IMPUTATION OF INCOME: A PROCEDURAL COMPARISON

Timothy F. Champney and Ralph Bell, The University of Chicago

1. Introduction

When a large number of questions are asked of a large number of survey respondents, it is inevitable that nonresponses due to refusals to answer specific questions, "don't knows", and other nonresponses will emerge in the data. Nonresponses may be handled in one of three ways (Frankel and Banks, 1979: 97-98): (1) calculations can be restricted to only those respondents with complete data, (2) calculations can be restricted to only those variables with reported values, or (3) a missing value imputation procedure can be applied to estimate what nonrespondents would have reported had they answered the questions. Although the third method appears, at first glance, to be the most radical approach to the nonresponse problem, an argument can be made that it is, in fact, the most conservative of the three alternatives.

By including only those cases with complete data or only those variables with reported values, a number of assumptions are implicitly made concerning the nature of nonresponse. Rubin (1978) has discussed differences in "ignorable" and "nonignorable" missing data processes. The missing data process is "ignorable" if the population distribution of the data being reported is identical to the unconditional distribution. For example, if the process generating the missing data is random, the process is "ignorable". If the process generating the missing data is "nonignorable", however, any statistical procedure that fails to take this condition into account will be biased. If, for example, the probability of reporting one's income is dependent on the actual value of that person's income, the process generating the missing values for income is "nonignorable" and estimates of the population mean and variance on income will be biased if they are simply calculated from the observed data.

A variety of procedures have been developed to impute missing data. Many of these procedures assume, to some degree, that the missing data generating process is "nonignorable". Several of the more traditional imputation procedures have been reviewed by Ford (1980). The procedures reviewed by Ford (1980) include: (1) a ratio estimation procedure, (2) a regression procedure, (3) the Current Population Survey (CPS) hot deck procedure (Nordbotten, 1963), and (4) the Statistics Canada procedure (College et al., 1978). The ratio procedure and the regression procedure assume that any variation in missing values beyond what can be accounted for by some functional relationship to one or more other variables reported in the data set is due to an "ignorable" process. The hot deck procedures also assume that the variation of missing values within "cells" of the imputation design is due to an "ignorable" process (Rubin, 1978).

Ford (1980) has outlined seven desirable attributes of an acceptable imputation procedure. The procedure should (1) reduce nonresponse bias, (2) use values from similar cases to compute imputed values, (3) produce a "clean data set", i.e., a complete, consistent data set that is readily analyzable using statistics programs, (4) preserve both joint and marginal population distributions as they are represented by the sample, (5) allow an assessment of the impact of the imputation procedure on the standard errors, (6) avoid using the same value many times, and (7) be practical to use with large data sets with several thousand or more cases.

Rubin (1978) has shown that, given the usual level of uncertainty about the missing data generating process and given the need to assess the impact of the imputation procedure on the standard errors, it is desirable to consider multiple imputed values for each missing value in the data. In practice, it is useful to consider at least two alternative estimates of a missing value since two is the minimum number of estimates necessary to calculate the component of variation due to the imputation procedure (Oh and Scheuren, 1980). If, however, as Rubin (1978) suggests, components of variation due to imputation, both within and between models, are to be assessed then more than two alternative estimates of the missing value must be computed (a minimum of two per model). Oh and Scheuren (1980) have provided a formula for estimating the impact of the imputation procedure on the standard error of proportions given two alternative estimates which will be discussed later in the paper.

2. Alternative Imputation Procedures

Non-Hot Deck Procedures. For comparison purposes, we considered several non-hot deck imputation procedures. The procedures selected for comparison were (1) the generation of random numbers with the same distribution as the observed values, (2) least squares multiple regression estimates, (3) additive cell mean estimates, and (4) least squares with a random number having the same distribution as the residuals from the regression model added.

Hot Deck Procedures. Several alternative imputation procedures possessing the qualities outlined by Ford (1980) were considered for use. Difficulties with the CPS hot deck procedure were realized when Frankel and Banks (1979) reported large differences in the mean estimates in imputed values depending on whether the data were sorted in the usual east to west geographic order or whether the data were sorted in the reverse order. Clearly, an imputation method that does not depend on the initial ordering of the data is more desirable. The Statistics Canada procedure (College et al., 1978) appeared to offer a viable solution to the ordering problem. With this procedure, some variable or combination of variables that are present for all cases are used to define a distance function for locating suitable "donors". In hot deck terminology, a "donor" refers to the case that contributes its value to a case or cases with missing values. The cases with missing values are referred to as "candidates". In the Statistics Canada procedure, the case closest to the candidate, as defined by the distance function, is selected as the donor for that case. In practice, this method can be accomplished by sorting the cases on a variable defining the distance function followed by a pass through the data employing a forward and backward spacing of the records to identify the closest donor for each candidate in the data set. This procedure can be handled by sequential access methods in a computer language such as FORTRAN, or by direct access methods such as the data manipulation capabilities provided by the SAS software package (SAS Institute, Inc., 1979; 1981: 1.3-1.4).

The procedure (5) selected for the present comparison varied somewhat from the Statistic Canada procedure in that the previous case and next following case in the data deck with valid data were selected as donors for imputing two alternative values.

Modified Hot Deck Procedures. As we began to examine the Statistics Canada procedure in detail, several difficulties became apparent. The identification of an appropriate distance function was not a trivial matter since it is the location of the appropriate donor that solves the "nonignorability" problem in the missing value process. A method for solving this problem has been suggested by Schieber (1978), however. This method involves using multiple regression to generate predicted values of the to-be-imputed variable for all cases in the data set. The predicted values are then used to define the distance function. Scheuren (personal communication) has pointed out that the use of this procedure may be problematic when the fit of the regression surface is poor or when the distance function between donors and candidates is large. In order to compensate for these problems, Schieber (1978) computed the imputed values as the predicted value of the candidate plus the residual of the donor. Note that when the donor and the candidate have the same predicted value, this procedure is equivalent to taking the observed value of the donor as the estimate for the candidate. Furthermore, this procedure also solves the common problem of failing to find donors within a cell of the hot deck design. The problem is solved because indicators of cell membership may be incorporated as additive terms in the multiple regression model, thereby eliminating the use of cells altogether. One drawback of Schieber's (1978) method is that it does not provide a way to assess the impact of the imputation procedure on the standard errors. If, however, two alternative estimates of each missing value are computed, it is possible to estimate the impact of the procedure. In the present application, (procedure 6) alternative estimates were computed by taking the residuals from the donors with the closest and next closest predicted values and adding each separately to the predicted value of the candidate. Alternatively, if the fit of the regression model is poor or the sample size is small, one may take the observed values for the nearest and next nearest donors as the imputed values for the candidate. This method was also selected for comparison (procedure 7).

The seven imputation procedures were compared using income data from telephone surveys conducted in five cities. The remainder of the paper discusses our methodology in more detail and elaborates on the findings of our comparisons.

3. The Problem

Given the relatively large number of alternative imputation procedures available, the problem becomes deciding which of the methods most accurately reproduces original data values and which of the methods provides the best estimates of the mean and of the impact of the imputation procedures on the variance when data are missing either as a result of an "ignorable" or "nonignorable" process. We have selected seven imputation methods to compare in these terms. The seven methods under consideration are (1) random numbers, (2) additive cell means, (3) least squares multiple regression, (4) least squares plus a random residual term, (5) a variant of the Statistics Canada hot deck imputation procedure, ("old hot deck") (6) the predicted values of the candidate plus the residual of the donor (Schieber's (1978) method), ("new Pred + Res")and (7) a variant of number 6 using the observed value of the donor ("new HD obs").

4. Data and Comparison of Methods

Data. To compare the efficiency of the seven imputation methods described above, we used baseline data from a longitudinal evaluation of the Municipal Health Services Program (MHSP) conducted by the Center for Health Administration Studies at the University of Chicago. These data were collected using a telephone survey of selected areas of five cities originally funded by the Robert Wood Johnson Foundation to initiate the program. The telephone interviews were conducted during the period covering late 1979 and early 1980. Target areas were determined on the basis of patient origin studies and were defined as the geographic area containing the residences of approximately 75% of the MHSP patients. The five cities funded by the Johnson Foundation were Baltimore (BALT), Cincinnati (CINC), Milwaukee (MILW), St. Louis (STLU), and San Jose (SANJ).

Blocks of telephone exchanges in service within the defined service areas were obtained from local telephone companies. Numbers were randomly selected from this pool of exchanges to draw the sample. When a working number was reached by the interviewer. a screening procedure was employed to insure that we would obtain sufficient numbers of MHSP users and non-users.

The questionnaire was designed to collect a wide range of information on the utilization and costs of health care by residents of the service areas. In addition, a wide range of social and demographic data were collected on the members of sampled households. Once a household was screened into the survey, the interviewer asked to speak to the person in the household who was most familiar with the health care of the other members of that household. This person then served as a proxy respondent for the other members of the household. For households with up to five members, data were collected on every household member. In households with six or more members, four members were randomly selected for inclusion. The main respondent was always an adult 17 years of age or older and was always included in the sample. The main respondent for the household was asked to provide the yearly income for each sampled member of the household. In order to simplify comparisons of imputation methods, the sampling design implied here is ignored.

Imputation Models. All of the imputation models tested involved modeling the distribution of income and, in most cases, the conditional distribution of income as well. Following other social science research, we have assumed that income has a log-normal distribution. Income was, therefore, transformed by taking the natural logarithm of income plus 1 (log(income=1)). We added a constant of one to the value of income in order to avoid taking the logarithm of zero.

Based on their statistically significant relationship with income, three categorical variables were selected as predictors. The three variables selected as predictors were: employment status, relationship to the main respondent, and sex. These three categorical predictors were used to define cells for the Statistics Canada-like method and for the additive cell means method. These variables were represented by a set of dummy coded variables in the regression methods. No attempt was made to adjust estimates of regression weights for truncation in the data. This could be employed as a further refinement of the above methods.

One continuous variable, number of hours worked during the previous year, was also incorporated within some of the methods. The number of hours worked during the previous year was used to compute the distance function in the Statistics Canada-like procedure and its log transformation was used in the regression models. The same variables were employed as predictors under the various methods where ever possible in order to minimize differences in results that may have been due to differences in the choice of predictors rather than to differences in the performance of the imputation procedures.

For the random number method, the estimated mean and standard deviation of the log transformation of income were computed separately for each of the five cities. Random numbers with these means and standard deviations were then generated using the pseudo-random normal variate generator available in the SAS statistics package. In order to estimate the component of variance due to imputation, two random number estimates were generated for each missing value. Random numbers for the regression plus random residual procedure were generated in a similar fashion but such that the distribution had a mean of zero and a standard deviation equal to the standard deviation of the residual from the regression models.

Evaluation Methods. In order to compare the performance of the seven imputation procedures tested, all cases with reported values for income were selected from the data set in each of the five cities. The procedures were run separately for each of the cities so that the cities could be regarded as replications. Two sets of missing values were generated for each site. The first set of cases were randomly selected using a random number generator to tag cases as missing for subsequent imputation. These cases were selected with a probability of .10. The second set of cases were based on the observed distribution of income. Cases were tagged as missing when log (income+1)/2 plus a standard normal deviate exceeded 5.1. This method assured that cases with higher income would be more likely to be tagged as missing but also assured that some cases with lower income would also be tagged. About 10% of all cases were tagged as missing using this method.

For each combination of imputation method, site, and missing value generation method, three performance statistics were computed: (1) the average mean square deviation of imputed values from reported values (2 in each case except for the cell mean and the least squares methods), (2) means calculated using all cases including the imputed values, and (3) an estimate of the variance including an adjustment for the component due to imputation. The formula used to compute an estimate of the variance was based on the formula for estimating the variance of proportions given by Oh and Scheuren (1980: 94) as:

$$\mathrm{VAR}\ (p) = \frac{n-m}{n}^{2} \cdot \frac{p_1^n - p_2^n}{2}^{2} + \frac{p^r(1-p^r)}{m}$$

ignoring the correction for cells with only one donor and where n is the total number of cases in the sample, m is the number of reported values, p^r is the proportion of positive responses among those who reported, and p_1^N and p_2^N are the alternative imputed proportions of positive responses among the cases with missing data. The right hand term of the formula is the squared sample standard deviation and the left hand term is an adjustment for the imputation. In the present case, the formula was modified so that the estimate of variance for a continuous variable (i.e., income) was

$$\mathrm{VAR}\ (\bar{Y}) = \frac{n - m}{n}^{2} \cdot \frac{\bar{y}_1^N - \bar{y}_2^N}{2}^{2} + \mathrm{VAR}\ (\bar{\bar{Y}})$$

where VAR $(\bar{Y})$ is the estimated population variance for the reported values (or the mean of each pair of imputed values for missing cases)[2] and $\bar{Y}_1^N$ and $\bar{Y}_2^N$ are means of the alternative estimates among cases with missing values.

5. Results of the Comparisons

Imputation Summary Statistics. The results of the missing value generation and the model fitting are summarized in Table 1. Under both the missing proportional to income and the missing at random conditions, the percentage of missing values was approximately 10%. The multiple R-squares for the regression models were between about .60 and .70 and the additive cell mean R-squares were between about .50 and .60. The R-square values appeared unaffected by the type of missing value generation process used. That is, the R-squares for models where missing values were generated at random are similar to the models where they were generated at a higher rate for higher income respondents.

Mean Squared Errors. The mean squared deviation of imputed values from the observed values is reported in Table 2 for income in its original metric (expressed in millions). The mean square deviations of imputed values from observed values in log units were also computed but are not tabled. Outliers, particularly in the San Jose (SANJ) data, tended to distort the results expressed in dollar income. The pattern of the results was otherwise fairly consistent and computation of mean square errors in log units removed the inconsistency. In the four remaining sites, the least squares regression method came closest to reproducing the original reported values, regardless of the process under which the missing values were generated. The cell means procedure also outperformed the hot deck procedures in most cases. When data were missing proportional to income, the Statistics Canada-like hot deck procedure performed slightly better than the newer methods which use a distance function based on the predicted values of income from the regression model. When data were missing at random the newer procedures performed better in two of the five sites Baltimore (BALT) and Milwaukee (MILW). The least accurate procedure was consistently the random number method. The least squares plus a random residual was consistently the next least accurate of the methods we compared.

Impact on Means. Means were computed using the next following case for the Statistics Canada-like method and the nearest neighbor for the "new" hot deck methods. In all cases a single imputed value rather than an average of two for each case was used. As indicated in Table 3, when data were missing at random the imputation method had very little impact on the sample means. When data were missing with a probability proportional to income a fairly consistent pattern of differences between methods emerged. When the cell mean or least squares methods were used, the sample means calculated using the original data were most severely underestimated. The random number method also resulted in underestimation of the original sample means. Although all methods resulted in underestimation of the sample means, the least squares method with a random residual added (random + LS) resulted in the smallest underestimation followed closely by the various hot deck procedures.

Impact on Variance. Population variance estimates (expressed in millions) are presented in Table 4 using the original observed values and the modification of Oh and Scheuren's (1980) formula presented above. The cell mean versus least squares (CM vs LS) estimate was computed using the average of the cell mean and least squares estimates and the differences between the means of these imputed values. When data were missing at random, all of the methods resulted in a variance estimate that was fairly close to the estimates for the observed values. The random number methods tended to overestimate the variance, however. The hot deck methods differed little from one another and resulted in a slight underestimation of the variance. The cell mean versus the least squares estimate resulted in the most severe underestimation of the variance.

When data were missing proportional to income, all of the methods tended to underestimate the variance. This situation was most exaggerated in the case of San Jose where three extreme outliers from the upper tail of the income distribution were considered missing and subsequently imputed. In three of the other four sites, the random number plus least squares (Random + LS) method resulted in an estimate slightly closer to the estimate based on observed values than the hot deck methods.

6. Discussion

The mean square errors indicate that the least squares method may result in the most accurate reproduction of the observed values for income. The generality of our conclusion to other data sets may be hampered by the unusually large R-squares we were able to generate for models based on our data. Further research on this method applied to hospital utilization and expenditure data (Champney and Bell, 1982), resulted in model R-squares ranging from .25 to .30. The same pattern of results were obtained for these models as those reported here for income. The least squares method suffers, however, from its failure to provide an accurate estimate of the population variance. The least squares method also suffers in that it leads to underestimation of the sample mean when data are missing proportional to income. In this case the hot deck methods and the least squares with random residual method provide somewhat more satisfactory estimates of the mean and variance. No particular hot deck method was found to be superior to the other hot deck methods, however. All methods were equally easy to implement using currently available statistical software. The multiple R-square may be considered an advantage of the newer hot deck methods (those using a distance function based on multiple regression predicted values), however, since the multiple R-square serves to describe the similarity of donors to candidates. With the exception of handling extreme outliers, both the least squares with a random residual term and the hot deck methods performed amazingly well when data were missing proportional to income. In the present case, the attenuation of the reported income distribution with missing income values was contrived to be rather moderate. Further work is needed to fully determine which imputation method performs most satisfactorily under extreme attenuation or complete truncation of the income distribution, however.

FOOTNOTES

[1] This research is supported by contract HCFA-500-78-0097 from the U.S. Department of Health and Human Services, Health Care Financing Administration and a grant from the Robert Wood Johnson Foundation. The research was based at The Center for Health Administration Studies, University of Chicago. Helpful comments were received from Martha J. Banks.

[2] This further modification of the formula was necessary because we found that the above formula led to underestimation of the population variance of a continuous variable when cases were missing proportional to income.

REFERENCES

[1] Champney. T. F. and R. Bell. "Imputation procedures: a comparison using hospital utilization and expenditure data". Paper to be presented at the annual meetings of the American Public Health Association in Montreal, November 1982.

References(continued)

[2] College, M.D., J. H. Johnson, R. Pare, and I.G. Sande. "Large scale imputation of survey data". American Statistical Association, Proceedings of the Section on Survey Research Methods, 1978: 431-435.

[3] Ford, B.L. Incomplete Data in Sample Surveys: the Theory of Current Practices. National Academy of Sciences, Panel on Incomplete Data, 1980.

[4] Frankel, M.R. and M.J.Banks. "Adjusting for nonresponse to specific questions". Pp. 97-103 in R. Andersen, J. Kasper, M.R. Frankel, and Associates (Eds.), Total Survey Error. San Francisco: Jossey Bass, 1978.

[5] Nordbotten, S. "Automatic editing of individual statistical observations". Conference on European Statisticians No. 2, United Nations, 1963.

[6] Oh, H.L. and F.J. Scheuren. "Estimating the variance impact of missing CPS income data". American Statistical Association, Proceedings of the Section on Survey Research Methods, 1980: 408-415.

[7] Rubin, D.B. "Multiple imputations in sample surveys-a phenomenological Baysean approach". American Statistical Association, Proceedings of the Section on Survey Research Methods, 1978:20-28.

[8] SAS Institute Inc. SAS User's Guide. Raleigh, North Carolina, 1979.

[9] SAS Institute Inc. SAS 79.5 Changes and Enhancements, Technical Report #P-115. Cary, North Carolina, 1981:1.3-1.4.

[10]Schieber, S.J. "A comparison of three alternative techniques for allocating unreported social security income on the survey the low income aged and disabled". American Statistical Association, Proceedings of the Section on Survey Research Methods, 1978: 212-218.

Table 1. Imputation and Model Summary.

		Missing at Random			Missing Proportional to Income		
Site	Total N	% Missing	LS R Square	CM R Square	% Missing	LS R Square	CM R Square
BALT	1017	10.0	.710	.579	9.3	.715	.585
CINC	1187	9.3	.602	.487	12.2	.611	.498
MILW	1017	9.9	.702	.496	11.0	.712	.500
STLU	848	11.2	.721	.564	9.9	.712	.571
SANJ	1087	10.8	.564	.543	1198	.674	.538

Table 2. Comparison of Imputation Mean Square Errors.*

Data Missing at Random

				Method			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Site	Random	Cell Mean	Least Squares	Random + LS	Old Hot Deck	New Pred+Res	New HD Obs
BALT	332.58	27.91	25.62	97.05	44.94	42.37	42.28
CINC	400.00	66.70	64.58	283.67	91.94	209.12	209.44
MILW	240.16	31.71	30.33	114.76	44.59	37.42	37.30
STLU	895.85	48.55	47.48	119.09	88.58	81.48	81.04
SANJ	442.54	45.55	38.71	123.49	158.15	182.36	182.30

Data Missing Proportional to Income

BALT	302.90	117.77	100.23	221.77	104.67	142.45	142.45
CINC	335.03	98.58	93.99	211.83	297.89	142.45	125.01
MILW	1008.12	50.25	42.97	117.61	64.48	57.63	57.67
STLU	248.08	56.12	48.90	98.69	69.96	72.39	72.17
SANJ	5782.83	5358.08	5420.93	5531.78	5374.89	5345.47	5346.06

*Results in millions.

Table 3. Means with Imputed Values Included

Data Missing at Random

		Method						
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Site	Observed Values	Random	Cell Mean	Least Squares	Random + LS	Old Hot Deck[1]	New[2] Pred+Res	New[2] HD Obs
BALT	11,030	11,467	10,946	10,969	11,226	11,087	11,125	11,124
CINC	12,296	12,601	12,146	12,151	12,601	12,259	12,328	12,328
MILW	10,595	10,569	10,388	10,463	10,547	10,603	10,611	10,612
STLU	9,723	9,745	9,552	9,553	9,784	9,705	9,726	9,725
SANJ	13,171	13,201	13,507	13,077	13,299	13,310	13,330	13,330

Data Missing Proportional to Income

Site	Observed Values	(1) Random	(2) Cell Mean	(3) Least Squares	(4) Random + LS	(5) Old Hot Deck	(6) New Pred+Res	(7) New HD Obs
BALT	11,030	10,542	10,599	10,640	10,865	10,798	10,786	10,785
CINC	12,296	11,742	11,715	11,743	12,073	11,892	11,922	11,923
MILW	10,595	10,240	10,141	10,190	10,433	10,388	10,391	10,389
STLU	9,723	9,395	9,300	9,359	9,580	9,563	9,458	9,458
SANJ	13,171	11,646	11,564	11,678	11,935	11,944	12,181	12,182

[1] Next case with value present used to impute

[2] Nearest neighbor used.

Table 4. Variance Estimates.*

Data Missing at Random

		(1)	(2 & 3)	(4)	(5)	(6)	(7)
	Observed	Random	CM vs. LS	Random	Old Hot	New	New
Site	Values			+ LS	Deck	Pred+Res	HD Obs
BALT	63.07	71.82	60.34	64.53	61.62	61.90	61.89
CINC	100.63	107.77	94.13	108.93	96.74	100.98	101.00
MILW	51.72	53.71	49.38	55.32	50.81	50.99	50.98
STLU	58.54	102.57	53.58	57.75	57.14	56.14	56.11
SANJ	722.35	734.00	718.71	723.43	726.77	728.36	728.37

Data Missing Proportional to Income

Site	Observed Values	(1) Random	(2 & 3) CM vs. LS	(4) Random + LS	(5) Old Hot Deck	(6) New Pred+Res	(7) New HD Obs
BALT	63.07	59.85	52.29	60.90	54.46	56.16	56.14
CINC	100.63	100.22	88.10	95.79	103.67	93.71	93.72
MILW	51.72	105.26	46.89	53.21	49.44	49.74	49.72
STLU	58.54	58.71	53.49	56.87	54.61	54.94	54.94
SANJ	722.35	85.37	74.21	84.20	84.91	84.84	84.75

*Results in Millions