IMPUTING FOR MISSING SURVEY RESPONSES

Graham Kalton, University of Michigan
Daniel Kasprzyk, Social Security Administration

1. Introduction

Nonobservation in sample surveys occurs in three ways: noncoverage, total nonresponse and item nonresponse. Noncoverage represents a failure to include some units of the target population in the sampling frame. Total nonresponse occurs when no information is collected from a sample unit, and item nonresponse occurs when some but not all the required information is collected from a sample unit. Compensation procedures are often employed to try to reduce the biasing effects of nonobservation on survey estimates. Compensation for noncoverage is typically implemented by making weighting adjustments based on an external data source. Compensation for total nonresponse is usually carried out by some form of weighting adjustment, while compensation for item nonresponse is commonly made by imputation, that is by assigning values for missing responses (Kalton, 1981). This paper reviews and evaluates several commonly used imputation procedures.

Item nonresponse may occur because a sample unit refuses or is unable to answer a particular question, because the interviewer fails to ask the question or to record the answer, or because an inconsistent response is deleted in editing. The extent of item nonresponse varies greatly between questions. Items such as race and sex usually have few nonresponses; on the other hand, receipts of various sources of income may have high nonresponse rates (Coder, 1978; Kalton, Kasprzyk and Santos, 1981). The multivariate nature of surveys, with all variables potentially subject to missing data, suggests the need for a general purpose strategy for handling item nonresponses. As such a strategy, imputation has three desirable features. First, like weighting adjustments for total nonresponse, it aims to reduce biases in survey estimates arising from missing data; the success of various imputation procedures in meeting this objective for various forms of estimates is discussed later. Second, by assigning values at the microlevel and thus allowing analyses to be conducted as if the data set were complete, imputation makes analyses easier to conduct and results easier to present. Complex algorithms to estimate population parameters in the presence of missing data (e.g. the EM algorithm of Dempster, Laird and Rubin, 1977) are not required. Third, the results obtained from different analyses are bound to be consistent, a feature which need not apply with an incomplete data set.

Imputation does, however, have its drawbacks. It does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set; indeed the biases could be much greater, depending on the imputation procedure and the form of estimate. There is also the risk that analysts may treat the completed data set as if all the data were actual responses, thereby overstating the precision of the survey estimates. Analysts working with a data set containing imputed values should proceed with caution, and be aware of the extent of imputation for the variables in their analyses as well as the details of the procedures used. Aspects of the imputation process which should be monitored to evaluate the possible impact of imputation on survey results are described by I. Sande (1979a,b). At a minimum, imputed values should be flagged so that analysts can distinguish between actual and imputed responses, and thus obtain an indication of the potential effect of imputation on their results. Providing imputed values are flagged, analysts are also in a position to ignore them and treat the incomplete data set in a way that is tailor-made for their particular needs.

The following sections describe a variety of imputation procedures and their properties. Practical considerations in their implementation and other issues are also discussed.

2. Imputation Methods

When item nonresponse occurs, substantial information about the nonrespondent is usually available from other items on the questionnaire. Most imputation methods use a selection of these items as auxiliary variables in assigning values for the missing responses. In general, the value imputed for the i-th nonrespondent for item y may be described by $y_{mi} = f(z_{1i}, z_{2i}, \ldots, z_{pi}) + e_{mi}$, where $f(z)$ is a function of the auxiliary variables (z) and e_{mi} is an estimated residual. Often $f(z)$ may be expressed as a linear function, $\beta_0 + \Sigma \beta_j z_{ji}$, and the β's may be estimated from the respondents' data as $b_{rj} (j = 0,1,\ldots,p)$ (Santos, 1981a,b).

The major consideration in choosing the auxiliary variables is their ability to predict the missing y-values. The use of techniques like regression, SEARCH, and log-linear models with the respondents' data can be helpful in determining a good set of auxiliary variables. If a sizeable amount of nonresponse is anticipated for a specific survey item, the inclusion of alternative questions aimed at providing auxiliary information for imputation purposes may be useful. Thus, for example, wage earners in the 1978 Income Survey Development Program Research Panel were asked to report not only their quarterly earnings from records (y), but also their hourly rates of pay (z_1), usual numbers of hours worked per week (z_2) and numbers of weeks worked in the quarter (z_3). In cases where they did not report their quarterly earnings, their missing y-values could be imputed using the function $f(z) = z_1 \cdot z_2 \cdot z_3$ (Kalton, Kasprzyk and Santos, 1981).

Imputation methods can be classified along two dimensions: (1) by their use of auxiliary variables, and (2) by the value assigned to the residuals. Some methods make no use of auxiliary variables. Other methods treat them as categorical, classifying the sample members into imputation classes according to their combination of responses to these variables; continuous auxiliary variables, such as age or income, are categorized for use with these methods. Still other methods treat all the variables as continuous, with any categorical variables being handled as dummy variables. The second dimension concerns whether or not a randomization process is used in assigning imputed values. We term an imputation method as stochastic when the residual

term e_{mi} is randomly assigned and deterministic when it is set to zero.

The paragraphs below briefly describe many of the widely used imputation procedures:

(a) Deductive imputation. This imputation method depends on some redundancy in the data so that a missing response can be deduced from the auxiliary information, i.e. $y_{mi} = f(z_i)$ exactly. For example, if a record should contain a series of amounts and their total but one of the amounts is missing, the missing value can be deduced by subtraction. The method can be extended to situations where the deduced value is highly likely to be the correct value or at least close to it; for instance, in a panel survey with a variable that remains almost constant over time, a missing response on one wave of the panel may be assigned the record's value for the item on the preceding or succeeding wave.

(b) Mean imputation overall (MO). This method assigns the overall respondent mean, $\bar{y}_r$, to all missing responses. It is the deterministic degenerate form of the linear function with no auxiliary variables, i.e. $y_{mi} = b_{ro} = \bar{y}_r$.

(c) Random imputation overall (RO). This method assigns each nonrespondent the y-value of a respondent selected at random from the total respondent sample. The method is the stochastic degenerate form of the linear function with no auxiliary variables, $y_{mi} = \bar{y}_r + e_{mi}$, with $e_{mi} = y_{rk} - \bar{y}_r$, which reduces to $y_{mi} = y_{rk}$. Given an epsem sample initially, the subsample of respondents to act as donors can be selected by any epsem sampling scheme (e.g. unrestricted sampling, SRS, proportionate stratified sampling, or systematic sampling).

(d) Mean imputation within classes (MC). This method divides the total sample into imputation classes according to values on the auxiliary variables. Within each class the respondent mean for the y-variable is assigned to all the nonrespondents in that class: $y_{mhi} = \bar{y}_{rh}$ for the i-th nonrespondent in class h ($h = 1,2,\ldots,H$). The classes may be defined as all the cells in the cross-tabulation of the (categorized) auxiliary variables, but this symmetry is not essential; instead, some auxiliary variables may be used for one part of the sample while others are used for another part, or groups of cells may be combined. If all the cells in the cross-tabulation are used, the linear function can be expressed as a model with the main effects and all levels of interaction for the auxiliary variables. In general, the model can be represented by $y_{mi} = b_{ro} + \Sigma\, b_{rj} z_{ji}$, where the z_{ji} are dummy variables, $z_{ji} = 1$ if the i-th nonrespondent is in class j, $z_{ji} = 0$ otherwise ($j = 1,2,\ldots,(H-1)$). Since $e_{mi} = 0$, the method is a deterministic one.

(e) Random imputation within classes (RC). This method corresponds to the random overall method except that it is applied within imputation classes. Each nonrespondent is assigned the y-value of a respondent randomly selected from the same imputation class. The method is the stochastic equivalent of the mean within class method, with $y_{mhi} = \bar{y}_{rh} + e_{mhi}$ and $e_{mhi} = y_{rhk} - \bar{y}_{rh}$, reducing to $y_{mhi} = y_{rhk}$. It may alternatively be expressed as $y_{mji} = b_{ro} + \Sigma\, b_{rj} z_{ji} + e_{mji}$, where e_{mji} is a respondent residual selected at random within imputation class j in which nonrespondent i is located.

(f) Hot-deck imputation. The term hot-deck imputation has a variety of meanings, but refers here to the sequential type of procedure used by the Bureau of the Census with the labor force items in the Current Population Survey (CPS)(Brooks and Bailar, 1978). This is sometimes known as the traditional hot-deck procedure. The procedure begins with the specification of imputation classes, and for each class the assignment of a single value for the y-variable to provide a starting point for the process. These starting values may, for instance, be obtained by taking a respondent value for each class or a representative value such as the class mean from a previous round of the survey. The records of the current survey are then treated sequentially. If a record has a response for the y-variable, that value replaces the value previously stored for its imputation class. If the record has a missing response, it is assigned the value currently stored for its imputation class. A major attraction of this procedure is its computing economy, since all imputations are made from a single pass through the data file.

The hot-deck method is similar to the random within class method in which donors are selected by unrestricted sampling (i.e. SRS with replacement). If the order of the records in the data file were random, the two methods would be equivalent, apart from the start-up process. The sequential hot-deck procedure generally benefits from the non-random order of the data file, since use of the preceding donor in the imputation class yields an additional degree of matching which is advantageous if the file order creates positive autocorrelation. This benefit is unlikely to be substantial, however, when the imputation classes are small and spread throughout the file - as is often the case.

A disadvantage of the hot-deck method is that it may easily give rise to multiple use of donors, a feature which leads to a loss of precision for the survey estimators. This occurs when within a given imputation class a record with a missing response is followed by one or more records with missing responses; all these records are then assigned the value from the last respondent in the class. The random within class method with unrestricted sampling of donors shares this disadvantage. With the random within class method, however, the multiple use of donors may be minimized by sampling donors without replacement.

It is impossible to develop a model-free theoretical evaluation for the hot-deck method because of its dependence on the order of the file and its lack of a probability mechanism. For this reason, it will not be examined in the subsequent sections; the results for the random within class method with unrestricted sampling should, however, provide a reasonable guide to its performance. Useful discussions of the hot-deck procedure are provided by Bailar, Bailey and Corby (1978), Bailar and Bailar (1978, 1979), Ford (1980), Oh and Scheuren (1980), Oh, Scheuren and Nisselson (1980) and I. Sande (1979a,b).

(g) Flexible matching imputation. The term flexible matching imputation is used here for the modified hot-deck procedure that has been used

since 1976 for the CPS March Income Supplement. The procedure sorts respondents and nonrespondents into a large number of imputation classes, constructed from a detailed categorization of a sizeable set of auxiliary variables. Nonrespondents are then matched with respondents on a hierarchical basis, in the sense that if a nonrespondent cannot be matched with a respondent in the initial imputation class, classes are collapsed and the match is made at a lower level. Three levels are used with the March Income Supplement, the lowest level being such that a match can always be made. The procedure enables closer matches to be secured for many nonrespondents than does the traditional hot-deck procedure. It also avoids the multiple use of respondents in classes where the number of nonrespondents does not exceed the number of respondents. Further details on the implementation and evaluation of the procedure are given by Coder (1978) and Welniak and Coder (1980).

(h) Predicted regression imputation (PR). This method uses respondent data to regress y on the auxiliary variables. Missing y-values are then imputed as the predicted values from the regression equation, $y_{mi} = b_{ro} + \Sigma b_{rj} z_{ji}$. This is a deterministic method with $e_{mi} = 0$. The auxiliary variables may be quantitative or qualitative, the latter being incorporated by means of dummy variables. If the y-variable is qualitative, log-linear or logistic models may be used. As in any regression analysis, specific interaction terms may be included in the regression equation, and transformations of the variables may be useful.

A special case of the regression model is the ratio model $y_{mi} = b_r z_i$ with a single auxiliary variable and an intercept of zero (Ford, Kleweno and Tortora, 1980). This model may be used in panel surveys with z representing the same variable as y measured on the previous wave.

(i) Random regression imputation (RR). This method is the stochastic version of the predicted regression method: the imputed values are the predicted values from the regression equation plus residual terms e_{mi}. Depending on the assumptions made, the residuals can be determined in various ways, including:

(i) If the residuals are assumed to be homoscedastic and normally distributed, a residual can be chosen at random from a normal distribution with zero mean and variance equal to the residual variance from the regression.

(ii) If the residuals are assumed to come from the same, unspecified distribution, they can be chosen at random from the respondents' residuals.

(iii) As a protection against non-linearity and non-additivity in the regression model, the residuals may be taken from respondents with similar values on the auxiliary variables. If the donor respondent has the identical set of z values as the nonrespondent, the procedure reduces to assigning the respondent's y-value to the nonrespondent. This point demonstrates the close relationship between this procedure and the random within class method.

Applications of regression and categorical data models for imputation are described by Schieber (1978), Herzog and Lancaster (1980) and Herzog (1980).

(j) Distance function matching. This method assigns the y-value of the nearest respondent to each nonrespondent, with "nearest" defined by a distance function of the auxiliary variables. The method is primarily concerned with quantitative variables; however, qualitative variables may be included either by using the distance function approach within imputation classes formed by qualitative auxiliary variables or by incorporating these variables into the distance function. With a single auxiliary variable, the sample may be ordered by the variable, and the nearest respondent (donor) to each nonrespondent is taken where "nearest" may be defined as the minimum absolute difference between the nonrespondent's and donor's values in the auxiliary variable or in some transformation of the auxiliary variable. When several auxiliary variables are used, the issue of transformations becomes more critical; one approach is to transform all auxiliary variables to their ranks. Thus, one distance function proposed is given by $D(i,k) = \text{Sup}_h w_h |R_{hi} - R_{hk}|$, where R_{hi} and R_{hk} are the ranks of the nonrespondent and potential donor on variable h, and w_h is a weight representing the importance of variable h in the distance function (I. Sande, 1979a). Another approach, based on the Mahalanobis distance, has been suggested by Vacek and Ashikaga (1980). The distance function can be constructed to reduce the multiple use of donors. For instance, distance may be defined as $D(1 + pd)$ where D is the basic distance, d is the number of times the donor has already been used and p is a penalty for each usage (Colledge et al., 1978).

A variant of this method assigns the nonrespondent the average value of neighboring respondents, for instance the average value of the two adjacent respondents (Ford, 1976). As with other averaging procedures, this procedure suffers the disadvantage of distorting distributions (see Section 3.2).

3. Properties of Various Imputation Methods

This section reviews the effects of the six imputation methods listed in Table 1 on estimates of means, distributions, variances, covariances, and regression and correlation coefficients. The stochastic methods encompass a number of variants depending on how the e_{mi} are obtained. With the random regression method, we consider only the version which selects the e_{mi}'s from the respondents' residuals by some form of epsem sampling.

In the following we make several simplifying assumptions. First, we assume that respondents to the item always respond over conceptually repeated applications of the survey and nonrespondents never do. This assumption, which divides the population into strata of respondents and nonrespondents, is an obvious oversimplification because, for some units, chance plays a role in whether they respond or not. However, the tractability of the simplified model leads to informative results, and therefore it is adopted for this discussion. A more complicated model, a probability response model, is developed by Platek, Singh and Tremblay (1978), and Platek and Gray (1978, 1979).

Table 1: Six Imputation Methods

Use of auxiliary variables	Deterministic	Stochastic
None	Mean overall (MO)	Random overall (RO)
Imputation classes	Mean within classes (MC)	Random within classes (RC)
Regression	Predicted regression (PR)	Random regression (RR)

Second, we often assume that the missing responses are missing at random in the total sample (which we denote by MAR). While this assumption is unrealistic, it does, nevertheless, lead to insights into the properties of the various methods. Santos (1981a,b) derived many of the results reported here and has also considered the more realistic assumption that the missing values are missing at random within specified subgroups of the population. Note that with the MAR assumption, the simple procedure of deleting all sample records with missing responses leads to unbiased estimators of the parameters considered here.

Third, we assume that the sample is large, that it is selected by SRS, and that the finite population correction factor may be ignored. Many of the results presented are large sample approximations.

This review is concerned mainly with the biases of the standard estimators when some values have been imputed, since with large samples sizeable biases will dominate mean square errors. Imputation does, however, also affect the variances of estimators; this is illustrated below by considering the effects of the mean and random overall imputation methods on the precision of the sample mean.

3.1 Sample Mean

With y_{rk} and y_{mi} denoting actual and imputed responses respectively, the mean of a SRS of size n may be expressed as

$$\bar{y} = (\Sigma y_{rk} + \Sigma y_{mi})/n = \bar{r}\bar{y}_r + \bar{m}\bar{y}_m$$

where $\bar{y}_r$ and $\bar{y}_m$ are the means, and $\bar{r} = r/n$ and $\bar{m} = m/n$ are the proportions, of actual and imputed responses. Under the MAR model, comparison of the biases of $\bar{y}$ computed with the six imputation methods given in Table 1 are fairly uninformative since all the methods lead to at least approximately unbiased estimators.

In general, the means based on the stochastic methods have the same biases as those based on their deterministic counterparts. This may be demonstrated by decomposing the expectation of $\bar{y}$ into two parts, $E = E_1E_2$, where E_1 denotes expectation over the initial sample and E_2 denotes the conditional expectation over the sampling of residuals given the initial sample. Then, providing respondent residuals are sampled by an epsem sampling scheme, $E_2(e_{mi}) = 0$. Thus $E_2(y_{mis}) = E_2(y_{mid} + e_{mi}) = y_{mid}$, where y_{mis} and y_{mid} are the imputed values for a stochastic and the corresponding deterministic method. It follows that the conditional expectation of the mean computed with a stochastic imputation method is equal to the mean under the corresponding deterministic method, and hence that the means computed with the two methods have the same bias. Thus, $B(\bar{y}_{MO}) = B(\bar{y}_{RO})$, $B(\bar{y}_{MC}) = B(\bar{y}_{RC})$ and $B(\bar{y}_{PR}) = B(\bar{y}_{RR})$, where B(x) denotes the bias of x, and the subscripts refer to the six imputation methods listed in Table 1.

Assuming that on conceptually repeated applications of the survey some elements always provide responses on y when sampled while the remainder never do, the general bias of $\bar{y}_{MC}$ and $\bar{y}_{RC}$ can be expressed as

$$B(\bar{y}_{MC}) = B(\bar{y}_{RC}) = \Sigma M_h(\bar{Y}_{rh} - \bar{Y}_{mh})/N = B$$

where in imputation class h, M_h is the number of nonrespondents, $\bar{Y}_{rh}$ and $\bar{Y}_{mh}$ are the means for respondents and nonrespondents respectively, and N is the population size. The general bias of $\bar{y}_{MO}$ and $\bar{y}_{RO}$ is given by

$$B(\bar{y}_{MO}) = B(\bar{y}_{RO}) = [\Sigma W_h(\bar{Y}_{mh} - \bar{Y}_r)(\bar{R}_h - \bar{R})/\bar{R}] + B$$
$$= A + B$$

where W_h is the proportion of the population in class h, $\bar{R}_h$ is the response rate in class h, $\bar{Y}_r$ is the overall respondent mean, and $\bar{R}$ is the overall response rate. Thus, if A and B have the same sign, imputation class methods produce means with less absolute bias than the overall methods by an amount $|A|$. However, if A and B have different signs, $\bar{y}_{MC}$ and $\bar{y}_{RC}$ can have greater absolute bias than $\bar{y}_{MO}$ and $\bar{y}_{RO}$; when A and B are of opposite signs, use of the imputation class methods produces a smaller absolute bias only when $|A| > 2|B|$ (Thomsen, 1973; Kalton, 1981).

We will examine the effect of imputation on the variance of $\bar{y}$ only for the methods that do not use auxiliary variables. With the mean overall imputation method, $y_{mi} = \bar{y}_r$, so that $\bar{y}_{MO}$ reduces to $\bar{y}_r$. With SRS, conditional on r and ignoring the fpc, $V(\bar{y}_{MO}) \doteq S_r^2/r$ where S_r^2 is the element variance of the respondents. The variance of the mean under the random overall imputation method is given by

$$V(\bar{y}_{RO}) = V_1E_2(\bar{y}_{RO}) + E_1V_2(\bar{y}_{RO})$$
$$= V_1(\bar{y}_{MO}) + E_1V_2(\bar{y}_{RO}).$$

The second term in this equation is termed the imputation variance; it represents the loss of precision in $\bar{y}_{RO}$ from using the stochastic imputation method. A useful index of this loss of precision is I, the proportionate increase in variance arising from the imputation variance, $I = E_1V_2(\bar{y}_{RO})/V_1(\bar{y}_{MO})$.

Kalton and Kish (1981) derive the value of I for several different epsem schemes for sampling donors. In the case of unrestricted sampling $I \doteq \bar{m}(1 - \bar{m})$, which attains a maximum value of 25% at $\bar{m} = 50\%$. With donors selected by SRS, $I \doteq \bar{m}(1 - 2\bar{m})$ for $m < r$, and this reaches a maximum value of 12.5% at $\bar{m} = 25\%$. The substantial reduction in the imputation variance

through using SRS rather than unrestricted sampling occurs because the SRS scheme avoids the multiple use of donors. The use of proportionate stratified sampling with respondents stratified by the y-variable, or systematic sampling with respondents ordered by the y-variable, can further substantially reduce the imputation variance.

The imputation variance may also be reduced by taking a larger sample of donors, i.e. using multiple imputations. Instead of taking a sample of m donors, a sample of size cm is taken (where c is a positive integer), and each nonrespondent is given c imputed values. One technique for handling these multiple imputations is to divide each nonrespondent's record into c parts, with each part being assigned a weight of 1/c; then each part receives the y-value from one of the c donors sampled for that nonrespondent. With unrestricted sampling of donors, the use of c imputations per donor leads to a proportionate increase of variance of $I \doteq \bar{m}(1 - \bar{m})/c$. When the donors are sampled by SRS, $I \doteq \bar{m}[1 - \bar{m}(1 + c)]/c$ with $cm < r$. Even a small number of multiple imputations can reduce the imputation variance to a minor concern. For instance, with $c = 2$, the maximum value of I with unrestricted sampling is 12.5% at $\bar{m} = 50\%$, and with SRS it is 4.2% at $\bar{m} = 16.7\%$. Other uses of multiple imputation are discussed in Section 4.

3.2 Distribution and Variance

If the survey analysis was concerned only with means, a deterministic imputation method would be preferred, because it avoids the introduction of the imputation variance. The main drawback to deterministic methods is that they distort the distribution and hence attenuate the element variance of the variable for which imputations are made. Since distributions are frequently presented in survey reports, this distortion is a serious concern.

The mean overall imputation method creates a spike in the y-distribution since all the missing values are assigned the same value, $\bar{y}_r$. Since $y_{mi} = \bar{y}_r = \bar{y}$, the effect of the mean overall method on the element variance is seen from

$$E(s^2_{MO}) = E\{\Sigma(y_{rk}-\bar{y})^2+\Sigma(y_{mi}-\bar{y})^2\}/(n-1)$$

$$= E\{\Sigma(y_{rk} - \bar{y}_r)^2/(n - 1)\} = (r - 1)S^2_r/(n - 1),$$

where the expectation is conditional on r and S^2_r is the respondent element variance. If the missing data are MAR, the relbias of s^2_{MO} as an estimator of the population variance S^2 is thus approximately $-\bar{M}$, where $\bar{M}$ is the expected nonresponse rate. The random overall method, on the other hand, retains the respondent distribution in expectation, and $E(s^2_{RO}) \doteq S^2_r$, with $S^2_r = S^2$ if the missing data are MAR.

The mean within classes method produces a series of spikes in the y-distribution at the means of the imputation classes, $\bar{y}_{rh}$. The random within classes method retains the respondent distributions within classes in expectation, and adjusts the overall distribution for differential response rates across the classes. The sample element variance with the mean within classes method may be expressed as

$$s^2_{MC} = \{\Sigma(y_{rk} - \bar{y})^2 + \Sigma m_h(\bar{y}_{rh} - \bar{y})^2\}/(n - 1).$$

If the missing data are MAR, the relbias of s^2_{MC} as an estimator of S^2 is approximately $-\bar{M}(1 - \eta^2)$, where η^2 is the proportion of variance explained by the imputation classes. Under the MAR model s^2_{RC} is approximately unbiased for S^2.

The predicted regression method curtails the spread of the y-distribution. Under the MAR model, the relbias of s^2_{PR} as an estimator of S^2 is $-\bar{M}(1 - R^2)$, where R^2 is the proportion of variance explained by the regression. The random regression method adjusts the y-distribution for the missing cases and retains the residual variability exhibited in the respondents' data. Under the MAR model, s^2_{RR} is approximately unbiased for S^2.

In summary, if the missing data are MAR, the stochastic imputation methods yield approximately unbiased estimates of distributions and element variances, whereas the deterministic methods distort distributions and attenuate variances.

3.3 Covariance

To describe the effects of the various imputation methods on element covariances, another variable x in addition to y needs to be specified. Initially we assume that x is known for all sampled elements.

In general, the sample covariance with actual and imputed responses may be expressed as

$$s_{xy} = \{\Sigma(x_{rk}-\bar{x})(y_{rk}-\bar{y})+\Sigma(x_{ri}-\bar{x})(y_{mi}-\bar{y})\}/(n-1). \quad (1)$$

For the stochastic imputation methods, the imputed values y_{mis} may be substituted for y_{mi} in (1). Then the conditional expectation of s_{xy}, the expectation over the stochastic imputation subsampling, is obtained by replacing y_{mis} by $E_2(y_{mis}) = y_{mid}$, the value for the corresponding deterministic method, in (1). This argument shows that the biases of s_{xy} under the stochastic and corresponding deterministic methods are the same, i.e. $B(s_{xyMO}) = B(s_{xyRO})$, $B(s_{xyMC}) = B(s_{xyRC})$ and $B(s_{xyPR}) = B(s_{xyRR})$.

The effect of the mean overall method on the covariance corresponds to its effect on the variance. With $y_{mi} = \bar{y}_r = \bar{y}$, s_{xy} in (1) reduces to

$$s_{xyMO} = (r - 1)s_{rxy}/(n - 1), \quad (2)$$

where s_{rxy} is the sample covariance between x and y for the respondents. The conditional expectation of s_{xyRO} is also given by (2). If the missing y-values are MAR, the relbiases of s_{xyMO} and s_{xyRO} as estimators of the population covariance S_{xy} are both approximately $-\bar{M}$.

From (1), the element covariance under the mean within class method becomes

$$s_{xyMC}=\{\Sigma(x_{rk}-\bar{x})(y_{rk}-\bar{y})+\Sigma m_h(\bar{x}_{rmh}-\bar{x})(\bar{y}_{mh}-\bar{y})\}/(n-1)$$

where $\bar{x}_{rmh}$ is the mean x-value for the m_h sampled elements in imputation class h with missing y-values. This formula also represents $E_2(s_{xyRC})$, and suggests that these methods fail to capture the within imputation class covariance for the elements with imputed y-values. In the case of the MAR model, these covariance estimators have a relbias of approximately $-\bar{M}(S_{xy.z}/S_{xy})$, where

$S_{xy.z} = \Sigma W_h S_{xyh}$ is the average within class covariance for classes formed by the auxiliary variable z and W_h is the proportion of the population in class h.

The two regression methods (PR and RR) produce estimators s_{xy} with the same bias in estimating S_{xy}. Under the MAR model their approximate relbias can be expressed in the same form as that for the imputation class methods, that is $-\bar{M}(S_{xy.z}/S_{xy})$ with $S_{xy.z}$ denoting the partial covariance of x and y given z. This relbias may also be expressed as $-\bar{M}[1 - (\rho_{xz}\rho_{yz}/\rho_{xy})]$, where ρ_{uv} denotes the correlation between u and v.

A disturbing feature of these results is that s_{xy} calculated with imputed values obtained from any of these imputation methods is potentially subject to substantial bias even under the MAR model. The estimates s_{xy} computed with the imputed values obtained from the imputation class and regression methods are unbiased only if the partial covariance $S_{xy.z}$ is zero. In general, there is no reason to assume uncritically that $S_{xy.z}$ is zero. Note, however, that if x = z, so that x is used as an auxiliary variable in the imputation scheme, $S_{xy.z}$ is zero. This result suggests that if the covariance between x and y is to play an important role in the survey analysis, x should, if possible, be used as an auxiliary variable in imputing for missing y-values.

We turn now to the case where x as well as y is subject to missing data. For simplicity we consider only the mean overall and random overall methods. By an extension of the approach used to derive (2), s_{xy} in (1) reduces with the mean overall imputation method to

$$s_{xyMO} = (r' - 1)s_{r'xy}/(n - 1), \qquad (3)$$

where r' is the subset of elements providing both x and y values. The conditional expectation of s_{xyRO} is also given by (3) if the missing x and y values are imputed independently.

Suppose now that all sampled elements either provide both x and y values or provide neither value, and that the random overall method is used to impute for the missing values, with a nonrespondent's x and y values both coming from the same respondent. In this case, $E_2(s_{xyRO})$, the expectation over the imputation subsampling, is approximately s_{rxy}, so that under the MAR model, s_{xyRO} is approximately unbiased for S_{xy}. When a record has several missing values, this result indicates that using the same donor for all the missing values retains the respondents' covariance structure for the variables involved (see Coder, 1978, on the use of joint imputation from the same donor in the CPS March Income Supplement). This benefit also suggests that it might sometimes be worthwhile to delete an x or y value when the other is missing in order to employ joint imputations for the pair of values from the same donor. Where feasible, it is clearly preferable not to delete values in this way but rather to use x as an auxiliary variable in imputing for y, or vice versa. However, when this strategy is not practicable, the deletion and joint imputation procedure does serve to retain the respondent covariance structure and to ensure that the x and y values for a record are not inconsistent with one another.

The effect of imputation on covariances has implications for multivariate analyses. In a simple regression of y on x, where x is not subject to missing data, attenuation in the estimated covariance through imputation also applies to the regression coefficient; to guard against possible attenuation, x ought to be used as an auxiliary variable in the imputation scheme. Some simulation results for multiple regressions in which the dependent variable y included imputed values while information on the independent variables x was complete are provided by Santos (1981a). As a rough guide, his results indicate that regression coefficients of x variables used in the imputation scheme were not attenuated, but those of x variables not used were attenuated. Thus, imputation may distort the picture of the relative importance of the independent variables.

The effect of imputation on the correlation coefficient between x and y is a combination of its effects on the covariance and the standard deviations of the two variables. To illustrate this point, consider the mean overall and random overall methods with two different patterns of missing data. When information on x is complete and only y includes imputed values, the sample correlations with the mean and random overall methods are $r_{xyMO} = [(r - 1)/(n - 1)]^{1/2} r_{rxy}$ and $E_2(r_{xyRO}) = [(r - 1)/(n - 1)]r_{rxy}$, where r_{rxy} is the respondent sample correlation. The attenuation of the sample correlation for the random overall method is the same as that for the covariance, since this method retains the respondent standard deviation for y approximately in expectation. The attenuation for the mean overall method is smaller because of a cancellation between the attenuations of the covariance in the numerator of r_{xyMO} and of the standard deviation of y in the denominator.

Now suppose that x and y are either both missing or both available. In this case, the mean overall method reproduces the respondent correlation, $r_{xyMO} = r_{rxy}$, because of a complete cancellation between the attenuations of the covariance and the standard deviations of x and y. With the random overall imputation method, $E_2(r_{xyRO}) = [(r - 1)/(n - 1)]r_{rxy}$ if the pairs of missing x and y values are imputed independently, or $E_2(r_{xyRO}) = r_{rxy}$ if they are imputed jointly from the same donors.

Finally, it should be noted that correlations may be overestimated with deterministic imputation methods which employ auxiliary information even when the missing data are MAR. This point may be illustrated by the regression prediction imputation method when x = z is used as the auxiliary variable. In this case, the imputed values are all placed on the regression line, so that the respondent correlation is inflated.

4. Standard Error Estimation

There is a risk with imputation that analysts may compute sampling errors from the completed data set as if all the data had been collected from respondents, thus attributing greater precision to the survey estimates than is warranted. Thus, the variance of the mean of a SRS might be estimated by the standard formula $v(\bar{y}) = s^2/n$, whereas the actual variance is $V(\bar{y}) = S_r^2(1 + I)/r$, conditional on r and ignoring

the fpc, with I the proportionate increase in variance arising from the imputation variance (see Section 3.1). Two components in the underestimation of $v(\bar{y})$ for $V(\bar{y})$ can be identified. In the first place, $v(\bar{y})$ treats the sample as one of size n, whereas there are only r responses. For this reason, $v(\bar{y})$ underestimates $V(\bar{y})$ by a factor of r/n. Secondly, s^2 underestimates $S_r^2(1 + I)$. With a deterministic imputation scheme $I = 0$, but s^2 underestimates S_r^2; with a stochastic scheme s^2 is asymptotically unbiased for S_r^2, but $I > 0$. Thus, for instance, with the mean overall imputation scheme, $E(s^2) = [(r - 1)/(n - 1)]S_r^2$ and $I = 0$, so that $v(\bar{y})$ underestimates $V(\bar{y})$ by a factor $[r/n][(r - 1)/(n - 1)]$. With the random overall imputation scheme, with unrestricted sampling of a large sample of donors, $E(s^2) \doteq S_r^2$ and $I = \bar{m}(1 - \bar{m})$. Thus, $v(\bar{y})$ underestimates $V(\bar{y})$ by $[r/n][1 + \bar{m}(1 - \bar{m}]^{-1}$. (It should be noted that this underestimation of standard errors may not apply to the same extent with multi-stage designs.)

One way to handle the general problem of sampling error estimation for statistics based on data sets with imputed values is by means of multiple imputations as advocated by Rubin (1978, 1979). With this method, the construction of a complete data set by imputing for the missing responses is conducted several (say c) times independently, each time according to the same stochastic imputation procedure. The sample estimates (z_i; i = 1,2,...c) can then be computed for each of the c replicates, and their average $\bar{z} = \Sigma z_i/c$ calculated. A variance estimator for $\bar{z}$ is then given by $\bar{v} + w$, where $\bar{v}$ is the average estimated variance of the z_i within the replicates and $w = \Sigma(z_i - \bar{z})^2/(c - 1)$. In order to make this variance estimator unbiased for $V(\bar{z})$, additional variability may be incorporated in w by adding a random variable to each imputed value, the variable having the same value for each imputed value in a replicate, but a different value for each replicate.

A major problem with the use of multiple imputations is the additional computer analysis needed, which increases as the number of replicates, c, increases. For this reason, a small value of c may be preferred; Rubin (1978, 1979) recommends c = 2. A serious limitation to a small value of c, however, is the low precision of the resulting variance estimator. Even with a small c, it is questionable whether the multiple imputation approach is feasible for routine analysis. It may be best reserved for special studies, such as that described by Herzog (1980) and Herzog and Lancaster (1980).

In passing two further uses of multiple imputations deserve comment. First, as noted in Section 3.1, the use of multiple imputations reduces the imputation variance. Second, multiple imputations may be generated from different imputation procedures, making different assumptions about the nonrespondents. Comparisons of the survey estimates then indicate the sensitivity of the results to the imputation procedures employed.

5. Issues of Practical Implementation

In reviewing imputation procedures for item nonresponse, it should be recognized that the typical survey collects a substantial amount of data for each sampled element, often covering as many as a hundred variables or more. Consequently, the task of forming a complete data set by imputing values for all the missing responses is sizeable, because all variables are likely to have some missing responses. It is generally not practicable to invest a substantial effort in developing a separate tailor-made imputation method for each variable; at best, this is possible for only a small selection of the most important survey variables.

When developing an imputation procedure for a variable, y, all the other survey variables are available to act as auxiliary variables. The choice of auxiliary variables may be guided by analyses of the relationships between y and the other variables; with a regression imputation procedure, regression analyses of y on the other variables may be useful, while with an imputation class procedure a technique like SEARCH - a successor to the Automatic Interaction Detector (AID) technique - may be used to identify classes of the sample that are homogeneous in y (Sonquist, Baker and Morgan, 1974).

The choice between an imputation class or regression imputation method is influenced in part by the nature of the auxiliary variables. Imputation class methods readily handle categorical auxiliary variables, but require quantitative variables to be categorized. Regression methods readily handle both quantitative and categorical variables (through dummy variables), but impose a linear, additive model (unless non-linear terms or interactions are specifically incorporated). By adopting a more restrictive model than the imputation class methods (which allow for all interactions), the regression methods can incorporate a wider range of auxiliary variables. However, regression methods depend on the construction of a suitable model, and if a seriously misspecified model is used the methods may generate poor, even impossible, imputed values. It seems best, therefore, to reserve their use for those important survey variables for which careful model development is warranted. As noted earlier, one way to reduce the reliance on the model with a random regression method is to take a residual from a "close" respondent to add to the predicted value. This method is fairly similar to a random imputation class method. An attraction of the random imputation and hot-deck type imputation methods is that they are less model dependent than regression methods. Since they impute respondents' values to nonrespondents, they cannot, for instance, generate impossible values.

The fact that every variable collected in a survey is potentially subject to missing data seriously complicates the imputation task. One difficulty it creates is that auxiliary variables used in imputation may themselves sometimes be missing. With random and hot-deck type imputation methods, it also raises the issue that when two or more items are missing on a record it is preferable, ceteris paribus, to impute them from the same donor; otherwise, as noted above, the

covariance between the items will be attenuated and inconsistent values may be imputed. Joint imputations may be implemented by using the same imputation classes for all the items concerned and then using a single donor for the missing items of a given nonrespondent. This procedure may, however, operate against the optimum choice of imputation classes for a specific item; instead of maximizing the proportion of variance explained in one item using a technique such as SEARCH, a multivariate version with several dependent variables may be used (Gillo and Shelly, 1974). A compromise solution is often necessary, making joint imputations for a group of closely-related items, but treating different groups of items separately. One approach is a sequential procedure used by the Bureau of the Census (Coder, 1978; Brooks and Bailar, 1978): first, fill in the "small holes" in basic items that are used in forming the initial imputation classes; second, impute for a group of closely-related items using one set of imputation classes; third, impute for another group of variables using a different set of imputation classes (which may be defined to include variables from the first group of variables); etc.

A special case of the sequential approach can be applied in the commonly encountered situation of a quantitative variable that has a zero value for, or does not apply to, many sample elements (e.g., interest income for a sample of persons). For such variables, imputation may be conducted in two steps: first to impute whether the variable is zero or not; and then, if not zero, to impute the amount. Herzog (1980) uses this approach with a regression imputation for the amount of Social Security benefit received. Ford, Kleweno and Tortora (1980) call the approach a zero spike procedure and use it with a ratio estimator when a non-zero imputation is made at the first step.

Another facet of the multivariate nature of survey data is that often many of the variables are highly interrelated. In the initial stages of processing survey data, numerous edit checks are commonly specified, and failures of certain responses to satisfy these checks leads to the deletion of some responses, with the consequent need for imputation. When many interrelated edit constraints are applied, the choice of which responses to delete when inconsistencies are found is a difficult one. A principle, such as minimizing the number of deletions, may be used (Greenberg, 1981; Fellegi and Holt, 1976).

Editing is also closely connected to imputation through the need for the imputed values to satisfy edit constraints. When many constraints are employed, the range of imputed values to satisfy the constraints may be severely limited. In theory, the proper use of the variables in the constraints as auxiliary variables should ensure that the imputed values satisfy the constraints. In practice, however, the complexity of multiple constraints often makes this impossible. Records in which imputations have been made ought to be re-edited after imputation, unless the imputation procedure itself guarantees that the edit constraints will be satisfied. If some records then fail the edit constraints, deletions and further imputations will be required. I. Sande (1979, 1982) brings out the close relationship between editing and imputation. Automatic edits and imputation with categorical edits are discussed by Hill (1978), and G. Sande (1979) describes a procedure for linear edits with continuous variables.

Sometimes transformations can be helpful in ensuring that imputed values satisfy edit constraints. A simple example is the imputation of a household's earnings, y, using a random regression imputation method. An impossible negative earnings amount could be imputed from the regression of y on the auxiliary variables. This outcome would be avoided if log y were imputed. As a second example, consider a hot-deck imputation of length of first marriage for persons married more than once, with the dates of first and second marriages being known. A matching of nonrespondents and respondents on the exact lengths of the time between the first and second marriages would ensure that the nonrespondents received a length of first marriage that was less than the time between marriages; however, an approximate match, which would have to be used in practice, would not guarantee this property. A way to avoid the potential inconsistency with the approximate match is to impute not for length of first marriage but for length as a proportion of the interval between the two marriages. A transformation of this type is often useful with quantitative variables in the presence of inequality constraints (I. Sande, 1979, 1982).

6. Concluding Remarks

A major attraction of imputation is that it generates a complete data set that may be readily used for many different forms of analysis. As the preceding sections have shown, however, caution is needed in analyzing a data set that includes imputed values. In the case of univariate analyses, deterministic imputation methods serve well for estimating means and totals, but they distort the distributional properties of the variable; stochastic methods are less efficient for estimating means and totals but they preserve the variability in the respondent data. All methods are likely to attenuate the covariances between the variable subject to imputation and other variables, except for those other variables that are used as auxiliary variables in the imputation scheme. In consequence, when a data set contains imputed values, special care is needed in studying the interrelationships between variables, whether the interrelationships are examined in terms of cross-tabulations, regression analyses or other forms of multivariate analysis.

Alternative ways of handling missing survey data include dropping cases with missing values on the relevant variables from the analysis, direct estimation of the population parameters from a modeling approach, and weighting adjustments. Dropping cases with missing values is a widely used procedure, sometimes adopted on the grounds that it avoids assumptions required in procedures which attempt to compensate for missing data. It should, however, be recognized that even this procedure employs an implicit assumption about the similarity of respondents and nonrespondents; for instance, with the response and nonresponse strata model employed in Section 2, the respondent mean from a SRS is unbiased for the overall population

mean only under the assumption that the respondent and nonrespondent stratum population means are equal. Since the dropping cases procedure is based on such an assumption, there seem good grounds for using a compensation procedure that employs a more suitable assumption than the implicit assumption when the latter is unrealistic. This reasoning justifies the use of an appropriate imputation procedure to compensate for item nonresponse for univariate analyses; however, the potential damaging effects of imputation on multivariate analyses may often make the dropping cases procedure a preferable choice.

The direct estimation of population parameters by a modeling approach that takes account of missing data has much to commend it. However, the labor and computing time to implement the approach preclude its use as a general purpose strategy for handling missing survey data in all the many analyses that are conducted with a survey data set. Rather, the approach seems best reserved for a small range of special analyses. In view of the dangers of imputation for multivariate analysis, there is a strong case for a greater use of the modeling approach. Little (1982) provides a useful review of this approach.

Weighting adjustments are commonly used to compensate for total nonresponse rather than item nonresponse. For univariate analyses there is a close correspondence between weighting and imputation. For such analyses any imputation procedure that assigns a respondent's value to a nonrespondent is equivalent to a weighting procedure that adds the nonrespondent's weight to that of the respondent. The widely-used weighting class procedure that increases the weights of the rj respondents in class j by a factor of $(rj + mj)/rj$, where there are mj nonrespondents in class j, can be viewed as equivalent to a multiple imputation procedure that divides each nonrespondent record into rj parts, and assigns the rj responses one to each part. Thus, within each class this weighting procedure is equivalent to the special case of the multiple imputation procedure with SRS sampling of respondents, where the number of sampled donors is an exact multiple of the number of respondents; this special case gives rise to no imputation variance (Kalton and Kish, 1981). Moreover the procedure retains the distributional properties of the respondents' data. This combination of features makes the weighting class procedure more attractive for univariate analysis than the random imputation within classes procedure. The weighting class procedure can be applied by associating a weight variable to each survey item. If no response is obtained to an item, the weight variable for that item is set equal to zero; for responses to the item in class j, the weight is set equal to $(rj + mj)/rj$. (As described, the scheme assumes that all sampled elements have unit weights; however, it can be readily adapted for unequal weights). The limitation of this scheme is that in general it cannot be employed in multivariate analyses, since each item has a different weight. The only case where all the items retain the same weight is when they are all missing or present together - i.e. the case of total nonresponse. Weighting adjustments for total nonresponse retain the covariance structure of the respondents, and hence - unlike imputation procedures - they are not harmful to multivariate analyses.

Finally, we should note that weighting adjustments and imputation are usually employed in combination, weighting adjustments to compensate for total nonresponse and imputation for item nonresponse. The use of weighting adjustments means that the survey data set to which imputation is applied is one with unequal weights; unequal weights may also arise because of unequal selection probabilities and post-stratification adjustments. The results presented in this paper relate to the use of imputation with self-weighting samples. In general little attention has been given to the issues that unequal weights raise for imputation, although recently some useful contributions have been made (Cox, 1980; Cox and Folsom, 1978, 1981). In this area, and indeed in many other areas, more research is needed on the use of imputation as a way of handling item nonresponses in surveys.

References

Bailar, B.A. and Bailar III, J.C. (1979). Comparison of the biases of the "hot-deck" imputation procedure with an "equal- weights" imputation procedure. Symposium on Incomplete Data: Preliminary Proceedings (Panel on Incomplete Data of the Committee on National Statistics/National Research Council), 422-447. U.S. Department of Health, Education, and Welfare, Washington, D.C.

Bailar, B.A., Bailey, L. and Corby, C.A. (1978). A comparison of some adjustment and weighting procedures for survey data. Survey Sampling and Measurement (Namboodiri, N.K. ed.), 175-198, Academic Press, New York.

Bailar III, J.C. and Bailar, B.A. (1978). Comparison of two procedures for imputing missing survey values. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 462-467.

Brooks, C.A. and Bailar, B.A. (1978). An Error Profile: Employment as Measured by the Current Population Survey. Statistical Policy Working Paper 3. U.S. Department of Commerce. U.S. Government Printing Office, Washington, D.C.

Chapman, D.W. (1976). A survey of nonresponse imputation procedures. Proc. Soc. Statist. Sect., Amer. Statist. Ass., 1976(1), 245-251.

Coder, J. (1978). Income data collection and processing from the March Income Supplement to the Current Population Survey. The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing, February 23-24, 1978 (D. Kasprzyk ed.), Chapter II. U.S. Department of Health, Education and Welfare, Washington, D.C.

Colledge, M.J., Johnson, J.H., Pare, R. and Sande, I.G. (1978). Large scale imputation of survey data. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 431-436.

Cox, B.G. (1980). The weighted sequential hot deck imputation procedure. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 721-726.

Cox, B.G. and Folsom, R.E. (1978). An empirical investigation of alternative item nonresponse adjustments. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 219-223.

Cox, B.G. and Folsom, R.E. (1981). An evaluation of weighted hot-deck imputations for unreported health care visits. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1981, 412-417.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc., B, 39, 1-38.

Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. J. Amer. Statist. Ass., 71, 17-35.

Ford, B. (1976). Missing data procedures: a comparative study. Proc. Soc. Statist. Sect., Amer. Statist. Ass., 1976, 324-329.

Ford, B. (1980). An overview of hot deck procedures. Draft paper for Panel on Incomplete Data, Committee on National Statistics, National Academy of Sciences.

Ford, B.L., Kleweno, D.G. and Tortora, R.D. (1980). The effects of procedures which impute for missing items: a simulation study using an agricultural survey. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 251-256.

Gillo, M.W. and Shelly, M.W. (1974). Predictive modeling of multivariable and multivariate data. J. Amer. Statist. Ass., 69, 646-653.

Greenberg, B. (1981). Developing an edit system for industry statistics. Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface, 11-16. Springer-Verlag, New York.

Herzog, T.N. (1980). Multiple imputation of individual Social Security amounts, Part II. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 404-407.

Herzog, T.N. and Lancaster, C. (1980). Multiple imputation of individual Social Security amounts, Part I. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 398-403.

Hill, C.J. (1978). A report on the application of a systematic method of automatic edit and imputation to the 1976 Canadian Census. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 474-479.

Kalton, G. (1981). Compensating for Missing Survey Data. Survey Research Center, University of Michigan, Ann Arbor, Michigan.

Kalton, G., Kasprzyk, D. and Santos, R. (1981). Issues of nonresponse and imputation in the Survey of Income and Program Participation. Current Topics in Survey Sampling. (D. Krewski, R. Platek and J.N.K. Rao, eds.) pp. 455-480. Academic Press, New York.

Kalton G. and Kish, L. (1981). Two efficient random imputation procedures. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1981, 146-151.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. J. Amer. Statist. Ass., 77, 237-250.

Oh, H.L. and Scheuren F. (1980). Estimating the variance impact of missing CPS income data. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 408-415.

Oh, H.L., Scheuren, F. and Nisselson, H. (1980). Differential bias impacts of alternative Census Bureau hot deck procedures for imputing missing CPS income data. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 416-420.

Platek, R. and Gray, G.B. (1978). Nonresponse and imputation. Survey Methodology, 4, 144-177.

Platek, R. and Gray, G.B. (1979). Methodology and application of adjustments for nonresponse. Bull. Int. Statist. Inst., 48.

Platek, R., Singh, M.P. and Tremblay, V. (1978). Adjustment for nonresponse in surveys. Survey Sampling and Measurement, (Namboodiri, N.K. ed.)., Chapter 11. Academic Press, New York.

Rubin, D.B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 20-34.

Rubin, D.B. (1979). Illustrating the use of multiple imputations to handle nonresponse in sample surveys. Bull. Int. Statist. Inst., 1979.

Sande, G. (1979). Numerical edit and imputation. Int. Ass. Statist. Computing, 42nd Session of Int. Statist. Inst., 1979.

Sande, I.G. (1979a). A personal view of hot deck imputation procedures. Survey Methodology, 5, 238-258.

Sande, I.G. (1979b). Hot deck imputation procedures. Symposium on Incomplete Data: Preliminary Proceedings (Panel on Incomplete Data of the Committee on National Statistics/ National Research Council), 484-498. U.S. Department of Health, Education, and Welfare, Washington, D.C.

Sande, I.G. (1982). Imputation in surveys: coping with reality. Amer. Statistician, 36(1), 145-152.

Santos, R.L. (1981a). Effects of Imputation on Complex Statistics, Survey Research Center, University of Michigan, Ann Arbor.

Santos, R.L. (1981b). Effects of imputation on regression coefficients. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1981, 140-145.

Scheiber, S.J. (1978). A comparison of three alternative techniques for allocating unreported Social Security Income on the Survey of the Low-Income Aged and Disabled. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1978, 212-218.

Sonquist, J.A., Baker, E.L. and Morgan, J.N. (1974, rev. ed.). Searching for Structure. Institute for Social Research, University of Michigan, Ann Arbor.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. Statistisk Tidskrift, 4, 278-283.

Vacek, P.M. and Ashikaga, T. (1980). An examination of the nearest neighbor rule for imputing missing values. Proc. Statist. Computing Sect., Amer. Statist. Ass., 1980, 326-331.

Welniak, E.J. and Coder, J.F. (1980). A measure of the bias in the March CPS earnings imputation system. Proc. Sect. Survey Res. Meth., Amer. Statist. Ass., 1980, 421-425.