Missing-Data Adjustments in Large Surveys

Roderick J. A. Little
Department of Biomathematics, School of Medicine, University of California, Los Angeles, CA 90024

Useful properties of a general-purpose imputation method for numerical data are suggested and discussed in the context of several large government surveys. Imputation based on predictive mean matching is proposed as a useful extension of methods in existing practice, and versions of the method are presented for unit nonresponse and item nonresponse with a general pattern of missingness. Extensions of the method to provide multiple imputations are also considered. Pros and cons of weighting adjustments are discussed, and weighting-based analogs to predictive mean matching are outlined.

KEY WORDS: Imputation; Incomplete data; Matching; Multiple imputation; Regression models; Weighting.

1. INTRODUCTION

Missing data is a pervasive problem in sample surveys. For a general review of the problem, see Madow, Nisselson, Olkin, and Rubin (1983). Three broad general strategies for dealing with the problem can be distinguished—direct analysis of the incomplete data, imputation, and weighting complete cases. 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 in which 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).

Both imputation and weighting create rectangular data sets convenient for subsequent analysis. Weighting is often used to handle *unit* nonresponse, in which entire questionnaires are missing because of noncontact or refusal to conduct the interview. Nonrespondents are dropped from the file, and weights are assigned to respondents that attempt to adjust for the selection bias created by the omission of nonrespondents. Imputation is often used to handle *item* nonresponse, in which an interview is conducted but responses to particular items are missing. Missing values are replaced by estimates based on the recorded information in the incomplete questionnaire. For discussions of imputation methods, see Kalton and Kasprzyk (1982, 1986), Sande (1982), Madow et al. (1983, vol. 2), and Little and Rubin (1987, chap. 4).

The main body of this article considers imputation as an adjustment strategy. Section 2 discusses whether imputation is a good idea, as opposed to direct analysis of the incomplete data. The answer, of course, depends greatly on how imputation is implemented; I outline general properties of a good imputation method, some clearly desirable and others more contentious. Section 3 develops the ideas of Section 2 in the context of some large government surveys. Section 4 focuses on a particular form of imputation, which I call *predictive mean matching*, that seems one useful tool for implementing the principles of imputation in Section 2. The method was first proposed by Rubin (1986) in the context of statistical matching, and it is extended here to handle multivariate nonresponse and multiple imputation.

Section 5 reviews recent approaches to weighting adjustment. Methods for forming weights based on the response propensity are discussed, and an extension of the method is described for monotone patterns such as those in panel surveys with attrition. The close relationship between weighting and certain forms of imputation for monotone missing data is examined and applied to derive weighting analogs to imputation based on predictive mean matching. Section 6 summarizes the perspective on missing-data adjustments adopted in the article.

2. WHY ADJUST?

Standard statistical methods are based on rectangular files. If data sets with missing values are passed to statistical packages, then a rectangular file is usually achieved by restricting attention to cases that are complete on the set of variables analyzed. Discarding incomplete cases is inefficient, but, more seriously, the complete cases may no longer be representative of the target population; consequently, estimates derived from them are subject to nonresponse bias. The main reason for weighting or imputation in large surveys is to produce a more representative rectangular file for analysis.

Critics of imputation argue that missing-data adjustments must be developed within the context of the specific analysis. Since different analysts are concerned with different contexts, no single set of imputations can satisfy all interests. In principle, I agree with this argument. Ideally the analyst should formulate a statistical model for the survey variables under study and the missing-data mechanism and then estimate parameters from the incomplete data by methods such as ML without bothering to fill in the missing values (Little 1982; Little and Rubin 1987).

In practice, however, this may place excessive demands on the analyst interested in the subject matter rather than in specialized statistical methodology. Methods for fitting, say, simultaneous-equations models to multivariate data with missing values are not highly developed in the literatrure. Such methods can be developed, using tools such as the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977), but few are willing to devote the energy to develop customized algorithms for their missing-data problems. They want to be able to forget the fact that data are missing. The practical question is whether all-purpose weights or imputed values can be developed that provide better answers than are obtained by simple, widely available missing-data methods such as analysis of complete cases. I think that *carefully constructed* adjustments can provide better answers for many questions. Nonresponse adjustments should be developed that approximate a full model-based analysis for as wide a range of the eventual applications as possible. The data producer should also communicate the operating characteristics of the adjustment procedure to the user so that its limitations are clear. Moreover, imputations should be flagged so that users have the option of developing their own adjustments.

Naive imputation can be worse than doing nothing; as a simple example, imputing the unconditional sample mean of a variable provides the same estimate of the overall mean as that obtained from discarding incomplete cases, with an associated standard error that is worse because it is based on an overstated sample size. With a view to avoiding such pitfalls, a statement of principles of imputation seems worthwhile. Here is my list of desirable imputation properties. Imputation is a form of prediction; therefore:

1. *Imputations should be based on the predictive distribution of the missing values, given the observed values for a case.* A model (implicit or explicit) underlies this distribution, so a systematic approach to imputation requires modeling the data. The quality of the imputations depends on the quality of the underlying model.

2. *In principle, all observed items for a case should be taken into account in developing imputations.* Use of all items present in a case may be unrealistic, particularly in large surveys, so in practice some selectivity is often needed, based on subject-matter knowledge. Principle 2 is a useful objective, however, because judicious use of this information yields more accurate predictions. Moreover, measures of association, such as covariances and slopes involving incomplete variables, are distorted for variables not included in the imputation model (see Kalton and Kasprzyk 1982; for expressions for bias in estimates of slopes, see Santos 1981).

3. *Models for prediction should take into account contextual knowledge about the variables being imputed.* Principles 1 and 3 argue for a modeling approach, in contrast to, say, traditional U.S. Bureau of the Census practice (see, e.g. Bailey, Chapman, and Kasprzyk 1986; Hanson 1978), which emphasizes design-based descriptive analysis and leaves model assumptions to external users of the data.

Another factor to consider is how to use knowledge in the form of editing constraints, such as the requirement that an aggregate variable equals the sum of its parts (for discussion and references, see Sande 1982). As noted by Little and Smith (1987), we lack a synthesis of methods that deal with complex editing constraints (e.g., Fellegi and Holt 1976; Greenberg and Surdi 1984; Sande 1979) with statistical methods for handling multivariate missing data. The relative emphasis given to these approaches seems to depend strongly on context.

4. *Models for prediction should avoid excessive extrapolation beyond the range of the data, unless objective evidence is available to substantiate these models.* Examples of models involving extrapolation include regression imputation when the imputed cases have covariate values beyond the range of the values for respondents and certain nonrandom nonresponse models that rely strongly on normality and structural assumptions (Greenlees, Reece, and Zieschang 1982; Lillard, Smith, and Welch 1982, 1986), discussed further in Section 3. Such models seem attractive simply because many nonresponse mechanisms are, in practice, nonrandom. In my view, however, the methods rely too much on untestable assumptions (Goldberger 1980; Little 1982, 1985; Rubin 1982) to be as highly valued as they appear to be considered by some authors (e.g., Winer 1983). Nonrandom nonresponse models are useful, particularly for studies of sensitivity of answers to alternative assumptions about the missing-data mechanism (e.g., Fay 1986). For all-purpose imputation, I prefer to adopt the more modest objective of making the best use of available information on incomplete cases.

5. *Imputations should be drawn from the predictive distribution in Principle 1, not means, to preserve the distribution of the variables in the filled-in data set.* That is, noise (in the form of residuals or random errors) should be added to the predicted means. The addition of noise is inefficient when the sole interest is in means and totals and hence may seem undesirable. Usually, however, other aspeacts of the distribution are also under study. Quantities such as percentiles and measures

of spread and covariation are distorted by mean imputation.

6. *A method should be provided for computing sampling errors of estimates that are appropriate in that they take into account the fact that some values have been imputed.* A well-known defect of imputation is the overestimation of precision that results when the filled-in data are treated as observed data. Multiple imputation (Rubin 1978, 1987), which is discussed in Section 3.6, seems to me the most promising tool for achieving this objective.

3. DISCUSSION

3.1 Introduction

In this section, I expand on the principles of imputation suggested in Section 2 in the context of imputation practice for some large government surveys. Principle 1 seems almost self-evident. Some, particularly survey samplers, might question whether the missing values can be regarded as random variables with a predictive distribution. An alternative formulation is to model the response probabilities. This approach seems insufficiently flexible for survey-item nonreponse (Little 1982; Little and Rubin 1987, chaps. 4 and 11), although it may have merit for some unit nonresponse problems (Oh and Scheuren 1983), as discussed in Section 5. I now turn to the other principles in Section 2.

3.2 Condition on Observed Items

Conditioning on observed items has the potential of reducing both nonresponse bias and variance (Little 1986). A practical illustration of reduction in nonresponse bias is contained in the simulation study by Colledge, Johnson, Pare, and Sande (1978). Three estimators, (a) the respondent mean, (b) a ratio estimator based on the variable *total expenses,* and (c) the imputed mean from the sequential hot-deck procedure used in practice, were compared on data from a Canadian survey of construction firms with values artificially deleted according to each of three deletion mechanisms. When deletion depended on the fully recorded variable *gross business income,* the respondent mean was severely biased, the ratio estimator was biased to a much lesser extent, and the sequential hot deck had the smallest bias. Study of the latter reveals that it effectively conditions on both total expenses and gross business income, so the estimators' biases are inversely related to their degree of conditioning. A regression estimator that conditions on both total expenses and gross business income might have proved more competitive with the sequential hot-deck method.

The sequential hot deck achieves a considerable degree of conditioning by computing a distance between responding and nonresponding cases based on recorded variables and then matching cases that are close together in this metric. On the other hand, the imputation scheme in the U.S. Annual Survey of Manufactures (ASM), described by Kusch and Clark (1979), essentially confines the conditioning to a single item x. If y is missing and x is present, the value $\hat{y} = \overline{R}x$ is imputed, where $\overline{R}$ is the average ratio of y to x among respondents. Better imputations may result from methods that make use of more than two recorded items simultaneously. Steps in this direction, whereby editing and imputation are based on the estimated mean and covariance matrix of the items measured on the log scale, were suggested by Little and Smith (1987). The parameters are estimated using a robust modification of the multivariate normal EM algorithm. Such methods can be refined to take into account logical constraints between the items (Little and Smith 1983), which are treated with care in the Kusch and Clark (1979) methodology.

A common approach to conditioning is to form *adjustment cells* based on the observed items and then to impute nonrespondent data using respondent information in the same cell. A highly developed example is the Current Population Survey (CPS) hot deck (Welniak and Coder 1980) for imputing missing income fields in the March income supplement of the CPS. Individuals are classified into adjustment cells on the basis of information available from the interview. Then missing values for an incomplete case are replaced by values from a donor individual in the same adjustment cell.

A large number of observed items are included as potential predictors. Consequently, the matrix of adjustment cells is very large and items are frequently dropped or categories are coarsened in order to find a matching respondent. In particular, Lillard et al. (1982, 1986) criticized the fact that the CPS hot deck can drop three-digit occupational detail, leading, for example, to underestimation of a nonresponding judge's salary. They noted that imputation within adjustment cells based on a set of variables S effectively assumes a model in which all of the main effects and interactions between the variables in S are included. For example, if adjustment cells are based on age, occupation, sex, and region, then the implicit imputation model effectively includes the effect of the age × occupation × sex × region interaction on income. In contrast, imputations based on a regression model for income can allow more items to be included by eliminating high-order interactions.

It is rather surprising, given this analysis, that the alternative regression model for income imputation of Lillard et al. (1982, 1986) compounds the failure to include sufficient occupational detail, by omitting occupation variables entirely! The model of Greenlees et al. (1982) includes occupation, but confines attention to the first digit of the code, which is clearly inadequate. David, Little, Samuhel, and Triest (1986) fitted a 90-variable regression model with considerable occupational detail and compared methods based on this model to the CPS hot deck using a confidential file that matches the CPS to Internal Revenue Service income data. The results did not indicate clear superiority for either approach. I prefer the regression approach, however, because I believe that main effects of variables are usually more important than interactions, and I do not think interactions should be included in the imputation model at the expense of main effects.

The theoretical advantage of regression methods over adjustment-cell methods seems even more pronounced in longitudinal imputation such as that arising in panel surveys like the Survey of Income and Program Participation (SIPP) (David 1985). Here information on other waves can be available to impute missing items in a particular interview, so the number of potential predictors can be extremely large. For discussions of imputation in the SIPP, see Kalton and Miller (1986), Heeringa and Lepkowski (1986), and Little and Su (in press).

3.3 Use Subject-Matter Knowledge

Obviously subject-matter knowledge has an impact on the CPS hot deck through the choice of adjustment cells and in collapsing rules adopted when the adjustment cells become too small. It seems preferable, however, to base imputations on a more formal regression model that builds on previous models in the literature, such as that of David et al. (1986). Note, however, that the imputation model should include *all* variables thought to be predictive of income, including variables that might not be considered exogenous to income in a behavioral econometric model such as that of Lillard et al. (1986). For example, suppose the variable size of residence was recorded in the CPS. The variable is predictive of income, but it might be excluded from behavioral models for income because it might not be considered exogenous. Questions of exogeneity are irrelevant in the context of imputation, however; the objective is prediction, not causal inference.

The imputations for the ASM involve substantial input from subject-matter specialists, in that the imputes must satisfy range checks and ratio edits based on historical knowledge about the industry (Greenberg and Surdi 1984). Input of this kind is indispensable, but it can be expensive. More automatic and empirically based edit/imputation schemes such as that of Little and Smith (1987) may be useful for limiting the cost of specialist input.

3.4 Avoid Excessive Extrapolation Beyond the Range of the Data

Hot-deck methods always impute respondent values and hence are conservative in that they avoid extrapolating nonrespondent values outside the range of the respondent data. In contrast, the models of Lillard et al. (1982) and Greenlees et al. (1982) include potentially large adjustments for nonrandom nonresponse based on the stochastic censoring models (Amemiya 1984; Hausman and Wise 1970; Heckman 1976). These adjustments are highly sensitive to distributional assumptions and to the choice of predictors in regression equations for income and response, a choice that is far from obvious. For example, the model of Lillard et al. (1982, 1986) suggests that the CPS hot deck underestimates the wages and salary of nonrespondents by an astonishing 73%. Comparisons by David et al. (1986) with IRS data do not indicate significant underestimation, however, and I believe that the 73% result displays the unreliability of the Lillard et al. (1982, 1986) model rather than the performance of the CPS hot deck.

The central issue here is not whether nonresponse to income is nonrandom, but whether nonresponse to income is nonrandom *after adjustment for all of the covariate information.* David et al. (1986) found no evidence of this when fitting their relatively detailed model for wages and salary.

An important redeeming feature of the CPS hot deck for critics is the fact that the imputations are flagged and, therefore, can be replaced if necessary; that is, the method is *reversible,* an important feature given the lack of a universal nonresponse adjustment scheme.

3.5 Draws, Not Means, From the Predictive Distribution

Hot-deck methods like the CPS hot deck and the sequential hot deck of Colledge et al. (1978) impute actual values of missing variables rather than means and hence tend to preserve distributions in the filled-in data. In contrast, methods that impute predicted values from a regression [see, e.g., the imputation methods in the BMDPAM program of Dixon (1983)] or ratios as in Kusch and Clark (1979), are imputing estimates of conditional means and thus need to be modified to preserve distributions and associations between variables. Modifications can take the form of adding a residual from a matched respondent (Scheuren 1976). David et al. (1986) and the "Row*Column" method of Little and Su (in press) contain applications of this idea to income nonresponse in the CPS and the SIPP.

3.6 Provide Valid Standard Errors From the Filled-In Data

All of the imputation methods discussed so far supply only one imputed value, so inferences based on the filled-in data do not reflect added uncertainty from the fact that values are missing. Rubin (1978, 1987) proposed multiple imputation as a convenient method for estimating the added variance from estimating the missing values and yielding approximately valid inferences. A set of $i = 1, \ldots, I$ imputations is supplied for each missing value, and I analyses are performed, where for analysis i the ith imputation of each missing value is substituted. For the ith analysis, let $\hat{\theta}_i$ be the estimate of a particular parameter θ of interest and let $\hat{v}_i$ be its estimated variance, ignoring the effect of imputation. The final estimate of θ is $\hat{\theta} = \Sigma \theta_i/I$, with estimated variance

$$\hat{v}^2 = s_w^2 + (1 + 1/I)s_b^2, \qquad (1)$$

where $s_w^2 = \sum \hat{v}_i/I$ is the average variance within imputed data sets, and $s_b^2 = \sum (\hat{\theta}_i - \hat{\theta})^2/(I - 1)$ represents between-imputation variance. Large-sample inference for θ is based on comparing $(\hat{\theta} - \theta)/\hat{v}$ with a t distribution with ν df, where

$$\nu = [1 + \{I/(I + 1)\}s_w^2/s_b^2]^2(I - 1)$$

is based on a Satterthwaite approximation (Rubin and Schenker 1986). For details and practical examples of the method, see Rubin (1987).

4. AN EXTENSION OF ADJUSTMENT CELL METHODS: PREDICTIVE MEAN MATCHING

4.1 Univariate Nonresponse

A key problem for imputation methods is how to incorporate information from observed items in a simple but relatively efficient way. The adjustment-cell approach seems useful for a relatively small set of predictors. For large problems, however, the number of cells rapidly becomes unmanageable, and methods for combining cells are somewhat arbitrary. Moreover, the method seems ill-suited when several continuous items are available for prediction.

No single method is appropriate for all imputation problems. As one example of a methodology that follows the guidelines proposed in Section 2, however, I would like to discuss and develop a method proposed by Rubin (1986) in the context of statistical matching, which I call *predictive mean matching*. I consider first the case in which nonresponse is confined to a single item y, and I then consider extensions to more general patterns of nonresponse.

The center of the predictive distribution y is estimated by regressing y on the set of observed items, say x. As noted in Principle 5, noise should be added to these predicted means to preserve distributions in the filled-in data. One might simply add normal random deviates with mean 0 and variance $s_{y \cdot x}^2$, the residual variance from the regression. This method depends heavily on normal linear model assumptions, however, and an approach less dependent on normality is to add respondent residuals. David et al. (1986) added a residual from a respondent with a predicted value of y similar to that of the nonrespondent, a method that protects against heteroscedasticity, where the residual variance depends on the predicted mean and provides some protection against (local) model failures. The resulting method successfully preserved the distribution of wages and salary in the filled-in data.

A closely related method is to match each nonrespondent to the respondent with the closest predicted mean and then impute that respondent's value directly (Rubin 1986)—that is, impute

$$\hat{y}_j = y_k, \qquad (2)$$

where $(\hat{\mu}_j - \hat{\mu}_k)^2 \le (\hat{\mu}_j - \hat{\mu}_l)^2$ for all respondents l, $\hat{\mu}_j$ is the predicted mean of Y for individual j, and y_k is the observed value of Y for respondent k. This is a form of distance function matching (Sande 1979), which I call predictive mean matching. Its advantage over the method of David et al. (1986) is that only eligible values of the missing variable are imputed (so, e.g., a binary variable is always assigned one of its two possible outcomes). Moreover, since the predictive mean is only used to define a match, the method is perhaps less sensitive to model misspecification.

Other metrics for distance-function matching have been proposed—for example, the Mahalanobis distance (Vacek and Ashikaga 1980). The metric based on the predictive mean seems most appropriate, however, given that our objective is to impute from the predictive distribution. In particular, for imputing y there is nothing to be gained by matching on variables that are not predictive of y. Such variables are ignored in the predictive mean metric but might dominate the Mahalanobis distance metric.

Hot-deck imputation within adjustment cells can be regarded as a special case of predictive mean matching, where predictors are categorical and all interactions between them are included in the regression, because then all respondents and nonrespondents in the same cell have the same predicted mean and hence are distance 0 apart in the predictive mean metric.

4.2 Multivariate Nonresponse

Suppose now that y is a $(p \times 1)$ vector of items that are missing as a block (i.e., for every case they are either all observed or all missing), and let x denote a $(q \times 1)$ set of fully observed items. The method of Section 4.1 could be applied separately to each component of y, but since the metric in (2) changes for each component of y, imputed components of y for an incomplete case can come from a variety of respondents, thus distorting associations between the y variables. If these associations are important, a better plan is to match each nonrespondent to a single respondent and assign the entire respondent y vector to the nonrespondent, as was done by Czajka (1986) and Hinkins and Scheuren (1986) and is generally done in the CPS hot deck. Logical constraints between the y's such as those in the ASM example, will tend to be better preserved by this approach.

The multivariate regression of y on x yields a $(p \times 1)$ vector of predicted means $\hat{\mu}_j = \hat{\mu}(x_j)$ for each case j. It is less clear how to form a single metric from these p linear combinations of x's. From a statistical standpoint, one might tolerate greater matching error for y variables that are subject to greater prediction error.

This suggests matching using the metric

$$d^2(j, k) = (\hat{\mu}_j - \hat{\mu}_k)^T S_{y \cdot x}^{-1}(\hat{\mu}_j - \hat{\mu}_k), \qquad (3)$$

where $S_{y \cdot x}$ is the residual covariance matrix of y from the regression of y on x. More generally, one might wish to assign weights to the variables, say w_v^2 for the variable y_v, based on an assessment of their relative importance in the analysis, and then modify the metric (3) by multiplying the (u, v)th term in S by $w_u^{-1}w_v^{-1}$. An obvious special case restricts (3) to a subset of y variables, which might be a sensible practical measure if the number of y variables is excessively large.

Now consider multivariate data with an arbitrary pattern of missing values. Let y represent q variables subject to missing values and x fully observed variables. Suppose that the cases are arranged so that the first m are complete on x and y. First, an estimate of the parameters of the multivariate regression of y on x is derived. The estimate might be found from the full data by finding ML estimates of the mean and covariance matrix of x and y under multivariate normality (Beale and Little 1975; Dixon 1983; Orchard and Woodbury 1972) and then sweeping on the x variables to convert them from dependent to independent variables. More simply, estimates may be obtained from complete cases. Second, for each incomplete case i, the multivariate regression of the missing y's for that case (say y_{mis}) on x and the observed y's for that case (say y_{obs}) is estimated by sweeping on the variables y_{obs}. Then predictions of y_{mis} are obtained for the incomplete case ($\hat{\mu}_{\text{mis},j}$) and the complete cases ($\hat{\mu}_{\text{mis},k}$, $1 \leq k \leq m$), and j is matched to a complete case using the criterion (3), with $\hat{\mu}$ replaced by $\hat{\mu}_{\text{mis}}$ and $S_{y \cdot x}$ replaced by the estimated residual covariance matrix of y_{mis} given y_{obs} and x. Values of y_{mis} from the matched case are used to fill in the missing values for case i.

This method reduces to the method based on (3) when the values of y are missing or present as a block. It may appear to be computer intensive, but the computation of the various regressions is easily accomplished using the sweep operator. The method requires calculating sets of predicted values for the complete cases that are different for incomplete cases with different sets of missing y variables. This is somewhat tedious when the number of complete cases is large; computations are simplified by ordering the cases by missingness pattern, minimizing the number of sweeps needed between successive patterns, and exploiting simple relationships between the predicted values from regressions with different predictors.

The general procedure described here computes the mean and covariance matrix assuming multivariate normality and hence may be sensitive to outliers. As noted previously, lack of normality is less important here, since the model is used only to supply a metric for matching. Judicious use of transformations, such as the logarithm transform used by Little and Smith (1987), however, will improve the method. That paper also discussed a robust modification of the normal EM algorithm that reduces the impact of outliers. Related robust ML procedures were given in Little and Rubin (1987, chap. 10).

4.3 Multiple Imputation

In Section 3.6, I outline multiple imputation, a method for providing valid inferences from filled-in data sets. The appropriate imputations in Rubin's theory are draws from the predictive distribution of the missing values that the predictive mean matching method approximates. Thus a straightforward implementation of the method with predictive mean matching is to find the K nearest respondents to a particular nonrespondent and sample from this set to provide I imputed values. Sampling with replacement is required to simulate the distribution correctly. The choice of K requires a compromise between being large enough to simulate the predictive distribution effectively, and small enough to maintain the quality of the matches. Rubin (1987) discussed the choice of I; even $I = 2$ imputations will provide better estimates and some idea of the added variance from imputation. See Oh and Scheuren (1980) for an application of this idea to the CPS hot deck.

Rubin (1987) showed that when multiple imputation is achieved by repeated application of a hot-deck method, (1) still underestimates the variance, since it effectively assumes that the parameters of the predictive distribution (here the mean and covariance matrix of the variables) are known. He provides a Bayesian refinement of the method that allows for uncertainty in estimating these parameters. I present analogous refinements for predictive mean matching.

First consider univariate nonresponse, as discussed in Section 4.1. Let $\beta_{y \cdot x}$ and $\sigma_{y \cdot x}^2$ denote the vector of regression coefficients and residual variance from the regression of y on $x = (x_1, \ldots, x_q)$. Let $\hat{\beta}_{y \cdot x}$ and $s_{y \cdot x}^2$ denote corresponding estimates from the regression on the m complete cases, and let $C = \text{cov}(\hat{\beta}_{y \cdot x}) = (X^TX)^{-1}s_{y \cdot x}^2$, where X is the design matrix. Our modification is to replace the metric in (2) for imputation i with

$$(\tilde{\mu}_j^{(i)} - \hat{\mu}_k)^2, \qquad (4)$$

where $\hat{\mu}_j$ has been replaced by $\tilde{\mu}_j^{(i)}$, a draw from the posterior distribution of the predicted mean for nonrespondent j. Under standard normal assumptions and noninformative priors,

$$\tilde{\mu}_j^{(i)} = \hat{\mu}_j + (x_j^T e_i)/k_i, \qquad (5)$$

where x_j is the vector of x values for nonrespondent j, e_i is a draw from the multivariate normal distribution with mean 0 and covariance matrix C, and $k_i^2 = u_i^2/(m - q - 1)$, where u_i^2 is a draw from the chi-squared distribution with $m - q - 1$ df. Note that the adjustment (5) is not applied to $\hat{\mu}_k$; if imputation were achieved by adding an empirical residual to the predicted mean, the appropriate draw would be $\mu_j^{(i)} + (y_k - \hat{\mu}_k)$, which corresponds to matching on the metric (4).

Suppose now that y is a $(p \times 1)$ vector of items that are missing as a block. We modify the metric (3). To avoid multivariate notation, (3) can be rewritten as

$$d^2(j, k) = (\hat{\mu}_{j1} - \hat{\mu}_{k1})^2/s_{11\cdot x} + (\hat{\mu}_{j2\cdot 1} - \hat{\mu}_{k2\cdot 1})^2/s_{22\cdot 1,x} + \cdots + (\hat{\mu}_{jp\cdot 12\cdots p-1} - \hat{\mu}_{kp\cdot 12\cdots p-1})^2/s_{pp\cdot 12\cdots p-1,x},$$

where for $r = 1, \ldots, p$, $\hat{\mu}_{jr\cdot 12\cdots r-1}$ is the predicted mean of variable y_r for case j from the regression of y_r on $y_1, \ldots, y_{r-1}$ and x, and $s_{rr\cdot 12\cdots r-1,x}$ is the associated residual variance. The modification replaces $\hat{\mu}_{jr\cdot 12\cdots r-1}$ by a draw from the posterior distribution of $\mu_{jr\cdot 12\cdots r-1}$, namely,

$$\tilde{\mu}^{(i)}_{jr\cdot 12\cdots r-1} = \hat{\mu}_{jr\cdot 12\cdots r-1} + (x_j^T e_{ri})/k_{ri}, \qquad (6)$$

where e_{ri} and k_{ri} are the analogous quantities to e_i and k_i in (5) for the regression of y_r on $y_1, \ldots, y_{r-1}$ and x. Moreover, $s_{rr\cdot 12\cdots r-1,x}$ is replaced by a draw from the posterior distribution of the residual variance, namely,

$$s^{(i)}_{rr\cdot 12\cdots r-1,x} = s_{rr\cdot 12\cdots r-1,x}/k^2_{ri}. \qquad (7)$$

Exact adjustments are not available for a general pattern of missing data. One possibility is to apply (6) and (7), with the covariance matrix of the regression coefficients approximated by the weighted sum of squares and cross-products matrix proposed by Beale and Little (1975). This approximation performed reasonably in simulations in Little (1979). The method yields the appropriate adjustments for the special case of monotone missing data, where for $r = 1, \ldots, p - 1$, y_r is more observed than y_{r+1} (see Rubin 1974).

My adjustments rely on normal theory, but they are valid asymptotically under broader conditions. Even if multiple draws are based on the unadjusted metrics in Sections 4.1 and 4.2, multiple imputation gives some idea of the added uncertainty from imputation and as such seems a useful improvement on current practice.

5. WEIGHTING ADJUSTMENTS

5.1 Response Propensity Weighting

An alternative method of nonresponse adjustment to imputation in some circumstances is to weight respondents. The method is usually applied to handle unit nonresponse, with weights proportional to the inverse of response rates computed within adjustment cells. For an application of this approach to the CPS, see Hanson (1978).

One can view this method as analogous to weighting for differential sample selection using the inverse of the sampling weights, as in the Horvitz–Thompson estimator for a population mean or total (Oh and Scheuren 1983). If a sizable number of variables x is available for respondents and nonrespondents, then a natural generalization of the method is to regress the binary nonresponse indicator r on x, using logistic or probit regression if necessary, and derive predicted response propensities $\hat{p}_i = \Pr(r_i = 1 \mid x_i)$ for respondents and nonrespondents. Weights can then be defined as proportional to the inverse propensities for respondents, or by forming adjustment cells based on the propensity score. I shall use the term *response propensity weighting* to describe both of these methods.

Rosenbaum and Rubin's (1983) theory of propensity scores shows that response propensity weighting effectively removes nonresponse bias when nonresponse is random within subpopulations with the same value of x (David et al. 1983). Little (1986) compared response propensity weighting with mean imputation within subclasses, for estimates of overall and subclass means of a variable subject to nonreponse. A successful application of propensity weighting was made by Czajka, Hirabayashi, Little, and Rubin (1987).

An important practical advantage of response propensity weighting over the methods discussed in Section 4 is that the adjustments are based on a *univariate* regression of r on x, whereas (for vector y subject to nonresponse) imputations are based on the *multivariate* regression of y on x. If y has a high dimension, then the latter may involve a prohibitive amount of work. For example, adjustments for wave nonresponse in the SIPP would involve modeling the vector of outcomes for a single interview, which is extremely large.

Although the logistic regression of r on x is univariate, estimating this regression is still a nontrivial task in panel surveys in which the set of x variables is very large; in practice, judicious selection from the available x variables, based on prior knowledge and preliminery analysis, may be necessary.

Four disadvantages of response propensity weighting can be cited. First, weighting as it is usually implemented is not reversible, since the nonresponding units are dropped. This, of course, could be corrected by appending the nonresponding units to the file, retaining their original sampling weights, and assigning them nonresponse weight 0 so that they do not interfere with analysis of respondents. This step would allow other adjustments for unit nonresponse to be tried and thus provide scope for improvements to existing practice.

Second, propensity score weighting can lead to estimates with large variance, as discussed in Little (1986). The variance of weighting adjustments for unit nonresponse in the CPS is controlled by not allowing the weights to exceed a fixed number (2). This method may reduce mean squared error, but it is ad hoc with little or no theory to substantiate the choice of cut-off. Little (1986) proposed smoothing the weights using empirical Bayes methods.

Third, correct inference based on the weighted sample may be problematical. In particular, standard errors of estimates computed from the weighted data by conventional methods will be inappropriate, as shown by Jones and Chromy (1982). They provided corrections to standard errors for estimating weighting nonresponse adjustments within adjustment cells for multistage stratified cluster designs. They did not appear to allow for correlations of estimates between primary sampling units (PSU's) introduced by the weighting method, however. The weighting unit nonresponse adjustment for the CPS provides a practical example of this problem—since the adjustment cells pool data across PSU's, the standard error computations are not technically valid and tend to underestimate the true values.

Finally, weighting methods do not generalize readily to an arbitrary pattern of missing values, such as that occurring with item nonresponse. Little and David (1983) showed how the weighting method can be extended to handle *monotone* patterns of nonresponse, such as those occurring with attrition from a panel study. Let y_0 denote the set of fully observed variables; the pattern of missing data is monotone if the other variables can be arranged into sets $y_1, y_2, \ldots, y_k$ such that (a) all variables in each set have the same pattern of missing data and (b) for $j = 1, \ldots, k - 1$, y_j is *more observed* than y_{j+1} (Rubin 1974); that is, y_j is observed for all cases for which y_{j+1} is observed. Let r_j be the missing-data indicator for y_j, with value 1 if y_j is present and 0 otherwise. The weighting scheme is based on the following set of regressions: (a) r_1 on y_0, using all cases; (b) r_2 on y_1 and y_0, using cases with $r_1 = 1$; (c) r_3 on y_2, y_1, and y_0, using cases with $r_2 = r_1 = 1$; and (d) r_k on $y_{k-1}, y_{k-2}, \ldots, y_0$, using cases with $r_{k-1} = r_{k-2} = \cdots = r_1 = 1$. As before, these regressions can be logistic or probit if necessary. Cases i with y_{1i} observed are assigned y_1-weights w_{1i}, where w_{1i}^{-1} is proportional to the predicted probability that $y_{1i} = 1$ from (a). Cases i with y_{2i} observed are assigned y_2-weights $w_{2i} = w_{1i}w_{2\cdot 1i}$, where $w_{2\cdot 1i}^{-1}$ is proportional to the predicted probability that $y_{2i} = 1$ from (b), and so on. Finally, cases i with y_{ki} observed are assigned y_k-weights $w_{ki} = w_{1i}w_{2\cdot 1i} \cdots w_{k\cdot 12\cdots k-1,i}$, where $w_{k\cdot 12\cdots k-1,i}^{-1}$ is proportional to the predicted probability that $y_{ki} = 1$ from (d). [The 1973 CPS–IRS–SSA Exact Match Study provides an alternative weighting application. In that case, Scheuren (1981) took an approach that led to 15 different weights.]

The cases and weights used for any analysis are those corresponding to the least observed y included in the analysis. For example, if the analysis involves variables in y_2, y_3, and y_4, then cases with y_4 observed are used, with the y_4 weights.

Discussion of the relative merits of imputation and weighting is clarified by considering the extent to which they can produce the same answers to questions of estimation and inference. In Section 5.2, I show that certain of the predictive mean matching procedures discussed in Section 4 can be recast as weighting methods. Thus the distinction between the methods is in some circumstances purely cosmetic.

5.2 Integer-Weighting Analogs of Predictive Mean Imputation Methods

We learn in elementary statistics that when values in a sample are repeated we can replace them by a sample of distinct values with associated frequencies. In a similar way, when the value of a missing variable y is imputed by substituting the value of a randomly chosen respondent, the data can be replaced by a weighted sample in which respondent i is weighted by 1 plus the number of times it appears as a donor. This analogy between hot-deck imputation and integer weighting is well known (Kalton and Kasprzyk 1986; Oh and Scheuren 1983). Suppose now that matching is achieved using a set of fully observed variables x. The weighted file still corresponds to the imputed file, provided that respondents and nonrespondents match exactly on x. If the matching is not exact on x, then some coarsening of the x information results from replacing the full data on x by the weighted data on x for respondents to y. If the effects of this coarsening are negligible, then estimates from diverse statistical analyses, including estimation of means and totals for any population subgroup, multiple regression, simultaneous equation modeling, log-linear models for discrete data, and so on, are the same for the imputed or the integer-weighted file.

By the same argument, the multiple imputation methods of Section 4.3 involving matches to more than one respondent have multiple weighting analogs, exact except for coarsening of the x information. Each set of imputations corresponds to a set of integer weights for respondents. If I multiple imputes are carried out, the weighted analog provides I columns of weights. Estimates from each imputed data set are replaced by weighted estimates using each set of weights in turn. Rubin's technology for combining these estimates can then be invoked in exactly the same way as for multiple imputation. Thus multiple integer weighting can provide a solution to the problem of inference from weighted samples noted in the previous section.

This argument generalizes without difficulty to the case in which y is a vector of variables missing for the same set of cases, as in unit nonresponse. It can also be generalized to handle a monotone pattern of nonresponse, with different sets of weights for each block of variables with the same pattern. Satisfactory weighting analogs of the matching methods in Section 4.2 have not been worked out for a general pattern of nonresponse, however.

6. CONCLUSION

This article has considered methodology for missing-data adjustments from a model-based perspective. It is argued that imputation and weighting can be useful methods of adjustment in applied settings. The performance of the methods depends on the quality of the statistical models (implicit or explicit) that underpin them, however.

Until recently, practical solutions to survey nonresponse like the CPS hot deck appeared divorced from missing-data methods based on more conventional model-based statistical theory. Now such methods appear closer to the mainstream of statistical thought. The hot deck is close to the bootstrap in spirit, and Rubin's multiple-imputation theory provides a formal justification of hot-deck methods by viewing the imputations as simulations from the Bayesian predictive distribution. Similarly, properties of weighting nonresponse adjustments have been illuminated by Rosenbaum and Rubin's (1983) propensity-score theory.

These developments may appear of academic interest to the practitioner faced with real and complex missing-data problems. I believe, however, that the introduction of model-based theory should greatly benefit future developments in applications. Statistical modeling of the data can guide the choice of predictors of missing values and cut down on ad hoc decisions. Refinements such as multiple imputation (or multiple weighting) provide measures of the added uncertainty from nonresponse. The assumptions underlying particular methods can be clarified, and properties of the nonresponse adjustment methods can be conveyed more systematically to the consumer.

ACKNOWLEDGMENTS

This research is supported by National Science Foundation Grant SES-84-11804. An earlier version of the article was presented at the Second Annual Research Conference of the U.S. Bureau of the Census. I thank the associate editor and referees for useful and constructive comments.

[*Received April 1986. Revised August 1987.*]

REFERENCES

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics,* 24, 3–61.

Bailey, L., Chapman, D. W., and Kasprzyk, D. (1986), "Nonresponse Adjustment Procedures at the U.S. Bureau of the Census," *Survey Methodology,* 12, 161–179.

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society,* Ser. B, 37, 129–146.

Colledge, M. J., Johnson, J. H., Pare, R., and Sande, I. G. (1978), "Large Scale Imputation of Survey Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 431–436.

Czajka, J. L. (1986), "Imputation of Selected Items in Corporate Tax Data: Improving Upon the Earlier Hot Deck," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 304–309.

Czajka, J. L., Hirabayashi, S. M., Little, R. J. A., and Rubin, D. B. (1987), "Evaluation of a New Procedure for Estimating Income and Tax Aggregates From Advance Data," in *Statistics of Income and Related Administrative Record Research: 1986–1987,* Washington, DC: U.S. Department of the Treasury.

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

David, M., Little, R. J. A., Samuhel, M. E., and Triest, R. K. (1983), "Nonrandom Nonresponse Models Based on the Propensity to Respond," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* pp. 168–173.

——— (1986), "Alternative Methods of CPS Income Imputation," *Journal of the American Statistical Association,* 81, 29–41.

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,* Ser. B, 39, 1–38.

Dixon, W. J. (ed.) (1983), *BMDP Statistical Software,* Los Angeles: University of California Press.

Fay, R. (1986), "Causal Models for Patterns of Nonresponse," *Journal of the American Statistical Association,* 81, 354–365.

Fellegi, I., and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation," *Journal of the American Statistical Association,* 71, 17–35.

Goldberger, A. S. (1980), "Abnormal Selection Bias," Discussion Paper 8066, University of Wisconsin, Social Systems Research Institute.

Greenberg, B., and Surdi, R. (1984), "A Flexible and Interactive Edit and Imputation System for Ratio Edits," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 421–426.

Greenlees, J. S., Reece, W. S., and Zieschang, K. O. (1982), "Imputation of Missing Values When the Probability of Response Depends on the Variable Being Imputed," *Journal of the American Statistical Association,* 77, 251–261.

Hanson, R. H. (1978), "The Current Population Survey, Design and Methodology," Technical Paper 40, U.S. Bureau of the Census, Washington, DC.

Hausman, J. A., and Wise, D. A. (1979), "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica,* 47, 455–473.

Heckman, J. (1976), "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement,* 5, 475–492.

Heeringa, S. G., and Lepkowski, J. M. (1986), "Longitudinal Imputation for the SIPP," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 206–210.

Hinkins, S., and Scheuren, F. (1986), "Hot Deck Imputation Procedure Applied to a Double Sampling Design," in *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 181–195.

Jones, S. M., and Chromy, J. R. (1982), "Improved Variance Estimators Using Weighting Class Adjustments for Sample Survey Nonresponse," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 105–110.

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

——— (1986), "The Treatment of Missing Survey Data," *Survey Methodology,* 12, 1–16.

Kalton, G., and Miller, M. (1986), "Effects of Adjustments for Wave Nonresponse on Panel Survey Estimates," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 194–199.

Kusch, G. L., and Clark, D. F. (1979), "Annual Survey of Manufactures General Statistics Edit," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association,* pp. 183–187.

Lillard, L., Smith, J. P., and Welch, F. (1982), "What Do We Really Know About Wages: The Importance of Nonreporting and Census Imputation," technical report, the Rand Corporation, 1700 Main Street, Santa Monica, California.

——— (1986), "What Do We Really Know About Wages: The Importance of Nonreporting and Census Imputation," *Journal of Political Economy,* 94, 489-506.

Little, R. J. A. (1979), "Maximum Likelihood Inference for Multiple Regression With Missing Values: a Simulation Study," *Journal of the Royal Statistical Society,* Ser. B, 41, 76-87.

——— (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association,* 77, 237-250.

——— (1985), "A Note About Models for Selectivity Bias," *Econometrica,* 53, 1469-1474.

——— (1986), "Survey Nonresponse Adjustments for Estimates of Means," *International Statistical Review,* 54, 139-157.

Little, R. J. A., and David, M. (1983), "Weighting Adjustments for Nonresponse in Panel Surveys," unpublished manuscript.

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

Little, R. J. A., and Smith, P. (1983), "Multivariate Editing and Imputation for Economic Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 415-420.

——— (1987), "Editing and Imputation for Quantitative Survey Data," *Journal of the American Statistical Association,* 82, 58-69.

Little, R. J. A., and Su, H. L. (in press), "Missing Data in Panel Surveys," in *Panel Surveys,* eds. G. Duncan, D. Kasprzyk and M. P. Singh, New York: John Wiley.

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.

Oh, H. L., and Scheuren, F. E. (1980), "Estimating the Variance Impact of Missing CPS Income Data," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 408-415.

——— (1983), "Weighting Adjustment for Unit Nonresponse," in *Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographies,* eds. W. G. Madow, I. Olkin, and D. B. Rubin, New York: Academic Press, pp. 143-184.

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

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* 70, 41-55.

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

——— (1978), "Multiple Imputations in Sample Surveys—a Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 20-34.

——— (1982), "Imputing Income in the CPS," in *The Measurement of Labor Cost,* ed. J. Triplett, Chicago: University of Chicago Press, pp. 333-343.

——— (1986), "Statistical Matching and File Concatenation With Adjusted Weights and Multiple Imputations," *Journal of Business & Economic Statistics,* 4, 87-94.

——— (1987), *Multiple Imputation in Sample Surveys and Censuses,* New York: John Wiley.

Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association,* 81, 366-374.

Sande, G. (1979), "Numerical Edit and Imputation," unpublished paper presented to the International Association for Statistical Computing at the 42nd Session of the International Statistical Institute, Manila, the Philippines.

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

Santos, R. L. (1981), "Effects of Imputation on Regression Coefficients," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 140-145.

Scheuren, Fritz (1976), "Preliminary Notes on the Partially Missing Data Problem—Some (Very) Elementary Considerations," working paper, Social Security Administration Methodology Group, Washington, DC.

——— (1981), "Methods of Estimation for the 1973 Exact Match Study," in *Studies From Interagency Data Linkages,* Report 10, Publication 13-11750, U.S. Department of Health and Human Services, Social Security Administration, pp. 1-122.

Vacek, P. M., and Ashikaga, T. (1980), "An Examination of the Nearest Neighbor Rule for Imputing Missing Values," in *Proceedings of the Section on Statistical Computing, American Statistical Association,* pp. 421-425.

Welniak, E. G., and Coder, J. F. (1980), "A Measure of the Bias in the March CPS Earnings Imputation Scheme," in *Proceedings of the Section on Survey Research Methods, American Statistical Association,* pp. 421-425.

Winer, R. S. (1983), "Attrition Bias in Econometric Models Estimated With Panel Data," *Journal of Marketing Research,* 20, 177-186.

Comment

I. G. Sande
Business Survey Methods Division, Statistics Canada, Ottawa, Ontario K1A 0T6, Canada

I always find Little's works stimulating. They oblige me to think about old problems in new ways and to extend my insight. This does not necessarily mean that I always agree with them, and, indeed, acceptance of what one sees in print is not the idea of reading.

In this article, Little discusses a method of imputation that is a mixture of model-based and nearest neighbor ideas—the matching is based on the predictions of missing values, but the actual imputations are from the matching record so identified. I have problems with this idea, because the match is done on the predictions, not the predictors, and implies a great deal of faith in the model that one is using.

I do not have such faith, especially as far as economic data are concerned. Certainly one would have to check out the predictive efficiency of a model before proceeding with this method. One positive feature is that some of the predictors could be nonnumeric, whereas