Missing Values in Multivariate Analysis

By E. M. L. Beale and R. J. A. Little

Scientific Control Systems Ltd. and Imperial College — *Imperial College*

[Received July 1973. Final revision May 1974]

Summary

This paper presents computational results for some alternative methods of analysing multivariate data with missing values. We recommend an algorithm due to Orchard and Woodbury (1972), which gives an estimator that is maximum likelihood when the data come from a multivariate normal population. We include a derivation of the estimator that does not assume a multivariate normal population, as an iterated form of Buck's (1960) method.

We derive an approximate method of assigning standard errors to regression coefficients estimated from incomplete observations, and quote supporting evidence from simulation studies.

A brief account is given of the application of these methods to some school examinations data.

Keywords: MISSING VALUES; MULTIPLE REGRESSION; MAXIMUM LIKELIHOOD

1. Introduction

Many multivariate analysis techniques, and in particular multiple regression, assume that one starts with an array of numbers x_{ij} representing the value of the jth variable in the ith observation. This will be for $j = 1, ..., n$ and $i = 1, ..., N$ if we have N observations and n variables. From these raw data one then forms a square matrix a_{jk} of sums of squares and products defined by the equation

$$a_{jk} = \sum_i x_{ij} x_{ik} \tag{1.1}$$

or, more usually, by the equation

$$a_{jk} = \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \tag{1.2}$$

where

$$\bar{x}_j = \sum_i x_{ij}/N. \tag{1.3}$$

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

But what should we do if there are gaps in the original data, that is to say if individual variables are missing in some observations? Sometimes the fact that the variable is missing indicates that its true value is probably unusual, and in these circumstances any mechanical method of analysis may be very misleading. But information about some variables may simply not be readily available, particularly if the relevance of this information is doubtful, as in exploratory regression work.

One natural approach to this problem is to omit all incomplete observations. This is unsatisfactory if many variables are known for an incomplete observation, particularly if the variables that are known prove on analysis to include all those that are important for the study.

The other standard approach is to estimate each $\bar{x}_j$ independently from those observations where the particular variable is known, and then to estimate each a_{jk} separately as the average value of $(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$, where averaging takes place over all observations where both x_{ij} and x_{ik} are known. This approach apparently makes more use of the available data, but it can give very poor results, as demonstrated in simulation studies reported by Haitovsky (1968).

Another approach is to substitute suitable guessed values for the unknown quantities. For example, one can assume that any unknown quantity equals the mean of all known values of the variable. This approach has been used in many practical analyses and given acceptable results. On the other hand, it can also give very poor results with highly correlated data.

Yet another approach is to assume that the data came from a multivariate normal distribution and to estimate the parameters of this distribution by maximum likelihood. Until recently this approach seemed to pose formidable mathematical and computational problems, but Orchard and Woodbury (1972) have shown that these parameters can be estimated using a sophisticated version of the method of fitting suitable approximations to the unknown values in incomplete observations.

Section 2 of this paper derives Orchard and Woodbury's Missing Information Principle. The argument follows theirs, but emphasizes that the effect of the principle is to replace a maximization problem by a fixed point problem. We give a formal definition of the principle, expressed in a way that reduces the possibility of finding stationary values of the likelihood other than the maximum. We believe that this clarifies the logic of the principle. We follow Orchard and Woodbury in showing that the principle leads to a simple iterative algorithm for finding estimators for our problem that are maximum likelihood when the population is multivariate normal.

Section 3 presents an alternative derivation of essentially the same estimators, as an iterated version of those proposed by Buck (1960). The only difference is that the adjusted sum of squares and products matrix is divided by $(N-1)$ instead of N to derive the estimated covariance matrix. This correction makes no practical difference in our simulation studies, but it brings the method into line with conventional practice when all observations are complete. Our derivation is more arbitrary than Orchard and Woodbury's since it appeals to a desire for unbiasedness. But we think it is of interest since it does not assume that the underlying population is multivariate normal. We consider the problem of bias in detail for the special case of one incomplete observation.

Section 4 reports the results of simulation studies comparing six estimators on artificial data generated from multivariate normal populations subjected to random deletions. The estimators are found by:

(1) Ordinary least squares using complete observations only.
(2) Buck's method.
(3) Iterated Buck, or corrected maximum likelihood.
(4) A method that estimates the means, variances and covariances of the independent variables only by corrected maximum likelihood, uses these to fit missing values of the independent variables and then uses ordinary least squares on all observations for which the dependent variable is present.

(5) Method 4, but with incomplete observations given appropriately reduced weights.

(6) A combination of Methods 3 and 5.

The conclusion is that corrected maximum likelihood is generally best.

Section 5 discusses the problems of assessing the value of incomplete observations, and the problem of assessing approximate standard errors to regression coefficients estimated from incomplete data. We have found by further simulation studies that ideas based on Method 5 can be used to derive reasonable approximations to the covariance matrix of regression coefficients estimated by corrected maximum likelihood.

Section 6 outlines the facilities that we recommend for a practical missing values program.

Section 7 describes the results of using this program on some data kindly supplied by Dr Robert Wood and Miss Carolyn Ballantyne of the Schools Examination Department of the University of London.

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 vectors $\mathbf{z}$ and $\mathbf{y}$ with a joint distribution depending on the vector $\boldsymbol{\theta}$ of parameters, where $\mathbf{y}$ has been observed but $\mathbf{z}$ has not been observed. In our application of the principle $\boldsymbol{\theta}$ represents the set of means and the covariance matrix for the multivariate normal distribution, $\mathbf{y}$ represents the complete observations and the known variables in the incomplete observations, while $\mathbf{z}$ represents the missing values in the incomplete observations.

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

Let $f(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})$ denote the probability density function for the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$, and let $L(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})$ denote $\ln f(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})$. Then we know that

$$L(\mathbf{z}, \mathbf{y};\, \boldsymbol{\theta}) = L(\mathbf{y};\, \boldsymbol{\theta}) + L(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta}). \tag{2.1}$$

Now take any assumed value $\boldsymbol{\theta}_A$ for $\boldsymbol{\theta}$. This, together with the observed value of $\mathbf{y}$, defines a distribution for $\mathbf{z}$, and we can now take expectations of both sides of (2.1), integrating out with respect to $\mathbf{z}$. This is expressed by the equation

$$E\{L(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})|\, \mathbf{y};\, \boldsymbol{\theta}_A\} = L(\mathbf{y};\, \boldsymbol{\theta}) + E\{L(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})|\, \mathbf{y};\, \boldsymbol{\theta}_A\}. \tag{2.2}$$

If the distribution of $\mathbf{z}$ has a probability density element $f(\mathbf{z}|\mathbf{y};\, \boldsymbol{\theta}_A)\, d\mathbf{z}$ then (2.2) can be equivalently written as

$$\int L(\mathbf{z}, \mathbf{y};\, \boldsymbol{\theta})\, f(\mathbf{z}|\, \mathbf{y}, \boldsymbol{\theta}_A)\, d\mathbf{z} = L(\mathbf{y};\, \boldsymbol{\theta}) + \int L(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta})\, f(\mathbf{z}|\, \mathbf{y};\, \boldsymbol{\theta}_A)\, d\mathbf{z}. \tag{2.3}$$

We can now find the value $\boldsymbol{\theta}_M$ of $\boldsymbol{\theta}$ that maximizes the left-hand side of (2.3). This may depend on $\boldsymbol{\theta}_A$, so we may write

$$\boldsymbol{\theta}_M = \boldsymbol{\phi}(\boldsymbol{\theta}_A). \tag{2.4}$$

Equation (2.4) represents a transformation from the vector $\boldsymbol{\theta}_A$ to the vector $\boldsymbol{\theta}_M$. We now define the Missing Information Principle.

The Missing Information Principle

Estimate $\boldsymbol{\theta}$ by a fixed point of the transformation $\boldsymbol{\phi}$, namely a value of $\boldsymbol{\theta}$ such that

$$\boldsymbol{\theta} = \boldsymbol{\phi}(\boldsymbol{\theta}). \tag{2.5}$$

We call (2.5) the "fixed point equation". We justify this approach by two theorems, which show that the maximum likelihood estimator of $\boldsymbol{\theta}$ is a root of the fixed point equation, and conversely every root of the fixed point equation is a maximum or stationary value of the likelihood.

Hence, if the likelihood function is differentiable, any solution of the fixed point equations automatically satisfies the "likelihood equations" found by setting the partial derivatives of the likelihood equal to zero. Some solutions of the likelihood equations may not be fixed points, since the fixed point method involves finding $\boldsymbol{\theta}_M$ giving a global maximum, and not merely a stationary value, of the left-hand side of (2.3). (In this respect our approach differs slightly from that of Orchard and Woodbury, who implicitly define $\boldsymbol{\phi}$ be setting the derivatives of (2.2) with respect to $\boldsymbol{\theta}$ equal to zero.)

Theorem 1. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ satisfies (2.5).

Theorem 2. If $L(\mathbf{z}|\mathbf{y};\boldsymbol{\theta})$ is a differentiable function of $\boldsymbol{\theta}$, then any other value of $\boldsymbol{\theta}$ satisfying (2.5) must represent either a maximum or a stationary value of $L(\mathbf{y};\boldsymbol{\theta})$.

To prove these theorems, consider the last term on the right-hand side of (2.3).

We have

$$\int L(\mathbf{z}|\mathbf{y};\boldsymbol{\theta})\, f(\mathbf{z}|\mathbf{y};\boldsymbol{\theta}_A)\, d\mathbf{z},$$

regarded as a function of $\boldsymbol{\theta}$, is maximized when $\boldsymbol{\theta} = \boldsymbol{\theta}_A$. This is simply Jensen's inequality. The proof is elementary: see, for example, Kendall and Stuart (1967), pp. 39–40.

The theorems now follow immediately. If $\boldsymbol{\theta}_A = \hat{\boldsymbol{\theta}}$, then the value $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ maximizes both terms on the right-hand side of (2.3) simultaneously. It therefore certainly maximizes their sum. This proves Theorem 1.

To prove Theorem 2, we note that $\boldsymbol{\theta} = \boldsymbol{\theta}_A$ maximizes the second term on the right-hand side of (2.3), and by hypothesis this is differentiable. It cannot then be a maximum of the left-hand side of (2.3) unless it is either a maximum or a stationary value of the first term on the right-hand side.

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

$$\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \boldsymbol{\theta}_A = (\boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A), \quad \boldsymbol{\theta}_M = \boldsymbol{\phi}(\boldsymbol{\theta}_A) = (\boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M).$$

The log-likelihood for the multivariate normal distribution is

$$L(\mathbf{X};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{n} (x_{ij}-\mu_j)\, \sigma^{jk} (x_{ik}-\mu_k) - \tfrac{1}{2} N \log(\det \boldsymbol{\Sigma}),$$

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

$$E\left\{L(\mathbf{X};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \middle| P_T;\, \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A\right) = -\tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{n} (\hat{x}_{ijA}-\mu_j)(\hat{x}_{ikA}-\mu_k) + \sigma_{jkA.P_i}\Big\}\, \sigma^{jk}$$
$$- \tfrac{1}{2} N \log(\det \boldsymbol{\Sigma}),$$

where

$$\hat{x}_{ijA} = E\{x_{ij} \mid P_i;\, \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A\}$$

and

$$\sigma_{jkA.P_i} = \operatorname{cov}\{x_{ij}, x_{ik} \mid P_i;\, \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A\}.$$

Maximizing with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ gives the analogue of (2.4):

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

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

for $1 \leqslant j, k \leqslant n$. Now set $\boldsymbol{\mu}_A = \boldsymbol{\mu}_M = \boldsymbol{\mu}$, $\boldsymbol{\Sigma}_A = \boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}$. The fixed point equations are

$$\hat{x}_{ij} = E(x_{ij} \mid P_i;\, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.6}$$

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_{ij}, \tag{2.7}$$

$$\sigma_{jk} = \frac{1}{N} \sum_{i=1}^{N} \{(\hat{x}_{ij}-\mu_j)(x_{ik}-\mu_k) + \sigma_{jk.P_i}\}, \tag{2.8}$$

$$\sigma_{jk.P_i} = \operatorname{cov}(x_{ij}, x_{ik} \mid P_i;\, \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.9}$$

These are the equations found by Orchard and Woodbury. To find the maximum likelihood estimates we obtain initial estimates of $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ and cycle through (2.6)–(2.9) until we find no significant changes in the estimates between successive iterations. Note that

$$\hat{x}_{ij} = x_{ij}, \quad \text{if } x_{ij} \text{ is observed,}$$
$$= \text{a linear combination of the variables in } P_i,$$
$$\text{if } x_{ij} \text{ is missing.}$$

At each iteration the data are completed by equation (2.6), and the means and a sum of squares and products matrix found for the variables. This matrix is adjusted by adding $\sigma_{jk.P_i}$ for every observation i to the *jk*th element. Note that this adjustment is zero unless both x_{ij} and x_{ik} are missing.

It seems reasonable to hope that in this case we have a unique fixed point of the transformation (2.4), since the effect of changing one component of $\boldsymbol{\theta}_A$ is to change the corresponding component of the $\boldsymbol{\theta}_M$ by a smaller amount in the same sense. Thus cycling through equations (2.6)–(2.9) represents an iterative procedure for finding the maximum likelihood estimates $\hat{\mu}_j, \hat{\sigma}_{jk}$, which is much simpler to carry out than working directly from the likelihood equations.

3. A DERIVATION *via* BUCK'S METHOD

We now derive essentially the same estimators using the approach due to Buck (1960).

Buck starts by using the complete observations to estimate the means of all the variables, and also the covariance matrix. These values can then be used to estimate any missing quantities x_{ij} as linear functions of the variables that are known for this observation. If we then substitute the estimates for the unknown variables, we can build up the vector $\bar{x}_j$ and the matrix (a_{jk}) defined by (1.2) and (1.3).

This is a useful improvement over the estimators found by setting unknown variables equal to their sample means. But the a_{jk} calculated in this way provide a biased estimate of the values that they would have taken if none of the data had been incomplete. Buck's method therefore estimates this bias, and subtracts it from the computed value of $(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)$ to derive the final assumed value of a_{jk}.

Let us express this in symbols. We write $\hat{x}_{ij}$ for the assumed value of the jth variable in the ith observation. If this value has been observed then $\hat{x}_{ij} = x_{ij}$. Otherwise it is a fitted value. We then modify (1.2) and (1.3) to read

$$a_{jk} = \sum_i (\hat{x}_{ij}-\bar{x}_j)(\hat{x}_{ik}-\bar{x}_k)+c_{ijk}, \tag{3.1}$$

$$\bar{x}_j = \sum_i \hat{x}_{ij}/N. \tag{3.2}$$

The problem remains to determine suitable formulae for the correction terms c_{ijk}. This problem is a subtle one, and is discussed rather briefly by Buck. The solution has been indicated at the end of Section 2, but it is of interest to explore it in more detail, discussing terms of order $N-1$ in a special case. Suppose that we have only one incomplete observation, where only the first $p(<n)$ variables are known. For notational convenience we assume that the incomplete observation has $i = 1$.

Let $\tilde{x}_l$ denote

$$\frac{1}{N-1}\sum_{i=2}^{N} x_{il},$$

i.e. the mean value of the lth variable over all complete observations. Define b_{jk} for $j = 1, \ldots, n$, $k = 1, \ldots, p$ as

$$\begin{aligned} b_{jk} &= \text{partial regression coefficient of } x_j \text{ on } x_k \text{, estimated from the complete data, if } j>p, \\ &= 0, \quad \text{if } j \leqslant p \text{ and } j \neq k, \\ &= 1, \quad \text{if } j \leqslant p \text{ and } j = k. \end{aligned}$$

Then

$$\hat{x}_{1j} = \tilde{x}_j + \sum_{l=1}^{p} b_{jl}(x_{il}-\hat{x}_l) \quad (j = 1, \ldots, n).$$

We are then interested in evaluating the expected value of

$$S_{jk} = (\hat{x}_{1j}-\bar{x}_j)(\hat{x}_{1k}-\bar{x}_k)+\sum_{i=2}^{N}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k),$$

where $N\bar{x}_j = \hat{x}_{1j}+(N-1)\tilde{x}_j$ and $N\bar{x}_k = \hat{x}_{1k}+(N-1)\tilde{x}_k$.

We can manipulate the expression for S_{jk} so that it reads

$$\begin{aligned} S_{jk} &= \hat{x}_{1j}\hat{x}_{1k}+\sum_{i=2}^{N}x_{ij}x_{ik}-\frac{1}{N}(\hat{x}_{1j}+(N-1)\tilde{x}_j)(\hat{x}_{1k}+(N-1)\tilde{x}_k) \\ &= \left\{\tilde{x}_j+\sum_{l} b_{jl}(x_{1l}-\tilde{x}_l)\right\}\left\{\tilde{x}_k+\sum_{l} b_{kl}(x_{1l}-\tilde{x}_l)\right\} \\ &\qquad +\sum_{i=2}^{N}x_{ij}x_{ik}-\frac{1}{N}\left\{N\tilde{x}_j+\sum_{l} b_{jl}(x_{1l}-\tilde{x}_l)\right\}\left\{N\tilde{x}_k+\sum_{l} b_{kl}(x_{1l}-\tilde{x}_l)\right\} \\ &= \sum_{i=2}^{N}x_{ij}x_{ik}-(N-1)\tilde{x}_j\tilde{x}_k+\frac{N-1}{N}\sum_{l_1}\sum_{l_2} b_{jl_1}b_{kl_2}(x_{1l_1}-\tilde{x}_{l_1})(x_{1l_2}-\tilde{x}_{l_2}). \end{aligned}$$

We are now concerned with the expected value of S_{jk}. We must therefore define some properties of the population from which the observations are drawn.

Without loss of generality we may assume that the true means of all variables are zero.

Let u_{jk} denote the covariance of x_j and x_k. Let v_{jk} denote the "partial covariance" of x_j and x_k, by which we mean the covariance of

$$\left(x_j-\sum_{l\leqslant p}\beta_{jl}x_l\right) \quad\text{and}\quad \left(x_k-\sum_{l\leqslant p}\beta_{kl}x_l\right),$$

where β_{jl} and β_{kl} are the (partial) regression coefficients defining the best linear approximations to x_j and x_k respectively in terms of the variables known in the first observation. Note that $v_{jk}=0$ unless x_j and x_k are unknown, i.e. $j>p$ and $k>p$.

Now

$$\begin{aligned} E\left\{\sum_{i=2}^{N}x_{ij}x_{ik}-(N-1)\tilde{x}_j\tilde{x}_k\right\} &= (N-2)\,u_{jk}, \\ E(x_{1l_1}-\tilde{x}_{l_1})(x_{1l_2}-\tilde{x}_{l_2}) &= \frac{N-1}{N}u_{l_1l_2}, \end{aligned}$$

and b_{jl_1} and b_{jl_2} are virtually independent of $(x_{1l_1}-\tilde{x}_{l_1})$ and $(x_{1l_2}-\tilde{x}_{l_2})$. This independence is exact if the population is multivariate normal. We therefore deduce that

$$E(S_{jk}) \simeq (N-2)\,u_{jk}+\sum_{l_1}\sum_{l_2}u_{l_1l_2}E(b_{jl_1}b_{kl_2}).$$

Now

$$E(b_{jl_1}b_{kl_2}) = \beta_{jl_1}\beta_{kl_2}+\operatorname{cov}(b_{jl_1},b_{kl_2})$$

and

$$\begin{aligned} \sum_{l_1}\sum_{l_2}u_{l_1l_2}\beta_{jl_1}\beta_{kl_2} &= E\left\{\left(\sum_{l_1}\beta_{jl_1}x_{l_1}\right)\left(\sum_{l_2}\beta_{kl_2}x_{l_2}\right)\right\} \\ &= u_{jk}-v_{jk}, \end{aligned}$$

while

$$\operatorname{cov}(b_{jl_1}, b_{kl_2}) = g^*_{l_1 l_2} v_{jk},$$

where $g^*_{l_1 l_2}$ is the (l_1, l_2)th element of the inverse of the $(p \times p)$ matrix $\mathbf{G}$ where

$$g_{l_1 l_2} = \sum_{i=2}^{N} (x_{il_1} - \bar{x}_{l_1})(x_{il_2} - \bar{x}_{l_2}).$$

Hence

$$E(S_{jk}) \simeq (N-2)\,u_{jk} + u_{jk} - v_{jk} + \sum_{l_1}\sum_{l_2} u_{l_1 l_2} g^*_{l_1 l_2} v_{jk}.$$

But now

$$E(g_{l_1 l_2}) = (N-2)\,u_{l_1 l_2},$$

so

$$E(g^*_{l_1 l_2}) \simeq \frac{1}{N-2} u^*_{l_1 l_2},$$

where $u^*_{l_1 l_2}$ denotes the $(l_1 l_2)$th element of the inverse of the $p \times p$ matrix $(u_{l_1 l_2})$. Hence

$$E(S_{jk}) \simeq (N-1)\,u_{jk} - v_{jk}\left(1 - \frac{1}{N-2}\sum_{l_1}\sum_{l_2} u_{l_1 l_2} u^*_{l_1 l_2}\right)$$

$$= (N-1)\,u_{jk} - v_{jk}\left(1 - \frac{p}{N-2}\right)$$

$$= (N-1)\,u_{jk} - \left(\frac{N-p-2}{N-2}\right) v_{jk}.$$

So that if we want to modify S_{jk} to obtain an approximately unbiased estimate of $(N-1)\,u_{jk}$, we must add an unbiased estimate of

$$\frac{N-p-2}{N-2} v_{jk}.$$

The computational requirements, to estimate v_{jk} and also the regression coefficients b_j and b_k, are not nearly as severe as one might think, because the same computation provides all these data simultaneously. This important computational point has been known for some time. It is discussed for example by Jowett (1963), and an expository account with more emphasis on computational aspects is contained in Beale (1970).

To form these quantities we may take the matrix $(\hat{u}_{jk})$ defined by

$$\hat{u}_{jk} = \frac{1}{N-2} \sum_{i=2}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k),$$

and form $\hat{v}_{jk}$ by pivoting on the first p diagonal elements. This gives a biased estimate of v_{jk}, since it ignores the degrees of freedom associated with the variables on which pivoting has taken place. Specifically,

$$E(\hat{v}_{jk}) = \frac{N-p-2}{N-2} v_{jk}.$$

It therefore follows that an appropriate formula for the correction term c_{ijk} in (4) is

$$\begin{aligned} c_{ijk} &= \hat{v}_{jk} \quad \text{if } x_{ij} \text{ and } x_{ik} \text{ are both unknown} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.3}$$

It is somewhat remarkable that all the various terms of order $1/N$ in the above analysis cancel, and that the resulting formula for c_{ijk} is the "naive" estimate for the partial covariance of x_j and x_k given the known variables. This cancellation does not happen exactly with more complicated deletion patterns, but there is no simple correction formula for the bias of order $1/N$.

The analysis implicitly assumes that the probability of a particular variable being missing is independent of the numerical values of any of the variables for this observation. This important assumption was noted in the Introduction. But the analysis does not assume that the underlying population is multivariate normal. This is of some practical significance, since multiple regression is widely applied to non-normal data. On the other hand, it is worth noting that if the population is not multivariate normal, then any unknown variable is not necessarily best estimated by a linear function of the known variables for the observation. So it may be possible to develop slightly more powerful estimators for particular non-normal populations.

This analysis has concentrated on the situation where we have $(N-1)$ complete observations and a single incomplete observation. We now consider what to do when more observations are incomplete. Buck's method uses only the complete observations to define the means $\tilde{x}_j$ and the estimated covariance matrix $\hat{u}_{jk}$. But our simulation studies suggest that an iterated version of Buck's method is generally superior. This method takes trial values for the $\tilde{x}_j$ and the $\hat{u}_{jk}$, uses them to compute the $\hat{x}_{ij}$ and c_{ijk} and hence a_{jk} and $\bar{x}_j$ from (3.1) and (3.2). We then set

$$\tilde{x}_j = \bar{x}_j, \tag{3.4}$$

$$\hat{u}_{jk} = a_{jk}/(N-1), \tag{3.5}$$

and repeat the process until there are no further changes on any $\tilde{x}_j$ or $\hat{u}_{jk}$.

We noted in Section 2 that Orchard and Woodbury have derived the same algorithm, with N substituted for $N-1$ in (3.5), as giving maximum likelihood estimates when the population is multivariate normal.

4. A SIMULATION STUDY COMPARING DIFFERENT ESTIMATORS

In this section we report briefly on a simulation study comparing estimators found by six different methods which are listed in Section 1. No further comment is necessary for the first three methods; we now describe methods 4–6 in more detail.

Methods 4 and 5 adopt a least squares approach. Suitable fitted values are found for the missing independent variables in every observation where y is observed, and a least squares analysis carried out on these completed observations. This approach has the intuitive appeal that the data on the dependent variable y are not used when missing independent variables are fitted, for the following reason: the best fitted value for a missing independent variable x_{ij}, prior to least squares analysis, is its conditional mean given the known *independent* variables in observation i. We thus estimate this best value by regressing the unknown on the known independent variables within each observation, using an estimated covariance matrix of the x's. This covariance matrix is found by using iterated Buck on the independent data, with the y's excluded.

After fitting missing values in this way, Method 4 proceeds with an ordinary least squares analysis. But this is inefficient since it amounts to giving the same weight to incomplete observations as that given to complete observations. Method 5 computes a weight w_i for each observation i, and carries out a weighted least squares analysis.

To find the weights, let

σ^2_{yi} denote the conditional variance of y given the known independent variables in observation i,

σ^2_y denote the residual variance of y when all the independent variables are fitted.

The effect of fitting missing values by Method 4 is to give "neutral" values to the missing independent variables given the known independent variables. The mean square error in the dependent variable is then the mean square error when this is fitted as a function only of those independent variables known for this observation, i.e. σ^2_{yi}.

Thus, if complete observations are given weight 1, the correct weight for observation i is

$$\sigma^2_y/\sigma^2_{yi}.$$

Given an estimated covariance matrix of all the variables, we therefore estimate σ^2_{yi} and σ^2_y by pivoting on the independent variables. Let s^2_{yi} and s^2_y be the corresponding estimates. Then define

$$\begin{aligned} w_i &= s^2_y/s^2_{yi} \quad \text{if the dependent variable } y_i \text{ is present} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{4.1}$$

In Method 5, the initial estimated covariance matrix is found by giving complete observations weight 1, and incomplete observations weight 0. The new weights are found by equation (4.1), and a new weighted sum of squares and products matrix formed. This, divided by the sum of the weights, gives a new estimated covariance matrix, from which new weights are found. The procedure is repeated until the weights do not change significantly.

The final method, Method 6, is a combination of Methods 3 and 5. An estimate of the covariance matrix of all the variables is found by Method 3; call it $\hat{\Sigma}$. 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. Then weights w_i are found directly from (4.1) by pivoting on $\hat{\Sigma}$, and a weighted least squares analysis carried out on the data with y present.

The six methods are thus:

Method 1: Ordinary least squares on complete observations only.

Method 2: Buck's (1960) method.

Method 3: Iterated Buck, i.e. corrected maximum likelihood.

Method 4: Ordinary least squares on observations with y present, after fitting missing values of the independent variables by modified maximum likelihood using the independent variables only.

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

Method 6: Method 5, but using a covariance matrix for all the variables, found by Method 3, to find the fitted values and estimate the weights.

For each method we must decide when an observation is so incomplete that it should be ignored. For all methods we ignore observations with all variables missing. For Method 4 we also ignore observations with all independent variables missing, since the inclusion of such observations with full weight was found to make the results significantly worse.

In all cases the data were generated from a multivariate normal population with 1 variable identified as the dependent variable, and between 2 and 4 independent variables. Some of the populations were the same as those studied by Haitovsky, but other had smaller values of R^2. We took 50, 100 or 200 observations and deleted either 5, 10, 20 or 40 per cent of the observed values of each variable. The values to be deleted were chosen randomly, and independently for each variable. Our criterion for judging the effectiveness of each estimator was the residual sum of squares of deviations of the observed and fitted values of the dependent variable when the deleted values were restored. In symbols we may write this as

$$S = \sum_{i=1}^{N} \left\{ y_i - b_0 - \sum_j b_j x_{ij} \right\}^2,$$

where b_0 and b_j are the constant term and regression coefficients estimated from incomplete data by one of the six methods, and x_{ij} and y_i are the true values of all variables without deletions.

Clearly a small value of S represents a successful method. It seems more sensible to judge a method by the overall success of the regression equation rather than by the closeness of individual regression coefficients to their true values.

We computed the average value of S for each of our 6 methods over 10 sets of random numbers for each covariance matrix and each number of observations and deletion pattern. The results are expressed in Table 1 as percentage increases over the absolute minimum possible value of S for each set of data. This minimum is obtained by ordinary least squares on the data before the missing values are deleted. Notice that these results are the same for Method 1 for Problems *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 $S/\min S$ is invariant under these transformations for Method 1, which uses only complete observations.

We draw the following conclusions from Table 1.

Methods 2 and 3 consistently beat the Standard Method 1. Method 3 always improves on Method 2, except for three very marginal cases with 5 per cent 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 two cases in Problem *E*; otherwise it is usually slightly worse, and much worse on Problems *F* and *G*, where $R^2 > 0{\cdot}98$. 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 neutral values are critical, and a better estimate of the covariance matrix of the independent variables used to fit these 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.

TABLE 1

Average percentage increase in residual sum of squares over best fit when all variables are known (averaged over 10 *runs)*

		5% *Deletions*		10% *Deletions*			20% *Deletions*			40% *Deletions*	
Problem	*Method*	100 *obs.*	200 *obs.*	50 *obs.*	100 *obs.*	200 *obs.*	50 *obs.*	100 *obs.*	200 *obs.*	200 *obs.*	*Av.*
A	1	0·4	0·3	2·7	1·4	0·5	3·9	3·9	1·3	6·4	2·3
3 var.	2	0·2	0·1	2·0	0·8	0·2	2·1	3·0	0·7	3·3	1·4
R^2 = 0·9516	3	0·2	0·1	1·9	0·8	0·2	1·9	2·4	0·7	1·9	1·1
	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
	6	0·3	0·2	2·0	1·0	0·3	2·3	2·8	0·8	2·2	1·3
B	1	0·9	0·4	4·5	2·5	0·7	8·6	4·7	3·1	30·6	6·2
4 var.	2	0·5	0·2	3·1	0·8	0·4	4·3	1·8	1·5	15·7	3·1
R^2 = 0·0888	3	0·6	0·2	3·0	0·8	0·4	3·8	1·4	1·2	3·1	1·6
	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
	6	0·6	0·2	3·0	0·8	0·4	3·8	1·4	1·2	3·6	1·7
C	1	1·6	0·8	7·7	3·3	2·4	36·2	12·1	7·3	37·4	12·1
5 var.	2	0·8	0·3	3·4	1·8	0·9	23·1	4·1	2·5	25·3	6·9
R^2 = 0·4402	3	0·8	0·3	2·6	1·7	0·8	9·5	2·9	1·5	6·8	3·0
	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
	6	0·8	0·3	2·9	1·8	0·8	10·4	3·0	1·4	6·8	3·1
D	1	1·6	0·8	7·7	3·3	2·4	36·2	12·1	7·3	37·4	12·1
5 var.	2	0·9	0·3	4·2	2·0	1·0	24·6	4·8	2·8	25·2	7·3
R^2 = 0·6339	3	0·9	0·3	3·2	1·8	0·9	11·2	3·4	1·9	6·5	3·3
	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
	6	1·0	0·3	3·6	2·0	1·0	12·9	3·4	1·8	8·0	3·8
E	1	1·6	0·8	7·7	3·3	2·4	36·2	12·1	7·3	37·4	12·1
5 var.	2	0·7	0·3	5·7	1·5	1·2	25·6	6·1	3·4	27·3	8·0
R^2 = 0·7173	3	0·7	0·3	5·2	1·3	1·1	16·3	5·8	2·5	9·7	4·8
	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
	6	0·8	0·3	5·8	1·3	1·2	12·8	5·2	2·3	14·6	4·9
F	1	1·6	0·8	7·7	3·3	2·4	36·2	12·1	7·3	37·4	12·1
5 var.	2	1·4	0·7	6·4	2·9	2·0	32·6	9·9	6·4	32·7	10·6
R^2 = 0·9866	3	1·5	0·7	5·3	3·0	1·9	27·0	8·7	5·5	23·5	8·5
	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
	6	1·4	0·6	5·6	3·1	2·0	25·3	8·5	5·5	25·8	8·6
G	1	1·6	0·8	7·7	3·3	2·4	36·2	12·1	7·3	37·4	12·1
5 var.	2	1·4	0·7	6·3	2·8	2·0	33·6	10·1	6·5	33·4	10·8
R^2 = 0·9904	3	1·5	0·7	5·3	3·0	2·0	30·9	8·4	5·8	24·4	9·1
	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
	6	1·4	0·6	5·6	3·1	2·1	28·2	8·3	5·8	26·7	9·1

TABLE 1 (*continued*)

Covariance matrices for the problems

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	$R^2 = 0·7173$
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	$R^2 = 0·9904$

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 in Methods 5 and 6 are used to derive approximate standard errors in the regression coefficients for Method 3. We return to this point in Section 5 below.

It is perhaps worth noting that we also tested the straight Maximum Likelihood Method of Orchard and Woodbury. The results are almost identical to those of Method 3. Mostly they are worse, but by less than 0·1 per cent. We therefore see no reason to use straight maximum likelihood, in preference to the conventional correction represented by Method 3.

5. THE VALUE OF INCOMPLETE OBSERVATIONS AND STANDARD ERRORS OF REGRESSION COEFFICIENTS

Now that we have a satisfactory method of analysing data with missing values, two important subsidiary questions arise. One, which is particularly relevant at the data-collection or design of experiment stage, is the value of incomplete data. The other is the assignment of approximate standard errors to regression coefficients estimated from incomplete data.

Fortunately, both these questions can be answered in the same way, using the ideas underlying Methods 5 and 6 as described in Section 4. There we found a weight w_i to associate with observation i when neutral values of the independent variables are fitted. This also measures the value of that observation. For the weighted least squares analysis we form

$$w_i, \quad \text{as in equation (4.1)},$$

$$\tilde{x}_j = \sum_{i=1}^{N} w_i \hat{x}_{ij} \Big/ \sum_{i=1}^{N} w_i, \tag{5.1}$$

$$s_{Wjk} = \sum_{i=1}^{N} w_i (x_{ij} - \tilde{x}_j)(x_{ik} - \tilde{x}_k). \tag{5.2}$$

Let $\mathbf{S}_W$ be the matrix formed from elements s_{Wjk} for all independent variables. Then put $\mathbf{C} = \mathbf{S}_W^{-1} s_y^2$. Then $\mathbf{C}$ represents the estimated covariance matrix for the regression coefficients found by Methods 5 and 6. This analysis assumes that we have enough data to estimate the means and covariance matrix, in order to derive both the fitted neutral values and the weights. The analysis is in that sense a large-sample analysis, but seems useful as such.

To estimate the covariance matrix of the regression coefficients estimated by Iterated Buck, $\mathbf{C}$ was found in exactly the same way as for Method 5, using equations (4.1), (5.1), (5.1) and (5.2). This obviously requires more computing than conventional least squares analysis with complete data, since it involves forming and inverting a new matrix $\mathbf{S}_W$. But it requires substantially less work than Method 5, since it does not require a second iterative loop to derive the weight w_i.

We obtain the fitted values $\hat{x}_{ij}$ by regression on all variables that are known for the ith observation. It would arguably be more logical, though less convenient computationally, to obtain them by regression only on the independent variables.

To test the validity of this approximate covariance matrix in the conditions of our stimulation study, 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 $r.p.$ degrees of freedom, where r is the number of replications (here 10) and p is the number of regression coefficients estimated. We do this by forming

$$\sum (\boldsymbol{\beta} - \mathbf{b})^{\mathrm{T}} \mathbf{S}_W (\boldsymbol{\beta} - \mathbf{b}) / s_y^2,$$

where the summation extends over all replications.

The results are tabulated in Table 2 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 of the standard errors, 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 about the precision of the estimates. But we should point out that we have not tested the theory with 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.

TABLE 2

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

	5% Deletions		*10% Deletions*			*20% Deletions*			*40% Deletions*	
Problem	100 *obs.*	200 *obs.*	50 *obs.*	100 *obs.*	200 *obs.*	50 *obs.*	100 *obs.*	200 *obs.*	200 *obs.*	*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

6. PRACTICAL CONSIDERATIONS IN A MISSING VALUES PROGRAM

In this section we outline some features that we think are important for practical programs for missing variable analysis.†

The input parameters should include a maximum number of iterations for the calculation of corrected maximum likelihood estimates. A default option of 100 is suggested for this parameter, because, as Orchard and Woodbury remark, the convergence can be quite slow. For the random deletion patterns in our simulation study 10 iterations were often enough, but one problem required 171 iterations.

The finishing tolerance T_F has a standard value of 0·01. Iterations stop if the changes in successive values of $\bar{x}_j$ are all less than $T_F\sqrt{\hat{u}_{jj}}$ and the changes in successive values of $\hat{u}_{jk}$ are all less than $T_F\sqrt{(\hat{u}_{jj}\,\hat{u}_{kk})}$.

It is characteristic of modern multivariate analysis that one wants to be able to consolidate the data before deciding on the precise form of the analysis to be carried out. This poses some problems. We suggest allowing each variable to be coded in one of four alternative ways:

I or blank An independent variable.

D A dependent variable, to be fitted as a linear function of all independent variables.

ID An independent variable, but one which is also required to be fitted as a linear function of all other independent variables.

N A variable not to be used as either an independent or a dependent variable, but which is to be used to build up the maximum likelihood estimate of the covariance matrix of all variables.

† These are included in the Scicon program.

The overall covariance matrix for all variables is then estimated by corrected maximum likelihood, and the appropriate submatrices are extracted for any desired regression analyses. The standard errors of the resulting regression coefficients can then be estimated by the method described in Section 5. This approach may underestimate precision when missing variables are highly correlated with known variables that are excluded from the regression analysis, but it seems a reasonably safe procedure.

Another feature is concerned with the problem discussed by Woodbury (1971). If two variables are never observed together, then the data give no evidence about their conditional correlation, given all other variables. Woodbury then recommends setting this conditional correlation equal to zero. The iterative procedure converges to this solution, but incredibly slowly.

7. AN ANALYSIS OF SCHOOL EXAMINATION DATA

Our program has been used on some data kindly supplied by Dr Robert Wood and Miss Carolyn Ballantyne of the School Examinations Department of the University of London.

The background to the data is described by Dr Wood as follows:

> "In testing a wide ability range of candidates, it is thought that a single test of conventional length (50–60 questions) may fail to provide adequate discrimination between individuals at certain parts of the range, notably the extremes. To rectify this the idea of administering different tests to different candidates has been proposed where one or more elements are common. If, for instance, there are three tests of increasing difficulty, with the middle one common, the able candidates are encouraged to choose the two harder tests while the less able are pointed towards the two easier tests. Having administered the tests, the problem becomes one of placing all candidates on the same scale for the purpose of awarding grades.
>
> Recently we experimented with a two-stage test which had the following choice structure:
>
> | | *Test* | | | | | |
> |---|---|---|---|---|---|---|
> | | 1 | 2 | 3 | 4 | 5 | 6 |
> | *Strategy* | (10) | (25) | (15) | (10) | (15) | (25) |
> | *A* | × | × | × | | | |
> | *B* | × | | × | × | × | |
> | *C* | × | | | | × | × |
>
> There were six parts, the number of items in each part appears in brackets. Three choice strategies *A*, *B* and *C* were available, and each strategy is picked out with crosses. Although 100 questions were presented, each candidate was only required to tackle 50. The problem, then, is to estimate what scores the candidates who went for strategy *A* would have obtained on parts 4, 5 and 6, and then to cumulate actual and estimated scores to obtain an overall mark."

There were 321 observations, 73 following Strategy *A*, 188 Strategy *B* and 60 Strategy *C*. Tests 2–6 are in increasing order of difficulty.

We analysed the data on two different bases: once applying the corrected maximum likelihood method to the data as presented, the other time after making angular transformations on all the scores. The final results were very similar. The calculations using angular transformations converged after 92 iterations and required

50 seconds of CPU time on the Univac 1108 computer. Without angular transformations the calculations terminated after 100 iterations and had nearly but not quite converged. This run required 70 seconds of CPU time. The structure of the problem is revealed by the way the assumed means of the 6 variables changed as the iteration proceeded. The initial estimates of the means, based on the observations for which each variable was obtained, were

$$5{\cdot}72,\quad 20{\cdot}10,\quad 10{\cdot}12,\quad 5{\cdot}51,\quad 7{\cdot}23 \quad\text{and}\quad 11{\cdot}18.$$

The final estimates were

$$5{\cdot}72,\quad 21{\cdot}35,\quad 10{\cdot}57,\quad 5{\cdot}47,\quad 6{\cdot}64 \quad\text{and}\quad 6{\cdot}74.$$

This indicates that the fitted scores on the easier Tests 2 and 3 for the candidates who did not take them were somewhat higher than the scores obtained by the candidates who did take them. But the opposite effect is seen with the more difficult Tests 5 and 6. This is as it should be.

ACKNOWLEDGEMENTS

This work has been based on earlier work on regression with missing values, carried out at Imperial College and at Scientific Control Systems Limited. In particular we should like to acknowledge the benefit we have gained from studying the Imperial College M.Sc. Project Reports by M. J. Wallace in 1969 and D. Chant in 1970.

One of us (R. J. A. Little) acknowledges the receipt of a Research Studentship from the Science Research Council.

Finally, we are most grateful for many constructive criticisms of earlier drafts of this paper by Professor D. E. Barton.

REFERENCES

BEALE, E. M. L. (1970). Computational methods in least squares. In *Integer and Nonlinear Programming* (J. Abadie, ed.), pp. 213–227. Amsterdam: North Holland.

BUCK, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Statist. Soc.* B, **22**, 302–306.

HAITOVSKY, Y. (1968). Missing data in regression analysis. *J. R. Statist. Soc.* B, **30**, 67–82.

JOWETT, G. H. (1963). Application of Jordan's procedure for matrix inversion in multiple regression and multivariate distance analysis. *J. R. Statist. Soc.* B, **25**, 352–357.

KENDALL, M. G. and STUART, A. (1967). *The Advanced Theory of Statistics*, Vol. II, 2nd ed. London: Griffin.

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: theory and applications. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.*, Vol. I, pp. 697–715.

WOODBURY, M. A. (1971). Contribution to the discussion of "The analysis of incomplete data" by H. O. Hartley and R. R. Hocking. *Biometrics*, **27**, 808–813.