# A Method of Estimation of Missing Values in Multivariate Data Suitable for use with an Electronic Computer

S. F. Buck

*Journal of the Royal Statistical Society. Series B (Methodological)*, Volume 22, Issue 2 (1960), 302-306.

Stable URL:
http://links.jstor.org/sici?sici=0035-9246%281960%2922%3A2%3C302%3AAMOEOM%3E2.0.CO%3B2-G



*Journal of the Royal Statistical Society. Series B (Methodological)*

# A Method of Estimation of Missing Values in Multivariate Data suitable for use with an Electronic Computer

By S. F. Buck

*Rothamsted Experimental Station*†

[Received October 1959. Revised March 1960]

Summary

Estimation of statistical parameters from multivariate data results in wasted information, if units with incomplete data are rejected entirely, and perhaps in inconsistencies in the variance–covariance matrix if the variances and correlation coefficients are estimated from all available data on individual variates and pairs of variates respectively. An alternative is to estimate the missing values by regression techniques and to calculate a revised variance–covariance matrix. This method is suitable for use with an electronic computer. It is shown that with this method the resultant covariances are unbiased, but that the variances require correction for bias. A numerical example is given.

## 1. Introduction

In practice we frequently have situations where we wish to estimate statistical parameters from multivariate data which are incomplete.

The elimination from the sample of all those units for which one or more values are missing results in wasted information. Alternatively estimation of variances and correlation coefficients from all available data on individual variates and pairs of variates respectively, may result in an inconsistent variance–covariance matrix.

We can represent the sample of $n$ units, on each of which it is desired to have measurements of $k$ variates, by expressing the measurements, $x_{ij}$ $(i = 1, 2, ..., n;$ $j = 1, 2, ..., k)$, in the form of an $n \times k$ matrix, $\mathbf{X}$, in which some of the elements are missing.

Estimation of the parameters of a normal bivariate population with missing values (i.e. $k = 2$) has been considered by Wilks (1932) and Rao (1956, pp. 161–165), whilst Matthai (1951) considered the general multivariate normal population and gave results for the trivariate case. Their method leads to simultaneous estimation of the parameters of the bivariate or trivariate normal population by maximum likelihood solution. Their solutions are not explicit, and Matthai mentions that estimates obtained may prove to be inconsistent, e.g. a correlation coefficient greater than unity. More recently, Edgett (1956) found the maximum likelihood solutions in explicit form for the parameters from a trivariate normal population in the special case where some of the sample observations for one of the variates only are missing.

Anderson (1957) and Nicholson (1957) have considered the same problem and have indicated the best mathematical method to use in calculating the maximum likelihood estimates of parameters from an incomplete multivariate normal distribution.

† Work done as an Agricultural Research Council Scholar.

For $k>2$ the maximum likelihood solution could become too involved for practical purposes. Apart from this, however, there is the more fundamental objection that most multivariate data cannot be regarded as samples from multivariate normal distributions.

It is our purpose in the present paper to give a method of estimating the variance–covariance matrix of any $k$-variate population. The method is particularly suitable for use on an electronic computer, and is applicable to any number of variates. The method consists of estimating the missing values in the sample by regression techniques, using these values to calculate a revised variance–covariance matrix.

It is worth noting that the maximum likelihood method and the present method agree for the special case (regression method), considered by Yates (1953, Section 6.8), of estimating the mean of a variate when there is supplementary information on another variate.

## 2. Units with One Missing Value

Suppose that $m$ of the $n$ units have the complete set of $k$ observations recorded on them; we can consider these as forming the first $m$ rows of $\mathbf{X}$.

We calculate the expected value of $x_{rj}(r = 1, 2, \ldots, m)$, by forming, for each value of $j$, the multiple regression of the $j$th variable on the other $k-1$ variables from our set of observations consisting of the first $m$ rows of $\mathbf{X}$. That is, we obtain $k$ equations which can be expressed in the form

$$E(x_{rj}) = f_j(x_{r1}, x_{r2}, \ldots, x_{rj-1}, x_{rj+1}, x_{rj+2}, \ldots, x_{rk}). \tag{1}$$

In this case, we take the $f_j$ to be linear functions.

We use equations (1) to estimate the missing values in the following way. If the $i$th unit has the $j$th observation missing, we can estimate its value $x_{ij}$, by using one of the equations (1) substituting $x_{ij}$ for $x_{rj}$, i.e.

$$E(x_{ij}) = f_j(x_{i1}, x_{i2}, \ldots, x_{ij-1}, x_{ij+1}, x_{ij+2}, \ldots, x_{ik}). \tag{2}$$

## 3. Units with More than One Missing Value

We can extend the method to the case in which units have more than one missing value, as follows. If $v$ variates are missing, then from the first $m$ rows of $\mathbf{X}$ we require to calculate the multiple regression formula for each missing variate on $k-v$ other variates. If any combination of $v$ variates may be missing,

$$k\binom{k-1}{v-1}$$

possible equations have to be calculated, and we estimate a missing value by its expected value obtained from the correctly chosen regression equation.

These regression equations can be obtained fairly simply on an electronic computer, from the inverse of the single $k \times k$ information matrix derived from the first $m$ rows of $\mathbf{X}$. The technique adopted, which is described by Woolf (1951), allows any of the $k$ variates to be chosen as the dependent variate, and it is possible to add, remove, or replace independent variables in the equations.

## 4. Bias in the Variance–Covariance Matrix

We discuss below whether any bias is introduced into the variance–covariance matrix of the corrected data.

We can denote the variance–covariance matrix for the $m$ complete individuals by

$$\mathbf{A} = ((a_{sj})) = \begin{pmatrix} a_{11} & \boldsymbol{\alpha}_1' \\ \boldsymbol{\alpha}_1 & \mathbf{C} \end{pmatrix} \quad (s = 1, 2, \ldots, k). \tag{3}$$

Then the regression coefficients of $x_1$ on $x_2, x_3, \ldots, x_k$ are given by

$$\mathbf{b} = \boldsymbol{\alpha}_1' \mathbf{C}^{-1}. \tag{4}$$

We estimate the value of $x_1$ for those individuals which have $x_1$ only missing, by writing $\mathbf{x}' = (x_2, \ldots, x_k)$ and taking

$$\begin{aligned} \hat{x}_1 &= \mathbf{bx}, \\ &= \boldsymbol{\alpha}_1' \mathbf{C}^{-1} \mathbf{x}. \end{aligned} \tag{5}$$

Now the covariance of $\hat{x}_1$ with $x_2, x_3, \ldots, x_k$ is given by

$$\begin{aligned} \operatorname{cov}(\hat{x}_1, \mathbf{x}') &= \boldsymbol{\alpha}_1' \mathbf{C}^{-1} \mathbf{C} \\ &= \boldsymbol{\alpha}_1' \end{aligned} \tag{6}$$

and hence $\operatorname{cov}(\hat{x}_1, \mathbf{x}') = \operatorname{cov}(x_1, \mathbf{x}')$, i.e. the expected values of the covariance elements in $\mathbf{A}$ are the same for both actual and predicted values of $x_1$.

Now
$$\begin{aligned} \operatorname{var}(\hat{x}_1) &= \boldsymbol{\alpha}_1' \mathbf{C}^{-1} (\operatorname{var} \mathbf{x}) \mathbf{C}^{-1} \boldsymbol{\alpha}_1 \\ &= \boldsymbol{\alpha}_1' \mathbf{C}^{-1} \mathbf{C} \mathbf{C}^{-1} \boldsymbol{\alpha}_1 \\ &= \boldsymbol{\alpha}_1' \mathbf{C}^{-1} \boldsymbol{\alpha}_1. \end{aligned} \tag{7}$$

If we write
$$\mathbf{A}^{-1} = \begin{pmatrix} c_{11} & \mathbf{e}_1' \\ \mathbf{e}_1 & \mathbf{F} \end{pmatrix},$$

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

$$\begin{aligned} a_{11} c_{11} + \boldsymbol{\alpha}_1' \mathbf{e}_1 &= 1 \\ \boldsymbol{\alpha}_1 c_{11} + \mathbf{C}\mathbf{e}_1 &= 0. \end{aligned}$$

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

and consequently
$$\operatorname{var}(\hat{x}_1) = \boldsymbol{\alpha}_1' \mathbf{C}^{-1} \boldsymbol{\alpha}_1 = a_{11} - 1/c_{11}. \tag{8}$$

Therefore, if the value $x_1$ is missing for a proportion $\lambda$ of all units, and the predicted values are substituted and a new variance–covariance matrix calculated, then the expectations in this matrix are the same as they would be if there were no missing values, except for the variance $a_{11}'$ of $x_1$ for which we have in terms of expectations

$$a_{11}' = a_{11} - \lambda/c_{11}, \tag{9}$$

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

The general correction to this bias follows; that if $x_j$ is missing for a proportion $\lambda_j$ of all units, then the estimated variance $a_{jj}'$ obtained from the corrected sample is adjusted to

$$a_{jj}' + \lambda_j / c_{jj}, \tag{10}$$

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

## 5. EXAMPLE

As an example of the method, we will consider some data collected by D. Hibbert of the Central Laboratory, British Sugar Corporation, to investigate the relationship between the purity of sugar beet with quantitative measurements of some of its chemical constituents. The data consists of 72 samples of sugar beet on which were measured the four variates: purity measure, $x_1$; sugar content, $x_2$; noxious nitrogen content, $x_3$; $K_2O$ content, $x_4$.

To provide the example, 35 observations were picked randomly from the total 288 ($4 \times 72$) observations and these were taken to be missing. We were left with (i) 43 complete samples, (ii) 6 samples with $x_1$ missing, (iii) 7 samples with $x_2$ missing, (iv) 5 samples with $x_3$ missing, (v) 6 samples with $x_4$ missing, (vi) 2 samples with $x_1$ and $x_3$ missing, (vii) 2 samples with $x_3$ and $x_4$ missing and (viii) 1 sample with $x_1, x_3$ and $x_4$ missing.

In our notation $n = 72, k = 4, m = 43$.

To estimate the missing values we require:

For (ii) the regression of $x_1$ on $x_2, x_3, x_4$;
For (iii) the regression of $x_2$ on $x_1, x_3, x_4$;
For (iv) the regression of $x_3$ on $x_1, x_2, x_4$;
For (v) the regression of $x_4$ on $x_1, x_2, x_3$;
For (vi) the regression of $x_1, x_3$ on $x_2, x_4$;
For (vii) the regression of $x_3, x_4$ on $x_1, x_2$;
For (viii) the regression of $x_1, x_3, x_4$ on $x_2$.

These regression equations were obtained on the electronic computer from the 43 complete samples, the missing values were estimated and inserted into the incomplete data.

We can compare results obtained (1) from the complete data (which in this artificial case we know), with (2) those from the incomplete data consisting of the 43 complete units, (3) those obtained by estimating variances and correlations from all the available data, and (4) those obtained by the method given above.

For method 3 the means and standard deviations were estimated from the data available on individual variates, whilst the correlations were estimated from data on pairs of variates. For method 4 the variance–covariance matrix of the substituted values was adjusted by equation (10) above.

The results are given in Table 1. In this example there is little to choose between methods 3 and 4, both of which give very similar results to those obtained from the complete data. Both methods are clearly very much superior to method 2.

With the low correlations existing in this material it is not to be expected that the potential inconsistencies in the variance–covariance matrix will cause trouble in the estimation of partial regressions, etc.

The partial regressions of $x_1$ on the other variates are as follows:

(1) $x_1$ (complete data) $= 774{\cdot}38 + 0{\cdot}165x_2 + 0{\cdot}087x_3 - 0{\cdot}462x_4$;

(2) $x_1$ (complete units) $= 789{\cdot}38 + 0{\cdot}192x_2 - 0{\cdot}261x_3 - 0{\cdot}280x_4$;

(3) $x_1$ (all available data) $= 770{\cdot}40 + 0{\cdot}167x_2 + 0{\cdot}140x_3 - 0{\cdot}471x_4$;

(4) $x_1$ (substituted values) $= 781{\cdot}08 + 0{\cdot}161x_2 - 0{\cdot}044x_3 - 0{\cdot}349x_4$.

It will be seen that method (3) has in fact given slightly better agreement with method (1) than has method (4). A final assessment of the value of the alternative methods in different circumstances can only be made by extensive tests on more highly correlated material.

TABLE 1

*Comparison of the results obtained from* (1) *complete data*, (2) *complete units*, (3) *all available data*, (4) *substituted values*

| | *Mean values and standard deviations of a single observation* | | | | *Correlation coefficients* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $r_{12}$ | $r_{13}$ | $r_{14}$ | $r_{23}$ | $r_{24}$ | $r_{34}$ |
| (1) Complete data: 72 samples | 801·2 ± 14·72 | 202·9 ± 58·55 | 72·8 ± 4·66 | 27·3 ± 3·65 | +·648 | −·055 | −·101 | −·121 | +·023 | +·031 |
| (2) Complete units: 43 samples | 803·3 ± 15·27 | 211·3 ± 56·44 | 72·9 ± 5·03 | 27·3 ± 3·67 | +·716 | −·178 | −·090 | −·098 | +·009 | +·342 |
| (3) All available data (single values and pairs) | 801·3 ± 14·80 | 203·8 ± 58·58 | 73·0 ± 4·83 | 27·3 ± 3·69 | +·644 | −·118 | −·071 | −·224 | +·061 | +·132 |
| (4) Substituted values | 801·2 ± 14·56 | 204·3 ± 58·23 | 73·0 ± 4·82 | 27·3 ± 3·67 | +·641 | −·121 | −·061 | −·146 | +·044 | +·145 |

## REFERENCES


ANDERSON, T. W. (1957), "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing", *J. Amer. Statist. Assoc.*, **52**, 200–204.

EDGETT, G. L. (1956), "Multiple regression with missing observations among the independent variables", *J. Amer. Statist. Assoc.*, **51**, 122–131.

MATTHAI, A. (1951), "Estimation of parameters from incomplete data with application to design of sample survey", *Sankhyā*, **11**, 145–152.

NICHOLSON, G. E. (1957), "Estimates of parameters from incomplete multivariate samples", *J. Amer. Statist. Assoc.*, **52**, 523–526.

RAO, C. R. (1956), *Advanced Statistical Methods in Biometric Research.* New York: Wiley.

WILKS, S. S. (1932), "Moments and distributions of estimates of population parameters from fragmentary samples", *Ann. Math. Statist.*, **3**, 163–195.

WOOLF, B. (1951), "Computation and interpretation of multiple regressions", *J. R. Statist. Soc.*, B, **13**, 100–119.

YATES, F. (1953), *Sampling Methods for Censuses and Surveys.* (2nd ed.) London: Griffin.