AN EVALUATION OF ALTERNATIVE IMPUTATION METHODS

Jill M. Montaquila, Westat, Inc., and Chester H. Ponikowski, Bureau of Labor Statistics
Chester H. Ponikowski, BLS, Postal Square Building Suite 3160, 2 Massachusetts Ave. NE,
Washington, DC 20212

KEY WORDS: Item Nonresponse, Hot Deck

I. INTRODUCTION

Imputation is a method of adjusting for missing data. Missing responses to data items is a common problem in sample survey settings. These missing responses often occur because the respondent refuses or is unable to provide data for a particular item or items. Missing data may also result from the interviewer failing to ask for or record the data items, from data entry clerks omitting or mis-keying the data item, or by an editing process that deletes inconsistent data.

In such cases, imputation is often the method of choice for adjusting for item nonresponse. Imputation replaces each missing data item with at least one possible response. The "completed" data set can then be used in subsequent analyses of the data. Kalton and Kasprzyk (1982) point out that imputation has three desirable features: "First ... it aims to reduce biases in survey estimates arising from missing data .... 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 ... are not required. Third, the results obtained from different analyses are bound to be consistent, a feature which not need to apply with an incomplete data set."

Several methods have been proposed for imputing missing item responses (Kalton and Kasprzyk, 1982; Rubin, 1978). Kalton and Kasprzyk (1982) describe a variety of imputation methods that are used and their properties.

This paper describes the Employment Cost Index survey, the establishment survey used to compare the performance of imputation methods (Section II); describes imputation methods studied (Section III); presents empirical analysis and results (Section IV); and proposes issues for further research (Section VI).

II. DESCRIPTION OF EMPLOYMENT COST INDEX SURVEY

The Employment Cost Index (ECI) survey is an establishment survey conducted by the Bureau of Labor Statistics (BLS). The goal of the survey is to produce estimates of the quarter-to-quarter and year-to-year change in the cost of wages, benefits, and total compensation. In addition, level estimates of cost of compensation per employee per hour worked are published annually. All state and local governments and private sector industries, except for farms and private households, are covered in the survey. All employees are covered except the self-employed.

The Universe Database (UDB) serves as the sampling frame for the ECI survey. The UDB is created from State Unemployment Insurance (UI) files of establishments, which are obtained through the cooperation of the individual state agencies.

The ECI sample is selected using a 2-stage stratified design with probability proportional to employment sampling at each stage. The first stage of sample selection is a probability sample of establishments, and the second stage of sample selection is a probability sample of occupations ('hits') within the sampled establishments. For a more detailed description of the ECI survey sample design, refer to the *BLS Handbook of Methods* (Bulletin 2414, September 1992).

The ECI survey collects wage data as well as benefit cost data for 22 benefit items, including health insurance, paid vacations, and pension and retirement. Occasionally, during the quarterly update, responding establishments refuse to provide or are unable to provide wage or benefit cost data for a given occupation. Thus, item nonresponse results. Ignoring the item nonresponse and using only complete data records could result in substantial bias in estimates and incorrect variance estimates.

In our study, we used the data from the December 1993 and March 1994 updates. For the March 1994 quarter, the ECI had a sample of 5,614 establishments which consisted of 25,823 sampled occupational observations. The dataset included auxiliary data from the frame as well as reported data obtained during collection.

III. IMPUTATION METHODS

The imputation methods studied are nearest neighbor within-cell hot-deck, random within-cell hot-deck, regression, and cell mean imputation. These were chosen for our study because they appear to be most commonly used in establishment surveys. For the purpose of this study, we are interested in evaluating methods for imputing missing benefit cost data. The data items subject to missingness are benefit cost levels; however, in order to obtain reasonable imputed levels, we impute the quarter-to-quarter benefit cost change and then apply the previous quarter cost level to obtain the imputed current quarter cost level. In cases where the previous quarter cost level is missing, the average cost level of respondents within a specified cell is imputed.

A. Random Within-Cell Hot-Deck. Imputation classes ("cells") are formed, based on auxiliary data that are known for all units. Within each cell, a unit that is missing the characteristic of interest (i.e., a "recipient") takes the value of the characteristic of a "usable" unit ("donor") that is selected at random without replacement within the same cell. In our application,

$$y^{*}_{iqbt} = y_{iqbt-1} \times r^{rd}_{iqbt}$$

where

y^{*}_{iqbt} = imputed current quarter cost per employee per hour worked of benefit b for occupation (hit) q in cell i

y_{iqbt-1} =previous quarter cost per employee per hour worked of benefit b for hit q in cell i

r^{rd}_{iqbt} = donor's ratio of current quarter to previous quarter cost of benefit b, where the donor is chosen at random within the same cell.

Imputation cells are formed based on the following characteristics: benefit item, ownership, industry (SIC), major occupational group (MOG), an indicator of whether the benefit cost is wage-related (WAGEREL), union status, and region.

An advantage of the random within-cell hot-deck method is that it retains the respondent distribution of quarter to quarter ratio of the characteristic within cell i (Kalton and Kasprzyk, 1982).

B. Nearest Neighbor Within-Cell Hot-Deck. The "unusable" takes the value of the characteristic of the "usable" unit within the same cell that is "nearest" to the unusable, where "nearness" is defined by a pre-specified distance function. In our application,

$$y^{*}_{iqbt} = y_{iqbt-1} \times r^{nn}_{iqbt}$$

where

y^{*}_{iqbt} and y_{iqbt-1} are defined as above

r^{nn}_{iqbt} = ratio of current to previous quarter cost for benefit b in hit q chosen from among all usable hits in cell i such that $\left| y_{iqbt-1_d} - y_{iqbt-1_r} \right|$ is minimized, where y_{iqbt-1_d} and y_{iqbt-1_r} are prior quarter costs of donor and recipient establishments, respectively.

Imputation cells are formed based on the same characteristics as for the random within-cell hot-deck.

This method allows for the use of additional auxiliary information that may be highly correlated with the characteristic of interest in choosing the donor.

C. Cell Mean. The "unusable" takes the mean of the characteristic among all "usables" within the same cell. Currently, the ECI uses the cell mean method to impute missing benefit cost ratios. In our application,

$$y^{*}_{iqbt} = y_{iqbt-1} \times r^{cm}_{iqbt}$$

where

y^{*}_{iqbt} and y_{iqbt-1} are defined as above

$r^{cm}_{iqbt} = (\sum_{q \in R} w_{iq} r_{iqbt}) / \sum_{q \in R} w_{iq}$ is the missing benefit cost ratio for hit q in cell i

w_{iq} = weight applied to hit q in cell i

r_{iqbt} = actual benefit cost ratio for quote q in cell i

R = set of all usable occupational hits in cell i

Imputation cells are formed based on the same characteristics as for the random within-cell hot-deck and nearest neighbor within-cell hot-deck.

Imputing the cell mean results in a spike in the conditional distribution of the characteristic, conditional on the cell-defining auxiliary variables, at the cell mean. This results in an attenuation of covariances (Kalton and Kasprzyk, 1986).

D. Regression Method. Regression methods involve regressing the benefit costs for usable cases on known auxiliary variables and using the estimated regression equation to "predict" values for unusables. West *et al* (1989) considered several regression models in an evaluation of imputation methods for employment data using data from the Current Employment Statistics

(CES) Survey of establishments. For their purposes, they found that a regression method appeared superior to other methods considered.

In our application, the quarter-to-quarter percent benefit cost change (BENRATIO) is regressed on various explanatory variables. The BENRATIO for each "unusable" occupational hit is predicted using the estimated regression equation. Models were fit separately for each benefit item. We selected three models to consider for this study:

Model 1: BENRATIO regressed on main effects for prior quarter benefit cost per hour (PQPERHR), major occupational group (MOG), region (REG), full-time/part-time status (FTPT), and time/incentive (TIMEINC), and all interaction effects involving MOG, REG, FTPT, and TIMEINC.

Model 2: BENRATIO regressed on main effects for PQPERHR, MOG, REG, FTPT, and TIMEINC.

Model 3: BENRATIO regressed on main effects and all interaction effects for MOG, REG, FTPT, and TIMEINC.

We also evaluated a regression imputation with repeated random residuals. For this alternative, we used the imputed values obtained from Model 3, and added a random residual; we repeated this process five times. Each random residual was obtained by taking one random draw from the normal distribution with mean 0 and variance equal to the variance of the (empirical) residuals, i.e., mean square error obtained from the model. A “completed” dataset was obtained from each set of imputed values, yielding five completed datasets. We performed separate analyses on each of the five completed datasets, and then combined the results in order to obtain overall estimates and standard errors, where the standard errors contain estimates of the “between imputation” variance. Our intention here was to see whether there were appreciable gains in reliability of variance estimates by using regression imputation with repeated random residuals.

Several of the imputation methods considered are based on the formation of disjoint imputation cells, and the subsequent collapsing of cells when necessary. We assume that the missing data are missing at random (MAR) within cells, and that there is an ignorable response mechanism within cells. That is, we assume that the conditional distribution of the characteristic of interest for unobserved units (which may or may not have been included in the sample) given the cell-defining auxiliary variables and the observed values is independent of the response mechanisms.

To maintain comparability between methods, the cells and collapse patterns used in this study are the same for each of the three cell-based methods under consideration. Analysis of variance results showed that, among occupations with usable benefit cost data, each of these variables have highly significant main effects on benefit cost levels. Thus, the predictive distribution of benefit cost level given these observed variables should have relatively small variance (Rubin, 1978).

For the random within-cell hot-deck and nearest neighbor within-cell hot-deck, we required that, whenever possible, usables be used at most once in imputing missing benefit costs for unusables, in order to minimize loss in precision that may result from using donors multiple times. If a cell had one or more unusables but no available usables, the cell was collapsed with other similar cells according to the predetermined collapsing pattern until a usable donor was found.

IV. EMPIRICAL ANALYSIS AND RESULTS

To perform the evaluation of the methods being considered, we induced missingness among the complete data cases. We focused on benefit cost level estimates for five different benefit items: Benefit items 02 (Vacations), 05 (Other Leave), 10 (Life Insurance), 11 (Health), and 14 (Pension). Our reason for choosing these benefit items in this study was that benefit cost estimates for these benefits are widely used, and these benefit items have high benefit cost item nonresponse rates relative to most other benefit items, as indicated in Table 1. The figures in Table 1 represent, for each benefit item collected by ECI, the proportion of responding occupations having missing benefit cost data for the March 1994 quarter.

Each estimate is of the following form:

$$\bar{y}_b^* = \sum_{i=1}^{n} \sum_{q=1}^{m_i} (w_{iq} y_{iqbt}^*) / \sum_{i=1}^{n} \sum_{q=1}^{m_i} w_{iq}$$

where

$\bar{y}_b^*$ = weighted estimate of cost per employee per hour worked for benefit b

y_{iqbt}^* = imputed current quarter cost for benefit b in hit q and cell i. This could be a value from any of the imputation methods.

w_{iq} = weight for occupational hit q in establishment i
n = number of establishments in sample
m_i = number of occupational hits selected in establishment i

The estimates and variance estimates were calculated using software for survey data analysis (SUDAAN Release 6.0) for multistage sample designs.

Regardless of the imputation method used, the usual variance estimator will underestimate the variance of $\bar{y}_b^*$, since it does not account for additional variability due to imputation, i.e., "imputation variance." Underestimation of true variance can be a very serious problem when the proportion of missing values for a particular characteristic of interest is high (Rao and Shao, 1992). Several methods have been proposed to account for imputation variance, including multiple imputation, Rao and Shao's adjusted jackknife variance estimator (Rao and Shao, 1992), and Fay's method (Fay, 1993).

Table 2 presents the estimates of employer costs per hour worked for selected employee benefits based on imputed data for observations with missing values for each of the imputation methods we considered. In addition estimates of employer costs based on the actual data for observations with missing values are presented. Standard errors of the corresponding estimates are provided in parentheses.

Table 1. Benefit Cost Item Nonresponse by Benefit Item, March 1994

Benefit Item	Benefit Cost Item Nonresponse Rate (%)
01 (Premium pay)	26.26
02 (Vacations)	**18.76**
03 (Holidays)	14.25
04 (Sick Leave)	26.57
05 (Other Leave)	**37.71**
06 (Shift Differential)	13.58
07 (Nonprod. Bonus)	15.60
08 (Severance Pay)	17.88
09 (Supp. Unemployment)	7.56
10 (Life Insurance)	**18.95**
11 (Health)	**20.69**
12 (Sickness and Accident)	19.62
13 (Long Term Disability)	18.63
14 (Pension)	**14.59**
15 (Social Security)	14.20
16 (Savings & Thrift)	7.22
17 (Railroad Retirement)	7.22
18 (Railroad Supp. Retirement)	7.22
19 (F.U.T.A.)	12.88
20 (S.U.I.)	21.81
21 (Workers' Comp.)	29.52
22 (Other Legally Required)	9.83

The pairwise comparisons of imputation methods were made using paired t-test and no significant differences were found at α=0.05 level. The similarity in these estimates is due in part to the cell definitions. Several auxiliary variables, many having several levels, were used in constructing the cells. This was done in an attempt to achieve homogeneity of benefit cost (item) response propensity within cells, and thus reduce item nonresponse bias.

Comparing estimates based on imputed data for observations with missing values to estimates based on the actual data, using paired t-test, shows for the most part that the differences are not significant at α=0.05 level. The estimates based on the regression approach tend to be slightly closer to estimates based on the actual data. This is due partly to the exclusion of outliers in the regression approach.

Comparing the standard errors shows that the cell mean method tends to have lower standard errors, as expected. Also, as expected, the random within-cell hot-deck and nearest neighbor hot-deck tend to have the highest standard errors.

V. CONCLUSIONS

We compared the performance of several imputation methods in imputing missing item values in the establishment survey. Estimates and standard errors were calculated for each method based on the dataset with missing values and dataset with no missing values, i.e., based on actual values. The results show that the choice of imputation method did not significantly affect the estimates. The similarities among the methods is due in part to the high degree of homogeneity within imputation cells. The variance estimates obtained did not appear to vary much across imputation methods and as expected the random within-cell hot-deck and nearest neighbor hot-deck tend to have the highest standard errors.

Also, estimates and standard errors were calculated from each of the five completed datasets obtained using the repeated regression with random residual method. In most cases, there were no differences at all among the estimates (carried to one tenth of a cent). This suggests that the between imputation variability is negligible. Thus, this indicates that no appreciable gain can be expected from doing multiple imputations.

There are other issues to consider in determining which imputation method should be used for a particular application. There are several practical issues that involve the ease of implementation, such as ease of programming, amount of collapsing, and cost of executing. For our particular implementation, all three cell-based methods appeared to be relatively equivalent in their difficulty to program; due to procedures available in SAS, the regression method was far simpler to implement. The nearest neighbor hot-deck and random hot-deck methods required more collapsing of cells than the cell mean method, since those methods attempted to use each usable benefit cost ratio at most once as a donor. The regression method was the least costly to implement in our case, mostly due to the fact that this method involved no explicit collapsing.

VI. ISSUES FOR FURTHER RESEARCH

In this study, we have compared four imputation methods commonly used in establishment surveys. However, there are other methods that are currently being used. For example, multiple imputation methods (Rubin, 1978) involve independently imputing J>1 values for each missing value. That is, for each missing benefit cost ratio, J benefit cost ratios would be drawn with replacement from the predictive distribution of the benefit cost ratios, given the observed values of the benefit cost ratios. This method enables the analyst to obtain valid variance estimates by incorporating into the variance estimate an estimate of the imputation variance.

Also, Rao and Shao (1992) have proposed an adjusted jackknife variance estimator for use with the random hot-deck imputation method. This variance estimator is said to be asymptotically unbiased. Fay (1993) developed a model-assisted approach for obtaining valid variance estimators when the random hot-deck is used.

We would like to test these other methods on the same dataset and compare the results with those presented in this paper. We would also like to compare multiple imputation and the Rao-Shao jackknife. However, since the between imputation variance appears to be negligible in this case, one would not expect to see different results when multiple imputation and the Rao-Shao jackknife are used.

Finally, we should note that the similarities among the methods is likely to be due mostly to the high degree of homogeneity within imputation cells. Under the ignorable nonresponse model, the homogeneity should reduce the bias due to item nonresponse; however, there is the potential for an increase in variance. In the future, we would like to investigate the gains in precision that might be attained by constructing less homogeneous imputation cells.

REFERENCES

BLS Handbook of Methods (Bulletin 2414,September 1992), Washington, D.C.: Bureau of Labor Statistics, 67-77.

Fay, R.E. (1992), "When Are Inferences from Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods,* Washington, D.C.: American Statistical Association.

_______ (1993), "Valid Inferences From Imputed Survey Data," *Proceedings of the Section on Survey Research Methods,* Washington, D.C.: American Statistical Association.

Kalton, G., and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 22-31.

Kalton, G., and Kish, L. (1981), "Two Efficient Random Imputation Procedures," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 146-151.

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

Rao, J.N.K., and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika,* 79, 811-822.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 20-34.

______ (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley & Sons, Inc.

Shah, B.V., Barnwell, B.G., Hunt, P.N., and LaVange, L.M. (1992), *SUDAAN User's Manual, Release 6.0,* Research Triangle Park, N.C.: Research Triangle Institute.

West, S.A., Butani, S., Witt, M., and Adkins, C. (1989), "Alternate Imputation Methods for Employment Data," in *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 227-232.

Table 2. Estimates of Employer Costs for Employee Benefits for Imputation Methods Considered

	NNHD	RAN	CM	REG1	REG2	REG3	"ACT"
Benefit Item 02							
Goods-Producing	0.917 (0.058)	0.953 (0.064)	0.947 (0.058)	0.979 (0.065)	0.977 (0.064)	0.983 (0.065)	0.962 (0.062)
Sales Occupations	0.410 (0.041)	0.419 (0.042)	0.416 (0.041)	0.417 (0.042)	0.417 (0.042)	0.385 (0.026)	0.420 (0.041)
Retail	0.305 (0.031)	0.310 (0.031)	0.307 (0.031)	0.309 (0.031)	0.308 (0.031)	0.292 (0.026)	0.309 (0.031)
Service Prod. Ind.: WC Occs: Ad Supp	0.596 (0.029)	0.596 (0.029)	0.598 (0.029)	0.594 (0.028)	0.594 (0.028)	0.595 (0.028)	0.595 (0.029)
Benefit Item 05							
Goods-Producing	0.091 (0.008)	0.091 (0.008)	0.091 (0.007)	0.090 (0.006)	0.090 (0.006)	0.090 (0.006)	0.087 (0.005)
Sales Occupations	0.078 (0.015)	0.075 (0.013)	0.071 (0.012)	0.075 (0.013)	0.075 (0.013)	0.075 (0.013)	0.074 (0.013)
Retail	0.044 (0.010)	0.047 (0.010)	0.040 (0.008)	0.048 (0.010)	0.048 (0.010)	0.048 (0.010)	0.047 (0.010)
Service Prod. Ind: WC Occs: Ad Supp	0.104 (0.010)	0.104 (0.010)	0.104 (0.010)	0.104 (0.010)	0.104 (0.010)	0.104 (0.010)	0.105 (0.010)
Benefit Item 10							
Goods-Producing	0.081 (0.004)	0.081 (0.004)	0.081 (0.004)	0.083 (0.004)	0.082 (0.004)	0.082 (0.004)	0.081 (0.004)
Sales Occupations	0.047 (0.005)	0.046 (0.004)	0.046 (0.004)	0.042 (0.004)	0.042 (0.004)	0.044 (0.004)	0.044 (0.004)
Retail	0.031 (0.003)	0.033 (0.004)	0.032 (0.003)	0.028 (0.003)	0.028 (0.003)	0.029 (0.003)	0.030 (0.003)
Service Prod. Ind: WC Occs: Ad Supp	0.048 (0.003)	0.048 (0.003)	0.049 (0.003)	0.049 (0.003)	0.049 (0.003)	0.049 (0.003)	0.050 (0.003)
Benefit Item 11							
Goods-Producing	2.210 (0.094)	2.242 (0.095)	2.208 (0.093)	2.184 (0.088)	2.180 (0.088)	2.205 (0.090)	2.236 (0.098)
Sales Occupations	1.005 (0.062)	1.032 (0.064)	1.010 (0.061)	1.007 (0.063)	1.007 (0.063)	1.012 (0.063)	1.010 (0.063)
Retail	0.994 (0.096)	1.009 (0.094)	0.962 (0.074)	0.992 (0.092)	0.991 (0.091)	1.004 (0.093)	0.969 (0.079)
Service Prod. Ind.: WC Occs: Ad Supp	1.612 (0.058)	1.620 (0.060)	1.658 (0.060)	1.608 (0.057)	1.611 (0.057)	1.621 (0.056)	1.598 (0.055)
Benefit Item 14							
Goods-Producing	0.462 (0.045)	0.462 (0.050)	0.439 (0.044)	0.453 (0.044)	0.452 (0.044)	0.452 (0.044)	0.434 (0.044)
Sales Occupations	0.191 (0.031)	0.189 (0.029)	0.191 (0.029)	0.187 (0.029)	0.188 (0.029)	0.195 (0.029)	0.204 (0.030)
Retail	0.117 (0.023)	0.121 (0.023)	0.114 (0.021)	0.111 (0.021)	0.111 (0.021)	0.114 (0.022)	0.115 (0.021)
Service Prod. Ind.: WC Occs: Ad Supp	0.339 (0.032)	0.326 (0.030)	0.338 (0.030)	0.335 (0.030)	0.336 (0.030)	0.338 (0.031)	0.348 (0.033)

The figures in parentheses are standard errors of the corresponding estimates.

Note: NNHD = the nearest neighbor hot-deck method
RAN = the random within-cell hot-deck method
CM = the cell mean method
REG1 = the regression method model 1
REG2 = the regression method model 2
REG3 = the regression method model 3
ACT = estimates based on the actual data