Statistical Methods in Medical Research 1996; **5**: 215–238

Handling missing data in survey research

JM Brick and **G Kalton** Westat Inc., Rockville, Maryland, and Joint Program in Survey Methodology, University of Maryland, College Park, Maryland, USA

Missing data occur in survey research because an element in the target population is not included on the survey's sampling frame (noncoverage), because a sampled element does not participate in the survey (total nonresponse) and because a responding sampled element fails to provide acceptable responses to one or more of the survey items (item nonresponse). A variety of methods have been developed to attempt to compensate for missing survey data in a general purpose way that enables the survey's data file to be analysed without regard for the missing data. Weighting adjustments are often used to compensate for noncoverage and total nonresponse. Imputation methods that assign values for missing responses are used to compensate for item nonresponses. This paper describes the various weighting and imputation methods that have been developed, and discusses their benefits and limitations.

1 Introduction

Missing data occur in survey research for a variety of reasons and are a cause of concern for survey analysis. A considerable amount of research has been conducted in the past 20 years developing and refining methods for compensating for missing survey data, and the research is ongoing. This paper reviews this research with an emphasis on methods that are in widespread current use.

This review focuses on general-purpose methods of weighting and imputation that deal with the missing data during the production of the survey's analysis file. The aim of these methods is to compensate for the missing data in such a manner that the analysis file may be subjected to any form of analysis without the need for further consideration of the missing data. As the review will indicate, that aim is sometimes overly ambitious. In some cases specific techniques for handling the missing data that are tailor-made for a particular method of analysis may be required. Such techniques, which are described by Little and Rubin,[1] are not treated here.

Missing survey data can be classified as arising from four main sources, with the form of compensation depending on the source. The most widely recognized source is total or unit nonresponse, which occurs when no survey data are collected for an element selected for the sample. (We use the term 'element' to refer to the unit of analysis; in most health surveys the element will be a person, but it could be, for example, a household or a hospital visit.) Total nonresponse results from refusals to participate in the survey, noncontacts (not-at-homes), and other reasons such as a language barrier, deafness or being too ill to participate. Compensation for total nonresponse is usually made by means of weighting adjustments in which respondents are assigned greater weight in the analysis in order to represent the nonrespondents.

Address for correspondence: JM Brick, Westat Inc., 1650 Research Boulevard, Rockville, MD 20850-3129, USA.

0962-2802(96)112RA

A second source of missing survey data is noncoverage, which occurs when some elements in the population of inference for the survey are not included in the survey's sampling frame. These missing elements have no chance of selection for the sample and hence go unrepresented. Like total nonresponse, compensation for noncoverage is usually made by means of weighting adjustments. However, the form of weighting adjustment is different. In the case of total nonresponse, the nonrespondents can be identified within the selected sample, and the weighting adjustment can be applied using sample data alone. In the case of noncoverage, the sample provides no information about the missing elements, so that the weighting adjustments need to be based on external data sources.

A third source of missing survey data is item nonresponse, which occurs when a sampled element participates in the survey but fails to provide acceptable responses to one or more of the survey items. Item nonresponse may arise because a respondent refuses to answer an item on the grounds that it is too sensitive, does not know the answer to the item, gives an answer that is inconsistent with answers to other items and hence is deleted in editing, or because the interviewer fails to ask the question or record the answer. The usual form of compensation for item nonresponse is imputation, which involves assigning a value for the missing response.

A fourth source of missing data is what may be termed partial nonresponse. Partial nonresponse falls between total and item nonresponse. Whereas total nonresponse relates to a failure to obtain any responses from a sampled element and item nonresponse usually implies the failure to obtain responses for only a small number of survey items, partial nonresponse involves a substantial number of item nonresponses. Partial nonresponse can occur, for instance, when a respondent cuts off the interview in the middle, when a respondent in a panel survey fails to provide data for one or more of the waves of the panel, or when a respondent in a multiphase survey provides data for some but not all phases of data collection. Partial nonresponse may be handled by either weighting or imputation. With the weighting approach, the partial nonrespondents are dropped from the analysis file, and weighting adjustments similar to those used for total nonrespondents are used in compensation. However, that approach involves discarding the responses that the partial nonrespondents did provide (except to the limited extent that these responses can be incorporated into the weighting adjustments). With the imputation approach, the partial nonrespondents are retained in the analysis file and imputation is used to fill in all their missing responses. However, with this approach it is extremely difficult to maintain the associations between all the survey variables (see Section 3). The choice between these two approaches is not clear cut, and depends on the circumstances. In many cases the weighting approach is preferred. Compensation for partial nonresponse will not be considered further here. Discussions of the choice between weighting and imputation for handling partial nonresponse occurring in the form of wave nonresponse in panel surveys are provided by Kalton[2] and Lepkowski.[3]

The next two sections review weighting and imputation methods for handling missing survey data. Section 2 discusses both weighting methods for total nonresponse based on data internal to the sample and weighting methods for noncoverage using data from external sources. Section 3 reviews the variety of imputation methods that have been developed to assign values for item nonresponses, with a particular emphasis on the widely used hot-deck and regression-based methods. The final section of the paper presents some concluding remarks.

2 Weighting

In most sample surveys, weights are attached to each respondent record and then used in analyses to produce approximately unbiased estimates of parameters of the target population. These weights compensate for the facts that sampled elements may be selected at unequal sampling rates and have different probabilities of responding to the survey, and that some population elements may not be included in the list or frame used for sampling. The main objective of weighting is to reduce bias in survey estimates by making each respondent represent a different fraction of the target population.

Analysing data using weights does have some disadvantages. One is that weighting may complicate analyses since many standard statistical software packages either do not recognize weights or else treat them as counts of identical observations. In the latter case, the sum of the weights is equated to the sample size, leading to incorrect estimates of the precision of the survey estimates. A second disadvantage is that weighting often increases the variances of the estimates. Kish[4] shows that in frequently occurring conditions differential weighting adjustments increase the variance of the estimate by $(1 + L)$, where L is the relative variance of the weights. In effect, weighting often reduces bias while increasing variance. Given the bias-variance trade off, it would be useful to be able to compare the mean square errors of alternative estimators, but this is seldom possible because the bias components cannot be estimated.

Weighting is typically performed in three stages. The first stage is to produce base weights that account for the unequal probabilities of selecting elements from the sampling frame. Base weights are set equal to, or proportional to, the inverses of the probabilities of selecting the elements ($w_i = \pi_i^{-1}$, where π_i is the probability of selecting element i for the sample). The second stage is to adjust the base weights of the respondents to compensate for sampled elements that did not respond to the survey. The most common method of doing this is to partition the sample into weighting classes and increase the base weights of the respondents in a given weighting class by the inverse of the response rate in that class. This procedure is called a weighting class adjustment. The third stage further adjusts the weights of the respondents so that estimates of certain population totals conform to known values for these totals. The primary purpose of this adjustment is to reduce the bias due to incomplete coverage of the target population. A well known and frequently used method of making this adjustment is poststratification, a method that is similar in form to the weighting class adjustment. The key difference between the two methods is that the weighting class adjustment makes the respondent data conform to sample totals (for respondents and nonrespondents combined) whereas the poststratification adjustment makes them conform to population totals from external data sources. To reflect this difference, Kalton and Kasprzyk[5] referred to the last two weighting stages as sample weighting adjustments and population weighting adjustments, respectively.

Our concern here is with the use of weighting adjustments to compensate for missing data, that is, the second and third stages in the previous paragraph. We will consider the effect of such adjustments in estimating a population mean. Our starting point is the unadjusted weighted sample mean using the base weights w_i that are proportional to the inverses of selection probabilities. The bias arising from using this mean to estimate the population mean in the presence of nonresponse can be expressed in a number of ways. One approach is to adopt a deterministic model for nonresponse, assuming that every element in the population can be classified into either a stratum of respondents who always respond or a stratum of nonrespondents who never respond to the survey. The bias of the unadjusted mean using data only from the respondents can then be expressed as

$$B(\bar{y}_r) = P_m(\overline{Y}_r - \overline{Y}_m) \tag{2.1}$$

where $\bar{y}_r$ is the respondent mean computed using the base weights, P_m is the nonresponse rate, $\overline{Y}_r$ is the mean of the responding elements in the population, and $\overline{Y}_m$ is the mean of the nonresponding elements.[6,7] Similar expressions can be written for other statistics, such as the difference between subclass means.[8]

It is clear from equation (2.1) that the bias of $\bar{y}_r$ is a function of both the nonresponse rate and the difference in the characteristics of the responding and nonresponding elements. Achieving a high response rate is the only way to be confident that the bias is negligible, since it is usually impossible to assess the differences in the characteristics of the respondents and nonrespondents. This is especially true in surveys with many items, in which the differences between respondents and nonrespondents may be small for some items and large for others.

An alternative way of modelling the nonresponse mechanism is to assume that the population elements have different (nonzero) probabilities of responding to the survey if sampled.[9–13] Under this model, the approximate bias of $\bar{y}_r$ can be expressed as

$$B(\bar{y}_r) \approx \frac{1}{N\overline{\phi}}\sum(Y_i - \overline{Y})(\phi_i - \overline{\phi}) \tag{2.2}$$

where ϕ_i is the probability that element i responds to the survey ($\phi_i > 0$), $\overline{\phi} = \Sigma\phi_i/N$, and N is the number of elements in the population.[13] This expression shows that the bias is a function of the relationship between the characteristic being estimated and the response probability, with $\bar{y}_r$ being unbiased only when Y_i and ϕ_i are uncorrelated.

If the response probabilities ϕ_i were known, the bias in $\bar{y}_r$ could be completely eliminated by applying ϕ_i^{-1} as a nonresponse weighting adjustment, leading to an adjusted weight of $w_i^* = w_i\phi_i^{-1}$. Unfortunately, the response probability is unknown and it must be estimated by using a model of the response mechanism.

A very simple model for the response probabilities is the missing completely at random (MCAR) model which assumes that the missing observations are a random subsample of the full sample.[1] Since this implies that $\phi_i = \overline{\phi}$ for all i, $B(\bar{y}_r)$ in equation (2.2) is zero and $\bar{y}_r$ is approximately unbiased. In most surveys, the MCAR model is not realistic; studies typically reveal substantial differences between respondents and nonrespondents. A somewhat more reasonable model is the missing at random (MAR) model that assumes that the population can be partitioned into classes such that in each class the missing observations are a random subsample of all the elements in the class. The unadjusted respondent mean is biased under this model.

Different models for response probabilities and methods for adjusting weights are presented below. The first two methods discussed are sample weighting adjustments. The remaining methods are commonly used as population weighting adjustments. Most methods can, however, also be used to make either sample or population weighting adjustments.

2.1 Weighting class adjustments

Sample weighting adjustments attempt to reduce nonresponse bias by allocating the base weights of nonrespondents to the respondents. A crude adjustment is to increase the base weights of respondents by the inverse of the overall weighted response rate

$$\hat{\phi} = \Sigma^r w_i / (\Sigma^r w_i + \Sigma^m w_i)$$

where $\sum^r w_i$ and $\sum^m w_i$ are the sums of the base weights for respondents and nonrespondents, respectively. Using the base weight, w_i, or the adjusted weight, $w_i^* = w_i \hat{\phi}^{-1}$, gives the same values for all estimates that are averages (e.g. means, proportions, regression coefficients). The adjustment affects only estimates of totals. The adjusted weight w_i^* leads to approximately unbiased estimates only when the MCAR assumption holds.

Other sample weighting adjustment procedures require data on the characteristics of both the respondents and the nonrespondents. For example, class membership must be available for both respondents and nonrespondents to form weighting class adjustments. The weighting class estimator of the mean, $\bar{y}_{wc}$, is formed by adjusting the weights of the respondents in each class by the inverse of the response rate in the class ($w_{ci}^* = w_{ci}\hat{\phi}_c^{-1}$, where $\hat{\phi}_c$ is the weighted response rate in class c). Under the response probability model, the approximate bias of this estimator is[13]

$$B(\bar{y}_{wc}) \approx \frac{1}{N}\sum\sum\frac{1}{\bar{\phi}_c}(Y_{ci} - \bar{Y}_c)(\phi_{ci} - \bar{\phi}_c) \tag{2.3}$$

As can be seen from equation (2.3), the weighting class estimator is unbiased if the MAR assumption holds within the classes, that is, if $\phi_{ci} = \bar{\phi}_c$ for all classes. The most common approach for forming weighting classes is to try to create classes based on the auxiliary variables available for both respondents and nonrespondents such that the MAR assumption is satisfied. Little[11] terms this response propensity stratification. The auxiliary variables are used to separate the sample into weighting classes with different response rates, since variation in response rates implies that the MAR assumption is not met for any combination of such classes. It is then hoped that the MAR assumption is satisfied within these classes.

A less stringent condition for the weighting class estimator to be unbiased is that Y_{ci} and ϕ_{ci} are independent within weighting classes. Assuming that Y_i and ϕ_i are independent conditional on the full set of auxiliary variables available, $\mathbf{x}$, Little[11] describes a method that he calls predicted mean stratification for forming weighting classes to satisfy the conditional independence assumption within classes. The method involves developing a regression equation for predicting Y from $\mathbf{x}$ for the respondents, computing the predicted Y values for the full sample, dividing these Y values into groups, and using these groups as the weighting classes. An attraction of predictive mean stratification is that it leads to a more precise estimate of $\overline{Y}$ than response propensity stratification. However, predictive mean stratification relates to a specific variable Y, and may not be appropriate for other survey variables. Since most surveys are multipurpose in nature, collecting data on large sets of variables, response propensity stratification is generally the preferred approach for constructing weighting classes.

The objective of the weighting class estimator is to eliminate, or at least reduce, the bias associated with the unadjusted mean, $\bar{y}_r$. To see the effect of the adjustment, the bias of $\bar{y}_r$ from equation (2.2) can be expressed using an analysis of covariance decomposition as

$$\begin{aligned} B(\bar{y}_r) &\approx \frac{1}{N\phi}\sum\sum(Y_{ci} - \overline{Y}_c)(\phi_{ci} - \phi_c) + \frac{1}{N\overline{\phi}}\sum N_c(\overline{Y}_c - \overline{Y})(\overline{\phi}_c - \overline{\phi}) \\ &\approx A + B \end{aligned} \tag{2.4}$$

where N_c is the number of elements in weighting class c. The first term, A, is comparable to the bias of the weighting class estimator given in equation (2.3). Thus the main effect of the nonresponse weighting class adjustment is to remove the term B. It may be noted that B is nonzero only if both the response rates and the means vary across the classes. Thomsen[14] derives a similar result using the deterministic nonresponse model. As Thomsen observes, the weighting class estimator does not necessarily reduce the absolute value of the bias. In particular, if A and B are of different signs, the bias of $\bar{y}_r$ may be less than that of $\bar{y}_{wc}$.

In many cases little is known about the nonrespondents. Often all that is known is their sample characteristics, such as the primary sampling units (PSUs) and the strata in which they are located. In this situation, the small number of auxiliary variables available for constructing weighting classes makes the choice of these classes straightforward. Weighting class adjustments based on sample characteristics are used in many health surveys, often as the first stage of adjustment to the base weights. Cox and Cohen[15] describe the use of such weighting class adjustments as part of the development of the weights for the US National Medical Care Expenditure Survey, a household survey that collected data on the use of health care services by individuals and families together with the associated costs. While such adjustments can be beneficial, it needs to be recognized that the paucity of data about the nonrespondents limits the ability of these adjustments to reduce the bias in the survey estimates.

In some cases, a large number of auxiliary variables are available for use in making nonresponse adjustments. This may happen when the sample is selected from a register that contains a good deal of information about those listed or with partial nonresponse, such as later wave nonresponse in a panel survey. Preliminary analyses may be called for to determine how the auxiliary variables may best be used in forming weighting classes. For example, logistic regression may be used to model response status (respondent/nonrespondent) as a function of the auxiliary variables, and then the weighting classes may be formed using the important auxiliary variables from the model. Alternatively a branching algorithm such as CART,[16] CHAID[17] or SEARCH[18] may be used to form the weighting classes directly. These methods are applied by Rizzo *et al.*[19] in an exploratory study of methods for adjusting for nonresponse in later waves of a household panel survey.

When many auxiliary variables are available, an alternative to weighting class adjustments is to base the adjustments directly on a logistic (or probit) regression model in which response status is regressed on the auxiliary variables available for respondents and nonrespondents.[19,20] The nonresponse adjustment for a respondent is then obtained from the inverse of that respondent's predicted response probability from the model. This adjustment results in an approximately unbiased estimator if the response model corresponds to the regression model used to estimate the predicted response probabilities. Typically the regression model will contain only main effects for the auxiliary variables with perhaps a few interactions. If all the auxiliary variables are categorical (or categorized) and if all the interactions between the auxiliary variables are included in the model, the model is, in effect, equivalent to one in which all the regressors are simply indicators of cell membership in the crosstabulation of the auxiliary variables. In this situation, the weighting adjustment from the model is essentially the same as that from the weighting class method. The attraction of the regression model is that, by excluding some interaction terms, more auxiliary variables can be included in the nonresponse adjustment.

2.2 Response probability adjustments

In the 1950s, Politz and Simmons[21,22] developed a method for nonresponse adjustment for surveys in which the interviewer made a single call on each sampled person. The method, which follows up on an idea suggested by Hartley,[23] involves asking respondents how many of the previous five evenings they were at home at about the time of the interview. Response probabilities are computed for each respondent as the proportion of the six evenings the respondent was at home (including the one in which he or she was interviewed), and the inverses of these probabilities are then used as nonresponse weighting adjustments. It should be noted that, in this simple form, the method does not require information about the nonrespondents. However, it attempts to compensate only for intermittent not-at-homes; it does not compensate for refusals and other types of nonresponse, or for those who were out on each of the six evenings.

Another approach to determining response probabilities involves making call-backs or sending additional mailings to nonrespondents at previous calls or to previous mailings, and then developing a response probability model for the successive calls. Thomsen and Siring,[24] for instance, develop a model in which the probability of response at the first call is p and after that it is $(1 - p - f)(1 - \Delta p - f)^{c-2}\Delta p$ at call $c \geq 2$, where f is the probability that the interviewer obtains no response and categorizes the case as a refusal and Δp is the probability of response at any call after the first. Note that this model assumes that f is constant across calls and that Δp is constant across repeat calls. The parameters of the model can be estimated by fitting the model to the observed pattern of calls. The survey estimates may then be computed taking into account the differing response probabilities of the respondents at different calls. A variety of response probability models of this type have been developed for use in forming particular estimates.[24–29] However, these methods are not applied in large-scale multipurpose surveys. This may be due to the lack of confidence in the assumptions of the response models, the complexity of computing the predicted probabilities for some of the methods, or the increase in variance associated with the differential weights.

A related approach that has been studied frequently is to adjust the weights of later respondents to account for all the nonrespondents. The basic model implied by this approach is that the nonrespondents are MCAR among the set of later respondents. Bartholomew,[30] for example, employs and investigates the assumption that the respondents at the second call are a random sample of all sample members not responding at the first call. An obvious extension of this model is to assume that the nonrespondents are MAR among the set of later respondents, that is, that they are missing at random within weighting classes based on information available for both later respondents and nonrespondents.

This last approach bears a resemblance to the basic theory developed by Hansen and Hurwitz[31] for two-phase sampling for nonresponse, in which a subsample of the sample of initial nonrespondents is selected for intensive follow-up. As an example, a subsample of 50% of the nonrespondents in the National Survey of Family Growth, Cycle IV,[32] was selected and their responses were weighted up by a factor of 2 to represent all nonrespondents. The difference between the two-phase sampling procedure and the above approach is that the former draws a probability sample of initial nonrespondents and hence does not need to rely on any model assumptions (provided that all elements in the second-phase sample respond) whereas the above approach depends on the assumption that the final nonrespondents are MCAR (or MAR) among the set of initial nonrespondents.

2.3 Poststratification adjustments

Thus far, the adjustments discussed have been sample weighting adjustments that do not address biases due to noncoverage of the target population. For example, in telephone surveys the methods do not compensate for persons excluded from the sampling frame because they live in households without telephones. Adjusting the weights so that the sample conforms to known population totals from an external source is a way of attempting to reduce noncoverage biases. In some circumstances, population weighting adjustments may also reduce the variances of the estimates; sample weighting adjustments nearly always increase variances.

The data requirements for population weighting adjustments differ from those of sample weighting adjustments. Population weighting adjustments do not need data for nonrespondents; instead they require known population totals for the target population. A concern with population weighting adjustments is that serious biases or inconsistencies may arise when the characteristics used to classify the population and the respondents to the survey are not measured in the same way.

Poststratification,[33–35] or ratio estimation, is one of the most frequently used population weighting adjustments. In poststratification, weights for elements in a class are multiplied by a factor so that the sum of the weights for the respondents in the class equals the population total for the class ($w^*_{ci} = w_{ci}(X_c/\hat{X}_c)$, where X_c is the known population total in class c and $\hat{X}_c$ is the estimate of this total based on the unadjusted weights). The only difference between the poststratified estimator and the weighting class estimator is that the adjustment factor for the poststratified estimator is based on the known rather than the estimated total for the class.

The approximate bias of the poststratified mean is given by equation (2.3), but in this case ϕ_i is the overall probability of the sampled individual being included in the sampling frame and responding to the survey. The poststratified estimator is approximately unbiased provided the MAR assumption holds and the poststratified classes and the MAR model classes coincide.

Poststratification is widely used in household surveys to control the weighted sample totals to known population totals for certain demographic subgroups. For example, in the US National Health Interview Survey,[36] poststratification by age, race and sex is employed. Poststratification is also used in other types of health surveys, including the US National Ambulatory Medical Care Survey,[37] a list sample of physicians used to estimate the provision and use of ambulatory medical care services. In that survey, the sample is poststratified to counts of the number of office-based physicians categorized by medical speciality.

2.4 Raking adjustments

When several auxiliary variables are available from the survey along with the corresponding external population distributions, poststratification adjustment, may not be possible or desirable. In some cases, the population totals of the complete crossclassification of the auxiliary variables may not be known, while the marginal distributions for each variable are available. Even if the full crossclassification is known, the number of respondents in each cell may be small or zero and this can lead to inconsistent and highly variable estimates.

Raking,[13,38–40] sometimes called raking ratio or rim estimation, is an alternative method of adjustment that ensures that the adjusted weights of the respondents conform to each of the marginal distributions of the auxiliary variables. Raking involves an iterative adjustment of the weights, using an iterative proportional fitting algorithm. First, the weights are ratio adjusted to conform to the marginal distribution of the first auxiliary variable. These adjusted weights are then ratio adjusted to conform to the marginal distribution of the second auxiliary variable, etc. The first iteration concludes when the last auxiliary variable is fitted. Subsequent iterations are performed until the weights conform to the marginal distributions of all the auxiliary variables. Under general conditions[41] the algorithm converges to a solution.

The conditions under which the raked mean is unbiased are rather complex. Consider the simple case of raking with two auxiliary variables, and assume that the MAR model holds within each of the cells of the crossclassification of these variables. As a general formulation, the response probability in cell (*cd*) may then be represented by $\phi_{cdi} = \alpha_c\beta_d + \gamma_{cd}$. In general, the raked mean is biased under this model unless $\gamma_{cd} = 0$, i.e. unless $\phi_{cdi} = \alpha_c\beta_d$.[13,40,42] An exception occurs, however, when there is no interaction in the study variable in the crossclassification; in this case the raked mean is unbiased even if $\gamma_{cd} \neq 0$. The poststratified estimator, with population totals for the cells of the crossclassification as the control totals, is of course unbiased under this model whether or not $\gamma_{cd} = 0$.

Raking is a widely used method of adjustment in health and other surveys when many population control totals are available. For example, raking was used in the 1991

General Social Survey in Canada, a random digit dial telephone survey that concentrated on health issues.[43] In that survey, province, age and sex control totals were used for raking.

2.5 Calibration adjustments

Poststratification and raking are two adjustment methods that fit within the broader framework of generalized raking or calibration estimation.[42,44] The general approach is to adjust the weights so that they satisfy the condition that their sum is equal to the population total for each of the auxiliary variables and that the distance between the unadjusted and adjusted weights is minimized. Distance may be defined according to a variety of different distance functions. The condition on the sum of the weights is called a calibration equation. For example, suppose that the population count for each class in the crossclassification of the auxiliary variables (X_c) is known and that these counts are used to define the calibration equations

$$\sum_{i\in c} w^*_{ci} = X_c \quad \text{for all } c$$

Deville *et al.*[42] show for this case that the adjusted weights that minimize the distance from the inverse of the selection probabilities are the poststratified weights, $w^*_{ci} = w_{ci}(X_c/\hat{X}_c)$, regardless of the distance function.

Different forms of calibration estimators can be obtained by choosing different calibration equations and distance functions. Even though the weighting adjustments that result by choosing different distance functions are similar when the sample sizes are large,[19,42] some of the choices of distance function result in estimators that are important from other perspectives.

A linear distance function leads to the generalized regression estimator, an extension of the ordinary regression estimator that is appropriate for use with many auxiliary variables. The generalized regression estimator has been studied and employed in a number of applications.[45–51] A drawback of the generalized regression estimator is that it may result in some negative adjusted weights.[51] Various modifications to the regression estimator have been proposed to avoid this drawback.[52] Fuller *et al.*[53] describe an application of regression weighting in the US 1987–1988 Nationwide Food Consumption Survey that includes 27 auxiliary variables and employs a modification to ensure positive weights.

Deville *et al.*[42] examine other choices of distance function and describe a computer program for computing the weighting adjustments for four distance functions. They also study the bias and variance of the poststratified and raking ratio estimators as members of the class of calibration estimators. One of the benefits of estimators being members of the class of calibration estimators is that they share the same asymptotic variance. Thus, it is possible to estimate the variance of estimators in the class that are difficult to derive analytically, such as that of the iterative raking ratio estimator, by using the simpler variance formula for the generalized regression estimator.

The calibration framework is also attractive because it is relatively simple to constrain the variability in the adjusted weights by the choice of the distance function. Two distance functions that have this property are the logit and truncated linear methods. For example, the logit distance function is given by

$$G(t) = \frac{(1-L)(U-1)}{U-L}\left[(t-L)\log\left(\frac{t-L}{1-L}\right) + (U-t)\log\left(\frac{U-t}{U-1}\right)\right] \quad \text{if } L < t < U$$

and $G(t) = \infty$ otherwise, where L and U are lower and upper limits on the adjustments that may be specified within certain constraints. Truncated distance functions may result in weight distributions with masses near the upper and lower limits, but they provide a simple and direct method for avoiding negative weights and for restricting the range of weights possible. Controlling the range of the weights is attractive because large variability in the weights can substantially lower the precision of the survey estimates.

2.6 Trimming adjustments

A serious concern with the use of weighting adjustments to compensate for total nonresponse and noncoverage is that the process may give rise to a wide variability in the final weights, with a resultant loss of precision in the survey estimates. This concern is of particular importance when many auxiliary variables are used in the adjustments and when the adjustments are made in several stages, since a wide variability in weights can easily arise in such cases. Various procedures have been developed to avoid this outcome. Often weighting classes are collapsed when the adjustments in some classes are too large.[13,35,54] Also the truncated distance functions described above were devised to limit the variability in the weighting adjustments.

When a weighting process has produced very variable weighting adjustments, the largest weights may be trimmed to reduce the weight variability. Potter[55–57] describes a variety of *ad hoc* procedures to trim the final weights. Some of these procedures operate by inspecting the distribution of the weights and trimming the most extreme weights so they do not exceed a specified limit. The excess weight is then redistributed to the observations that are not trimmed. These procedures are designed to reduce the variances of the survey estimates but do not address the biases that the trimming may introduce. In order to take account of the bias, other trimming procedures use the values of the characteristics observed in the survey in addition to the weights to determine the cut-points for trimming. Two related methods that address the problem of extreme weights are weight shrinkage[58] and weight smoothing.[59]

The trimming and weight shrinkage adjustment methods are in line with the argument advanced by Kish[4] that the mean square error of the resultant estimates should be examined. Adjustments that add substantially to the variances without making sufficient reductions to the biases of the survey estimates may increase their mean square errors, thus lowering the overall quality of the estimates.

3 Imputation

As with total nonresponse and noncoverage, missing data arising from item nonresponse can lead to biased survey estimates if they are simply ignored, that is, if the analysis is restricted to the records with responses for the items in question. If the nonresponse rate for an item is low, as often applies, the amount of bias in univariate analyses for that item will be small, so that dropping the records with missing records may appear to be a reasonable procedure. However, much survey analysis is multivariate in nature and low item nonresponse rates for several items together may result in a sizeable proportion of records with missing data for one or more of the items involved in a particular analysis. Thus, for example, a substantial proportion of records may be dropped when a multiple regression analysis with many variables is restricted to records with complete data for all the variables involved, often known as the 'complete case' solution for missing data. By assigning values for the missing responses, imputation enables all relevant records to be retained in every analysis. The main objective of imputation is to reduce and ideally eliminate the bias in survey estimates caused by ignoring records with missing data. We will discuss later its success in achieving this objective.

In addition to addressing nonresponse bias, there are some other attractive features of imputation. First, the creation of a complete dataset makes the analyst's task easier since the awkward problems of handling complex patterns of missingness across many variables are eliminated. Second, the results of all analyses conducted with the dataset are consistent with one another. This consistency may well not hold when the complete case approach is used for each analysis, since then often different subsets of records will be involved in different analyses. Third, the presentation of the survey results is simpler. For example, in a two-way table there is no need to have an extra row and column to provide counts of the item nonrespondents. Fourth, unlike the complete case approach, imputation retains all the reported data in multivariate analyses. It may therefore produce more precise estimates of multivariate parameters.

As well as its benefits, imputation also has its costs. One is that it may distort the association between variables. Imputation schemes can be chosen to maintain the associations of the variable subject to imputation and certain key variables, but associations with other variables may be attenuated. Another cost is that imputation fabricates data to some degree. Thus, standard analyses applied to a dataset completed by imputation will overestimate the precision of the survey estimates. These costs of imputation are discussed later, after the most common imputation methods have been described.

There is one type of imputation, deductive imputation, for which these costs do not apply. Deductive imputation is applicable when a missing response can be deduced from responses to other items. For example, a person aged under 16 years may be imputed to be single and a university teacher may be imputed to have a college degree. Deductive imputation is often considered to be editing rather than imputation. As these examples indicate, deductive imputation assumes a high degree of certainty about the missing response (but not necessarily absolute certainty – there may be a few university teachers without degrees). With less certainty, other imputation methods are needed.

When imputation is used, the imputed values need to be identifiable in the data file, so that analysts can, when necessary, distinguish imputed values from actual responses. Analysts may want to identify imputed values when examining unusual findings, when they want to employ their own procedures for handling item nonrespondents, or for computing variances of survey estimates. Often imputed values are identified by flags in the data file. Such flagging should be standard practice. In addition, analysts may need information about the imputation classes for purposes of variance estimation (see Section 3.2).

3.1 Imputation methods

The aim of imputation is to fill in the missing responses with appropriate values given the responses to the other items for the respondent involved and any other information known for that respondent. A diverse set of methods has been developed to carry out this task. As noted by Kalton and Kasprzyk,[5] however, most of the methods can be expressed in a multiple regression framework.

Let y be the variable for which missing values are to be imputed and let $\mathbf{z} = (z_1, z_2, ..., z_p)$ be the set of other variables – the auxiliary variables – that are to be used in imputing for the missing y values. At this point, we assume that y is a continuous variable and that there are no missing values for the z variables. Let y_{rk} denote the value of y for record k for which the value of y is reported and let y_{mi} and $\hat{y}_{mi}$ denote the actual and imputed values of y for record i for which the value of y is missing. Many imputation methods can then be represented by the regression equation

$$\hat{y}_{mi} = b_{r0} + \sum b_{rj} z_{mij} + \hat{e}_{mi} \tag{3.1}$$

where b_{r0} is the intercept and b_{rj} are the estimated regression coefficients for the regression of y on $\mathbf{z}$ obtained from the records with y values reported, z_{mij} is the value of z_j for record i with a missing y value, and $\hat{e}_{mi}$ is a residual term that is discussed further below. Most of the common imputation method can be represented by equation (3.1) with the appropriate definitions of $\mathbf{z}$ and $\hat{e}_{mi}$.

One basic distinction between imputation methods is that between those that set $\hat{e}_{mi} = 0$ and those that do not. The former may be termed deterministic methods and the latter stochastic methods. From the perspective of producing the best predictions for the missing values, deterministic methods are to be preferred because the inclusion of the residual term with the stochastic methods simply adds random noise. However, deterministic methods place all the imputed values on the regression line and hence they do not reflect the residual variability. As a result, the variance of the $\hat{y}_{mi}$ is attenuated as compared with the variance of the unobserved y_{mi}, and the shape of the distribution of the y values is distorted. Deterministic methods produce more precise estimates of means (e.g. mean blood pressure) but produce biased estimates of shape parameters (e.g. the proportion of the population with systolic blood pressure above 140 mm). Given the aim of producing a dataset that may be analysed in any way the analyst desires, and the importance of shape parameters in many analyses, stochastic imputation methods are generally preferred.

A second distinction between imputation methods concerns the specification of the auxiliary variables in the regression model. As an illustration, suppose that the variables age, race and gender are to be used in imputing for systolic blood pressure. One approach would be to use dummy variables for race and gender, and take these dummy variables and the age variable as the auxiliary variables $\mathbf{z}$ in the regression. If useful, additional auxiliary variables, such as age squared and certain interaction variables, could also be included. We term methods following this general approach as regression imputation methods.

An alternative approach treats all the auxiliary variables as categorical. Thus age would need to be categorized into a set of classes (e.g. under 30, 30–39, 40–54, 55–64, 65 and over). Then imputation classes are formed in terms of combinations of the auxiliary variables. For example, the imputation classes may be formed as the cells in the crosstabulation of age, race and gender, or some sets of these cells may be combined into single imputation classes. This approach fits into the regression framework by defining the auxiliary variables $\mathbf{z}$ to be dummy variables that index the imputation classes. We term methods following this general approach as imputation class methods.

Sometimes no auxiliary variables are used. In this case, the deterministic version of equation (3.1) reduces to assigning the overall respondent mean, $\bar{y}_r$, for all missing responses. The stochastic version adds a residual term to $\bar{y}_r$. A natural choice of residual is to take a residual from one of the respondents at random, say respondent k with estimated residual $y_{rk} - \bar{y}_r = \hat{e}_{rk}$. The imputed value for nonrespondent i with this stochastic method is then $\hat{y}_{mi} = \bar{y}_r + \hat{e}_{rk} = y_{rk}$, the value for respondent k. Respondent k is thus the 'donor' of the y value to the 'recipient', nonrespondent i.

Imputation methods that take no account of auxiliary data are simple to apply and are sometimes used for that reason. Their use should, however, be restricted to variables with only a minimal number of missing values and without highly related auxiliary variables. Otherwise, the associations between the variable subject to imputation and other variables will be attenuated and inconsistencies may occur.

The above discussion has focused on the imputation of missing values for a continuous variable like blood pressure. If y is a discrete variable (e.g. number of children) or a categorical variable (e.g. marital status), modifications may be needed in order to assign feasible values. For example, with the deterministic imputation method that uses no auxiliary information, the overall mean may be replaced by the overall median or mode. By assigning actual respondents' values, the stochastic version of this method avoids this problem.

3.1.1 Imputation class methods

With imputation class methods, the auxiliary variables are used to divide the sample into a set of classes and the imputations are performed within the classes. The classes may be formed based on subject-matter expertise or on statistical analyses that determine which set of variables are predictive of the variable for which imputations are required. Sometimes branching algorithms like CART,[16] CHAID[17] or SEARCH[18] are used to determine the imputation classes.

With deterministic imputation, the imputation class method involves assigning the class mean (or class median or class mode) to all nonrespondents in each class. This method is sometimes used. However, as has been noted earlier, deterministic methods distort distributions and hence stochastic methods are generally preferred.

With a stochastic imputation class method, nonrespondents are assigned values from respondents in the same class. This is the basic idea underlying the widely used hot deck imputation methods. Some of these methods are described below.

An early version of hot deck imputation is often known as a sequential hot deck method. With this method a set of imputation classes is defined, and for each imputation class a computer location is created in which a value of the variable to be imputed (y) is to be stored. The survey records are then considered sequentially throughout the data file. If a record has a value for y, that value is stored in the location for the record's imputation class, replacing the value currently residing in that location. If a record has a missing y value, it is assigned the y value currently stored in the location for its imputation class.

At the time it was developed, the computing economy of the sequential hot deck method, with the imputations for a variable being performed in a single pass of the data, was a major attraction. However, with the increases in computer processing power that have occurred in recent years, that attraction is nowadays less significant. The method has two main disadvantages. First, the number of imputation classes has to be restricted to ensure that each class has at least one record with a y value to donate to records with missing y values in that class. Second, when two or more records with missing values occur in sequence in a given imputation class, these records receive the same y value, the value taken from the previous responding record (the donor). This 'multiple use of donors' causes a loss in precision in the survey estimates as compared with using different donors for different recipients (see Section 3.1.4 below).

Various alternative hot deck methods have been devised to address these disadvantages. One relatively simple one stores, say, three donor values in each imputation class location so that, when a run of nonresponding records is encountered, the imputations can be spread across the different donors.

The hierarchical hot deck method involves a more major change. With this method, many imputation classes may be used initially. At the first step, all the survey records are separated into respondent records and nonrespondent records within the imputation classes. Provided that there are respondents in every nonempty class and that the ratio of the number of nonrespondents to respondents is not too large (e.g. not greater than 3), then a sample of donors can be selected from the respondents to provide values for the nonrespondents in a way that ensures that no respondent serves as a donor more than once more than any other respondent. If there are imputation classes with nonrespondents but without respondents or if the ratio of nonrespondents to respondents in some classes is deemed too great, then some collapsing of classes may be employed. The term 'hierarchical' is used to reflect this collapsing, with an initially detailed matching of respondents and nonrespondents, then collapsing the level of detail where necessary to ensure that donors are found for all nonrespondents. Coder,[60] Welniak and Coder,[61] and Welniak[62] describe the hierarchical hot deck imputation methods that have been used at different times with the US Current Population Survey.

An example of the use of hot deck imputation in a health survey is provided by the US National Health and Nutrition Examination Survey I (NHANES I). In that survey missing dental findings were imputed within imputation classes formed by age, race, sex and income group, and missing systolic and diastolic blood pressures were imputed within imputation classes based on age, sex, race, arm girth, weight and height.[63] Cox and Cohen[15] provide a detailed description of the hot deck imputation methods used to impute for a range of variables in the US National Medical Care Utilization and Expenditure Survey. Further discussions of the properties of hot deck imputation methods include Ford,[64] Little,[65] Sande,[66] Kalton,[8] and Kalton and Kasprzyk.[5]

3.1.2 Regression imputation

Regression imputation is a straightforward application of equation (3.1). Deterministic regression imputation is simply the predicted value from the regression. Although often used, it has the disadvantage of distorting distributions as already noted.

There are various alternative procedures for choosing the residual for stochastic regression imputation.[8] One is to select a respondent at random and to take that respondent's residual. Another is to take the residual from a respondent who has similar values for the auxiliary variables to the nonrespondent. The advantage of this matching of nonrespondents to similar respondents is that it reduces the reliance on the validity of the regression model.

One method for matching nonrespondents to respondents is to match them in terms of their predicted values. A nonrespondent is then assigned the residual from the respondent with the closest predicted value. If there is more than one such respondent, one of them may be selected at random. To reduce the multiple use of donors, when a respondent donates a residual to a nonrespondent, the distance between the predicted value of that respondent and other nonrespondents still requiring the allocation of residuals can be increased. Such a procedure is, for instance, used in a related imputation method of distance function matching described by Colledge *et al.*[67]

A variant of the above procedure has been termed predictive mean matching by Little.[65] With predictive mean matching, the nonrespondent is assigned the *actual* value from the matched respondent, rather than just the residual. The difference between the two imputations is thus the difference in the predicted values of the nonrespondent and the matched respondent. Often that difference will be small, but it may be appreciable in sparse areas of the dataset when the regression model fits well. The attraction of predictive mean matching is that the imputed values are all feasible values since they are respondents' actual values. When respondents' residuals are added to nonrespondents' predicted values, nonfeasible imputed values (such as negative incomes) may occur.

Predictive mean matching may be viewed as a generalization of hot deck imputation. If the auxiliary variables are all dummy variables indexing imputation classes, predictive mean matching involves selecting a donor from the same imputation class where possible. Where that is not possible, the donor is selected from the imputation class with the closest predicted value.

Ezzati-Rice *et al.*[68] describe an investigation of the use of two forms of regression imputation for imputing for missing values of height, weight, systolic and diastolic blood pressures, and total serum and high-density lipoprotein cholesterol in the NHANES III. They compare predictive mean matching with a procedure that added a residual obtained by hot deck imputation to the predicted value from the regression. In their study the two forms of imputation produced extremely similar results.

3.1.3 Multivariate imputation

Suppose that an imputation class method is used to assign the missing values for y. The sample mean from the imputed dataset is an unbiased estimate of the population mean provided that the missing values are MAR within the imputation classes under either a deterministic or stochastic imputation method. The sample distribution provides an unbiased estimate of the population distribution when the missing values are MAR within the imputation classes and a stochastic imputation method is used. The associations of y with other variables x are, however, generally attenuated unless the x variables are used as auxiliary variables in forming the imputation classes.

Since most survey analyses involve interrelationships between the survey variables, it is important to include many variables as auxiliary variables in the imputation process. The difficulty here is that these variables may themselves be subject to missing data. Moreover, there is seldom a small number of missing data patterns across variables; rather there are usually many diverse patterns. Not only does this complicate the imputation for y but also imputations are needed for each of the other variables. The imputation for a set of variables in a way that maintains the associations between all of them can be straightforwardly achieved when the pattern of missingness across variables is a monotone or nested one in which, when the variables are placed in increasing order in terms of their item nonresponse rates, records with missing values on any given variable also have missing values on all subsequent variables (see Little[69] for a discussion of the monotone missing data pattern). With such a pattern, imputation can proceed serially, imputing for the variables in order of their item nonresponse rates using all the prior variables as auxiliary variables, including the imputed values of prior variables where applicable.

In practice, however, the item nonresponses for a set of variables rarely conform to the monotone pattern. When there are many different patterns of response/nonresponse across the variables, it is far more difficult to devise an effective imputation scheme for each of the variables that takes account of the data available for all other variables and that maintains all the associations between the variables.

Some form of iterative procedure may be used to address this problem. Suppose that the variables $\mathbf{y} = (y_1, y_2, ..., y_p)$ are to be imputed as a set. First, a provisional imputation is made for y_1, perhaps using auxiliary variables $\mathbf{z}$ that have no missing data. Then provisional imputations are made for: y_2 using y_1 (including imputed values) and $\mathbf{z}$ as auxiliary variables; y_3, using y_1, y_2 and $\mathbf{z}$ as auxiliary variables; ...; and y_p, using $y_1, y_2, ..., y_{p-1}$ and $\mathbf{z}$ as auxiliary variables. Next, y_1 is re-imputed using $y_2, y_3, ..., y_p$ and $\mathbf{z}$ as auxiliary variables; y_2 is re-imputed using y_1 (with revised imputations), $y_3, ..., y_p$ and $\mathbf{z}$, etc. The procedure continues until some form of convergence is satisfied. Judkins *et al.*[70] describe the use of this approach with a hot deck imputation procedure.

Another form of iterative procedure is based on a multivariate model for $\mathbf{y}$. Schafer *et al.*[71] describe the development of such a model for imputing for body measurements (four measures), blood pressures (six measures), and lipids (two measures) in the NHANES III. The model included a range of auxiliary variables, which were also subject to missing data. Since some variables were continuous and others categorical, a model for mixed continuous and categorical variables was employed, and the parameters were estimated using an iterative algorithm for maximum likelihood estimation with incomplete data given by Little and Schlucter.[72] The imputations were performed using this modelling together with multiple imputations (see Section 3.2). Similar approaches are described by Kennickell[73] and Heeringa.[74,75]

3.1.4 Reducing imputation variance

Consider hot deck imputation with respondents chosen at random within classes to donate their y values to nonrespondents. For univariate analysis of y, hot deck imputation is equivalent to deleting the nonrespondents' records and compensating for them by adding the nonrespondents' weights to their donors' records. Thus, for such analyses, hot deck imputation is in effect a form of nonresponse weighting adjustment. With a usual weighting adjustment, the weights for all respondents' records in a class are inflated by a uniform amount to compensate for the nonrespondents. However, with hot deck imputation, a random sample of respondents is chosen to compensate for the nonrespondents. This sampling process introduces what may be termed imputation variance, and this imputation variance reduces the precision of the survey estimates.

There are two main methods for reducing imputation variance. One is through the sample design for selecting donors within each imputation class. For instance, selecting donors by simple random sampling without replacement is preferable to simple random sampling with replacement. By minimizing the multiple use of donors, the without replacement design leads to a lower imputation variance. Donors may also be selected by stratified sampling within an imputation class, or by systematic sampling from an ordered list, provided that there is more than one nonrespondent record in the class. Since the respondents' records can be stratified by y, such a procedure can be very effective in reducing imputation variance.[76]

A second approach is to use fractional imputation, which involves dividing nonrespondents' records into parts and imputing separately to each part. For example, each nonrespondent might be divided into three parts, each of which is allocated a weight of one-third of the nonrespondent's original weight, and then separate donors are chosen for each part. Fractional imputation is discussed by Kalton and Kish[76] and Fay.[77–79] The different, but related, approach of multiple imputation is discussed in the next section.

3.2 Variance estimation with an imputed dataset

An attraction of filling in missing values by imputation is that complete data methods of analyses can be applied straightforwardly to the resultant dataset. However, since imputation fabricates data to some degree, treating the imputed values as actual values in estimating the variances of survey estimates leads to an overstatement of the precision of the estimates.

One approach for obtaining valid variance estimates from imputed datasets is by means of multiple imputation. With multiple imputation, a model is developed for assigning values for missing responses, and the dataset is completed not once but several, say c, times. The survey estimate $\widehat{\theta}_\gamma$ $(\gamma = 1, 2, ..., c)$ is computed separately for each data set, and the overall estimate is the average of these estimates $\widehat{\theta} = \Sigma\widehat{\theta}_\gamma/c$. The variance of $\widehat{\theta}$ is then as given by

$$v(\widehat{\theta}) = \frac{\Sigma\hat{v}_\gamma}{c} + \frac{c+1}{c}\left[\frac{1}{(c-1)}\sum(\widehat{\theta}_\gamma - \widehat{\theta})^2\right]$$

where $\hat{v}_\gamma$ is the estimated variance of $\widehat{\theta}_\gamma$ obtained by treating the imputed data as real data. In practice a small value of c, say $c = 3$ or $c = 5$, is often adequate.

A requirement for $v(\widehat{\theta})$ to provide a valid estimate of $V(\widehat{\theta})$ is that a 'proper' method of imputation is used. A proper imputation method needs to reflect the uncertainty in the imputation model. Hot deck imputation is, for instance, not a proper imputation method. Whether an imputation method is proper depends on the estimate $\hat{\theta}$ under consideration: a method that is proper for one estimate may not be proper for another.

A method known as the approximate Bayesian bootstrap (ABB) may be used to give proper imputations for estimating the population mean with simple random sampling. Consider r respondents and m nonrespondents with the missing data MCAR. Then the ABB comprises the following three steps:

1) Draw a sample of r values at random with replacement from the r respondents.
2) Draw a sample of m values with replacement from the r values generated at step 1.
3) Repeat the process for each of the c datasets.

If the missing data are MAR within imputation classes rather than MCAR, the ABB can be applied separately in each imputation class. The use of proper imputations adds to imputation variance, but this effect is counteracted by the use of multiple imputations.

There is a substantial literature on multiple imputation; see, for example, the extensive bibliography cited by Rubin.[80] This literature includes both theoretical contributions[77–84] and applications.[71,73,85–89]

Although there has been a considerable amount of research conducted on multiple imputation, the method is not widely applied in multipurpose surveys for three main reasons. First, despite the recent advances in computing power, the need to analyse several datasets remains a deterrent. Secondly, the need to identify and apply different 'proper' imputation methods for different estimates is a serious drawback to the use of the method as a general purpose strategy for variance estimation. Thirdly, the method's applicability with complex sample designs still needs to be clearly demonstrated within the design-based mode of inference used by most survey statisticians.

Recently, research has been conducted on methods of variance estimation with imputed datasets with single imputations using the design-based mode of inference.[79,90–93] Rao and Shao,[90] for instance, develop a jackknife variance estimator for estimates of means or totals for stratified multistage samples when item nonresponses are imputed using a weighted hot deck procedure, under the MAR assumption within imputation classes. The procedure involves modifying the imputed values for each jackknife replicate to adjust for the difference between the replicate respondent mean and the full sample mean in each imputation class. The imputed values need to be flagged and the imputation classes need to be identified for the use of this procedure. Fay[78,79] has developed a computer program that applies this procedure together with fractional imputation. To date, the Rao–Shao procedure is limited to variance estimators for means and totals. Further research is needed to produce variance estimation procedures for estimates that involve several variables each of which may be subject to item nonresponse and that are derived from complex sample designs.

4 Concluding remarks

Missing data arising from total nonresponse, noncoverage and item nonresponse are an important issue in health surveys, as in all surveys. Simply ignoring the missing data can lead to serious biases in the survey estimates. The methods of weighting adjustments and imputation described in this paper attempt to compensate for the missing data, and hence to reduce the biases in the survey estimates. While these methods are generally beneficial, it needs to be recognized that they are inevitably imperfect. They rely on assumptions, such as the MAR assumption, that are questionable and difficult to verify empirically. Thus these compensation methods should not be viewed as a substitute for obtaining valid responses. The best solution for missing data is to focus resources in the data collection stage on obtaining as complete data as possible. This may sound trite, but design and data collection efforts to maximize coverage and response rates are sometimes neglected under the false assumption that post-survey adjustments can be relied upon to compensate for missing data.

Although weighting adjustments are widely used in health surveys, there are many surveys in which no such adjustments are made. Researchers sometimes argue that they do not employ weighting adjustments because the adjustments require assumptions to be made about the nature of the missing data, and they do not want to make these assumptions. However, whenever there are missing data, assumptions need to be made. Ignoring the missing data implicitly invokes the MCAR assumption, an assumption that is far more dubious than the MAR or other assumptions explicitly made with the adjustment procedures.

Imputation is less widely used than weighting adjustments, and is not applied in many health surveys. Often analyses are restricted to records with reported values for the variables involved. This strategy is also implicitly adopting the MCAR assumption. The use of imputation can improve on this strategy, but care is needed to maintain the associations between relevant variables in performing the imputations. Since there is a risk that imputation may distort the results of some analyses, analysts needs to be in a position to conduct specific analyses with tailor-made methods for treating the missing data, rather than rely on the general-purpose imputations. One of the reasons that imputed values must be identified by flags in the microdata files is to enable these analyses to be carried out.

We should note that the missing data adjustment procedures can become much more intricate than indicated here when a complex survey with several phases of data collection is involved. In essence, this is the partial nonresponse situation discussed in Section 1. A number of health surveys are complex ones. For example, the NHANES comprises screening interviews, personal interviews, and medical examinations in a medical examination centre.[94] Partial nonresponse occurs when a respondent completes the screening interview only, completes the screening and personal interviews only, or completes the screening and personal interviews and only certain parts of the medical examination. The US National Medical Examination Survey, a panel survey that collects data at several waves, provides another example.[15] Partial nonresponse occurs when one or more waves of data are missing. In such situations, choices need to be made between weighting adjustments and imputation for handling different types of missing data. Several stages of weighting adjustment are often used, and sometimes several different sets of final weights may be developed for different types of analyses. The choice of compensation procedures to be used for these complex surveys involves a trade-off between operational considerations and analytic benefits.

Finally, we should note that all the general purpose adjustment procedures described in this paper rely on the MAR assumption that the missing data mechanism is ignorable after controlling for relevant auxiliary variables measured in the survey. When the missing data depend on Y even after controlling for the auxiliary variables, the missing data mechanism is said to be nonignorable. Various procedures have been developed for handling nonignorable nonresponse (see, for example, Little and Rubin,[1] Greenlees *et al.*[95] and Rubin *et al.*[96]), but these procedures are highly dependent on strong model assumptions. Since these procedures are not used for general purpose adjustments in multipurpose surveys, they have not been included here.

References

1 Little RJA, Rubin DB. *Statistical analysis with missing data*. New York: John Wiley, 1987.

2 Kalton G. Handling wave nonresponse in panel surveys. *Journal of Official Statistics* 1986; **2**: 303–14.

3 Lepkowski JM. Treatment of wave nonresponse in panel surveys. In: Kasprzyk D, Duncan G, Kalton G, Singh MP eds. *Panel surveys*. New York: John Wiley, 1989: 348–74.

4 Kish L. Weighting for unequal P_i. *Journal of Official Statistics* 1992; **8**: 183–200.

5 Kalton G, Kasprzyk D. The treatment of missing survey data. *Survey Methodology* 1986; **12**: 1–16.

6 Cochran WG. *Sampling techniques*. New York: John Wiley, 1977.

7 Lessler JT, Kalsbeek WD. *Nonsampling error in surveys*. New York: John Wiley, 1992.

8 Kalton G. *Compensating for missing survey data*. Ann Arbor, MI: Institute for Social Research, 1983.

9 Platek R, Singh MP, Tremblay V. Adjustment for nonresponse in surveys. In: Namboodiri NK ed. *Survey sampling and measurement*. New York: Academic Press, 1978: 157–74.

10 Oh HL, Scheuren F. Weighting adjustments for unit nonresponse. In: Madow WG, Olkin I, Rubin DB eds. *Incomplete data in sample surveys, Volume 2: Theory and bibliographies*. New York: Academic Press, 1983: 143–84.

11 Little RJA. Survey nonresponse adjustments for estimates of means. *International Statistical Review* 1986; **54**: 139–57.

12 Särndal C-E, Swensson B. A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review* 1987; **55**: 279–94.

13 Kalton G, Maligalig DS. A comparison of methods of weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 annual research conference*, 1991; 409–28.

14 Thomsen I. A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift* 1973; **4**: 278–83.

15 Cox BG, Cohen SB. *Methodological issues for health care surveys*. New York: Marcel Dekker, 1985.

16 Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees*. New York: Chapman & Hall, 1993.

17 Kass GV. An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* 1980; **29**: 119–27.

18 Sonquist JA, Baker EL, Morgan JN. *Searching for structure*. Ann Arbor, MI: Institute for Social Research, 1973.

19 Rizzo L, Kalton G, Brick JM. A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology* 1996; **22**: 43–53.

20 Iannacchione VG, Milne JG, Folsom RE. Response probability weight adjustments using logistic regression. *Proceedings of the section on survey research methods, American Statistical Association*, 1991; 637–42.

21 Politz A, Simmons W. An attempt to get the not-at-homes into the sample without callbacks. *Journal of the American Statistical Association* 1949; **44**: 9–31.

22 Politz A, Simmons W. Note on 'An attempt to get the not-at-homes into the sample without callbacks.' *Journal of the American Statistical Association* 1950; **45**: 136–7.

23 Hartley HO. Discussion on 'A review of recent statistical developments in sampling and sampling surveys' by F. Yates. *Journal of the Royal Statistical Society A* 1946; **109**: 37–8.

24 Thomsen I, Siring E. On the causes and effects of nonresponse: Norwegian experiences. In: Madow WG, Olkin I eds. *Incomplete data in*

sample surveys, Volume 3, Proceedings of the symposium. New York: Academic Press, 1983; 25–9.

25 Simmons WR. A plan to account for 'not-at-homes' by combining weighting and callbacks. *Journal of Marketing* 1954; **19**: 42–53.

26 Drew JH, Fuller WA. Modeling nonresponse in surveys with callbacks. *Proceedings of the section on survey research methods, American Statistical Association*, 1980; 639–42.

27 Drew JH, Fuller WA. Nonresponse in complex multiphase surveys. *Proceedings of the section on survey research methods, American Statistical Association*, 1981; 623–8.

28 Ward JC, Russick B, Rudelius W. A test of reducing callbacks and not-at-home bias in personal interviews by weighting at-home respondents. *Journal of Marketing Research* 1985; **22**: 66–73.

29 Potthoff RE, Manton KG, Woodbury MA. Correcting for nonavailability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association* 1993; **88**: 1197–207.

30 Bartholomew DJ. A method of allowing for 'not-at-home' bias in sample surveys. *Journal of the Royal Statistical Society* 1961; **10**: 52–9.

31 Hansen MH, Hurwitz WN. The problem of nonresponse in sample surveys. *Journal of the American Statistical Association* 1946; **41**: 517–29.

32 Waksberg J, Sperry S, Judkins D, Smith V. *National survey of family growth cycle IV, Evaluation of Linked Design*. National Center for Health Statistics. Vital and Health Statistics **2**(117). Washington, DC: US Government Printing Office, 1993.

33 Kish L. *Survey sampling*. New York: John Wiley, 1965.

34 Holt D, Smith TMF. Post stratification. *Journal of the Royal Statistical Society, A* 1979; **142**: 33–46.

35 Little RJA. Post-stratification: a modeler's perspective. *Journal of the American Statistical Association* 1993; **88**: 1001–12

36 Massey JT, Moore TF, Parsons VL, Tadros W. *Design and estimation for the National Health Interview Survey*, 1985-94. National Center for Health Statistics. Vital and Health Statistics **2**(110). Washington, DC: US Government Printing Office, 1989.

37 Bryant E, Shimizu I. *Sample design, sampling variance, and estimation procedures for the National Ambulatory Medical Care Survey*. National Center for Health Statistics. Vital and Health Statistics **2**(108). Washington, DC: US Government Printing Office, 1988.

38 Deming WE, Stephan FF. On a least squares adjustment of a sample frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* 1940; **11**: 427–44.

39 Brackstone GJ, Rao JNK. An investigation of raking ratio estimators. *Sankhya C* 1979; **41**: 97–114.

40 Oh HL, Scheuren F. Modified raking ratio estimation. *Survey Methodology* 1987; **13**: 209–19.

41 Ireland CT, Kullback, S. Contingency tables with given marginals. *Biometrika* 1968; **55**: 179–88.

42 Deville J-C, Särndal C-E, Sautory O. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 1993; **88**: 1013–20.

43 Statistics Canada. *Health status of Canadians: Report of the 1991 General Social Survey*. Ottawa: Statistics Canada, 1994.

44 Deville J-C, Särndal C-E. Calibration estimators in survey sampling. *Journal of the American Statistical Association* 1992; **87**: 376–82.

45 Cassel CH, Särndal C-E, Wretman JH. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 1976; **63**: 615–20.

46 Särndal C-E. On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 1980; **67**: 639–50.

47 Isaki CT, Fuller WA. Survey designs under the regression superpopulation model. *Journal of the American Statistical Association* 1982; **77**: 89–96.

48 Armstrong J, St-Jean H. Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology* 1994; **20**: 97–106.

49 Estevao V, Hidiroglou MA, Särndal C-E. Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics* 1995; **11**: 181–204.

50 Bethlehem JG, Keller WJ. Linear weighting of sample survey data. *Journal of Official Statistics* 1987; **3**: 141–53.

51 Bethlehem JG. Reduction of nonresponse bias through regression estimation. *Journal of Official Statistics* 1988; **4**: 251–60.

52 Huang ET, Fuller WA. Nonnegative regression estimation for survey data. *Proceedings of the social statistics section, American Statistical Association* 1978; 300–3.

53 Fuller WA, McLoughlin MM, Baker HD. Regression weighting in the presence of nonresponse with application to the 1987–1988 Nationwide Food Consumption Survey. *Survey Methodology* 1994; **20**: 75–85.

54 Tremblay V. Practical criteria for definition of weighting classes. *Survey Methodology* 1986; **12**: 85–97.

55 Potter FJ. Survey of procedures to control extreme sampling weights. *Proceedings of the section on survey research methods, American Statistical Association* 1988; 453–8.

56 Potter FJ. A study of procedures to identify and trim extreme sampling weights. *Proceedings of the section on survey research methods, American Statistical Association* 1990; 225–30.

57 Potter FJ. The effect of weight trimming on nonlinear survey estimates. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 758–63.

58 Cohen T, Spencer BD. Shrinkage weights for unequal probability samples. *Proceedings of the section on survey research methods, American Statistical Association* 1991; 625–30.

59 Lazzeroni LC, Little RJA. Models for smoothing post-stratification weights. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 764–69.

60 Coder J. Income data collection and processing from the March Income Supplement to the Current Population Survey. In: Kasprzyk D ed. *The survey of income and program participation: proceedings of the workshop on data processing,* Chapter II. Washington, DC: Income Survey Development Program, US Department of Health, Education and Welfare, 1978.

61 Welniak EJ, Coder JF. A measure of the bias in the March CPS earnings imputation scheme. *Proceedings of the section on survey research methods, American Statistical Association* 1980; 421–5.

62 Welniak EJ. Effects of the March Current Population Survey's new processing system on estimates of income and poverty. *Proceedings of the business and economics section, American Statistical Association* 1990; 144–51.

63 Landis J, Lepkowski J, Eklund S, Stehouwer S. *A statistical methodology for analyzing data from a complex survey, the first National Health and Nutrition Examination Survey.* National Center for Health Statistics. Vital and Health Statistics 2(92). Washington DC: US Government Printing Office, 1982.

64 Ford BL. An overview of hot-deck procedures. In: Madow WG, Olkin I, Rubin DB eds. *Incomplete data in sample surveys,* Volume 2: *Theory and bibliographies.* New York: Academic Press, 1983: 185–207.

65 Little RJA. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics* 1988; **6**: 287–96.

66 Sande IG. Imputation in surveys: coping with reality. *American Statistician* 1982; **36**: 145–52.

67 Colledge MJ, Johnson JH, Pare R, Sande IG. Large scale imputation of survey data. *Survey Methodology* 1978; **4**: 203–24.

68 Ezzati-Rice TM, Fahimi M, Judkins D, Khare M. Serial imputation of NHANES III with mixed regression and hot-deck techniques. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 292–6.

69 Little RJA. Models for nonresponse in sample surveys. *Journal of the American Statistical Association* 1982; **77**: 237–250.

70 Judkins D, Hubbell KA, England AM. The imputation of compositional data. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 458–62.

71 Schafer JL, Khare M, Ezzati-Rice TM. Multiple imputation of missing data in NHANES III. *Proceedings of the US Bureau of the Census 1993 Annual Research Conference,* 1993; 459–87.

72 Little RJA, Schlucter MD. Maximum-likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* 1985; **72**: 492–512.

73 Kennickell AB. Imputation of the 1989 Survey of Consumer Finances: Stochastic relaxation and multiple imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association* 1991; 1–9.

74 Heeringa SG. Imputation of item missing data in the Health and Retirement survey. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 107–16.

75 Heeringa SG. Application of generalized iterative Bayesian simulation methods to estimation and inference for coarsened household income and asset data. *Proceedings of the section on survey research methods, American Statistical Association* 1995; 42–51.

76 Kalton G, Kish L. Some efficient random imputation methods. *Communications in Statistics* 1984; **13**(16): 1919–39.

77 Fay RE. When are inferences from multiple imputation valid? *Proceedings of the section on survey research methods, American Statistical Association* 1992; 227–32.

78 Fay RE. Valid inferences from imputed survey data. *Proceedings of the section on survey research methods, American Statistical Association* 1993; 41–8.

79 Fay RE. Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association* 1996; **91**: 490–98.

80 Rubin DB. Multiple imputation after 18+ years. *Journal of the American Statistical Association* 1996; **91**: 473–89.

81 Rubin DB. Basic ideas of multiple imputation for nonresponse. *Survey Methodology* 1986; **12**: 37–47.

82 Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: John Wiley, 1987.

83 Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with nonignorable nonresponse. *Journal of the American Statistical Association* 1986; **81**: 361–74.

84 Meng X-L. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science* 1994; **9**: 538–73.

85 Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association* 1991; **86**: 68–78.

86 Heitjan DF, Little RJA. Multiple imputation for the Fatal Accident Reporting System. *Applied Statistics* 1991; **40**: 13–29.

87 Khare M, Little RJA, Rubin DB, Schafer JL. Multiple imputation of NHANES III, *Proceedings of the section on survey research methods, American Statistical Association* 1993; 297–302.

88 Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. *Statistics in Medicine* 1991; **10**: 585–98.

89 Schenker N, Treiman DJ, Weidman L. Analyses of public use decennial census data with multiply imputed industry and occupation codes. *Applied Statistics* 1993; **42**: 545–56.

90 Rao JNK, Shao J. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* 1992; **79**: 811–22.

91 Rao JNK. On variance estimation with imputed survey data. *Journal of the American Statistical Association* 1996; **91**: 499–506.

92 Särndal C-E. Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology* 1992; **18**: 241–52.

93 Kovar JG, Chen EJ. Jackknife variance estimation of imputed survey data. *Survey Methodology* 1994; **20**: 45–52.

94 Ezzati TM, Massey JT, Waksberg J, Chu A, Maurer KR. *Sample design: Third National Health and Nutrition Examination Survey*. National Center for Health Statistics. Vital and Health Statistics 2(113), 1992.

95 Greenlees WS, Reece JS, Zieschang KD. Imputation of missing values when the probability of response depends on the variable being imputed. *Journal of the American Statistical Association* 1982; **77**: 251–61.

96 Rubin DB, Stern HS, Vehovar V. Handling 'don't know' survey responses: the case of the Slovenian plebiscite. *Journal of the American Statistical Association* 1995; **90**: 822–28.