 , \tag{6.3.1}$$

where $\mathbf{S}_x$ is the (p-1) x (p-1) matrix with (j,k)th element

$$\sum_{i=1}^{N}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k) .$$

For incomplete data we replace $\mathbf{S}_x$ in (6.3.1) by $\mathbf{S}_w$, with (j,k)th element

$$\sum_{i=1}^{N}\hat{w}_i(\hat{x}_{ij}-\hat{\bar{x}}_j)(\hat{x}_{ik}-\hat{\bar{x}}_k) , \tag{6.3.2}$$

where $\hat{x}_{ij}$ is the final observed or fitted value in Iterated Buck, and $\hat{\omega}_i$ is an estimate of the weight ω_i of equation (6.2.6), found by pivoting on $\hat{\Sigma}$. Observations with Y missing are given weight zero. These weights also appeared in the weighted least squares method of §5.5, and although the fitted values of (6.3.2) are different from the fitted values of that method, this approximate way of estimating $\mathrm{Cov}\,\hat{\beta}$ was developed from considering the weighted least squares approach.

First let us apply this approximation to the special case of two variables. Ignoring terms of order we have

$$E(\hat{\omega}_i) = \begin{cases} 1 & , \quad i=1,2,\ldots,n_c \quad ; \\ 0 & , \quad i=n_c+1, n_c+2, \ldots, n_c+n_1 \quad ; \\ 1-\rho^2 & , \quad i=n_c+n_1+1, n_c+n_1+1, \ldots, N , \end{cases}$$

and
$$E(\hat{x}_{i1}-\hat{\bar{x}}_1)^2 = \begin{cases} \sigma_{11} & , \quad i=1,2,\ldots,n_c+n_1 \quad ; \\ \rho^2\sigma_{11} & , \quad i=n_c+n_1+1, n_c+n_1+2, \ldots, N . \end{cases}$$

Hence the approximate method gives

$$\mathrm{Var}\,\hat{\beta}_1 \doteq \frac{\sigma_Y^2}{\sum_{i=1}^{N} E(\hat{\omega}_i)E(\hat{x}_{i1}-\hat{\bar{x}}_1)^2} ,$$

i.e.
$$\mathrm{Var}\,\hat{\beta}_1 \doteq \frac{\sigma_Y^2}{N\sigma_{11}[\lambda_c+\lambda_2\rho^2(1-\rho^2)]} . \qquad (6.3.3)$$

Compare this with the "naive" estimate,

$$\mathrm{Var}\,\hat{\beta}_1 \doteq [J_P(\beta_1,\beta_1)]^{-1} = \frac{\sigma_Y^2}{N\sigma_{11}[\lambda_c+2\lambda_2\rho^2(1-\rho^2)]} .$$

(6.3.4)

The latter is smaller by the factor 2 in the denominator, but it is an underestimate of $\mathrm{Var}\,\hat{\beta}_1$, since no allowance is made in estimating σ_{11} and σ_Y^2, and (6.3.4) is strictly less than the asymptotically exact estimate (6.2.14). Comparison of (6.3.3) and (6.2.14) is less clear, but an encouraging feature is the common factor $\rho^2(1-\rho^2)$ which appears in both expressions. Our approximate method gives no weight to observations on X_1 alone, but for the regression of X_2 on X_1 such observations are of limited value, as noted above. For the special case $\lambda_1=0$, writing $K=\rho^2(1-\rho^2)$, we have with a little rearrangement:

(i) asymptotic exact estimate $V_1 = \frac{\sigma_Y^2}{N\lambda_c\sigma_{11}}(1-2\lambda_2 K)$;

(ii) approximate weights estimate $V_2 = \frac{\sigma_Y^2}{N\lambda_c\sigma_{11}}\left(1-\frac{\lambda_2 K}{\lambda_c+\lambda_2 K}\right)$,

where $\lambda_c+\lambda_2=1$. If $K=0.2$, a fairly representative value, then

$V_1=V_2$ if $\lambda_2=\frac{5}{8}$, $\lambda_c=\frac{3}{8}$,

$V_1<V_2$ if $\lambda_2<\frac{5}{8}$, $\lambda_c>\frac{3}{8}$,

$V_1>V_2$ if $\lambda_2>\frac{5}{8}$, $\lambda_c<\frac{3}{8}$;

Hence the approximate estimate V_2 is conservative unless more than 5/8ths of the data are observations on X_2 alone.

The approximate method seems reasonable in this case, but its practical advantage is when p is large. As before, let V_1 and V_2 be the two estimates in the general case. Then the expected value of the contribution of the ith observation to the denominator of V_2 is to order$\left(\frac{1}{N}\right)$

$$E\left[\frac{\hat{\omega}_i}{\hat{\sigma}_Y^2}\left(\hat{x}_{ij}-\hat{\bar{x}}_j\right)\left(\hat{x}_{ik}-\hat{\bar{x}}_k\right)\right] \doteq \frac{\omega_i}{\sigma_Y^2}\left(\sigma_{jk}-\sigma_{jk\cdot p_i}\right) . \tag{6.3.5}$$

The corresponding contribution to $J_p(\beta_j,\beta_k)$ is

$$J_{p_i}(\beta_j,\beta_k) = \frac{\omega_i}{\sigma_Y^2}\left(\sigma_{jk}+\sigma_{jk\cdot p_{x_i}}-2\sigma_{jk\cdot p_i}\right) , \tag{6.3.6}$$

which is always larger than (6.3.5) for the diagonal terms j=k, since

$$\sigma_{jj\cdot p_{x_i}} - \sigma_{jj\cdot p_i} \geqslant 0 .$$

However as in the two variable case, these terms are reduced when the matrix $J_p(\boldsymbol{\sigma},\boldsymbol{\sigma})$ is inverted, which allows for the estimation of the other parameters. This fact, taken with the similarity of (6.3.5) and (6.3.6), suggests that the approximate weighting method will produce a reasonable estimate of $\mathrm{Cov}\,\hat{\boldsymbol{\beta}}$ in the general case, particularly if the data are not too sparse. This conclusion is supported by the results of the simulation study of Chapter 7.

7. SIMULATION STUDIES, AND A PRACTICAL APPLICATION.

7.1. Introduction.

We report on two simulation studies to compare various methods of estimating the linear regression of Y on $X_1, X_2, \ldots, X_{p-1}$, when the data are incomplete.

In the first study, also reported in Beale and Little (1973), a set of N independent observations from the $MN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution were generated from random numbers, where $\boldsymbol{\mu} = \mathbf{0}$, and $\boldsymbol{\Sigma}$ was one of a set of predetermined covariance matrices, including some matrices considered by Haitovsky (1968) in a similar simulation. Then for each variable X_j a predetermined proportion π_j of the values x_{ij} was deleted, independently for j=1,2,...,p, to produce a set of incomplete data.

The regression coefficients and residual error for the regression

$$E(Y) = \sum_{j=0}^{p-1} \beta_j X_j$$

were calculated by six different methods, and the resulting estimators compared with the estimators found by ordinary least squares on the undeleted sample. Our criterion for judging the effectiveness of each method was the residual sum of squares of deviations of the observed and fitted values of the dependent variable, when the deleted values are restored. That is,

$$S = \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{p-1} \hat{b}_j x_{ij} \right)^2 , \qquad (7.1.1)$$

where $(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{p-1})$ are the constant term and regression coefficients,estimated from the incomplete data by one of the six methods, and (x_{ij}, y_i) are the true values of the variables without deletion.

Clearly a small value of S represents a successful method. We decided to judge a method by the overall success of the regression equation, rather than the closeness of individual regression coefficients to their true values. Therefore we computed the average value of S for each of the six methods, over ten sets of random numbers, for each covariance matrix, and each number of observations and deletion pattern $(\pi_1, \pi_2, \ldots, \pi_p)$. The results are expressed as percentage increases over S_{min}, the minimum possible value of S for each set of data. This minimum value is obtained by ordinary least squares on the undeleted data.

For the problems considered in this study, Iterated Buck is corrected maximum likelihood, and consequently it does well in the comparisons. Of particular interest is how much the general procedure of Iterated Buck loses by ignoring some special structure in a set of non-normal data, as compared with a method which exploits this structure.

This question was considered in the second simulation study, which was based on Example (5.3.3). Data were generated for the model (5.3.12), i.e.

$$\left.\begin{array}{lll} y_i & = \sum_{j=0}^{3} \beta_j x_{ij} + \epsilon_i & ; \; x_{i3} \equiv x_{i1}^2 , \\ (x_{i1}, x_{i2})^T & \overset{iid}{\sim} MN_2(\mu_x, \Sigma_x) \; , & \\ \epsilon_i & \overset{i.i.d}{\sim} N(0, \sigma_Y^2) \quad . & \end{array}\right\} \; . \qquad (7.1.2)$$

The special structure of this problem is that $X_3 = X_1^2$. A certain number of values were deleted from Y, X_1 and X_2, and X_1 and X_3 were deleted together. Then for predetermined choices of the parameters, three methods of estimating the regression equation were compared by the criterion S of (7.1.1). These methods were

(a) to reject the incomplete data, and analyse the complete observations by ordinary least squares;

(b) Iterated Buck;

(c) Maximum Likelihood,as described in §5.3.

The results of this study are discussed in §7.3.

In Chapter 6 we proposed an approximate way of finding the covariance matrix of estimators of the regression coefficients found by Iterated Buck. This procedure is tested in the first simulation study, and found to be a reasonable approximation.

In §7.4 we report briefly on an application of Iterated Buck to a Discriminant Analysis of incomplete medical data, kindly supplied by Dr. C.C. Spicer of the Medical Research Council Computer Unit. The data consisted of 181 observations on 61 variables, and none of the observations were complete. The analysis exploited the formal equivalence of discriminant analysis and linear regression, and a reduced set of variables was found for calculation of the discriminant function.

The general conclusion is that Iterated Buck is a useful technique in a wide class of problems, and that the method works well for non-normal as well as normal populations. However there is scope for further work to find out how much it loses against a more specialized technique, which exploits the special structure of a particular problem. Such a technique can be constructed by solving the fixed point equations for a model which incorporates this structure.

7.2: The First Simulation Study: Methods, Results and Conclusions.

For incomplete data from $MN_p(\mu, \Sigma)$ the following methods were compared:-

Method 1: Ordinary Least Squares on the subset of complete observations.

Method 2: Buck's (1960) Method.

Method 3: Iterated Buck, or Corrected Maximum Likelihood.

Method 4: Ordinary Least Squares on observations with y present, after fitting missing values of the independent variables by Iterated Buck on the independent variables only.

Method 5: Method 4, but with incomplete observations given fractional weights.

Method 6: Method 5, but using the estimate of the covariance matrix from Method 3 to find the fitted values and weights.

Method 1 requires no explanation, and Methods 2 and 3 are described in Chapter 3. For Methods 4 and 5 the missing independent variables are estimated by Iterated Buck on the independent data P_x, prior to a least squares analysis. Method 4 is inefficient, since it amounts to giving incomplete observations the same weight as complete observations. Thus Method 5 computes a weight ω_i for each observation i, and carries out a weighted least squares analysis. This method is M.I.L. with a normal prior, and it was mentioned in §5.5. The weights depend on the estimates of the parameters, and hence require an iterative calculation.

We contrast the fitting procedure of the least squares approaches with that of Methods 2 and 3. In the latter methods, linear combinations of all the variables, P_i, present in the ith observation are fitted, whilst in

Methods 4 and 5 linear combinations of the known independent variables $, P_{x_i},$ are fitted. In fact the conditional mean

$$E(x_{ij} | P_{x_i})$$

is the best fitted value of a missing variable x_{ij} for use in a least squares analysis. The methods of Buck and Iterated Buck fit

$$E(x_{ij} | P_i) ,$$

and then <u>correct</u> the least squares analysis of the completed data for bias, by adding adjustments to the covariance matrix before pivoting. This summarizes the difference in the approaches.

Finally Method 6 is a combination of Methods 3 and 5. An estimate $\hat{\Sigma}$ of the covariance matrix of all the variables, Σ, is found by Method 3. Then missing values for the independent variables are found as in Methods 4 and 5, using the submatrix of $\hat{\Sigma}$ corresponding to the independent variables. Weights w_i are found directly by pivoting on $\hat{\Sigma}$, and a weighted least squares analysis is carried out on the completed observations for which y is observed.

The results of the simulations are given in Table 2, at the end of §7.3. The statistic

$$100\left(\frac{S_M}{S_{min}} - 1\right)\% \tag{7.2.1}$$

is calculated for the methods M=1,2,...6, averaged over 10 sets of data, and for 7 matrices A - G. Notice that the statistic (7.2.1) is the same for Method 1 for matrices C to G. This arises because the data for each of these cases are generated by transforming the same set of uncorrelated data. The statistic $\frac{S_1}{S_{min}}$ is invariant under these transformations, since Method 1 uses only complete observations.

We draw the following conclusions from Table 2. Method 3 always improves on Method 2, except for three very marginal cases with 5% deletions. The improvement is often considerable, for example in problems C and D Method 3 requires more computing than Method 2, but it can be used when there are no complete observations, and is therefore a more general method.

Method 4 is only appreciably better than Method 3 for 2 cases in problem E; otherwise it is usually slightly worse, and much worse on problems F and G, where the multiple correlation coefficient $R^2 > 0.98$. In these problems the method performs badly, because relatively useless observations are given the same weight as complete observations. Thus we do not recommend this method.

Method 5 is an improvement on Method 4, but is generally less effective than Method 3, and is sometimes beaten by Method 1 in problems F and G. In these problems the value of the fitted variables are critical, and a better estimate of the covariance matrix of the independent variables used to fit those values produces a considerably better fit. In Method 6 all the data are used in finding this covariance matrix, and the results are seen to be an improvement on Method 5.

It remains to compare the best of the least squares approaches, Method 6, with Iterated Buck, Method 3. There is not much to choose between the methods, but Method 3 is marginally better in a large majority of the cases considered. From a computing point of view the methods are very similar, and the weighting procedures of Methods 5 and 6 are used to derive approximate standard errors in the regression coefficients for Method 3.

It is perhaps worth noting that we also tested the straight Maximum Likelihood Method M.L.N. of Orchard and Woodbury. The results are almost identical to those of Iterated Buck. Mostly they are worse, but by less than 0.1% We therefore see no reason to use M.L.N. in preference to the conventional correction represented by Iterated Buck.

For Iterated Buck we formed an estimate of the covariance matrix of the estimates of the regression coefficients, by the approximate weighting procedure of

Chapter 6. To test the validity of the approximation, we could have taken each regression coefficient individually and formed an approximate χ^2 variable from the sum of squares of the deviations of the estimated regression coefficients from their true values, each divided by its estimated variance. But it seems preferable to form a single χ^2 variate on $r \times (p-1)$ degrees of freedom, where r is the number of replications, here 10, and $p-1$ is the number of regression coefficients estimated. We do this by forming

$$\frac{1}{\hat{\sigma}_y^2} \sum (\beta - \hat{\beta})^T S_w (\beta - \hat{\beta}) ,$$

where $\beta = (\beta_1, \beta_2, \dots \beta_{p-1})$, S_w is the estimated covariance matrix of $\hat{\beta}$, given by (6.3.2), and the summation extends over all replications.

The results are tabulated in Table 3 as multiples of the corresponding χ^2 statistic obtained from ordinary least squares on the complete data before deletions. Hence values >1 suggest an underestimate and values <1 an overestimate, compared with those found from the complete data. The results suggest that the approximate theory is adequate to give general guidance of $\mathrm{Cov}\,\hat{\beta}$. But we should point out that we have not tested the theory for more systematic deletion patterns. Such systematic patterns of missing data often arise in practice, and may not be quite as well covered by our approximate theory; the asymptotic exact theory of Chapter 6 is still correct for such cases, provided the deletion mechanism is random in the sense explained in Chapter 2.

7.3. The Second Study: Regression with a Quadratic Term.

Three methods were compared in this simulation study. Method 1 and Method 3 are the same as the previous study. Method 7 is corrected maximum likelihood for the model (7.1.2), and a description of the fitted values and adjustments is given in Example 5.3.3 of §5.3. Observations with only one variable present were avoided in the deletion process, since the solution of the fixed point equations involves double numerical integrations when a single variable other than X_1 is observed. In practice these observations would carry little weight, and could with some justification be rejected before applying Method 7.

Numerical integration is required for observations with X_1 missing; the integrals were reparametrized to depend on four distinct parameters and calculated at each iteration by a straight forward "halving" algorithm, with stopping values calculated to cover the distribution of $(X_1|X_2,Y)$, which could be unimodal or bimodal according to the values of X_2, Y and the parameters.

The coefficients of the model were chosen to illuminate differences between the methods. For all the problems,

$$\mu_x = 0 \ , \quad \Sigma_x = \begin{pmatrix} 1 & \rho_x \\ \rho_x & 1 \end{pmatrix} , \quad \beta_0 = 0 , \ \beta = (\beta_1, \beta_2, \beta_3)^T , \ \sigma_Y^2 = 0{\cdot}5 ,$$

and for the chosen values of β and ρ_x, the following correlations were calculated:-

R^2 = multiple correlation coefficient of Y with X_1, X_2 and X_3.

R^2_{12} = multiple correlation coefficient of Y with X_1 and X_2.

R^2_2 = multiple correlation coefficient of Y with X_2.

R^2_{13} = multiple correlation coefficient of Y with X_1 and X_3.

We now compare the results of Table 4, with the help of these correlations.

The results for Method 1 are the same for all the problems. The reason is the same as for the first simulation, i.e. the data for each problem are generated by transforming the same set of random numbers. This also allows for a more direct comparison of the methods between problems. For the second deletion pattern (20%, 20%, 40%) there are only 10 complete observations, which explains the high increase in residual sum of squares for Method 1 for this pattern.

One would expect Method 7 to improve on Method 3 most clearly when the quadratic term is highly significant in the regression equation. One measure of this significance is $R^2 - R^2_{12}$, which is high for problems D and E, low for problems A and intermediate for problems B and C. We see that Method 7 beats Method 3 clearly in problems D and E, whilst in problem A Method 3 is slightly better than Method 7. Here the β_3 term is low, and for the second deletion pattern Method 7 failed in one problem, because the assumed value of β_3 went to zero. In future this difficulty will be avoided by an alternative calculation of the fitted values and adjustments when $|\beta_3|$ falls below a certain tolerance. (In fact if $\beta_{3A} = 0$ the distribution of $(X_1 | Y, X_2)$ is normal, and so the alternative calculation is much simpler). However in this study the starred result is calculated over 19 sets of data, whilst all the other results are calculated over 20 sets of data.

For the intermediate values of $R^2 - R^2_{12}$, problems B and C, Method 7 beats Method 3, but not always by as large an amount as in problems D and E.

The results can be considered from other viewpoints. For example R^2_2 is in a sense a measure of the information in observations with X_1 missing, for Method 3, but it underestimates the information in these observations for Method 7. Thus Method 1 and Method 3 should be similar for the first patterns in problems D and E, where $R^2_2 = 0$. In fact Method 3 is marginally better in the results.

However detailed comparison of the methods is not practical with these results, since the percentage increases in residual sums of squares were very variable between problems, and the size of the study was limited by the large amount of computing required for Method 7. In all runs the iterative process was terminated after 20 iterations; some trial runs at 50 iterations did not affect the results.noticeably.

Overall the results indicate that the Maximum Likelihood method is slightly better than the general method of Iterated Buck when the quadratic term is reasonably significant in the regression equation.

Table 2

Average Percentage Increase in Residual Sum of Squares over best fit when all variables are known. Averaged over 10 runs.

Problem	Method	Percentage deletions from each variable, and number of observations.									
		(5%)	(5%)	(10%)	(10%)	(10%)	(20%)	(20%)	(20%)	(40%)	Av.
		(100)	(200)	(50)	(100)	(200)	(50)	(100)	(200)	(200)	
	1	0.4	0.3	2.7	1.4	0.5	3.9	3.9	1.3	6.4	2.3
A	2	0.2	0.1	2.0	0.8	0.2	2.1	3.0	0.7	3.3	1.4
	3	0.2	0.1	1.9	0.8	0.2	1.9	2.4	0.7	1.9	1.1
3 var.	4	0.3	0.2	2.1	0.9	0.3	2.4	2.9	0.9	2.3	1.4
	5	0.3	0.2	2.0	1.0	0.3	2.4	2.8	0.9	2.2	1.3
R^2 = 0.95	6	0.3	0.2	2.0	1.0	0.3	2.3	2.8	0.8	2.2	1.3
	1	0.9	0.4	4.5	2.5	0.7	8.6	4.7	3.1	30.6	6.2
B	2	0.5	0.2	3.1	0.8	0.4	4.3	1.8	1.5	15.7	3.1
	3	0.6	0.2	3.0	0.8	0.4	3.8	1.4	1.2	3.1	1.6
4 var.	4	0.6	0.2	3.0	0.8	0.4	3.8	1.3	1.2	3.3	1.6
	5	0.6	0.2	3.0	0.8	0.4	3.8	1.4	1.2	3.4	1.6
R^2= 0.09	6	0.6	0.2	3.0	0.8	0.4	3.8	1.4	1.2	3.6	1.7
	1	1.6	0.8	7.7	3.3	2.4	36.2	12.1	7.3	37.4	12.1
C	2	0.8	0.3	3.4	1.8	0.9	23.1	4.1	2.5	25.3	6.9
	3	0.8	0.3	2.6	1.7	0.8	9.5	2.9	1.5	6.8	3.0
5 var.	4	0.9	0.3	2.9	1.8	0.7	11.0	3.0	1.3	7.1	3.3
	5	0.8	0.3	2.9	1.8	0.8	10.7	2.9	1.4	6.8	3.2
R^2 = 0.44	6	0.8	0.3	2.9	1.8	0.8	10.4	3.0	1.4	6.8	3.1
	1	1.6	0.8	7.7	3.3	2.4	36.2	12.1	7.3	37.4	12.1
D	2	0.9	0.3	4.2	2.0	1.0	24.6	4.8	2.8	25.2	7.3
	3	0.9	0.3	3.2	1.8	0.9	11.2	3.4	1.9	6.5	3.3
5 var.	4	1.1	0.4	3.9	2.2	0.9	15.1	3.4	1.6	8.6	4.1
	5	1.0	0.3	3.6	2.0	0.9	13.9	3.2	1.6	8.1	3.8
R^2 = 0.63	6	1.0	0.3	3.6	2.0	1.0	12.9	3.4	1.8	8.0	3.8
	1	1.6	0.8	7.7	3.3	2.4	36.2	12.1	7.3	37.4	12.1
E	2	0.7	0.3	5.7	1.5	1.2	25.6	6.1	3.4	27.3	8.0
	3	0.7	0.3	5.2	1.3	1.1	16.3	5.8	2.5	9.7	4.8
5 var.	4	0.8	0.3	7.4	1.4	1.2	14.2	4.7	2.6	18.8	5.7
	5	0.8	0.3	6.1	1.4	1.2	12.1	4.8	2.3	17.7	5.2
R^2 = 0.71	6	0.8	0.3	5.8	1.3	1.2	12.8	5.2	2.3	14.6	4.9

Table 2. Continued.

Problem	Method	Percentage deletions from each variable, and number of observations.									Av.
		(5%)	(5%)	(10%)	(10%)	(10%)	(20%)	(20%)	(20%)	(40%)	
		(100)	(200)	(50)	(100)	(200)	(50)	(100)	(200)	(200)	
	1	1.6	0.8	7.7	3.3	2.4	36.2	12.1	7.3	37.4	12.1
F	2	1.4	0.7	6.4	2.9	2.0	32.6	9.9	6.4	32.7	10.6
	3	1.5	0.7	5.3	3.0	1.9	27.0	8.7	5.5	23.5	8.5
5 var.	4	15.9	4.2	77.9	33.2	13.0	245.4	65.5	26.4	118.2	66.6
	5	1.6	0.6	13.5	4.0	2.2	78.4	15.4	5.7	77.6	22.1
R^2 = 0.98	6	1.4	0.6	5.6	3.1	2.0	25.3	8.5	5.5	25.8	8.6
	1	1.6	0.8	7.7	3.3	2.4	36.2	12.1	7.3	37.4	12.1
G	2	1.4	0.7	6.3	2.8	2.0	33.6	10.1	6.5	33.4	10.8
	3	1.5	0.7	5.3	3.0	2.0	30.9	8.4	5.8	24.4	9.1
5 var.	4	21.5	5.5	104.2	47.8	20.1	372.9	96.6	37.2	178.3	8.2
	5	1.6	0.6	10.5	3.9	2.2	112.1	18.3	6.8	119.5	30.6
R^2 = 0.99	6	1.4	0.6	5.6	3.1	2.1	28.2	8.3	5.8	26.7	9.1

Table 2 Continued

Covariance Matrices for Problem:

A

	x_1	x_2	y
x_1	1.0000		
x_2	0.9817	1.0000	
y	0.9722	0.9697	1.0000

$R^2 = 0.9516$

B

	x_1	x_2	x_3	y
x_1	1.0000			
x_2	0.9128	1.0000		
x_3	0.8730	0.9529	1.0000	
y	0.2570	0.2851	0.2977	1.0000

$R^2 = 0.0888$

C

	x_1	x_2	x_3	x_4	y
x_1	1.0000				
x_2	0.8385	1.0000			
x_3	0.4596	0.6077	1.0000		
x_4	0.3618	0.4706	0.7962	1.0000	
y	0.7522	0.5958	0.6979	0.8232	2.2500

$R^2 = 0.4402$

D as C except that Var(y) = 1.5625 $R^2 = 0.6339$

E

	x_1	x_2	x_3	x_4	y
x_1	1.0000				
x_2	0.8743	1.0000			
x_3	0.4570	0.8255	1.0000		
x_4	0.3765	0.5181	0.6080	1.0000	
y	0.3705	0.4575	0.5039	0.8261	1.0000

$R^2 = 0.7173$

Table 2 Continued

Covariance Matrices for Problem:

F	x_1	x_2	x_3	x_4	y	
x_1	1.0000					
x_2	0.8738	1.0000				
x_3	0.5166	0.6314	1.0000			
x_4	0.4267	0.4650	0.7119	1.0000		
y	0.7852	0.6137	0.6389	0.8283	1.0000	$R^2 = 0.9866$

G	x_1	x_2	x_3	x_4	y	
x_1	1.0000					
x_2	0.8385	1.0000				
x_3	0.4596	0.6077	1.0000			
x_4	0.3618	0.4706	0.7962	1.0000		
y	0.7522	0.5958	0.6979	0.8232	1.0000	$R^2 = 0.9904$

Table 3

Approximate χ^2 statistic for covariances of regression coefficients estimated by modified maximum likelihood as a multiple of the χ^2 statistic for covariances of regression coefficients estimated from complete data before deletions:

Problem:	Method	Percentage deletions from each variable, and number of observations.									
		(5%) (100)	(5%) (200)	(10%) (50)	(10%) (100)	(10%) (200)	(20%) (50)	(20%) (100)	(20%) (200)	(40%) (200)	Av.
A		1.14	0.98	1.13	1.50	0.75	0.88	0.77	0.52	1.14	0.98
B		0.90	1.02	1.21	0.76	1.10	1.03	1.09	1.07	1.71	1.10
C		0.95	1.12	1.10	1.44	0.67	1.11	1.02	1.20	0.94	1.06
D		0.97	1.11	1.13	1.41	0.67	1.17	1.02	1.26	0.91	1.07
E		0.95	1.05	1.14	1.33	0.84	1.40	1.43	1.41	0.91	1.16
F		1.04	1.07	1.04	1.44	0.88	1.48	1.19	1.33	1.19	1.18
G		1.06	1.09	1.02	1.43	0.90	1.69	1.16	1.34	1.20	1.21

Table 4

Average Percentage increase in Residual Sum of Squares over best fit when all variables are known. Averaged over 20 runs.

Model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \epsilon$. 50 Observations.

Problem	Correlations	Method	Percentage deletions from (X_1,X_2,Y)	
			(40%, 0,0)	(20%; 20%, 40%)
A	$R^2 = 0.83$	1	9.4	91.5
$\beta = (1,.5,.4)$	$R_{12}^2 = 0.71$	3	8.5	32.1
$\rho_x = 0.8$	$R_{13}^2 = 0.79$	7	11.0	33.3*
$\sigma_y^2 = 0.5$	$R_2^2 = 0.59$			
B				
$\beta = (1,0,0.7)$	$R^2 = 0.8$	1	9.4	91.5
$\rho_x = 0.8$	$R_{12}^2 = 0.4$	3	9.7	46.1
$\sigma_y^2 = 0.5$	$R_{13}^2 = 0.8$	7	8.6	33.0
	$R_2^2 = 0.26$			
C	$R^2 = 0.8$	1	9.4	91.5
$\beta = (0,1,0.7)$	$R_{12}^2 = 0.4$	3	8.0	35.6
$\rho_x = 0.8$	$R_{13}^2 = 0.66$	7	5.6	30.8
$\sigma_y^2 = 0.5$	$R_2^2 = 0.4$			
D	$R^2 = 0.8$	1	9.4	91.5
$\beta = (0,0,1)$	$R_{12}^2 = 0$	3	8.5	37.1
$\rho_x = 0.2$	$R_{13}^2 = 0.8$	7	5.6	24.8
$\sigma_y^2 = 0.5$	$R_2^2 = 0$			

Table 4 Continued

Problem	Correlations	Method	Percentage deletions from (X_1,X_2,Y)	
			(40%, 0,0)	(20%, 20%, 40%)
E	$R^2 = 0.8$	1	9.4	91.5
$\beta = (0,0,1)$	$R_{12}^{3} = 0$	3	8.7	44.8
$\rho_x = 0.8$	$R_{13}^{2} = 0.8$	7	6.1	30.3
$\sigma_Y^2 = 0.5$	$R_2^2 = 0$			

*See text.

7.4. An Application of Iterated Buck in Discriminant Analysis.

Patients admitted to hospital on suspicion of having a certain disease are examined by a specialist over a period of about 4 days, and for a certain proportion an operation is advised. It is desirable to find out how much the decision to operate or not operate can be explained by the patient's history, and by tests which do not require the specialist's attention. Data on 60 explanatory variables, and the response variable

$$Y = \begin{cases} 1, & \text{for a decision to operate,} \\ 0, & \text{for a decision not to operate,} \end{cases}$$

were collected for a sample of 181 patients. These variables were continuous,

e.g. X_{11} = "Maximum Fever, in °C, during first day of admission",

discrete and ordered,

e.g. X_{19} = "Abdominal Tenderness", "None" $(X_{19}=0)$, "Slight" $(X_{19}=1)$, or "Severe" $(X_{19}=2)$,

or discrete and unordered,

e.g. X_6 = "Type of disease" (3 categories)

In all the observations some of the variables were not recorded, so if all the variables were to be treated simultaneously some missing values technique was required.

The means and covariances of the variables were estimated by 20 iterations of Iterated Buck on the data. A forward stepwise regression of $(X_1, X_2, \ldots, X_{60})$ on Y was then carried out. The standard stepwise regression programme consisted of 3 subroutines: (1) to form the sample means and S.S.C.P. matrix of a set of complete data; (2) to perform a forward stepwise regression to introduce q variables into the regression equations, by pivoting on the S.S.C.P. matrix found in (1); (3) to calculate various estimates of precision. For our purposes it was sufficient to use only the second subroutine, using the means and adjusted S.S.C.P. matrix found by Iterated Buck.

Most of the variation was explained by 6 variables. In Table 5 we list these variables in order of their introduction into the equation, and the multiple R^2 of Y with the new variable and its predecessors. We also give the estimated Mahalanobis distance D^2, where

$$D^2 = \frac{(n_0+n_1-2)}{\lambda} \frac{R^2}{1-R^2} \quad ; \qquad (7.4.1)$$

$\lambda = \frac{n_0 n_1}{n_0+n_1}$; n_j = number of observations with $Y=j$. This gives an idea of the separation between the two populations.

Variable entering Equation		R^2	D^2
1	X_{54}	0.38	3.7
2	X_{14}	0.49	5.4
3	X_{52}	0.55	7.0
4	X_{33}	0.59	8.2
5	X_{21}	0.61	8.9
6	X_{16}	0.63	9.8

Table 5. Results of Stepwise Regression.

Here n_1 = 38, n_0 = 153. With 20 variables in the equation, the estimated R^2 was 0.71. The results suggest a reasonable separation between the populations with 6 variables included in the discriminant function. This function is proportional to the expression given by the stepwise regression.

This rather crude analysis could be refined in some respects. Firstly, the unordered qualitative variables were ordered in a rough way prior to the analysis. A less arbitrary way of handling these variables would be to split a k-chotomous variable into k binary variables, indicating to which of the categories each observation belongs. Another procedure which avoids the generation of additional variables, is to rank the categories according to their observed relationship with the response variable. Form a linear discriminant based on the other variables, and find a score for each level of the unordered response X_r, by averaging the values of the linear discriminant over the observations for which X_r takes that level. Then rank the levels by these scores, or code the variable using these scores directly.

A second refinement concerns the fact that 14 variables, including the response, were binary. The response variable was never missing, but the fitted values for some of the explanatory variables, taking values 0 or 1, were occasionally outside the range (0,1). The robustness considerations of Chapter 3 are reassuring as far as the final estimates of the means and covariances are concerned, but the question arises whether a better estimate of a missing binary variable can be obtained by considering some form of multivariate logistic model. This suffers from the drawback that the introduction of correlation between the variables in a multivariate logistic-type distribution results in complicated marginal and conditional distributions, required for applying M.I.P. Thus for simplicity independence between missing binary independent variables will be a necessary approximation. Then fitted values and adjustments could be worked out by an application of M.I.P. This would be a worthwhile exercise for a smaller problem, although in this case the number of variables is rather prohibitive.

Finally, a practical consideration with large numbers of variables is to rearrange the data to minimize the computing time. Observations are grouped according to the deletion pattern, and the groups arranged to minimize the number of pivoting steps in each iteration. If p_i is the number of pivots required when the ith observation is introduced, we seek to minimize

$$t = \sum_{i=1}^{N} p_i .$$

Now $$p_i = \sum_{j=1}^{p} \left(d_{ij} - d_{i-1,j}\right)^2 ,$$

where $$d_{ij} = \begin{cases} 1 & , x_{ij} \text{ missing} \\ 0 & , x_{ij} \text{ present} \end{cases}$$

A simple ordering procedure is to choose observation i to minimize p_i, subject to the new observation not belonging to the set of observations $(1,2,\ldots,i-1)$ already chosen. Repeat the process for $i=2,3,\ldots,N$. The result is an ordering which does not necessarily minimize t, but produces great savings over an arbitrary arrangement, particularly when only a small number of the 2^p possible patterns are present in the data, as was the case in this problem.

8. MISSING VALUES IN TIME SERIES.

8.1. Introduction.

So far we have considered incomplete data where the observations are independent. In this chapter we take a topic in time series where unknown parameters can be estimated by maximum likelihood, and apply the Missing Information Principle to obtain m.l.e's when the data are incomplete. The subject is autoregression, and in the next section we estimate the parameters of an autoregressive series with lag one (AR1), when there are gaps in the series. We assume that the pattern of deletions is random, in the sense made precise in §2.3. In §8.3 we consider the generalization to higher order series, and the general conclusion is that m.l.e's can be found for any pattern of missing values without much difficulty.

8.2. The AR1 Series with Missing Values.

Suppose $(y_1, y_2, \ldots, y_N)$ is an AR1 series, so that

$$\left.\begin{array}{ll} y_n - \mu = \lambda(y_{n-1} - \mu) + \epsilon_n & , (n = 1, 2, \ldots, N) \\ \epsilon_n \overset{i.i.d}{\sim} N(0, \sigma^2) & \end{array}\right\} , \qquad (8.2.1)$$

and that some of the y_i's are randomly missing. We wish to find m.l.e.'s of λ, μ, and σ^2, under the simplest boundary condition, $y_1 - \mu = \lambda(y_0 - \mu) + \epsilon_1$, for some known constant y_0. Then the log-likelihood of a complete set of data is

$$\ell = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\left[y_n - \mu - \lambda(y_{n-1} - \mu)\right]^2 - \frac{N}{2}\log\sigma^2 . \tag{8.2.2}$$

The scores are:

$$\frac{\partial\ell}{\partial\mu} = \frac{1-\lambda}{\sigma^2}\sum_{n=1}^{N}\left[y_n - \mu - \lambda(y_{n-1} - \mu)\right] ,$$

$$\frac{\partial\ell}{\partial\lambda} = \frac{1}{\sigma^2}\sum_{n=1}^{N}\left[(y_n - \mu)(y_{n-1} - \mu) - \lambda(y_{n-1} - \mu)^2\right] ,$$

$$\frac{\partial\ell}{\partial\sigma^2} = -\frac{N}{2\sigma^4}\left\{\frac{1}{N}\sum_{n=1}^{N}\left[y_n - \mu - \lambda(y_{n-1} - \mu)\right]^2 - \sigma^2\right\} .$$

Let
$$\hat{y}_n = E(y_n \mid P ; \mu, \lambda, \sigma^2) , \tag{8.2.3}$$

$$c_{rs} = \mathrm{Cov}\left[(y_r, y_s) \mid P ; \mu, \lambda, \sigma^2\right] , \tag{8.2.4}$$

where P stands for the set of y_n's present. Then if $\theta = (\mu, \lambda, \sigma^2)^T$,

$$E\left(\frac{\partial \ell}{\partial \mu} \,\middle|\, P;\theta\right) = \frac{1-\lambda}{\sigma^2} \sum_{n=1}^{N} \left[\hat{y}_n - \mu - \lambda(\hat{y}_{n-1} - \mu) \right] ,$$

$$E\left(\frac{\partial \ell}{\partial \lambda} \,\middle|\, P;\theta\right) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left\{ \left[(\hat{y}_n - \mu)(\hat{y}_{n-1} - \mu) - \lambda(\hat{y}_{n-1} - \mu)^2 \right] + \left[c_{n,n-1} - \lambda c_{n-1,n-1} \right] \right\} ,$$

$$E\left(\frac{\partial \ell}{\partial \sigma^2} \,\middle|\, P;\theta\right) = -\frac{N}{\sigma^4} \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[(\hat{y}_n - \mu) - \lambda(\hat{y}_{n-1} - \mu) \right]^2 - \sigma^2 + \frac{1}{N} \sum_{n=1}^{N} \left[c_{nn} - 2\lambda_{n,n-1} + \lambda^2 c_{n-1,n-1} \right] \right\} .$$

These expressions, equated to zero, form the fixed point equations. They are solved iteratively to find m.l.e's of (μ, λ, σ^2). It remains to find $\hat{y}_n$ and c_{rs} for assumed values of the parameters, and for various deletion patterns.

Intuitively speaking, the model (8.2.1) implies that the distribution of y_n given (y_{n-1}, y_{n+1}) and any other set of y's, depends only on y_{n-1}, y_{n+1} and θ, and is independent of the position of (y_{n-1}, y_n, y_{n+1}) in the series. Similarly if the data includes the sequence

$$P_{n-k} , M_{n-k+1} , M_{n-k+2}, \ldots , M_{n-2}, M_{n-1}, P_n ,$$

where P stands for "present" and M stands for "missing", then the conditional distribution of the missing $(y_{n-k+1}, y_{n-k+2}, \ldots, y_{n-1})$ given the data depends only on y_{n-k}

and y_n. Thus the joint distribution of the missing y's given the data factorizes into blocks corresponding to groups of missing values. These blocks are multivariate normally distributed, with parameters which depend only on θ and the bounding y's at each end. We now express this in symbols.

For assumed values $\theta_A = (\mu_A, \lambda_A, \sigma_A^2)$ of the parameters, we write $z_n = y_n - \mu_A$, and find the distribution of the missing z's. The model

$$z_n = \lambda_A z_{n-1} + \epsilon_n \qquad (n = 1, 2, \ldots, N)$$

defines a linear transformation from a set of variables $(z_1, z_2, \cdots, z_N)$ to the set of variables $(\epsilon_1, \epsilon_2, \cdots, \epsilon_N)$, the latter being distributed as $MN_N(0, I_N \sigma_A^2)$, where I_N is the $(N \times N)$ identity matrix. Thus the z's have a joint density proportional to

$$\prod_{n=1}^{N} \exp\left[-\frac{(z_n - \lambda_A z_{n-1})^2}{2\sigma_A^2}\right] .$$

Hence the conditional distribution of the missing data M given the data P is

$$f(M|P) = \left\{\prod_{n=1}^{N} \exp\left[-\frac{(z_n - \lambda_A z_{n-1})^2}{2\sigma_A^2}\right]\right\} \left\{\int_{m, z_m \in M} \prod_{n=1}^{N} \exp\left[-\frac{(z_n - \lambda_A z_{n-1})^2}{2\sigma_A^2}\right] \prod_m dz_m\right\}^{-1} .$$

(8.2.5)

Inspection of (8.2.5) shows that (a) M is normally distributed, and (b) $f(M|P)$ factorizes into a product of joint densities, corresponding to "blocks" of successive missing values. For example if y_n is missing, y_{n-1} and y_{n+1} present, then

$$f(M|P) = f_1(y_n \,|\, y_{n-1}, y_{n+1})\, f_2(M_{\backslash n}|P),$$

where $M_{\backslash n}$ is M with y_n deleted. Also,

$$\begin{aligned} f_1(y_n \,|\, y_{n-1}, y_{n+1}) &\propto \exp\left\{ -\frac{1}{2\sigma_A^2}\left[(z_n - \lambda_A z_{n-1})^2 + (z_{n+1} - \lambda_A z_n)^2 \right] \right\} \\ &\propto \exp\left[-\frac{(z_n - \xi^{(1)})^2}{V_1} \right], \end{aligned}$$

where

$$\left.\begin{aligned} \xi_n^{(1)} &= E(z_n|P) = -\frac{\lambda_A}{1+\lambda_A^2}(z_{n-1} + z_{n+1}) \\ V_1 &= Var(z_n|P) = \frac{\sigma_A^2}{1+\lambda_A^2} \end{aligned}\right\}. \tag{8.2.6}$$

From (8.2.6) we have the following fitted values and adjustments:

$$\begin{aligned} E(y_n|P) &= \hat{y}_n = \xi_n^{(1)} + \mu_A, \\ Var(y_n|P) &= c_{nn} = v_1, \\ Cov(y_n, y_m|P) &= c_{nm} = 0, \quad n \neq m. \end{aligned}$$

Next consider a block of k missing values, $k \geqslant 2$. Suppose $y_{n+1}, y_{n+2}, \ldots, y_{n+k}$ are missing, and y_n, y_{n+k+1}

present. Let $\zeta^{(k)}$ and V_k denote the mean and covariance matrix of the k-variate normal distribution of $z_{n+1}, z_{n+2}, \dots, z_{n+k}$, given P. Writing $z^{(k)} = (z_{n+1}, z_{n+2}, \dots, z_{n+k})^T$, from (8.2.5).

$$-\frac{1}{2}\left(z^{(k)}-\zeta^{(k)}\right)^T V_k^{-1}\left(z^{(k)}-\zeta^{(k)}\right) = -\frac{1}{2\sigma_A^2}\left[\sum_{r=n}^{n+k}\left(z_r-\lambda_A z_{r+1}\right)^2\right]. \qquad (8.2.7)$$

Equating coefficients of $z^{(k)}$, we have

$$V_k^{-1} = \frac{1}{\sigma_A^2}\begin{bmatrix} 1+\lambda_A^2 & -\lambda_A & & & 0's \\ -\lambda_A & 1+\lambda_A^2 & -\lambda_A & & \\ & -\lambda_A & 1+\lambda_A^2 & \ddots & \\ & & \ddots & \ddots & -\lambda_A \\ 0's & & & -\lambda_A & 1+\lambda_A^2 \end{bmatrix} ; \quad \zeta^{(k)} = V_k \begin{bmatrix} -\lambda_A z_n \\ 0 \\ 0 \\ \vdots \\ 0 \\ -\lambda_A z_{n+k+1} \end{bmatrix}. \qquad (8.2.8)$$

For low values of k, the elements of V_k can be found as functions of λ_A by inversion. Unfortunately for high values of k, the elements are not simple functions of λ_A, and the matrix must be inverted for each assumed numerical value of λ_A. Let k be the largest number of consecutive missing values in the series. Then the matrix V_r required for the fitted values and adjustments for a block of r missing values, $r \leq k$ is found by pivoting on the first r diagonal elements of V_k^{-1}. So the computing involved is not excessive, involving the inversion of one $(k \times k)$ matrix at each iteration.

We see how the distribution of the missing values depends only on the parameters and the adjacent data. Thus the calculation of fitted values and adjustments is not affected by imposing a different boundary condition for y_1, which is present by definition. For example, the stationary boundary condition

$$y_1 \sim N(\mu, \tfrac{\sigma^2}{1-\lambda^2})$$

produces the same additional term to the log-likelihood (8.2.1) whether the data are complete or incomplete. The likelihood and fixed point equations are modified accordingly.

8.3. Higher Order Autoregressive Series.

The m.l.e's of the parameters for higher order autoregressive series are found in the same way as for AR1. The scores for a complete set of data are calculated, and modified by the fitted values and adjustments for the missing y's.

To find the fitted values and adjustments, the distribution of $(M|P)$ does not necessarily factorize into blocks of missing values. For example the sequence

$$\ldots, M_n, P_{n+1}, M_{n+2}, \ldots$$

in an AR2 series will not produce such a factorization. However in the AR2 case we can split the series into blocks with two variables present at each end. The distribution of the missing variables in such a block depends on the two bounding values at each end, and also on the values of any isolated present values within the block. The means and covariances of such a distribution can always be found by pivoting on the covariance matrix obtained if all the intermediate values were missing. This in turn is found by the obvious generalization of the right hand side of (8.2.7). Similar remarks can be made about higher order autoregressive series.

Appendix 1. The Pivot Operator.

For any symmetric $p \times p$ matrix A, define the operator $PIV(j)$ as follows:

$PIV(j)[A] = A^*$, a $(p \times p)$ symmetric matrix, such that

$$a^*_{jj} = -\frac{1}{a_{jj}} , \qquad \text{(A1)}$$

$$a^*_{jk} = \frac{a_{jk}}{a_{jj}} , (k \neq j), \qquad \text{(A2)}$$

$$a^*_{k\ell} = a_{k\ell} - \frac{a_{kj} a_{j\ell}}{a_{jj}} , (k \neq j, \ell \neq j) . \qquad \text{(A3)}$$

Also define

$$PIV(j,k,\ell,\ldots,r) = PIV(j) PIV(k) PIV(\ell) \ldots PIV(r) .$$

It can be shown $PIV(j)$ and $PIV(k)$ commute. Also the operator inverse to $PIV(j)$ is $RPIV(j)$, defined by the same equations as (A1) - (A3), except that (A2) is replaced by

$$a^*_{jk} = -\frac{a_{jk}}{a_{jj}} , (k \neq j) . \qquad \text{(A4)}$$

Now suppose $\Sigma_{(p \times p)}$ is a true or estimated covariance matrix of p random variables $X_1, X_2, \ldots, X_p$. If we split the variables into two groups (P, M), we can consider the linear regression of the variables in M on the variables in P. The regression coefficients and residual covariance matrix are then found from the matrix

$$\mathrm{PIV}(P)[\Sigma] ,$$

where $\mathrm{PIV}(P)$ denotes pivoting with respect to the subscripts of the variables in P. Then for $X_j \epsilon M$, $\mathrm{PIV}(j)$ corresponds to introducing X_j into the regression equations as an independent variable, and for $X_k \epsilon P$, $\mathrm{RPIV}(k)$ corresponds to removing X_k from the regression equations, i.e. changing X_k from an independent variable to a dependent variable. This correspondence makes pivoting a powerful tool in multiple regression.

Finally,

$$\mathrm{PIV}(1,2,\ldots,p)A = -A^{-1},$$

and applying the pivot operator is a satisfactory way of computing the inverse of a matrix. If a diagonal element a_{jj} becomes zero before $\mathrm{PIV}(j)$ is applied, the matrix is singular.

REFERENCES.

A.A. AFIFI and R.M. ELASHOFF (1966). Missing observations in multivariate statistics I: review of the literature. J. Amer. Statist. Assoc. 61 pp 595-604.

T.W. ANDERSON (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Amer. Statist. Assoc. 52 pp 200-203.

E.M.L. BEALE (1970). Computational methods in least squares. In Integer and Nonlinear Programming. Ed. J. Abadie (North Holland Amsterdam) pp 213-227.

E.M.L. BEALE and R.J.A. LITTLE (1973). Missing Values in Multivariate Analysis. Paper read to Royal Statistical Society Multivariate Analysis conference, Hull, 1973.

S.F. BUCK (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. J. Roy. Statist. Soc. (B) 22 pp 302-306.

J. BERKSON (1950). Are there two regressions? J. Amer. Statist. Assoc. 45 pp 164-180.

Y. HAITOVSKY (1968). Missing data in regression analysis. J. Roy. Statist. Soc. (B) 30 pp 67-82.

H.O. HARTLEY and R.R. HOCKING (1971). The analysis of incomplete data. Biometrics 27 pp 783-823.

G.H. JOWETT (1963). Application of Jordan's Procedure for matrix inversion in multiple regression and multivariate distance analysis. J. Roy. Statist. Soc. (B) 25 pp 352-357.

J.D. KALBFLEISCH and D.A. SPROTT (1970). Application of likelihoo methods to models involving large numbers of parameters. J. Roy. Statist. Soc. (B) 32 pp 175-208.

M.G. KENDALL and A. STUART (1967). The Advanced Theory of Statistics Vol. II Second edition. C. Griffin. London!

L. McDONALD (1971). On the estimation of missing data in the multivariate linear model. Biometrics 27 pp 535-543.

P.E. LIN and L.E. STIVERS (1974). On difference of means with incomplete data. Biometrika 61 2 pp 325-334.

D.M. MORRISON (1972). Analysis of a single sample of repeated measurements. Biometrics 28 pp 55-71.

D.M. MORRISON (1973). A test of equality of means of correlated variates with missing data in one response. Biometrika 60 pp 101-105.

J.A. NELDER and R.W.M. WEDDERBURN (1972). Generalized Linear Models. J. Roy. Statist. Soc. (A) 135 pp 370-383.

T. ORCHARD and M.A. WOODBURY (1972). A missing information principle: theory and applications. Proc. Sixth Berkeley Symp. Math. Statist. and Prob.1. pp 697-715.

K.D. TOCHER (1952). The design and analysis of block experiments. J. Roy. Statist. Soc. (B) 14 pp 45-100.

B.L. WELCH (1947). The generalization of 'Student's' problem when several different population variances are involved. Biometrika 34 pp 28-35.

F. YATES (1933). The analysis of replicated experiments when field results are incomplete. Emp. J. Expt. Agric. Vol.1 pp 129-142.