MULTIVARIATE EDIT AND IMPUTATION FOR ECONOMIC DATA

R.J.A. Little and Philip J. Smith, U.S. Bureau of the Census

1. Introduction

The quality control of data has become an increasingly important aspect of survey work. Non-sampling errors affecting the integrity of the data are possible at virtually every junction in a survey where data are communicated or transcribed from one person or devise to another. Without quality control of survey information, data intended for final analysis or tabulation and publication can be spurious or missing. In this case, analysis and publication of such information may be of dubious value and may jeopardize the credibility of the organization conducting the survey and preparing the analysis and report: bad data must be edited and values imputed when they are missing or have been deleted during the editing process.

The next section describes and edit/imputation procedure based on an extension of the method developed by Frane (1978) for identifying outlying cases, which is available in the BMDPAM Procedure in the BMDP statistical computer package (BMDP, 1979). The method is illustrated using selected data from the Annual Survey of Manufactures (ASM), with variables measured on the logarithmic scale. An important limitation of the method as applied to business surveys is the failure to take account of linear constraints between the variables measured on the raw scale, such as the requirement that a set of component items sum to their aggregate. Section 3 outlines a modification of the basic method which takes into account linear constraints, and discusses further work required to implement the procedure in a realistic setting.

2. The Basic Method

Our method is designed to analyze a (nxp) rectangular data matrix containing n observations on p variables $X_i = (X_{i1},\ldots,X_{ip})$, $i=1,\ldots,n$. Some values in the matrix are assumed to be missing because they were not recorded or have been edited out by other edit routines.

Figure 1 gives a simple flow chart for our prototype edit/imputation procedure for incomplete multivariate data. In the first step the missing data pattern is analyzed and observations and variables are rearranged according to the pattern of missing and observed values in each case. In the second step the program identifies cases which are apparently outliers. The third step identifies outlying variables within the outlying cases obtained from the previous step. Finally, imputed values are calculated for the missing or outlying values, and edited estimates of the mean and covariance matrix of the variables are produced.

We illustrate our procedure using data from the 1982 Annual Survey of Manufactures (ASM). It should be noted that the procedure we detail is not a comprehensive solution to all editing and imputation requirements of the ASM. Rather, our method is a general statistical tool which requires adaptation to the specific editing and imputation requirements of a survey. In Section 3 we describe of the ASM's particular editing and imputation requirements and sketch how our method might be adapted to meet them. The data we consider consists of a sample of 155 cases from a particular industry surveyed by the ASM.[1] Fourteen variables were selected from the data, seven current year variables and the seven corresponding variables from the preceding year. These variables are the number of production workers (PW), the number of all other employees (OE), legally required fringe benefits paid (LE), voluntary payments to fringe benefit programs (VP), total man hours worked by production workers (MH), production workers' wages (WW), and all other salaries and wages (OW). When these mnemonic names are prefaced by an 'A' they denote current year data. If they are prefaced by a 'B' they denote preceding year data.

Table 1 illustrates the first step of the program. Rows and columns of the data are rearranged to group similar patterns of missing data together. This not only clarifies the pattern of missingness, but also reduces the computation required to perform the editing and imputation.

A useful preliminary to outlier detection is to display the marginal distributions of the observed values of the variables. One might consider applying methods for univariable outlier detection to the marginal distributions of the variables. However, there are two reasons why this is not appropriate for industry data. Firstly, the distributions of all of the variables are known to be skewed. Thus, methods based upon normality cannot be applied without a preliminary transformation of the data. Secondly, even if an appropriate transformation can be applied, outliers based on transformed data are plausibly valid members of the underlying population, at least in the context of ASM data. Existing criteria for outlier detection in the ASM are based on relationships between variables, through the requirement that ratios of pairs of variables lie within specified limits. Thus a statistical method of outlier detection needs to focus on multivariate relationships rather than on marginal characteristics of the variables.

Step 2 of the flow diagram in Figure 1 proceeds to estimate the mean and covariance matrix $(\mu_c, \sum_c)$ of the variables. Because the data contain missing information, the EM algorithm is used to obtain maximum likelihood estimates of μ and $\sum$, assuming the observations are multivariate normally distributed[2] (Orchard and Woodbury, 1972; Beale and Little 1975; Dempster, Laird and Rubin, 1977). However, because these estimates are contaminated by outlying data, we subscript them with a "c" to denote this characteristic.

In our illustration, were transformed to the natural logarithm (ℓn) scale to remove the skewness of the marginal distributions. A more fundamental reason for the log transformation is that imputation procedures for business data are often based on ratio estimates. For example, if it is required to impute for A on the basis of a correlated field B and prior year values A' and B' of A and B, then we may impute for A as follows: A

= (A'/B') B. Taking logarithms, this imputation relation becomes linear in the logarithms of A, A', B' and B. Our imputation procedure can be viewed as a generalization of this kind of edit where a linear relationship between ℓn A, ℓn A', ℓn B', and ℓn B and other available predictors are empirically determined by regressions based on available data.

The estimated means and covariance matrix (μ_C, $\sum_C$) are presented in Table 2. The EM algorithm also supplies imputed values X_{ij} for missing values in the data matrix during the final E step. These imputed values are predicted conditional means from the regression of the missing variables on the observed variables in each case, and they are useful in subsequent steps of the algorithm.

The next step of the program identifies outlying cases. For the ith case, the Mahalanobis distance

$$D_i = (X_i - \mu_C)^T \sum_C^{-1} (X_i - \mu_C)$$

is calculated, where X_i is the vector of observed or imputed values for the ith case. D_i represents the distance of case i from the centroid of observations and accounts for distance for variables that are present, only.

In the absence of missing values and contaminated data, and under the multivariate normal assumption, $(n-p)nD_i/[(n-1)(n+1)p]$ has an F distribution with p and (n-p) degrees of freedom (Anderson, 1958; Hawkins, 1974). If the data are incomplete the exact distribution of D_i is unknown, but replacing μ_C and $\sum_C$ by μ and $\sum$, respectively, it can be shown that D_i is asymptotically chi-squared with p_i degrees of freedom under the model, where p_i is the number of present variables in unit i. To take some account of the fact that μ and $\sum$ are estimated, we conjecture that approximately $(n_C-p_i)n_C D_i/[(n_C-1)(n_C+1)p_i]$ has an F distribution with p_i and (n_C-p_i) degrees of freedom, where n_C is the number of complete observations. This conjecture has little theoretical basis, but does well in simulation studies for a related problem discussed by Little (1979). Case i may be identified as being outlying if D_i is too large. The program computes D_i for each case and using the above distributional results determines if case i is outlying.

This identification step sorts cases into two different groups: those that are outlying and those that are less likely to be outlying. A p-value of 0.01 or less for the Mahalanobis distance was used as a criterion for determining the outlying cases.

All outlying cases are not necessarily found by this procedure, since as noted above, estimates of μ_C and $\sum_C$ are contaminated. However, we expect that the most influential cases are identified. Therefore, better estimates of μ and $\sum$ can be obtained from the data cases identified as being inlying. Using the EM algorithm we obtain new estimates $\tilde{\mu}$ and $\tilde{\sum}$ which are contaminated to a lesser degree than μ_C and $\sum_C$. Also, new imputations for missing values are computed from these estimates.

In the next step the "improved" estimates are used to identify outlying variables within outlying cases. To do this, the following algorithm is implemented:

STEP 1: For each outlying case i, compute for each observed variable k

$$D_i^{(k)} = (X_{i(k)} - \tilde{\mu}_{(k)})^T \tilde{\sum}_{(k)}^{-1} (X_{i(k)} - \tilde{\mu}_{(k)}),$$

the Mahalanobis distance with variable k omitted. This distance shows the effect of eliminating k in computing the distance of the observation from the mean. If variable k is the only outlying value in this observation, then $D_i^{(k)}$ will be significantly smaller than D_i.

STEP 2: $\min_k D_i^{(k)}$ is determined:

The single most influential variable contributing to the extremity of observation i is found. Let us call this variable j_1. By removing j_1, the contribution of observation i to the probability of $X_{i(k)}$ is increased the most.

STEP 3: Compute $D_i^{(kj_1)}$, the Mahalanobis distance with both variable j_1 and k removed, for all present variables $k \neq j_1$.

STEP 4: Determine $\min_k D_i^{(kj_1)}$

The variable minimizing $D_i^{(kj_1)}$ is the next most influential variable, conditional on the removal of variable j_1 in step 2. Let j_2 denote this variable. The algorithm then proceeds to find j_3, the next most influential variable, conditional on the prior removal of variables j_2 and j_1, and so on until all the present variables in observation i are exhausted.

Table 2 gives a summarization of this algorithm for one of the outlying cases in Table 6. The total distance computed using $\tilde{\mu}$ and $\tilde{\sum}$ for this case is $D^2 = 119.3$ which corresponds to a p-value much less than 0.001.

Removal of the most influential variable, BPW reduces the distance to 53.79, a 54.92% decrease. If BPW was the only outlying value then by removing it the P-value associated with $D_{65}^{(BPW)}$ would be, at least, moderately large. However, on removing BPW, the p-value does not improve greatly (< 0.001) and consequently we are led to search for other outlying variables in the case.

Conditional on removing BPW, BWW is the next most influential variable. Removing it yields a remaining distance of $D_{65}^{(BPW,BWW)}$ with an insignificant p-value ($p > 0.01$). Consequently, we stop our search here having identified two outlying variables, BPW and BNW.

The next step in the program described in Figure 1 is to "edit" these outlying variables in outlying cases: we now treat them as if that data had been missing.

In the final step μ and $\sum$ are re-estimated via the EM algorithm and the Beaton Sweep operator is used in conjunction with these estimates to produce regressions of missing variables on non-missing variables for each case. Missing

values are then imputed from these regressions and a clean data set is produced.[3]

3. Discussion

As a general statistical tool, the procedure presented in Section 2 is deficient in two main respects. Firstly, the procedure for identifying outlying cases is based on contaminated estimates $(\mu_c, \sum_c)$ of the parameters. An iterative version of the method may appear justified, but by analogy with univariate procedures, we expect such a method to be unstable unless the p-value for identifying outlying cases is carefully chosen. We have developed (and will report elsewhere) a robust procedure for estimating $(\mu, \sum)$ from all available data which alleviates this problem.

A second difficulty with the procedure is the choice of p-values for selecting outlying cases and outlying values within cases. Rather conservative p-values are suggested, to avoid excessive editing of the data and consequent under-estimation of the variances of the variables. However, rules for the choice of p-value as a function of the amount of data do not appear to be easily derivable. Furthermore, it could be argued that the p-value is nothing more than a poor proxy for alternative criteria based on the distributions of the partial Mahalanobis distances described in the variable selection part of the procedure. These distributions are intractable, but could be approximated by simulation methods.

Considerable modifications of the basic method would be necessary to provide a flexible editing tool for ASM data. In particular, it should be recognized that the imputation procedure we describe is essentially empirical, and does not take into account *a priori* information about relationships between variables. Also, existing procedures for editing ASM data exploit exact linear relationships such as identities relating an aggregate variable to the sum of its parts. For example, the total wages paid, SW, must equal WW + OW, the sum of the wages paid to production workers and wages paid to other workers. The modifications required in our program to solve this important problem should reflect the nature of the linear constraints in a particular problem. Specifically, two issues need particular attention: (a) Does the fact that the linear constraint is satisfied by the recorded values increase one's confidence in their validity? If two or more independent data sources are involved in the recorded values, then the answer to this question is probably yes. On the other hand, if the total is obtained by summing the individual components, or one of the components is found by subtracting the other components from the total, then the satisfaction of the constraint simply confirms the arithmetic and does not confer any particular validity to the recorded values; (b) Is the total more reliably recorded than its components? In the ASM context, the SW variable is obtained from official records and is regarded as more trustworthy than the values of other variables.

Regardless of the answers to (a) and (b), we suggest that one of the variables in each constraint is dropped from our algorithm to avoid problems of near-collinearity. At the conclusion of the algorithm, the value of the omitted variable is changed if necessary to satisfy the linear edit constraint. If the linear edit constraint is not satisfied by the unedited variables, we propose an additional editing procedure before the variable selection algorithm for identifying outlying variables in cases. In this procedure the variable in the linear constraint is changed that results in the smallest Mahalanobis distance for the case. The total might change in this procedure; this would not be allowed to happen if the answer to (b) is yes.

If the answer to (a) or (b) are yes, then further improvements to the algorithm can be achieved by assigning priority levels to the variables in the stepwise variable selection procedure. If (a) is answered as yes and the linear constraint is satisfied by the unedited values then the variables in the constraint are assigned lower priority for selection than other variables. If (b) is answered as yes, then the total should replace one of its components as a variable in the algorithm and assigned low priority for selection. These rules require straightforward modifications to handle data where some of the components of the linear constraint are missing.

Strong *a priori* knowledge of limits for ratios of variables determined by ASM subject matter specialists is not exploited in our procedure. The stepwise selection of variables could be modified to give variables that fail to satisfy these *a priori* constraints higher priority for editing than others. Such modifications might be viewed as approximations to a Bayesian analysis. The development of more formal Bayesian procedures, which incorporate prior information into prior distributions for the parameters, appears to be a challenging statistical problem worthy of attention.

A final practical concern is the utility of imputations derived by the method. The imputed value for a missing or deleted variable in a particular observation is the conditional mean given the present variables in that observation, which is a linear function of the present values. As such, it has the important property of capturing relationships between the missing and present variables in the observation, as represented by their estimated covariance matrix. Alternative imputed values could be developed which add a random quantity to the mean to preserve distributional characteristics. (Schieber, 1978, and Little and Samuhel, 1983, discuss implementations for a single missing variable.) Imputed values may require modification to satisfy constraints between the variables. Also, the option of retaining outlying values which are considered valid by subject matter specialists would also be a feature of an operational edit/imputation system. For example, prior year data in the ASM might be treated as correct if they have passed editing constraints when the data was collected.

As noted in footnote 1, another aspect of ASM data not covered by the proposed procedure is the presence of zeros for some variables, corresponding to the absence of that variable in particular cases. The logarithmic transformation is not appropriate for such cases,

and they do not fall within the multivariate normal framework which underlies the method. Special procedures are required for editing these cases.

In short, further work is required to resolve these deficiencies. Nevertheless, the method as presented appears to be a useful, if crude, tool for editing incomplete multivariate data.

FIGURE 1

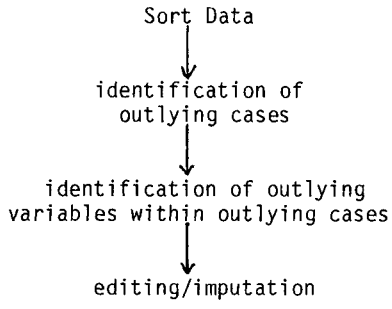

TABLE 1. Rearrangement of Data and Missing Data Pattern Analysis

Code	Variable Name	Case Number	Pattern ABCDEFGHIJKLMN
A	BPW	31	--------M-----
B	BWW	121	---------M----
C	BLE	92	----------M---
D	BVP	44	----------M---
E	BMH	78	----------M---
F	APW	133	----------M---
G	AWW	35	-----------M--
H	BOW	153	--------MM----
I	ALE	70	--------MM----
J	AVP	42	-----M--MM--M-
K	AMH	43	-------M---MMM
L	BOE	53	-------M---MMM
M	AOE	40	-------M---MMM
N	AOW	55	-------M---MMM
		41	-------M---MMM
		29	------M-MMM--M
		25	------M-MMM--M
		152	MMMMM--M---M--
		146	-----MM-MMM-MM

TABLE 2. Estimated Means and Covariances of an Transformed Data from Estimation in Step 2, ASM Data

ESTIMATED MEANS

BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
5.13	7.63	5.72	5.48	5.78	5.08	7.65	6.15	5.77	5.57	5.71	3.19	3.15	6.23

ESTIMATED COVARIANCE MATRIX

	BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
BPW	.97													
BWW	.97	1.09												
BLE	.94	1.08	1.20											
BVP	1.03	1.24	1.28	1.81										
BMH	.87	.90	.91	.98	.85									
APW	.96	.97	.97	1.05	.87	1.04								
AWW	.91	1.02	1.02	1.17	.85	.95	1.01							
BOW	.63	.74	.86	1.05	.63	.64	.72	1.13						
ALE	.89	1.00	1.08	1.18	.84	.94	.98	.81	1.05					
AVP	1.03	1.26	1.31	1.79	.98	1.09	1.23	1.07	1.22	1.95				
AMH	.89	.92	.96	1.03	.86	.98	.92	.66	.91	1.07	1.01			
BOE	.69	.72	.83	.95	.66	.70	.69	1.00	.80	.97	.69	1.18		
AOE	.72	.78	.89	1.03	.71	.74	.75	1.05	.86	1.07	.73	1.16	1.27	
AOW	.71	.82	.93	1.11	.69	.73	.80	1.13	.90	1.18	.72	1.06	1.19	1.32

TABLE 3

CASE NUMBER 65 TOTAL DISTANCE = 119.32 P-VALUE = .000

VARIABLE	RECORDED VALUE(RAW SCALE)	IMPUTED VALUE	RANK	INCREMENTAL DECREASE IN DISTANCE	DISTANCE REMAINING	P VALUE
BPW	143	74	1	54.92	53.79	.000
BWW	473	828	2	75.94	28.71	.014
ALE	209		3	85.83	16.91	.166
BOE	3		4	90.01	11.92	.355
BOW	58		5	95.86	4.94	.860
BVP	36		6	97.20	3.34	.921
AMH	140		7	97.84	2.58	.928
AVP	49		8	98.13	2.23	.905
BLE	99		9	98.34	1.98	.859
AWW	887		10	98.44	1.86	.769
AOW	117		11	98.64	1.62	.662
AOE	6		12	99.48	.61	.739
BMH	162		13	99.51	.59	.446
APW	81		14	100.00	.00	1.000

REFERENCES

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons: New York.

Beale, E.M.L. and Little, R.J.A. (1975). Missing Values in Multivariate Analysis. Journal of the Royal Statistical Society, Series B, (b) 37, pp. 129-146.

Beaton, A. E. (1968). The Use of Special Matrix Operators in Statistical Calculus. Research Bulletin RB-64-51. Education Testing Service: Princeton, New Jersey.

BMDP (1981) BMDP Statistical Software, 1981 edition. Los Angeles: University of California Press.

Dempster, A.P., Laird, M., and Rubin, D.D. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39, pp. 1-38.

Frane, J.W. (1978). Methods in BMDP for Dealing with Ill-Conditioned Data-Multicollinearity and Multivariate Outliers. Proceedings of the Computer Science and Statistics, A.R. Galland and T.M. Gerig, editors. North Carolina State University, P.O. Box 5457, Raleigh, North Carolina.

Hawkins, D.M. (1974). Detection of Errors in Multivariate Data Using Principal Components. Journal of the American Statistical Society, 69 (346), pp. 340-344.

Little, R.J.A. (1979). Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study. Journal of the Royal Statistical Society, Series B, 41, pp. 76-87.

Little, R.J.A. and Samuhel, M.E. (1983). Alternative Models for CPS Income Imputation. Paper for American Statistical Association Meetings, Toronto, Canada, August 1983.

Orchard, T. and Woodbury, M.A. (1972). Missing Information Principle: Theory and Application. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. pp. 697-715.

Schieber, S.J. (1978). A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of Low-Income Aged and Disabled. American Statistical Association 1978, Proceedings of the Survey Research Methods Section, 212-218.

FOOTNOTES

1 Our procedure only applies to cases which have positive values for the variables under study. Three cases were excluded that include zeros for one or more items. Special editing procedures are required for these cases.

2 The estimates are consistent for μ and Σ under any underlying distribution with finite fourth moments (Beale and Little, 1975). Thus the multivariate normality assumption is not essential for the utility of the method.

3 To limit instability in the regression estimates caused by near collinearity of the predictors, a potential predictor is not swept into the regression if its variance conditional on current predictors falls below a certain tolerance, chosen to be one percent of its unconditional variance. This practice is standard in most modern regression algorithms.