MULTIVARIATE ANALYSIS OF INCOMPLETE DATA:

ONE MISSING VALUE

by

W. K. Sitonik and W. T. Federer

BU-770-M* May 1982

Abstract

Missing observations are a common occurrence in any data set, and even more so in multiresponse observations. It is natural, therefore, that methods to facilitate meaningful analysis of incomplete multiresponse data be developed. Most of the methods so far developed by various investigators to alleviate this problem have been those which utilize the regression analysis techniques and the maximum likelihood estimation. The most commonly used of these methods are reviewed and their corresponding shortcomings pointed out. Alternative methods are indicated and the circumstances under which they perform reasonably noted.

Some of the regression analysis techniques and maximum likelihood estimation are employed in bivariate and trivariate sets of data to estimate a missing observation. The amount of bias introduced by including an estimated value in the analysis of the observed values is also computed for each data set and method of estimation.

---

* BU-770-M in the Biometrics Unit Series, Cornell University, Ithaca, NY 14853.

MULTIVARIATE ANALYSIS OF INCOMPLETE OBSERVATIONS: ONE MISSING VALUE

by

W. K. Sitonik and W. T. Federer

BU-770-M May 1982

Missing values bedevil any data set, and yet this is a common occurrence in any data collection phenomena. An otherwise simple analysis, with straightforward interpretations, on a data set can become extremely complicated when there are some missing values. This problem is so extensive that a lot of effort has been directed into research on new techniques to handle missing observations. The missing observations case can be roughly categorized into any of the following groups:

1. Randomly missing values.
2. Planned missing values.
3. Missing values due to the observed values being discarded as spurious (outliers).
4. Missing values because the measured materials became mixed up.
5. Missing values because the required material is sensitive (sensitive questions in a questionnaire).

From the diversity of the missing value situations, it can be seen that any taxonomy for incomplete data cannot encompass all possible situations of missing values. In this study the first, and possibly the third cases, will be considered although no attempt will be made on methods of determining when an observation is an outlier, and can thus be discarded. Interest will center only on the case when an observation has been declared missing.

Procedures for treating missing data are well developed for the univariate experiments (Federer, 1955). The methods developed for the uniresponse

observations are mainly for two or more samples. Analysis techniques for the one-sample cases are sufficient to handle unequal numbers of observations per class; no techniques need to be developed for the one-sample uniresponse observations with a missing value.

In the multiresponse case, however, discarding a whole individual because of one missing value may not be desirable, so there is need to estimate the missing value and retain the individual with the completed measurements in the analysis (Helwig and Council, 1979).

Several cases of missing values have been investigated by different investigators in an attempt to develop a universally acceptable technique for computing missing values in the one-sample case. These efforts have not been conclusive due to the complex and varying relationship between observations on any given individual. Nonetheless, methods have been developed which can be used "satisfactorily" to compute the estimates of missing values.

A very common technique of dealing with a missing value in a set of multiresponse observations has been to remove any individual with a missing value from the analysis. This approach is surprisingly good for small values of p ($p \leq 4$) responses and near-singular correlation coefficient matrices ($|R|$ near zero). For increasing values of $|R|$, the efficiency of the method decreases (Chan et al., 1976). It should be noted, however, that this method is analogous to the classical procedure in experimental design of inserting "neutral" values in place of the missing ones (Haitovsky, 1968).

This method, however, is obviously unsatisfactory if many values are known for an incomplete observation and if, in particular, the variables known prove on analysis to be important for the study.

Alternatively, one can replace the missing value by the mean of all the available observations. Such an estimate would be more accurate than any

other estimate from a sub-sample of the observations. The only main discrepancy is that the responses may not be in the same units. So the only reasonable estimate would be the sample mean of the particular response with the missing value. Although this method might bias the estimates badly, it is argued that the gain in precision might over-compensate the bias (Haitovsky, 1968).

Another approach which has received a lot of attention in literature, resulting in several modifications, is the regression analysis approach. Afifi and Elashoff (1966) point out the possibility of regressing the response variable(s) on each of the remaining complete response variables; the resulting regression equation being used to estimate the missing value(s). The estimated values of the missing observation are then averaged to give the "true" estimate. This method, although it uses all the available information, becomes intractable when the efficiency of the method needs to be computed.

A more appealing regression technique was initiated by Buck (1960) and further developed and extended by Beale and Little (1975). The technique is to carry out a multiple regression of the response variable with a missing value on the remaining variables. The multiple regression function so obtained is then used to compute an estimate of the missing value. Chan, et al. (1976), in a simulation study, showed that the method is efficient only for correlated response variables. When the response variables are not highly correlated, the method does not do any better than omitting the incomplete observations.

The approach which is the main focus of this study assumes that the observed values come from some multivariate normal distribution whose parameters are to be estimated by the maximum likelihood method. Although this method is considered far more efficient than all the others, when the observations are normally distributed, the amount of computation involved is prohibitive (Anderson,

1957). For one or two missing observations the method is fairly tractable, and will be applied to one missing observation in the study.

Let there be n individuals observed, $\underset{\sim}{Y}_{n\times p}$, each with p response variables, with a common and unknown covariance matrix, $\underset{\sim}{\Sigma}_{p\times p}$ and mean $\underset{\sim}{\mu}$. Then, the joint density function for the n observations is given by

$$f(\underset{\sim}{Y}:\underset{\sim}{\mu}\underset{\sim}{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp - \tfrac{1}{2}(\underset{\sim}{Y}-\underset{\sim}{1}\underset{\sim}{\mu}')\underset{\sim}{\Sigma}^{-1}(\underset{\sim}{Y}-\underset{\sim}{1}\underset{\sim}{\mu}')'$$

when measurements are made on each individual. Estimates of the parameters in the likelihood function are obtained by equating the partial equations to zero and solving for the parameters. Since this is found in any statistics book treating maximum likelihood estimation, the derivation for the complete observation case is not included.

First, let us consider a bivariate case (X,Y), where X and Y are two response variables measured on each of n individuals. Let (X,Y) be bivariate normal with mean (μ_1,μ_2) and covariance matrix $\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$. The density function of the n pairs of observations is given by

$$f(\underset{\sim}{X},\underset{\sim}{Y};\underset{\sim}{\mu}_1\underset{\sim}{\Sigma}) = \frac{1}{(2\pi)^{n}|\Sigma|^{n/2}} \exp - \tfrac{1}{2}\begin{bmatrix} \underset{\sim}{X} - \mu_1\underset{\sim}{1} \\ \underset{\sim}{Y} - \mu_2\underset{\sim}{1} \end{bmatrix}' \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{X} - \mu_1\underset{\sim}{1} \\ \underset{\sim}{Y} - \mu_2\underset{\sim}{1} \end{bmatrix},$$

which can also be expressed as the product of the marginal density of X and the conditional density of Y given that X has been assigned a known value x. Thus the joint density function becomes

$$f(X,Y;\underset{\sim}{\mu}_1\Sigma) = \frac{1}{(2\pi\sigma_{11})^{n/2}} \exp - \frac{1}{2\sigma_{11}} (\underset{\sim}{X}-\mu_1\underset{\sim}{1})'(\underset{\sim}{X}-\mu_1\underset{\sim}{1}) \cdot \frac{1}{(2\pi(1-\rho^2)\sigma_{22})^{n/2}} \exp$$

$$- \frac{1}{2(1-\rho^2)\sigma_{22}} (\underset{\sim}{Y}-\mu_2^*\underset{\sim}{1})'(\underset{\sim}{Y}-\mu_2^*\underset{\sim}{1})$$

where

$$\underset{\sim}{\mu} = (\mu_1,\mu_2)' \quad \text{and} \quad \mu_2^* = \mu_2 + \hat{\rho}\,\frac{\hat{\sigma}_2}{\hat{\sigma}_1}(X-\mu_1) .$$

Note also that σ_{11} and σ_{22} are estimated from n and n-1 observations respectively.

Suppose that a random sample of observations from the above bivariate normal distribution is as given below, with one missing value on the random variable Y:

X	36	51	53	23	19	34	24	65	44	31	29	58	37	46	50	44	56
Y	54	99	64	60	71	61	54	77	81	93	93	*	76	96	77	93	95

The maximum likelihood estimates of $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$ are:

$$\hat{\underset{\sim}{\mu}} = (41.77,77.75)' \quad \text{and} \quad \hat{\underset{\sim}{\Sigma}} = \begin{bmatrix} 180.529 & 80.500 \\ 80.500 & 250.200 \end{bmatrix}$$

respectively. From the above equations, it can be seen that the estimate of the missing y value is given by:

$$\hat{y} = \hat{\mu}_2 + \hat{\rho}\,\frac{\hat{\sigma}_2}{\hat{\sigma}_1}(X-\hat{\mu}_1) = 77.750 + 0.388 \times \frac{15.315}{13.144} \times 16.824 = 85.356 ,$$

that is the adjusted mean of the y values. This can be compared with the value of 86.091 from the simple linear regression or the actual observed value of 51.

It should be pointed out that when determining the estimates of the unknown parameters in μ and Σ for the maximum likelihood method, the rest of the observations on the incomplete individual are used. This is not the case in any of the regression methods except when computing the actual estimate.

As pointed out earlier, the inclusion of an estimated value in the analysis of the data set introduces some bias which is highly dependent on the method of estimation. In the present data set, the correlation coefficient, r of the complete observations is 0.21. When the same value is computed with one observation treated as missing and an estimated value substituted, it was shown that $r = 0.36$ and 0.41 for the simple linear regression and the maximum likelihood methods respectively.

An attempt to quantify the amount of bias thus introduced was made by Buck (1960) for one missing value and the regression method of estimation. This method will be used to determine the amount of bias introduced by the two estimated values in our example.

Let the variance matrix of the observed values be given as Σ above. Then, using Jordan's method of solving simultaneous equations and inverting matrices (Fox, 1954), the regression coefficient of Y on X is given by

$$b = \sigma_{12}/\sigma_{22}$$

so that the missing value, $\hat{Y} = bX$ and the covariance of Y and X remains unchanged for both observed predicted values.

Denote the inverse of the variance-covariance matrix of X and Y by

$$\Sigma^{-1} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} .$$

Then, using the fact that $\Sigma\Sigma^{-1} = I$ we have $\sigma_{21}c_{11} + \sigma_{22}c_{21} = 0$ and $\sigma_{21}c_{12} + \sigma_{22}c_{22} = 1$; thus, $c_{22} = (\sigma_{22} - \sigma_{21}\sigma_{11}^{-1}\sigma_{12})^{-1}$. Note that $\text{Var}(\hat{y}) = \sigma_{21}^2/\sigma_{11} = \sigma_{22} - \frac{1}{c_{22}} = \sigma_{22}'$.

Thus when an estimated value is substituted for a missing value y, the computed variance-covariance matrix would be the same as if there were no missing values except for the necessary adjustment on the variance of Y given above. For the estimates computed in the bivariate normal set of data, the estimated bias in the covariance matrices are 15.63 and 11.55 for the simple linear regression and maximum likelihood estimates respectively. All the computed estimates of the variance of the response variable Y underestimated the variance, with the observed value, (Var(y) = 276.65) substantially.

Similar computations, as those done for bivariate random variables, can be done for cases where more than two response variables are measured. A trivariate case will be considered to show a passible generalization of the above computation techniques to p-responses per individual with one missing value. The same data as in the foregoing example will be used, except an additional response variable will be introduced. Trivariate data is as follow:

X	53	23	19	34	24	65	44	31	29	58	37	46	50	44	56	36	51
Y	64	60	71	61	54	77	81	93	93	*	76	96	77	93	95	54	99
Z	0.4	0.4	3.1	0.6	4.7	1.7	9.4	10.1	11.6	12.6	10.9	23.1	23.1	21.6	23.1	1.9	29.9

The trivariate normal distribution for the three variables is given by:

$$f_{\underset{\sim}{\theta}}(X,Y,Z) = \frac{1}{(2\pi)^{2n/2}|\Sigma|^{n/2}} \exp - \tfrac{1}{2} \begin{bmatrix} \underset{\sim}{X} - \mu_1\underset{\sim}{1} \\ \underset{\sim}{Y} - \mu_2\underset{\sim}{1} \\ \underset{\sim}{Z} - \mu_3\underset{\sim}{1} \end{bmatrix}' \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{X} - \mu_1\underset{\sim}{1} \\ \underset{\sim}{Y} - \mu_2\underset{\sim}{1} \\ \underset{\sim}{Z} - \mu_3\underset{\sim}{1} \end{bmatrix}$$

$$= f(\underset{\sim}{X},\underset{\sim}{Z})f(\underset{\sim}{Y}|\underset{\sim}{X},\underset{\sim}{Z})$$

$$= f(\underset{\sim}{X},\underset{\sim}{Z}) \cdot \frac{1}{(2\pi)^{3n/2}\sigma_{Y\cdot XZ}^{n}} \exp - \frac{1}{2\sigma_{Y\cdot XZ}^{2}} (\underset{\sim}{Y}-\mu_{2\sim}^{*1})'(\underset{\sim}{Y}-\mu_{2\sim}^{*1})$$

where

$$\mu_2^* = \mu_2 + \beta_{YZ\cdot X}(Z-\mu_3) + \beta_{YX\cdot Z}(X-\mu_1)$$

$$\sigma_{23} = \sigma_{33}\beta_{YZ\cdot X} + \sigma_{13}\beta_{XY\cdot Z}$$

$$\sigma_{12} = \sigma_{22}\beta_{YZ\cdot X} + \sigma_{11}\beta_{YX\cdot Z}$$

$$\sigma_{Y\cdot XZ}^2 = \sigma_{22} - (\beta_{YZ\cdot X}^2\sigma_{33} + 2\beta_{YZ\cdot X}\beta_{YX\cdot Z}\sigma_{13} + \beta_{YZ\cdot X}^2\sigma_{11}) \ .$$

Therefore, the required estimated value of the missing observation can be computed as the conditional mean of y as given below:

$$\hat{y} = \hat{\mu}_2 + \hat{\beta}_{YZ\cdot X}(Z-\hat{\mu}_3) + \hat{\beta}_{YX\cdot Z}(X-\hat{\mu}_1)$$

$$= 77.75 + 0.5088(12.6-10.4824) + 0.1549(58-41.1765)$$

$$= 81.43 \ .$$

The variance of y, Var(y) = 250.20 and with the estimated value of the missing observation Var(y) = 235.36. Thus the bias introduced by inclusion of the estimate in the data analysis is 83.17 and the adjusted variance of y is 167.03. In most cases, inclusion of the estimate in the data analysis results in a non-positive definite covariance matrix. Engelman (1981) has developed a procedure which can convert such a matrix to a positive definite matrix. In the article, however, it is not indicated whether the adjustment so imposed on the covariance matrix has some adverse effects on the results.

It can be seen that for large values of responses, p the amount of computation in using the maximum likelihood estimation becomes extremely large and may lead to results with large rounding errors. An alternative method has been developed by Orchard and Woodbury (1972) called the Missing Information Principle. The method, which does not assume any distributional properties, gives the same estimates as the maximum likelihood estimates when the observations are multivariate normal. This procedure has been adapted to a computer program in the MULTIMISS subroutine of GENSTAT package (GENSTAT, 1978).

REFERENCES

Afifi, A. A. and Elashoff, R. M. (1966). Missing observations in multivariate statistics. I. Review of the literature. J. Amer. Statist. Assoc. 61, 595-604.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. J. Amer. Statist. Assoc. 52, 200-203.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. J. Roy. Statist. Soc. Ser. B, 37, 129-145.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electric computer. J. Roy. Statist. Soc. Ser. B, 22, 302-307.

Chan, L. S., Gilman, J. A. and Dunn, O. J. (1976). Alternative approaches to missing values in discriminant analysis. J. Amer. Statist. Assoc. 71, 842-844.

Engelman, L. (1981). An efficient algorithm for computing covariance matrices from data with missing values. Communications in Statistics, Simulation and Computation 11, 113-121.

Federer, W. T. (1955). Experimental Design - Theory and Application. McMillan, New York (Reprinted by the Oxford and IBH Publishing Company, Calcutta, 1967).

Fox, L. (1954). Practical solution of linear equations and inversion of matrices. Nat. Bur. Stand. App. Math. Ser. 39, 1-54.

GENSTAT (1977). A general statistical program. Rothamsted Experimental Station.

Haitovsky, Y. (1968). Missing data in regression analysis. J. Roy. Statist. Soc. Ser B, 30, 67-82.

Helwig, J. T. and Council, K. A. (ed.). (1979). SAS User's Guide.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle. Theory and Applications. Theory of Statistics 1. Sixth Berkeley Symp. on Math. Statist. Prob., University of California Press.