MISSING-DATA ADJUSTMENTS FOR PARTIALLY-SCALED VARIABLES

Roderick J.A. Little and Hong-Lin Su
Department of Biomathematics, U.C.L.A.

1. INTRODUCTION

Missing data is a pervasive problem in sample surveys. For a general review of the problem, see Madow et al. (1983). Two common strategies for dealing with the problem are direct analysis of the incomplete data and imputation. In the first approach, the missing values are left as gaps in the data set, identified by special missing data codes, and the treatment of missing data is deferred to the analysis stage. Given data in this form, most statistical analysis packages discard cases that contain incomplete information (complete-case analysis) or restrict attention to cases where the variable of interest is observed (available-case analysis). More elaborate approaches model the incomplete data and apply methods such as maximum likelihood (ML) (Little 1982; Little and Rubin 1987, Part II)).

Imputation creates a rectangular data set convenient for subsequent analysis, by replacing missing values by estimates based on the recorded information in the incomplete record. For reviews of imputation methods see Kalton and Kasprzyk (1982, 1986), Sande (1982), Madow et al. (1983, Vol 2), Little (1987) and Little and Rubin (1987, Chapters 4 and 12).

Missing-data methods for continuous variables and for categorical variables have an extensive literature: see Little and Rubin (1987) for review and references. In this paper we consider imputation and ML methods for handling multivariate missing data for a particular type of mixed continuous and categorical variable that has received little attention in the literature. Pregibon (1977) thought the variables we consider so common he called them "typical"; in an attempt at a more descriptive name, we call them "partially-scaled".

Partially-scaled variables consist of a binary variable indicating presence or absence of an attribute, and a continuous variable (usually positive) indicating an amount if the attribute is present. In symbols, Z = (R,A) where R=1 or 0, and A=0 if R=0, A = real number if R=1. For example, Z = income from a job, where R indicates job status and A indicates earnings. Other income types (such as social security), cost of time off from illness, crime victimizations where A is a measure of crime severity, are other examples of partially-scaled variables. For convenience we call R the recipiency variable and A the amount variable, although this descriptive terminology is not appropriate for all settings.

Nonresponse adjustments for a single partially-scaled variable has been considered in the literature; a common approach is considered in Section 2 below. Our purpose is to develop methods for a set of incomplete partially-scaled variables, as occurs in the following example.

Example 1. Monthly Income Data from a Panel Survey.

We consider item nonresponse to income in a large panel survey conducted by the U.S. Bureau of the Census, the Survey of Income and Program Participation (SIPP). For an overview of the SIPP see David (1985). One panel consists of about 30,000 individuals interviewed nine times over a period of two and a half years, at four month intervals. Detailed information on income and wealth is collected, yielding a large data base with over a thousand variables. We restrict attention to the variable WS = monthly wages and salary of first job. If interest is in annual information, then 12 partially scaled variables $WS_1,\ldots,WS_{12}$ are involved, where WS_j is the value of WS for month j, and consists of the recipiency indicator R_j and the amount A_j. The variables are recorded over three waves of the survey, WS_1-WS_4 in wave 1, WS_5-WS_8 in wave 2 and WS_9-WS_{12} in wave 3.

Missing data arise through item nonresponse, where an interview is conducted but recipiency and/or amount of one or more of the monthly WS variables are missing for a wave, and through wave nonresponse, where recipiency and amount are missing for all four months in a wave because the interview for that wave was not conducted. To describe the pattern of missing data, define for WS_j the missing-data flag M_j:

$$M_j = \begin{cases} 0, & \text{if Rj=1 and Aj is known,} \\ 1, & \text{if Rj=1 and Aj is missing,} \\ 2, & \text{if Rj=0 and Aj=0,} \\ 3, & \text{if Rj is missing.} \end{cases} \quad (1)$$

Since 0 and 2 both signify that WS_j is fully recorded, three codes are sufficient to identify the missing-data pattern for this month. Repeated over the 12 months, this yields 3^{12} = 531,441 possible patterns of missing data!

Data for our study were confined to the first three waves of the 1984 SIPP Panel, and concerned 30004 panel members studied by Kalton and Millar (1986), who were aged 15 or over at the first wave, who were respondents at that wave and who remained in the survey population (but were not necessarily respondents) throughout the next two waves. Table 1 summarizes the distribution of cases by missing data pattern for the first wave, that is, classified by the values of the variable (1) for the first 4 months. In an earlier paper, we discussed methods for filling in the amount variables A_j for cases with R_j known to be 1 for j=1,...,12. that is for cases with M_j equal to 0 or 1 for

all j (Little and Su 1986). In Section 3 we tackle the more difficult problem of analyzing data with different patterns of recipiency and

TABLE 1. Distribution of Cases by Missing-Data Pattern for the First Wave.

PATTERN	FREQ	PCT	PATTERN	FREQ	PCT
0000	10550	35.2	2000	21	0.1
0001	20	0.1	2002	92	0.3
0002	448	1.5	2022	83	0.3
0003	44	0.1	2200	542	1.8
0020	38	0.1	2202	76	0.3
0022	498	1.7	2211	26	0.1
0100	546	1.8	2220	484	1.6
0200	50	0.2	2221	24	0.1
0220	74	0.2	2222	11219	37.4
0222	367	1.2	2223	30	0.1
1000	21	0.1	2233	43	0.1
1001	529	1.8	2333	23	0.1
1010	520	1.7	3000	23	0.1
1111	804	2.7	3322	27	0.1
1122	25	0.1	3332	24	0.1
1222	21	0.1	3333	1897	6.3
Other*	338	1.1	Total	30003	100.0

* Includes 72 rare patterns with <20 cases.

missing information on both amounts and recipiencies. But first we consider the simpler problem where missing data are confined to a single partially-scaled variable.

2. MONOTONE MISSING-DATA PATTERNS FOR PARTIALLY-SCALED VARIABLES

Let Z =(R,A) denote an incompletely-observed partially-scaled variable, and let B be a set of fully recorded variables. A natural strategy for analyzing the data is to

i) estimate the distribution of B using all the observations;

ii) estimate the distribution of R given B using the observations for which R is recorded;

iii) estimate the distribution of A given B and R using the observations for which A is recorded and R=1.

Under certain assumptions this approach is efficient in that it makes full use of the available data. To see this, note that R is more observed than A in that R is observed for all cases where A is observed. Also B is assumed more observed than R. Thus the data have the monotone data pattern

$$B > R > A,$$

where > stands for "more observed than".

Following Rubin (1974), suppose we model the joint distribution of B,R,A as

$$p(B,R,A|\gamma,\theta,\Psi) = p(B|\gamma)\, p(R|B,\theta)\, p(A|B,R,\Psi),$$

where γ,θ and Ψ are sets of parameters. Then the likelihood of γ,θ,Ψ given the data d factorizes into three complete-data components:

$$L(\gamma,\theta,\Psi|d) = L_1(\gamma|d)\, L_2(\theta|d)\, L_3(\Psi|d), \quad (2)$$

where L_1 is the likelihood of γ from the distribution of B based on all the observations, L_2 is the likelihood of θ from the distribution of R given B based on all observations with R and B observed, and L_3 is the likelihood of Ψ from the distribution of A given B, R based on all observations with A,B and R observed. If γ, θ and Ψ are distinct sets of parameters, a natural restriction in many models, then maximization of L reduces to complete-data maximizations of L_1, L_2 and L_3, which correspond respectively to steps i), ii) and iii) in the process outlined at the start of this section.

The method of implementation of i) need not concern us here. Since R is binary, ii) may be implemented by methods such as logistic or probit regression of R on B. Since A is continuous, iii) may be accomplished by linear regression of A (or a suitable transformation such as the logarithm) on R and B. Since A=0 when R=0, the latter regression is restricted to observations with R=1. The usual modeling considerations concerning the choice of variables, inclusion of interactions, and so on, apply to the regressions in ii) and iii).

The regressions in ii) and iii) can also form the basis for imputation of the missing values of R and A, if this is desired. For observation i with r_i missing and value b_i of B, let $\hat{p}(b_i)$ be the predicted probability of recipiency from the regression of R on B. Then $\hat{r}_i=1$ is imputed with probability $\hat{p}(b_i)$ and $\hat{r}_i=0$ is imputed otherwise. In the latter case $\hat{a}_i$ is also imputed as zero. Observations i with a_i missing and $r_i=1$ or $\hat{r}_i=1$ are imputed amounts based on the regression of A on B and R=1, using a continuous variable imputation method. Herzog and Rubin (1983) apply this strategy to impute Social Security income in the Current Population Survey, using multiple imputation to estimate the imputation variance of estimates from the filled-in data.

We can extend this approach to two or more incomplete partially-scaled variables if the missing-data pattern has a convenient monotonic

form. For example if $Z_1=(R_1,A_1)$ and $Z_2=(R_2,A_2)$ are two partially-scaled variables with

$$B > R_1 > A_1 > R_2 > A_2,$$

the natural factored likelihood analysis would estimate the following sequence of regressions, using the set of cases for which each dependent variable is observed: R_1 on B; A_1 on B given R_1=1; R_2 on B, R_1 and A_1; and A_2 on B, R_1, A_1 given R_2=1. As before, logistic regression might be used for the recipiencies and linear regression for the amounts.

If on the other hand the data are such that

$$B > R_1 > R_2 > A_1 > A_2,$$

then the natural factored-likelihood analysis would carry out a different sequence of regressions: R_1 on B; R_2 on R_1 and B; A_1 on B and R_2 given R_1=1; and A_2 on B, R_1 and A_1 given R_2=1.

Methods based on other factorizations of the likelihood can be found for the case of two partially-scaled variables, and with more than two the number of possibilities expand considerably. However the chance of obtaining such convenient missing-data patterns in practice seems small. In particular the missing-data pattern in Example 1 is far from monotone. Hence we now concentrate on methods that can be applied to any pattern of missing data.

3. METHODS FOR GENERAL MISSING-DATA PATTERNS.

3.1. A Model for Mixed Continuous and Categorical Data.

Consider a complete random sample of size n on K continuous variables (X) and V categorical variables (Y). Categorical variable j has I_j levels, so that the categorical variables form a V-way contingency table with $C = \Pi_{j=1}^{V} I_j$ cells. For observation i, let x_i be the (1×K) vector of continuous variables and y_i be the (1×V) vector of categorical variables. Also construct from y_i the (1×C) vector w_i, which equals E_m if observation i belongs to cell m of the contingency table, where E_m is a (1×C) vector with 1 in the m^{th} entry and 0 elsewhere. Olkin and Tate (1961) define the general location model for the distribution of (x_i,w_i) in terms of the marginal distribution of w_i and the conditional distribution of x_i given w_i:

1. (x_i,w_i) are independently distributed over observations.

2. The w_i are multinomial with cell probabilities

$$Pr(w_i=E_m) = \pi_m,\ m=1,\ldots,C;\ \Sigma\pi_m = 1.$$

3. Given $w_i=E_m$, x_i has a K-variate normal distribution with mean $\mu_m = (\mu_{m1},\ldots,\mu_{mK})$ and covariance matrix Ω.

We write $\Pi = (\pi_1,\ldots,\pi_C)$ for the (1×C) vector of cell probabilities and $\Gamma = \{\mu_{mk}\}$ for the (C×K) matrix of cell means. There are C-1+KC+K(K+1)/2 parameters (Π,Γ,Ω) in the model. The complete-data ML estimates are easily shown to be

$$\hat{\Pi} = n^{-1}\ \Sigma w_i,$$
$$\hat{\Gamma}^T = (\Sigma x_i^T w_i)(\Sigma w_i^T w_i)^{-1},$$
$$\hat{\Omega} = n^{-1}\ \Sigma(x_i-w_i\hat{\Gamma})^T(x_i-w_i\hat{\Gamma}),$$

which are respectively the observed cell proportions, the observed cell means, and the pooled within-cell covariance matrix of X. Little and Schluchter (1985) show how to find ML estimates when values of X and W are missing, with an arbitrary pattern of missing values. Computation is via an iterative EM algorithm (Dempster, Laird and Rubin 1977), and involves only standard computational tools such as Sweep (Little and Rubin 1987, chapters 6 and 10). In particular, numerical integration is not required.

A full description of the algorithm is omitted here, but an informal description is as follows. The M (maximization) Step at each iteration is essentially the same as for complete-data, with sufficient statistics replaced by estimates from the E-Step. The E (expectation) step of the algorithm fills in missing values of the categorical variables for case i by estimated probabilities of falling in each cell in the set S_i of cells of the contingency table consistent with the observed components of y_i. These estimated probabilities take into account known values of the continuous variables; specifically the logits of the estimated probabilities are linear combinations of the observed values of the continuous variables in the case i. Missing values of continuous variables are estimated by weighted combinations of their conditional means within each cell in S_i given the observed continuous variables for that case, estimated by sweeping on the within-cell covariance matrix. The swept matrix also supplies adjustments to the estimated within-cell covariance matrix analogous to those in the multivariate normal EM algorithm (see for example Little and Rubin 1987, Section 8.2).

Since partially-scaled variables are mixtures of continuous and categorical variables, one might consider applying this model to incomplete data involving such variables, including recipiency variables in Y and amount variables in X. However, note that when a recipiency variable is zero the corresponding amount variable is identically 0. This constraint is not consistent with the general location model, which assumes the same covariance matrix for amount (X) variables in the recipient and nonrecipient cells.

This problem can be overcome by the following trick: for each amount variable A_j define a variable X_j in the general location model which takes the value of A_j for recipients, but which is treated as missing (rather than zero) for nonrecipients. Proceed with the EM algorithm, and at the conclusion replace values of X_j for nonrecipients (R_j=0) by 0. Note that since there are no data on X_j for cells with R_j=0, the mean of X_j in those cells is inestimable. However these means are of no interest and can be ignored; the algorithm still provides useful estimates of amounts for recipients, which are the only amounts that matter. One might expect the presence of inestimable parameters to impede the convergence of the algorithm, but the algorithm has converged satisfactorily in our applications to real and simulated data.

3.2. Alternative Approaches.

The EM algorithm of Section 3.1 seems a promising tool, but it is iterative, and may prove expensive for large data sets. It is also dependent on the model assumptions, in particular that the covariance matrix of the continuous variables is the same in each cell. Thus alternative procedures deserve study.

1) One might apply methods for a single partially-scaled variable in Section 2 independently to each partially-scaled variable. This may be adequate when correlations between the partially-scaled variables are not of interest. For example, aggregate amounts for a set of partially-scaled variables measured over time can be consistently estimated without attention to the correlation structure; however aggregate amounts restricted to recipients at all time points do require attention to the correlations of the recipiencies over time, and hence may be distorted by procedures that treat each partially--scaled variable independently.

2) If the iterative nature of the EM algorithm is a concern, then one iteration, starting from estimates based on complete cases, may yield satisfactory estimates with one pass through the data.

3) A broad class of methods that merit consideration are hot-deck imputation methods that match an incomplete case to a complete donor case based on some metric, and then impute the donor's values.

In applying 3), the main question is the appropriate choice of metric. It appears desirable to match nonrespondents to respondents that have the same recipiency pattern with respect to recorded recipiency variables. However, this requirement may severely limit the number of donors if the number of partially-scaled variables is large. If a suitable match cannot be found, then matches may be allowed where for some variables the incomplete case is a nonrecipient and the donor case is a recipient. The imputed amounts for such variables can be subsequently set to zero. Matches should be avoided that result in a zero amount from a nonrecipient being imputed for a non-zero missing amount of a known recipient.

Hopefully a number of complete cases will match each incomplete case on observed recipiencies. A metric is needed to choose between them, the choice of which will depend on context. For the income data in Example 1, we chose as metric the mean wages and salary (over the 12 months for the complete donor cases, and over the recorded months for the incomplete cases). Specifically, we matched to the first complete case that had a mean WS within 5% of the mean WS for the incomplete case. Refinements might also match on some other characteristic of the income amounts (such as their variability across months) or on observed covariates (for example, occupation or age). If a close match on mean WS is not achieved, a ratio adjustment of the imputed amounts may be worthwhile.

3.3 Creation of a Data File for Empirical Comparisons of Methods.

To provide an empirical comparison of the methods of Sections 3.2 and 3.3, a smaller data set was constructed from the SIPP data of Example 1, and values deleted in a somewhat realistic manner. Various missing-data methods were then applied to the deleted data, and the answers compared with the "true" estimates obtained from the data before deletion. The reduced data set was created as follows:

1) Fourteen hundred of the 30004 cases in Example 1 had unknown recipiency (M_j=3) for all 12 months. These cases were deleted from the file. The remaining 28604 split into 21467 complete cases (M_j=0 or 2 for all j), and 7137 incomplete cases.

2) The 21467 complete cases were split randomly into a sample of size 2385 (the C sample) and a sample of size 19082 (the $\bar{C}$ sample). A random sample of 2379 cases (the I sample) was selected from the 7137 incomplete cases.

3) Each case in the I sample was matched to a case in the $\bar{C}$ sample, using the matching method

described in Section 3.2, that is, matching first on recipiency pattern in observed months and then on the mean WS of recorded months. In a few cases, an exact match on recipiency pattern could not be achieved, and the matching criterion was relaxed in the manner also mentioned in Section 3.2.

4) The subset of matched cases in the $\bar{C}$ sample (say the J sample) was then combined with the C sample to obtain a complete data file with 4764 cases. Values in the J sample were deleted according to the missing data pattern of the corresponding matched I-sample cases to obtain an incomplete-data file for analysis.

The resulting file has the following properties:

1) The proportion of missing data is considerably higher than the proportion in the original data set. This increases the power to discriminate between alternative missing-data methods.

2) Since the incomplete cases in the J-sample were constructed via a match to the incomplete cases in the I-sample, they reflect characteristics of the I-sample cases used in the match, namely the pattern of incompleteness and the mean WS of recorded months.

3) The matching method used to determine which values are missing is also used to supply imputations in the "match" method described below. This may somewhat bias the comparisons with other adjustment procedures in favor of the match method. Biases like this are unavoidable when the investigator rather than real life creates the missing values.

3.4 Empirical Comparison of Methods

The ML method of section 3.1 could not be applied immediately to the data set described in Section 3.3, because of the relatively large space requirements. In particular, the 12 categorical variables $R_1,\ldots,R_{12}$ yielded a contingency table with over 4000 cells, which was too large for the version of the program available to us. In future we hope to be able to modify the program to deal with 12 months of data, but in the preliminary work described here we limited comparisons to the first wave (that is, the first four months) of data.

Of the 2379 cases in the J sample, only 1054 were incomplete in the first four months of data. To maintain a roughly 50/50 split between complete and incomplete cases, these 1054 cases were combined with a 50% sample of 1172 cases from the C sample, to create a data set with 2226 cases for analysis.

Preliminary analysis suggested that the cube root transformation symmetrized the distributions of WS amounts. Thus our analysis focused on the transformed variables $X_j = \sqrt[3]{WS_j}$. Specifically, the means of $X_1,\ldots,X_4$ were estimated, for the following sample bases: 1) All individuals in the sample, recipients and nonrecipients; 2) Restricted to recipients for the month of interest, that is, month j for X_j; and 3) Restricted to recipients for all four months in the wave. The following methods were used to estimate these quantities:

A) True: Sample means of the complete data before deletion.

B) Complete Cases (CC): Sample means of the 1172 complete cases, discarding the 1054 cases with values deleted.

C) Available Cases (AC): Sample means for month j based on the cases recorded for month j.

D) Maximum Likelihood (ML): Estimates found from the ML method described in Section 3.1, applied to the cube root transformed WS amounts.

E) Match: Sample means from data filled in by the imputation method of Section 3.2.

Results from the five methods are shown in Table 2. The performance of incomplete-data methods (B-E) can be assessed by how close their estimates are to the estimates from the data before deletion (A). Summary conclusions are as follows:

1) The ML and Match methods yield similar estimates, and both come very close to the estimates before deletion.

2) The CC and AC methods severely underestimate the means that include nonrecipients (panel A). The reason is that these methods overestimate the proportion of nonrecipients in the sample, since nonrecipients are less likely to be missing than recipients.

3) The CC and AC are quite close to the estimates for before deletion when nonrecipients are omitted, although not as close as the estimates from the ML and Match methods.

4. CONCLUSIONS

We have considered missing data for a particularly common type of survey variable, which we call partially-scaled. Two methods for handling missing data on a set of partially-scaled variables are proposed, one based on maximum likelihood for a general model for mixed continuous and categorical variables, and one based on imputation from a matched complete record. preliminary empirical work based on data from the SIPP shows that both of these methods have promise. Future work will develop the theoretical and empirical properties of these methods more completely.

ACKNOWLEDGMENTS

This research was supported by grant SES 84 11804 from the National Science Foundation. We wish to thank Graham Kalton for making available the data set which formed the basis for the empirical work in this article.

References

David, M. (1985), " Introduction: the Design and Development of SIPP," Journal of Economic and Social Measurement, 13, 215-224.

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

Herzog, T.N. and Rubin, D.B. (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," in Incomplete Data in Sample Surveys, Vol. 2: Theory and Annotated Bibliography (W.G. Madow, I. Olkin and D.B. Rubin, eds.), New York: Academic Press.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," Proceedings of the Survey Research Methods Section, American Statistical Association 1982, 22-31.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data," Survey Methodology, 12, 1-16.

Kalton, G. and Millar, M. (1986), "Effects of Wave Nonresponse on Panel Survey Estimates," Proceedings of the Survey Research Methods Section, American Statistical Association 1986.

Little, R.J.A. (1982), "Models for nonresponse in sample surveys", Journal of the American Statistical Association, 77, 237-250.

Little, R.J.A. (1987), "Missing Data in Large Surveys," to appear in Journal of Business and Economic Statistics,

Little, R.J.A. and Rubin, D.B. (1987), Statistical Analysis with Missing Data, New York: John Wiley and Sons.

Little, R.J.A. and Schluchter, M.D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data with Missing Values," Biometrika, 72, 497-512.

Little, R.J.A. and Su, H-L. (1986), "Item Nonresponse in Panel Surveys", to appear in Proceedings of the International Symposium on Panel Surveys, New York: John Wiley and Sons.

Madow, W.G., Nisselson, H., Olkin, I. and Rubin, D.B. (eds.), (1983), Incomplete Data in Sample Surveys, Vols. 1-3, New York: Academic Press.

Pregibon, D. (1977), "Typical Survey Data: Estimation and Imputation," Survey Methodology, 2, 79-102.

Olkin, I. and Tate, R.F. (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," Annals of Mathematical Statistics, 32, 448-465.

Rubin, D.B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," Journal of the American Statistical Association, 69, 467-474.

Rubin, D.B. (1976), "Inference and Missing Data", Biometrika, 63, 581-592.

Sande, I.G. (1982), "Imputation in Surveys: Coping with Reality," The American Statistician, 36, 145-152.

Table 2. Estimates of mean $\sqrt[3]{WS}$ by Month, Method, and Sample Base(Sample Size).

Method	Mean $\sqrt[3]{WS}$ for Month 1	2	3	4
(A) All individuals(recipients and nonrecipients)				
True	6.87(2226)	6.90(2226)	6.90(2226)	6.89(2226)
CC	4.83(1172)	4.89(1172)	4.92(1172)	4.96(1172)
AC	5.46(1499)	5.95(1678)	6.14(1703)	6.02(1657)
ML	6.85	6.88	6.88	6.87
Match	6.82	6.86	6.88	6.88
(B) Restricted to recipients for the month of interest				
True	10.18(1502)	10.20(1505)	10.19(1508)	10.20(1504)
CC	10.30(550)	10.30(556)	10.28(561)	10.27(566)
AC	10.24(799)	10.12(987)	10.21(1024)	10.24(974)
ML	10.16	10.19	10.17	10.17
Match	10.17	10.21	10.18	10.19
(C) Restricted to recipients for all fourth months				
True	10.39(1383)	10.46(1383)	10.43(1383)	10.41(1383)
CC	10.59(487)	10.73(487)	10.67(487)	10.67(487)
AC	10.57(697)	10.45(876)	10.50(907)	10.51(870)
ML	10.33	10.43	10.38	10.37
Match	10.37	10.45	10.40	10.39