Journal of Data Science **1**(2003), 275-292

Imputation Allowing Standard Variance Formulas

Michael P. Cohen
U.S. Bureau of Transportation Statistics

Abstract: Although deletion of cases is still a common method of dealing with item nonresponse, imputation is a major alternative. With traditional methods of imputation, though, the usual variance formulas understate the variance of estimates. This paper proposes that items be imputed from distributions more diffuse than those of the real data, thereby compensating for the underestimation of variance by the usual formulas. The impact on covariances is considered in the design of the method. The method is intended for use by data analysts applying techniques based on functions of first and second moments of means only.

Key words: Academic libraries, covariances, item nonresponse.

1. Introduction

1.1 Item nonresponse

Most surveys have item nonresponse no matter how well planned they may be. These missing data become a problem when it comes time to analyze the dataset. There are three main methods for dealing with item-level missing data: (i) delete complete cases whenever there are missing data for *any* variable being analyzed, (ii) delete cases but only as necessary for a particular family of estimates, and (iii) impute ("fill in values for") the missing data. Methods (i) and (ii) are still widely used in the social and behavioral sciences. Method (iii), though, has been demonstrated to be superior in previous research (Chan and Dunn, 1972; Beale and Little, 1975; Kim and Curry, 1977; Little, 1988; and Bello, 1995).

The problems with methods (i) and (ii) are not hard to ascertain. Method (i) may result in a substantial loss of cases, especially when many variables are being analyzed. The cases retained, moreover, may not be representative of those deleted, resulting in a bias. Method (ii) has the problems of method (i) but to a lesser degree. It has the serious additional problem of inconsistencies in the values of estimates. For instance, if x is being analyzed in conjunction with y, then the estimated mean of x will be based on cases where neither x nor y is missing. If, in another analysis, x is analyzed in conjunction with z instead, the estimated mean of x will in general be based on different cases so we get two different estimates of the same quantity. These inconsistencies can be very confusing to careful readers, resulting in a loss of confidence in the research.

Method (iii), called the *imputation technique*, solves the problems alluded to above. After imputation, one can use complete data methods of analysis without any need to discard cases. Another advantage is that the data can be imputed "in house," thus bringing the additional knowledge of the data collection people to bear on the missing data problem. This is not to say that imputation does not have its own drawbacks. Chief among these is the underestimation of standard errors (if standard variance formulas are used) — this happens essentially because the amount of "real" data is less than it appears to be. Although the reason is less obvious, covariance estimates undergo shrinkage toward zero (that is, *attenuation*). For general discussion of imputation, we recommend Kalton (1983), Kalton and Kasprzyk (1986), and Rubin (1987).

1.2 Standard formulas

The objective of this paper is to explore methods of imputation that permit the use of standard variance formulas. The results apply to estimates based on functions of first and second moments of means only. The question naturally arises, is this a direction worth pursuing? Survey methodologists, in particular, may regard this work as a step backward to the time when some recent developments (see the latter part of Subsection 2.1 were not yet available. But for social science data, one must consider the tremendous investment of training and experience that the social science analyst has in working with certain statistical analysis software systems, particularly SPSS and SAS. Quantitative methods courses in universities typically focus on the use of these software products. Thus it is common to treat item nonresponse by deleting cases and to estimate variances by adjusting by a design effect, by no adjustment at all, or by some *ad hoc* technique (e.g. use of a significance level of .01 instead of .05 in hypothesis testing with no variance adjustment).

Throughout this article, we assume that there is an indicator ("imputation flag") to show whether the item was a response or imputed.

The outline of this paper is as follows: Section 1 is this introduction. In Section 2 we discuss the one-variable case. The section consists of a subsection on the problems with the traditional approach followed by a subsection on the alternative approach. Section 3 expands the coverage to the multivariate case and, in particular, to the difficult problem of covariances. In the last section we make some final remarks.

2. The one variable case

2.1 Problems with the traditional approach

We begin by assuming the sample has been divided into groups of observations called *imputation classes* (Kalton, 1983, p. 67). Within each imputation class, we assume for now that the responding units for item y are a random subsample of all sampled units. Let the sample size in imputation class k be n_k with r_k responding and $m_k = n_k - r_k$ missing. We can number the units so that units $i = 1, 2, \ldots, r_k$ responded to item y and units $i = r_k + 1, \ldots, n_k$ did not. At this point we shall for simplicity drop the subscript k, denoting the imputation class, from the notation; however, it should be borne in mind that all calculations are within the imputation class. The best estimate (in many respects) of the mean of y within the imputation class is $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$ and the best estimate of the variance of the mean is $s^2_{\bar{y}_r} = \frac{1}{r(r-1)}\sum_{i=1}^{r}(y_i - \bar{y})^2$. For simplicity we are ignoring the sample weights in this discussion, but they could be incorporated. A finite population correction could also be included.

It is tempting to impute the missing values by $\bar{y}_r$. In fact, this choice has good "first order" properties in that $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}_r$. On the other hand,

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y}_r)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2 = \frac{r(r-1)}{n(n-1)} s^2_{\bar{y}_r},$$

so the variance of the mean will be underestimated. This perhaps should not be surprising in that we have chosen to impute the value that minimizes the variance expression. Here, as elsewhere in the paper, $y_{r+1}, \ldots, y_n$ refers to the imputed values (not the "true" values) of the missing y-values for units $r+1, \ldots, n$.

To combat the problem of underestimation of variances, imputation methods have been proposed that attempt to impute values drawn from the distribution of the observed y's. Although an improvement on mean imputation, this approach is also doomed to failure when it comes to estimation of variances (by standard formulas) as we shall see. If the imputed $y_{r+1}, \ldots, y_n$ are chosen to have about the same mean and deviations about the mean as the observed $y_1, \ldots, y_r$, then $\frac{1}{n-r}\sum_{i=r+1}^{n} y_i \approx \bar{y}_r$ so that $\bar{y} \approx \bar{y}_r$ where $\approx$ denotes "approximately equal" and $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ is the overall sample mean (in the imputation class). So, like mean imputation, the "first order" properties of these methods are good. Furthermore, $\frac{1}{r-1}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2 \approx \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. The variance of the mean is estimated by

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 \approx \frac{1}{n(n-1)}\frac{n-1}{r-1}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2 = \frac{r}{n} s^2_{\bar{y}_r}.$$

The variance of the mean is still underestimated although not so badly as with mean imputation. The problem is that the variance formulas are designed for n "real" observations, not $r < n$ observations and $m = n - r$ imputed values.

What, then, can be done? One promising approach is to alter the variance formula used (Rao and Shao, 1992; Särndal, 1992; Fay, 1996b; and Rao, 1996) but impute only once. Another idea, *multiple imputation*, makes use of several imputations to capture the missing variance component in variance estimates when missing data are present (Rubin, 1978, 1996; and Fay, 1992). Fay (1996a) and Kaufman (1996) investigate methods that are mixtures of these two approaches. The challenge is to find a method that is reasonably appealing to social science analysts who are inclined to delete cases to avoid the complications caused by missing data.

We consider in this paper single (as opposed to multiple) imputation methods that are intended for use with the standard variance formulas. The imputed values will be *more* dispersed than the observed values. Clearly this method will not work for estimating all features of the distribution; for example, it is not suited for estimating percentiles or histograms. But many statistical procedures depend on only the first two moments of means (and functions thereof), and it is for these procedures the imputation method is intended.

2.2 The alternative approach

Let us try to find imputed values $y_{r+1}, \ldots, y_n$ so that

$$\bar{y} = \bar{y}_r \tag{2.1}$$

and

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \frac{1}{r(r-1)} \sum_{i=1}^{r} (y_i - \bar{y})^2. \tag{2.2}$$

Let $D_r^2 = \frac{1}{r}\sum_{i=1}^{r}(y_i - \bar{y})^2$ and $D_m^2 = \frac{1}{m}\sum_{i=r+1}^{n}(y_i - \bar{y})^2$. Then D_r^2 and D_m^2 are respectively the average squared deviation of the observed and imputed values about their (common) mean. Rewriting (2.2) in terms of D_r^2 and D_m^2, we have

$$\frac{1}{n(n-1)}(rD_r^2 + mD_m^2) = \frac{1}{r-1}D_r^2. \tag{2.3}$$

Simplifying (2.3), we get

$$D_m^2 = \frac{n+r-1}{r-1}D_r^2. \tag{2.4}$$

There are many solutions to (2.1) and (2.4), but, if $m = n - r$ is even, there is one particularly simple solution:

$$y_i = \bar{y} \pm \sqrt{\frac{n+r-1}{r-1}} D_r \text{ for } i = r+1, \ldots, n, \tag{2.5}$$

where $m/2$ imputed values have the + sign and $m/2$ have the − sign. Note that if m is small so that $r \approx n$ then (2.5) reduces to $y_i \approx \bar{y} \pm 2^{1/2} D_r$ for $i = r+1, \ldots, n$; that is, the distance of an imputed value from the mean is about $2^{1/2}$ times the root mean squared deviation of the observed values. If $m \approx r \approx n/2$, representing a large amount of imputation, we have $y_i \approx \bar{y} \pm 3^{1/2} D_r$ for $i = r+1, \ldots, n$; in this case the distance of an imputed value from the mean is about $3^{1/2}$ times the root mean squared deviation of the observed values. Certainly the imputed values are much more dispersed than the observed values.

A result like (2.5), but with a finite population correction, was obtained by Lanke (1983) and discussed by Sedransk (1985).

If $m > 1$ is odd, a solution is available that is almost as simple as the one for m even. Set $y_{r+1} = \bar{y}$ and

$$y_i = \bar{y} \pm \sqrt{\frac{(n-r)(n+r-1)}{(n-r-1)(r-1)}} D_r \text{ for } i = r+2, \ldots, n, \tag{2.6}$$

where half the values imputed by (2.6) have the + sign and half have the − sign.

It ought to be mentioned that the imputed values will not generally satisfy the edit checks that the observed values had to satisfy, nor even necessarily be feasible values. But if the signs of the deviations from the mean of the imputed values are assigned randomly (or according to an appropriate pattern), the chance an estimated mean over a reasonably large domain will be outside the variables's range will be very small.

2.3 Example I

For illustrative purposes, a sample was selected from the 1994 Academic Library Survey of the U.S. National Center for Education Statistics. The

Table 1: Example for the One Variable Case: Academic Library Transactions

	T^*	$\bar{T}_D$	$\bar{T}_O$	$\bar{T}_{IC}$	$\hat{T}$
Imp. Class 1					
mean est.	13281	13771	15046	13771	13771
std. err. est.	1598	1778	1381	1304	1778
Imp. Class 2					
mean est.	19914	21174	20902	21174	21174
std. err. est.	2008	2099	1874	1869	2099
Overall					
mean est.	17371	18657	18657	18336	18336
std. err. est.	1435	1582	1316	1333	1526

sample of 60 institutions was selected by simple random sampling from the population of community colleges with academic libraries and where the number of faculty was in the range 25–124. In order to facilitate comparisons, only institutions with non-missing values for certain items (such as academic library transactions considered here) were included in the population.

The 60 institutions in the sample were divided into two imputation classes: Imputation Class 1 consists of 23 institutions with 25–49 faculty. Imputation Class 2 consists of 37 institutions with 50–124 faculty. For the variable *academic library transactions*, the values for 6 institutions in Imputation Class 1 and 4 institutions in Imputation Class 2, randomly selected, were set to missing.

Suppose we want to estimate the mean number of academic library transactions. The estimator $\bar{T}_D$ ("deletion") of Table 1 uses the non-missing cases only. The estimators $\bar{T}_O$ and $\bar{T}_{IC}$ both impute a mean value for the missing cases: $\bar{T}_O$ uses the overall mean whereas $\bar{T}_{IC}$ uses the mean within the imputation class. The estimator $\hat{T}$ is based on the imputations defined by (2.5). The mean using the true values for the missing cases is denoted by T^*. It is not really an estimator because the true values would not ordinarily be known, but it is included for comparison purposes.

Let r_k and n_k denote respectively the number of respondents and the sample size in Imputation Class k, $k = 1, 2$. Note that $\bar{T}_D$ and $\bar{T}_{IC}$ give the same mean estimates for each imputation class, but different mean estimates overall (that is, for the two imputation classes combined). This occurs because $\bar{T}_D$ is based on r_k values in Imputation Class k whereas $\bar{T}_{IC}$ is based on n_k values, some imputed. A problem with $\bar{T}_D$ is that it does not give each imputation class its proper weight.

In terms of estimating the mean at the imputation class level, $\bar{T}_D$, $\bar{T}_{IC}$, and $\hat{T}$ all do well. At the overall level, $\bar{T}_D$ and $\bar{T}_O$ do not account for having a sample size in each imputation class of n_k rather than r_k. As expected, the two mean imputation estimators $\bar{T}_O$ and $\bar{T}_{IC}$ underestimate the standard errors. The estimator $\hat{T}$ does well for both means and standard errors.

3 The multivariate case

Of course, in major surveys we almost always have many variables available to use as covariates but themselves possibly having missing values. Let us begin with the simplest such case.

3.1 Two variables, same units with missing values

We assume again the sample has been divided into imputation classes. Within the imputation class, suppose the responding units for items x and y are a random subsample of all sampled units. We further suppose in this subsection that x and y are observed for the same units and missing for the same units. Let the sample size in the imputation class be n with r units responding to the two items and $m = n - r$ missing the two items. We number the units so that units $i = 1, 2, \ldots, r$ responded to items x and y whereas units $i = r + 1, \ldots, n$ did not.

We seek to impute so that the means of x and y within the imputation class are $\bar{x}_r = \frac{1}{r}\sum_{i=1}^{r} x_i$ and $\bar{y}_r = \frac{1}{r}\sum_{i=1}^{r} y_i$. We also want

$$\begin{aligned} s^2_{\bar{x}} \equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{r(r-1)}\sum_{i=1}^{r}(x_i - \bar{x})^2 \text{ and} \\ s^2_{\bar{y}} \equiv \frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \frac{1}{r(r-1)}\sum_{i=1}^{r}(y_i - \bar{y})^2. \end{aligned}$$

Lastly, we would like to preserve the correlation of the means:

$$\rho_{\bar{x},\bar{y}} \equiv \frac{1}{n(n-1)} \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_{\bar{x}}s_{\bar{y}}} = \frac{1}{r(r-1)} \frac{\sum_{i=1}^{r}(x_i-\bar{x})(y_i-\bar{y})}{s_{\bar{x}}s_{\bar{y}}}.$$

Let

$$\begin{aligned} D_{x,r}^2 &= \frac{1}{r}\sum_{i=1}^{r}(x_i-\bar{x})^2 \\ D_{y,r}^2 &= \frac{1}{r}\sum_{i=1}^{r}(y_i-\bar{y})^2, \text{ and} \\ C_{x,y,r} &= \frac{1}{r}\sum_{i=1}^{r}(x_i-\bar{x})(y_i-\bar{y}). \end{aligned}$$

We concentrate first on the case $m = 4$, that is, four pairs of missing values. Let $\dot{x}_j = x_{r+j} - \bar{x}$ and $\dot{y}_j = y_{r+j} - \bar{y}$ for $j = 1, 2, 3, 4$ denote the differences of the imputed values from the appropriate mean. Then, by the argument used to get (2.4), we have

$$\begin{aligned} \dot{x}_1 + \dot{x}_2 + \dot{x}_3 + \dot{x}_4 &= 0, \\ \dot{y}_1 + \dot{y}_2 + \dot{y}_3 + \dot{y}_4 &= 0, \\ \dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2 + \dot{x}_4^2 &= \frac{4(n+r-1)}{r-1}D_{x,r}^2, \\ \dot{y}_1^2 + \dot{y}_2^2 + \dot{y}_3^2 + \dot{y}_4^2 &= \frac{4(n+r-1)}{r-1}D_{y,r}^2, \text{ and} \\ \dot{x}_1\dot{y}_1 + \dot{x}_2\dot{y}_2 + \dot{x}_3\dot{y}_3 + \dot{x}_4\dot{y}_4 &= \frac{4(n+r-1)}{r-1}C_{x,y,r}. \end{aligned}$$

To solve, let's try the trigonometric substitutions

$$\begin{aligned} \dot{x}_1 \;=\; -\dot{x}_3 &= \sqrt{\frac{2(n+r-1)}{r-1}}D_{x,r}\sin\theta, \\ \dot{x}_2 \;=\; -\dot{x}_4 &= \sqrt{\frac{2(n+r-1)}{r-1}}D_{x,r}\cos\theta, \\ \dot{y}_1 \;=\; -\dot{y}_3 &= \sqrt{\frac{2(n+r-1)}{r-1}}D_{y,r}\cos\phi, \text{ and} \end{aligned}$$

$$\dot{y}_2 \;=\; -\dot{y}_4 \;=\; \sqrt{\frac{2(n+r-1)}{r-1}} D_{y,r} \sin\phi.$$

One can verify that all equations are satisfied provided that

$$D_{x,r} D_{y,r} \left(\sin\theta\cos\phi + \cos\theta\sin\phi\right) = D_{x,r} D_{y,r} \sin(\theta+\phi) = C_{x,y,r}.$$

So

$$\theta + \phi = \arcsin\left(\frac{C_{x,y,r}}{D_{x,r} D_{y,r}}\right).$$

It is easy to check that the argument of the arcsin function is at most 1 in absolute value so $\theta + \phi$ is well defined. So long as the constraint on their sum is satisfied, θ and ϕ may take on a range of values, each corresponding to a solution to the original equations.

This family of solutions provides the convenience of being expressible in closed form. To choose an approximate "best" solution within the family, one could impose an additional desirable constraint.

Now let's turn to the harder case (because it is less symmetric): $m = 3$. The equations for this case are

$$\begin{aligned}
\dot{x}_1 + \dot{x}_2 + \dot{x}_3 &= 0, \\
\dot{y}_1 + \dot{y}_2 + \dot{y}_3 &= 0, \\
\dot{x}_1^2 + \dot{x}_2^2 + \dot{x}_3^2 &= \frac{3(n+r-1)}{r-1} D_{x,r}^2, \\
\dot{y}_1^2 + \dot{y}_2^2 + \dot{y}_3^2 &= \frac{3(n+r-1)}{r-1} D_{y,r}^2, \text{ and} \\
\dot{x}_1\dot{y}_1 + \dot{x}_2\dot{y}_2 + \dot{x}_3\dot{y}_3 &= \frac{3(n+r-1)}{r-1} C_{x,y,r}.
\end{aligned}$$

A particular solution, obtained through the use of substitutions and a careful examination of the solution to certain quadratic equations, is given by

$$\begin{aligned}
\dot{x}_1 &= -\sqrt{\frac{n+r-1}{2(r-1)}} D_{x,r}, \\
\dot{x}_2 &= \sqrt{2\frac{n+r-1}{r-1}} D_{x,r},
\end{aligned}$$

$$\begin{aligned}
\dot{x}_3 &= -\dot{x}_1 - \dot{x}_2, \\
\dot{y}_1 &= -\sqrt{\frac{n+r-1}{2(r-1)}}\left(\frac{C_{x,y,r}}{D_{x,r}}\right) \\
&\qquad + \sqrt{\frac{3(n+r-1)}{2(r-1)}\left(1-\frac{C_{x,y,r}^2}{D_{x,r}^2 D_{y,r}^2}\right)}\; D_{y,r}, \\
\dot{y}_2 &= \sqrt{\frac{2(n+r-1)}{r-1}}\left(\frac{C_{x,y,r}}{D_{x,r}}\right), \text{ and} \\
\dot{y}_3 &= -\dot{y}_1 - \dot{y}_2.
\end{aligned}$$

We have obtained, in fact, a family of six solutions, one for each permutation of the subscripts $1, 2, 3$. As for the $m = 4$ case, one could select among the closed form solutions here by imposing another constraint.

The cases $m > 4$ are easier (because the constraints are less restrictive) and can be handled in a variety of ways. For example, one way to treat $m = 5$, although probably not the best way, is to set $\dot{x}_5 = \dot{y}_5 = 0$ and then apply the solution for $m = 4$.

For $m = 1$ and $m = 2$, no exact solutions can be obtained. If the correlation between $\bar{x}$ and $\bar{y}$ is important, we recommend dealing with $m = 1$ or $m = 2$ by making a random choice among the solutions for $m = 3$ or $m = 4$.

3.2 Example II

The is a continuation of Example I from Subsection 2.3. We now focus on two variables: academic library transactions and total expenditures (more precisely, total operating expenditures of the academic library). The variable academic library transactions is missing for the same cases as in Subsection 2.3, and for exactly these cases the variable total expenditures is also missing.

In Table 2, the estimators $\bar{T}_D$, $\bar{T}_O$, and $\bar{T}_{IC}$ are the same as before, and $\bar{E}_D$, $\bar{E}_O$, and $\bar{E}_{IC}$ are the analogous estimators for total expenditures. The estimators $\hat{T}$ and $\hat{E}$ are defined as for $m = 4$ in the previous subsection. For Imputation Class I (with six missing values), the final two missing values were imputed at the mean, and the first four were expanded away from the mean by a factor of $\sqrt{3/2}$. The estimators $\bar{C}_D$, $\bar{C}_O$, $\bar{C}_{IC}$, and $\hat{C}$ are the corresponding estimators of the covariance of the means. The means T^* and E^* and the covariance of the means C^* are computed with no missing values for purposes of comparison.

The estimators $\hat{T}$ and $\hat{E}$ do well as in the one variable situation, and the covariance estimate $\hat{C}$ is not attenuated.

3.3 Two variables, only one variable missing

Within each imputation class, suppose now that item x is observed for all n units. Item y, on the other hand, is missing for $m \geq 1$ units and observed for the other $r = n - m$ units. We assume the missing y's are missing at random but not necessarily missing *completely at random*; that is, the missingness may depend on the observed x's and y's. The units are numbered so that units $i = 1, 2, \ldots, r$ responded to item y whereas units $i = r + 1, \ldots, n$ did not. The objective in this subsection is to impute the missing y's.

This situation introduces an important new feature: It is no longer appropriate to assume that $\bar{y}_r$ is the "best" estimate of the population mean of the y's. We can do better by making use of the x's corresponding to the missing y's.

Consider

$$e_i = \frac{y_i}{x_i}, \qquad i = 1, \ldots, n.$$

We shall explore the assumption that the e_i can be modelled as independent, identically distributed, and independent of the x's, within the imputation class. This assumption is reasonable in many circumstances and the reasoning can be extended to other situations.

We can apply the results of Subsection 2.2 to impute the "missing" e_i $(i = r + 1, \ldots, n)$ to satisfy:

$$\begin{aligned} \bar{e} = \bar{e}_r &\equiv \frac{1}{r}\sum_{i=1}^{r} e_i, \text{ and} \\ s^2_{\bar{e}} = s^2_{\bar{e}_r} &\equiv \frac{1}{r(r-1)}\sum_{i=1}^{r}(e_i - \bar{e})^2. \end{aligned}$$

Table 2: Example for the Two Variable Case:
Academic Library Transactions and Total Expenditures

	Academic Library Transactions				
	T^*	$\bar{T}_D$	$\bar{T}_O$	$\bar{T}_{IC}$	$\hat{T}$
Imp. Class 1					
mean est.	13281	13771	15046	13771	13771
std. err. est.	1598	1778	1381	1304	1778
Imp. Class 2					
mean est.	19914	21174	20902	21174	21174
std. err. est.	2008	2099	1874	1869	2099
Overall					
mean est.	17371	18657	18657	18336	18336
std. err. est.	1435	1582	1316	1333	1526
	Total Expenditures				
	E^*	$\bar{E}_D$	$\bar{E}_O$	$\bar{E}_{IC}$	$\hat{E}$
Imp. Class 1					
mean est.	222482	245393	186245	184970	245393
std. err. est.	17062	19830	25729	26107	19830
Imp. Class 2					
mean est.	419627	424144	380307	380579	424144
std. err. est.	35277	38028	39835	39767	38028
Overall					
mean est.	344054	363368	305917	305596	355623
std. err. est.	25799	28518	29027	29082	26988
	Covariance				
	C^*	$\bar{C}_D$	$\bar{C}_O$	$\bar{C}_{IC}$	$\hat{C}$
Imp. Class 1					
covariance est.	-628	-679	-790	-565	-679
Imp. Class 2					
covariance est.	-686	-709	-644	-676	-709
Overall					
covariance est.	-436	-472	-440	-422	-462

Table 3: Estimating Total Expenditures with Number of Staff as Auxiliary Variable

	E^*	$\bar{E}_D$	$\bar{E}_O$	$\bar{E}_{IC}$	$\hat{E}$	$\hat{E}_{AUX}$
Imp. Class 1						
mean est.	222482	245393	276169	245393	245393	233817
std. err. est.	17062	19830	18258	14539	19830	17797
Imp. Class 2						
mean est.	419627	424144	417573	424144	424144	423022
std. err. est.	35277	38028	34005	33860	38028	38640
Overall						
mean est.	344054	363368	363368	355623	355623	350494
std. err. est.	25799	28518	23725	24282	26988	27395

From the imputed e_i, we get imputed y_i by $y_i = x_i e_i$.

3.4 Example III

This example treats the estimation of total expenditures introduced in Example II of Subsection 3.2, but now we add an auxiliary variable *number of staff* (number of full-time equivalent academic library staff). The values of number of staff are available for all institutions. We shall form an estimator of total expenditures ($\hat{E}_{AUX}$) that uses the number of staff as an auxiliary variable in the manner described in Subsection 3.3. A rough plot of total expenditures versus number of staff revealed an intercept at around 100000. For this reason, 100000 was subtracted from the total expenditures before forming ratios and added back later. No attempt was made to fine-tune the model. In Table 3, $\hat{E}_{AUX}$ is the new estimator; the others were defined in Subsection 3.2.

Although the improvement is not dramatic, $\hat{E}_{AUX}$ seems to outperform the deletion estimator $\bar{E}_D$ without having distorted standard errors.

3.5 More general situations

We have discussed but a small subset of the multitude of missing data situations that arise in practice. In this subsection we shall just briefly touch upon three aspects needing more serious investigation.

1. We have only considered imputing one or two variables, but there will almost always be more than that, often hundreds. If there are k variables to be imputed, the number of pairwise correlations to consider is $k(k-1)/2$. Clearly we will reach a point where the equations for the correlations cannot be solved exactly. At least two ways of treating this problem come to mind.

   (a) The variables can be divided into blocks of variables thought to be closely related. We can then try to control only for the correlations between variables within the same block. The presumption is that this will account for most of the correlation.

   (b) As an alternative to trying to control certain correlations exactly, we might only seek to control them on average by randomizing among solutions to the equations for the correlations. A related idea would be to seek approximate solutions that minimize the distance (based on some distance function) to the solutions of the individual equations.

2. Even for two variables, we have only considered the two simplest patterns of missingness for the data: either only one of two variables has missing values, or the two variables have missing values for the same units. The hope, of course, is that we can solve more general problems by an iterative procedure, perhaps first imputing values when one variable is missing but not the other, then the reverse, and finally when both are missing.

3. We have treated imputation within imputation classes, implicitly assuming that the imputation will have good properties for means and variances and correlations of means across imputation classes. If the data for each imputation class are (at least approximately) independent from each other, then the assumption is justified. Otherwise, the results presented here can be extended, but only if we know what the

variances and correlations of the means of the observed values across imputation classes *should be.*

4. Final comment

Deletion of cases still seems to be the most common way that data analysts in the social and behavioral sciences cope with item nonresponse. There is therefore value in searching for techniques for handling missing data that are easy to use yet have desirable statistical properties.

This paper is just a beginning exploration of an approach to imputation that makes use of imputed values distributed more diffusely than the observed data. The approach is not intended for all statistical applications, only those based on functions of the first two moments of means. For many problems we hope it will develop into a reliable technique not requiring multiple imputations or special variance formulas.

Acknowledgements

The author thanks Steven Kaufman for invaluable discussions. He also thanks Graham Kalton for pointing out the relevance of Lanke (1983) and Sedransk (1985).

In developing the results of Subsection 3.1, extensive use was made of the symbolic mathematical software Derive.

The author appreciates the insightful remarks of the Editor and Referee.

References

Beale, E. M. L. and R. J. A. Little (1975). Missing values in multivariate analysis, *J. Roy. Statist. Soc. Ser. B*, **37** 129-146.

Bello, A. L. (1995). Imputation techniques in regression analysis: Looking closely at their implementation, *Comput. Statist. Data Anal.*, **20** 45-57.

Chan, L. S. and O. J. Dunn (1972). The treatment of missing values in discriminant analysis - I. The sampling experiment, *J. Amer. Statist. Assoc.*, **67** 473-477.

Fay, R. E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods, Amer. Statist. Assoc.*, 227-232.

Fay, R. E. (1996a). Alternative paradigms for the analysis of imputed survey data, *J. Amer. Statist. Assoc.*, **91** 490-498.

Fay, R. E. (1996b). Rejoinder, *J. Amer. Statist. Assoc.*, **91** (1996b) 517-519.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Ann Arbor: University of Michigan.

Kalton, G. and D. Kasprzyk (1986). The treatment of missing survey data, *Survey Methodology*, **12** 1-16.

Kaufman, S. (1996). Estimating the variance in the presence of imputation using a residual *Proceedings of the Section on Survey Research Methods, Amer. Statist. Assoc.*, 423-428.

Kim, J. O. and J. Curry (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, **6** 215-240.

Lanke, J. (1983). Hot deck imputation techniques that permit standard methods for assessing precision of estimates. *Statist. Review*, **21** 105-110.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Appl. Statist.*, **37** 23-38.

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

Rao, J. N. K. (1996). On variance estimation with imputed survey data. *J. Amer. Statist. Assoc.*, **91** 499-506.

Rao, J. N. K. and J. Shao (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79** 811-822.

Rubin, D. B., (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *J. Amer. Statist. Assoc.*, **91** 473-489.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, **18** 241-252.

Sedransk, J. (1985). The objectives and practice of imputation. *Proceedings of the First Annual Research Conference.* Washington, DC: United States Bureau of the Census 445-452.

Received July 14, 2002; accepted September 22, 2002

Michael P. Cohen
U.S. Bureau of Transportation Statistics
400 Seventh Street SW #4432
Washington DC 20590 USA
michael.cohen@bts.gov