ALTERNATIVE IMPUTATION METHODS FOR EMPLOYMENT DATA

Sandra West, Shail Butani, Michael Witt, Craig Adkins
Sandra West, Bureau of Labor Statistics,
441 G St. NW, Room 2127, Washington, DC 20212

KEYWORDS: Regression, Bayesian Model, Multiple Imputation, Hot Deck

1. Introduction

In this paper the results of an empirical investigation of different imputation methods for employment data are presented. The investigation began in connection with a revision project for the Bureau of Labor Statistics (BLS) program that maintains the BLS Universe Data Base (UDB). The UDB is a sampling frame of business establishments that is constructed from the State's ES-202 microdata file. The information used to maintain this file is obtained from quarterly unemployment insurance (UI) reports which each covered employer is required to submit. These quarterly reports contain, among other things, information on employment for each month of the quarter as well as a standard industrial classification (SIC) code for the establishment. Although the filing of the contribution report is mandatory under the current UI laws, each quarter there are always some reports that are filed late, delinquent accounts, as well as returns with partial data.

The goal of this project was to develop a single imputation procedure that would work reasonably well for all SIC groups within each State. The main objective of this investigation was to compare the current ES-202 method for imputing establishment employment data with alternative procedures based on regression models.

The employment data used in this study are discussed in Section 2. This section also includes a discussion of whether or not the nonrespondents are missing at random. Section 3 presents the notation used in this paper and the evaluation criteria that are used to compare the various imputation methods. Section 4 provides a description of the ES-202 Method of imputation, two hot deck procedures, and the mean imputation procedure. In Section 5, eight regression models for imputing employment are presented. One problem with a "best" regression-based prediction method is that all imputed values will fall on the estimated regression line and therefore, will lead to biases in estimates that involve the residual variance for nonrespondents. Simple methods that attend to this problem draw random residuals which are added to the model predictions. Details of such methods are given in Section 6. In Section 7, imputations are created under an explicit Bayesian model and multiple imputations are developed in Section 8. In a multiple imputation context, several imputed values would be created for each missing value, where ideally, uncertainty due to the estimation of the regression itself would be reflected across the imputations. Section 9 compares the results from the various imputation methods and summarizes the findings of this study.

2. Data

The purpose of this project was to develop a methodology to impute missing employment values for the ES-202 microdata file. Due to various reasons, it was not possible for any State to provide ES-202 microdata of the type needed. Consequently, an alternative data source, the Current Employment Statistics (CES) Survey of establishments, conducted monthly by BLS, was used for this study. The CES Survey, among other things, provides information on the monthly employment, SIC, and the closing for each establishment. The closing indicates the time frame in which the establishment responded to the survey in relation to the reference week, which is the calendar week that includes the twelfth day of the month. The first, second, and third closings normally fall, respectively, on the second, fifth, and eighth Friday following the reference week.

Most imputation procedures that are used and developed in survey sampling assume the missing data mechanism is ignorable (Little and Rubin, 1987). This issue was examined with mixed results for employment data on the CES database. Three industries were chosen and a comparison was made between those units that reported data in first or second closing (these are the respondents for this study) against those units that reported data in third closing (these are the nonrespondents for this study). The results show that there is not a significant difference in mean employment between the respondents and nonrespondents in SIC 373, but there is a difference in SICs 508 and 121. (For a definition of SICs, see Table I). Although this finding contradicts the underlying assumption of an ignorable response mechanism that is required for most of the imputation procedures examined in this paper, it does not necessarily imply that these imputation procedures are inappropriate for imputing employment values. The effectiveness of any given method is evaluated by two error measures which are discussed in the next section. Perhaps, the models could be further improved by modeling the nonresponse mechanism; this work is left for a future study.

3. Notation and Evaluation Criteria

The imputation procedures will be applied to predict the nonrespondents, by SIC group, employment size class and by month. The twelve month period ranging from November 1987 to October 1988 was considered. One, three and eight size class partitions were constructed to examine the size class effect, if any (see Table I). SIC groups 121, 373 and 508 were studied but due to the limitation of space, results are presented only for SICs 121 and 373.

Let the indices:

t = month
i = establishment
m = imputation procedure.

Let the variables:

$ES_{t,i}$ = Establishment i, in month t

$S_{12,t,m}$ = Set of establishments that responded by second closing, reported a value for month (t-1), and are in the domain of procedure m

$S_{3,t,m}$ = Set of establishments that responded in third closing, reported a value for month (t-1), and are in the domain of procedure m

$Y_{t,i}$ = Reported employment of $ES_{t,i}$

$Y^p_{t,i,m}$ = Predicted employment of $ES_{t,i}$

$N_{t,m}$ = Number of units in $S_{12,t,m}$

$N^p_{t,m}$ = Number of units in $S_{3,t,m}$

$E_{t,i,m}$ = Error in the prediction = $(Y^p_{t,i,m} - Y_{t,i})$

$AE_{t,i,m}$ = Absolute error in the prediction
= $|Y^p_{t,i,m} - Y_{t,i}|$

Evaluation Criteria

a. Mean Unit Error:

$$ME_m = \sum_{\substack{\text{size}\\ \text{class}}} \sum_t \sum_i E_{t,i,m} \Big/ \sum_{\substack{\text{size}\\ \text{class}}} \sum_t N^p_{t,m}$$

b. Mean Unit Absolute Error:

$$MAE_m = \sum_{\substack{\text{size}\\ \text{class}}} \sum_t \sum_i AE_{t,i,m} \Big/ \sum_{\substack{\text{size}\\ \text{class}}} \sum_t N^p_{t,m}$$

Note that ME_m represents a macro level statistic that indicates the effect that the imputation procedure has on total employment, while MAE_m is a micro level statistic that indicates the effect on the unit. The corresponding relative errors were also computed but are not presented in this paper due to space.

4. ES-202 Imputation Procedure and Other Standard Methods

ES: ES-202 Method of Imputation

Under this method, each nonrespondent's employment is imputed using its own history. The predicted value is therefore independent of size class and industry. It is computed as follows:

If $Y_{t-13,i}$, $Y_{t-12,i}$, $Y_{t-1,i}$ are nonmissing and $Y_{t-13,i} > 0$ then

$$Y^p_{t,i,ES} = (Y_{t-1,i})(Y_{t-12,i}) \Big/ (Y_{t-13,i}) .$$

Otherwise, if $Y_{t-13,i}$ or $Y_{t-12,i}$ are missing then $Y^p_{t,i,ES}$ is set equal to the most current, nonmissing $Y_{t-T,i}$ for $1 \le T \le 6$.

Otherwise, a predicted value was not computed for $ES_{t,i}$.

This method of imputation is based on the assumption that the current monthly change in employment for the unit, will be approximately the same as it was 12 months prior to the current month. An advantage to using this method is that it incorporates the seasonality of the unit's reported employment into the predicted value, although it does not take into account any nonseasonal industry drifts in employment. Another disadvantage to this method is that it will incorporate an atypical change in employment that the establishment experienced a year ago.

MN: Mean Imputation Method

The mean imputation method is a common method of imputation in many surveys, especially for those surveys with a high response rate, because it will not alter an estimate of the stratum mean if applied to each of the original sampling stratum. If the response rate is low for a survey, then this method of imputation would not be desirable because it adversely affects the distribution of the sample units by skewing the distribution toward the mean. The mean imputation method was applied as follows.

For any fixed SIC group, employment size class and month t and for all $ES_{t,i} \in S_{3,t,MN}$

$$Y^p_{t,i,MN} = \sum_{(ES_{t,i} \in S_{12,t,MN})} Y_{t,i} \Big/ N_{t,MN} .$$

Thus $Y^p_{t,i,MN}$ is equal to the average employment of the respondents in the stratum.

HD1: Hot Deck Imputation Method - Random Selection

For any fixed SIC group, employment size class and month t:

$$Y^p_{t,i,HD1} = Y^*_{t,j}$$

where $Y^*_{t,j}$ is the employment of a randomly selected respondent from $S_{12,t,HD1}$. Selection was done independently within strata and with replacement.

HD2: Hot Deck Imputation Method - Nearest Neighbor

The Nearest Neighbor hot deck method is desirable because for any particular nonrespondent, it selects the respondent that appears closest to the nonrespondent in an ordered list, and substitutes the respondent's employment value for the nonrespondent's. As with the ES-202 method, this method is independent of employment size class.

Within any fixed SIC group and for each month t, all establishments that reported employment in closing 1, 2 or 3 in month t, were ordered by $Y_{t-1,i}$ by $Y_{t-2,i}$ by state. For this ordering procedure, missing values for $Y_{t-1,i}$ and $Y_{t-2,i}$, were considered -1.

For all $ES_{t,j} \in S_{3,t,HD2}$, let $Y^{(1)}_{t,i}$ be the employment value for the first establishment $ES^{(1)}_{t,i} \in S_{12,t,HD2}$ that precedes $ES_{t,j}$ on the ordered list and $Y^{(2)}_{t,k}$ be the employment value for the first establishment $ES^{(2)}_{t,k} \in S_{12,t,HD2}$ that succeeds $ES_{t,j}$ on the ordered list. If

$$|Y^{(1)}_{t-1,i} - Y_{t-1,j}| \le |Y^{(2)}_{t-1,k} - Y_{t-1,j}|$$

then $Y^p_{t,j,HD2}$ is set equal to $Y^{(1)}_{t,i}$. Otherwise, $Y^p_{t,j,HD2}$ is set equal to $Y^{(2)}_{t,k}$.

5. Modeling Employment by Regression

A common method for imputing missing values is via least squares regression (Afifi and Elaskoff, 1969). The following section discusses regression models for employment.

Regression Models

In two papers on estimators for total employment (West 1982, 1983), it was discovered that the most promising

models for employment were the proportional regression models. These models specify that the expected employment for establishment i in the t^{th} month, given the following vector of y - values for month t-1:

$$\underline{Y}_{t-1} = [Y_{t-1,1}, Y_{t-1,2}, \ldots Y_{t-1,n}]$$

is proportional to establishment i's previous month's employment, $Y_{t-1,i}$. That is,

$$E(Y_{t,i} \mid \underline{Y}_{t-1} = \underline{y}_{t-1}) = \beta y_{t-1,i}$$

where β is some constant depending on t.

It was further assumed that the y's are conditionally uncorrelated. That is,

$$\text{cov}(Y_{t,i}, Y_{t,j} \mid \underline{Y}_{t-1} = \underline{y}_{t-1}) = \begin{cases} \nu_{t,i} & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases}$$

where $\nu_{t,i}$ represents the conditional variance of $Y_{t,i}$ which in general will depend on $Y_{t-1,i}$. Choosing a specific simple function to represent the variance $\nu_{t,i}$ accurately is difficult. Fortunately, knowledge of the precise form of $\nu_{t,i}$ is not essential, (see Royal, 1978).

The model can be rewritten as:

$$Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$$

where $E\{\varepsilon_{t,i}\} = 0$, and

$$E\{\varepsilon_{t,i}, \varepsilon_{t,j}\} = \begin{cases} \nu_{t,i} & \text{if } i = j \\ 0 & \text{Otherwise} \end{cases}$$

In West (1982), $\nu_{t,i} = \sigma^2 Y_{t-1,i}$ and $\nu_{t,i} = \sigma^2$ were considered. The model extended to two independent variables was also considered in that paper and it was found that the additional variable, Y_{t-2}, in the model was not necessary.

For the current CES data set, the following eight models were considered.

Models 1 - 4 assume $\nu_{t,i} = \sigma^2$:

Model 1: $Y_{t,i} = \alpha + \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 2: $Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 3: $\text{Ln}(Y_{t,i}) = \alpha + \beta \text{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Model 4: $\text{Ln}(Y_{t,i}) = \beta \text{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Models 5 - 8 are similar to models 1 - 4 respectively, except it is now assumed that $\nu_{t,i} = \sigma^2 Y_{t-1,i}$ for models 5 and 6, and $\nu_{t,i} = \sigma^2 \text{Ln}(Y_{t-1,i})$ for models 7 and 8:

Model 5: $Y_{t,i} = \alpha + \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 6: $Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i}$

Model 7: $\text{Ln}(Y_{t,i}) = \alpha + \beta \text{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Model 8: $\text{Ln}(Y_{t,i}) = \beta \text{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$

Let the indice m = RMr = Regression Model r, r = 1,...,8. Then the regression model parameters were estimated using the establishments in the corresponding set $S_{12,t,m}$ and an imputed value was calculated for those establishments in the set $S_{3,t,m}$. For clarity, the subscripts t and m were not used in conjunction with the parameters σ, α and β.

Models were fitted for the three SIC groups, twelve months of data, and three types of sample designs (1, 3 and 8 employment size classes). Based on R-squared values and other analyses, it was decided to omit models 1, 3, 5 and 7 from consideration.

Example Using Model 6

From model 6:

$$Y_{t,i} = \beta Y_{t-1,i} + \varepsilon_{t,i} \qquad \text{with } \nu_{t,i} = \sigma^2 Y_{t-1,i}$$

and β is estimated as:

$$\beta^p = \sum_{i \in S_{12,t,RM6}} Y_{t,i} \Big/ \sum_{i \in S_{12,t,RM6}} Y_{t-1,i} .$$

For any establishment j in $S_{3,t,RM6}$, the establishment's predicted employment value at time t is:

$$Y^p_{t,j,RM6} = \beta^p Y_{t-1,j} .$$

Adjustments for Models 4 and 8

Considering models 4 and 8, if it is assumed that $\varepsilon_{t,i}$ is normally distributed then $Y_{t,i}$ has a lognormal distribution with

$$\text{Mean: } \exp\{\beta \text{Ln}(Y_{t-1,i}) + .5\text{Var}(\varepsilon_{t,i})\}$$

$$\text{Variance: } \{\exp[\text{Var}(\varepsilon_{t,i})]-1\} \times \exp\{2\beta \text{Ln}(Y_{t-1,i}) + \text{Var}(\varepsilon_{t,i})\}.$$

Therefore, an unbiased estimator of $Y_{t,k}$ is:

$$\exp\{\beta \text{Ln}(Y_{t-1,k}) + .5\text{Var}(\varepsilon_{t,k})\} .$$

As an estimate of $\text{Var}(\varepsilon_{t,k})$, the residual mean square error, MSE, from the regression was used, and the first adjustments to the regression models (A1RMr, r = 4 and 8) are:

$$Y^p_{t,j,A1RMr} = \exp\{\beta^p_r \text{Ln}(Y_{t-1,j}) + .5\text{MSE}_r\}$$

Let $Z_{t-1,i} = \text{Ln}(Y_{t-1,i})$ then

$$\beta^p_4 = \sum_i Z_{t-1,i} Z_{t,i} \Big/ \sum_i Z^2_{t-1,i}$$

$$\beta^p_8 = \sum_i Z_{t,i} \Big/ \sum_i Z_{t-1,i} .$$

A second alternative adjustment to the logarithmic regression models, used by David (1986), led to the following unbiased prediction of $Y_{t,k}$:

$$\exp\{\beta^p Z_{t-1,k} + .5[\text{Var}(\varepsilon_{t,k}) + Z^2_{t-1,k}]\text{Var}(\beta^p)\}.$$

For models 4 and 8:

$$Y^{p}_{t,j,A2RMr} = \exp\{\beta^{p}_{r} Z_{t-1,j} + .5(MSE_r)(EMP_r)\}$$

where r = 4 and 8, $Z_{t-1,i}$ and β^{p}_{r} are defined as above, and

$$EMP_4 = 1 - \{Z^2_{t-1,j} / \sum_i Z^2_{t-1,i}\}$$

$$EMP_8 = 1 - \{Z_{t-1,j} / \sum_i Z_{t-1,i}\} .$$

6. Adding Residuals to the Regression Models

The methods discussed in the previous section could be thought of as imputing for missing employment by using the mean of the predicted Y_t distribution, conditional on the predictors, Y_{t-1}. As a result, the distribution of the imputed values has a smaller variance than the distribution of the true values, even if the assumptions of the model are valid. A simple strategy of adjusting for this problem is to add random errors to the predictive means, that is, draw residuals r_k, with mean zero, to add to $Y^{p}_{t,j,RMr}$.

In this project, it was decided to consider this imputation procedure with the residuals, r_i, equalling:

1. A random normal deviate using model r (m=RNDMr).
2. A randomly selected respondent's residual using model r (m=RSRMr).
3. A randomly selected respondent's residual using model r from redefined strata (m=NSRMr). Within each SIC group, all establishments (respondents and nonrespondents), were restratified by $Y^{p}_{t,i,RMr}$ using the same employment size class definitions depicted in Table I. A respondent's residual r_i was then randomly selected from the newly formed strata.

For each of the four models, residuals were added to the model predictions by the above three methods. For example, using model 6 and the first method described above, a prediction of $Y_{t,j}$ is:

$$Y^{p}_{t,j,RNDM6} = \beta^{p} Y_{t-1,j} + s\delta_j \qquad (6.1)$$

where δ_j is a random number from a $\mathcal{N}(0,1)$ distribution and s^2 is equal to the mean square error of the regression.

Alternatively, using the second or third method described above:

$$Y^{p}_{t,j,m} = \beta^{p} Y_{t-1,j} + r_k$$

where r_k is the residual from a randomly selected respondent k from the original employment stratum (m = RSRM6) or from the redefined employment stratum (m = NSRM6).

7. Bayesian Model

In creating imputed values under an explicit Bayesian model, three formal tasks can be defined: modeling, estimation and imputation. The modeling task chooses a specific model for the data. The estimation task formulates the posterior distribution of the parameters of that model so that a random draw can be made from it. The imputation task takes one random draw from the posterior distribution of y missing, denoted by $Y_{t,BAY}$, by first drawing a parameter from the posterior distribution obtained in the estimation task and then drawing $Y_{t,BAY}$ from its conditional posterior distribution given the drawn value of the parameter.

For the modeling task, consider model 2 and $Y_{t,i}$ having a $\mathcal{N}(\beta Y_{t-1,i}, \sigma^2)$ distribution. This is the specification for the conditional density $f(Y_{t,i} \mid Y_{t-1,i}, \theta)$ where $\theta = (\beta, \sigma)$. In order to complete the modeling task, the conventional improper prior for θ, Prob(θ) proportional to a constant, is assumed.

For the estimation task, the posterior distribution of θ is needed. Standard Bayesian calculations show that:

$$f(\sigma^2 \mid Y_{t,i}) = (\sigma^{p}_{1})^2[n-1] / \chi^2_{n-1}$$

$$f(\beta \mid \sigma^2) = \mathcal{N}(\beta^{p}_{1}, \sigma^2 v)$$

where

$$(\sigma^{p}_{1})^2 = \sum_i \{Y_{t,i} - \beta^{p}_{1} Y_{t-1,i}\}^2 / (n-1) = MSE$$

$$\beta^{p}_{1} = \sum_i Y_{t,i} Y_{t-1,i} / \sum_i Y^2_{t-1,i}$$

$$v = 1 / \sum_i Y^2_{t-1,i}$$

n = number of respondents.

Since the posterior distribution of θ is in terms of standard distributions, random draws can easily be computed.

The imputation task for this model is as follows:

1. Estimate σ^2 by a χ^2_{n-1} random variable, say h, and let
$$\sigma^2_2 = (\sigma^{p}_{1})^2 (n-1)(h)^{-1}$$
2. Estimate β by drawing one independent $\mathcal{N}(0,1)$ variate, say Z_o, and let
$$\beta_2 = \beta^{p}_{1} + \sigma_2 (v)^{.5} (Z_o)$$
3. Let n_o be the number of values that are missing, that is, the size of $S_{3,t,BAY}$. Draw n_o values of $Y_{t,BAY}$ as
$$Y^{p}_{t,k,BAY} = \beta_2 Y_{t-1,k} + \sigma_2 Z_k \qquad (7.1)$$
where the n_o normal deviates, Z_k are drawn independently.

Equation (7.1) can be rewritten as:

$$Y^{p}_{t,k,BAY} = \beta^{p}_{1} Y_{t-1,k} + \frac{(MSE)^{.5}(n-1)^{.5}}{(h)^{.5}} [(v)^{.5} Z_o Y_{t-1,k} + Z_k].$$

For model 6 an analagous Bayesian argument can be used to compute a $Y^{p}_{t,k,BAY}$. The result will be similar, except in this case:

$$\beta^{p}_{1} = \sum_i Y_{t,i} / \sum_i Y_{t-1,i} \quad \text{and}$$

$$v = [\Sigma_i Y_{t-1,i}]^{-1} .$$

8. Multiple Imputation

Multiple imputation is the technique that replaces each missing value with two or more acceptable values from a distribution of possibilities. The idea was originally proposed by Rubin. The main disadvantage that multiple imputation overcomes is that the resultant imputed values will account for sampling variability associated with the particular nonresponse model.

Multiple imputation can be obtained from the Bayesian Method by repeating the above three steps. Five repeated independent imputations were obtained by repeating the three steps. The average of these five values was taken as the imputed value.

Multiple imputation could also be obtained by using equation (6.1), adding $\mathcal{N}(0,s^2)$ residuals to the predictive mean. The error measures associated with using the average of five such repeated imputations were also considered.

9. Comparison of Imputation Methods and Conclusions

Mean Error (ME) and Mean Absolute Error (MAE) measures were generated for the three SIC groups, each imputation method and each size class combination. However, due to space limitations, Table II presents results only for SICs 121 and 373.

Intuitively, it would seem that by increasing the number of size classes, greater homogeneity would be obtained and thus smaller errors would result. The data, however, showed that little or no gain in accuracy was obtained by increasing the number of size classes. This was perhaps due to the smaller number of observations within each stratum. Also, the imputation technique chosen is to be implemented for the ES-202 microdata at the state level, as opposed to the national level, such as the CES data used for this paper. This means that many state/SIC cells will have only a small number of observations. It is therefore recommended that regardless of which imputation technique is chosen, it should be employed with no more than three size classes.

Since the error measures for many of the imputation methods differ by only .01, it is very difficult to say that a Mean Error (ME) of .01 is superior to an ME of .02. While some methods, such as the Mean Imputation, can be eliminated as being the "best" imputation method, the data show that there is no one method that always yields the smallest error measures. Consequently, it was decided to search for a method that performed well on both measures and for each SIC group. As a starting point, the 96 methods, (the 32 imputation methods considered in this paper with the 3 different size class partitions) were ranked according to MAE and ME, and the top ten in each category were investigated.

For SIC 121, there were four methods that were in the top ten in both categories; three of these four involved model 6. For SIC 373, there is no method that is among the top ten in both ME and MAE. For SIC 508, the three methods that are among the top methods in ME and MAE involve logarithmic models. Next the top ten methods were examined across SIC groups for MAE and ME. According to MAE there were three methods in the top ten of each SIC. Multiple Imputation, Bayesian Model 6; Multiple Imputation, Random Normal Residual Model 6; and Regression Model 6. With respect to ME, there was no intersection of methods.

Noting the robustness of model 6, and the simplicity and intuitive appeal of Regression Model 6, it is recommended that Regression Model 6 with one size class be used.

Future work will include applying some of these methods to the ES-202 microdata and modeling the nonrespondents. A Generalized Bayesian procedure for multiple imputations using belief functions will be developed. Also a study of estimators for total employment with a nonresponse procedure will be done.

References

1. David, M., Little, R., Samuel, M. and Triest, R., (1986), "Alternative Methods for CPS Income Imputation", Journal of the American Statistical Association, vol. 81, pp. 29-41.

2. Little, R. J. A. and Rubin, D. B., (1987), Statistical Analysis With Missing Data, John Wiley & Sons Inc., New York.

3. Royall, R. M. and Cumberland, W. G., (1978), "Variance Estimation in Finite Population Sampling", Journal of the American Statistical Association, vol. 73, pp. 351-358.

4. Rubin, D., (1987), Multiple Imputation for Nonresponse in Surveys, John Wiley and Sons Inc., New York.

5. West, S. A., (1982), "Linear Models for Monthly All Employment Data", Bureau of Labor Statistics Report.

6. West, S. A., (1983), "A Comparison of Different Ratio and Regression Type Estimators for the Total of a Finite Population", ASA Proceedings of the Section in Survey Research Methods.

7. West, S., Butani, S., Witt, M., Adkins, C., (1989), "Alternate Imputation Methods for Employment Data", US Bureau of Labor Statistics Report.

Table I: SIC Group and Employment Size Class Definitions

Employment Size Class Definitions

Size class is determined by the establishment's first nonmissing employment during the time period: October 1987 to October 1988. The definition of one, three and eight size classes are as follows (table entries indicate number of employees):

ONE	THREE	EIGHT	
0 and above	0 - 49	0 - 9	100 - 249
	50 - 249	10 - 19	250 - 499
	250 and above	20 - 49	500 - 999
		50 - 99	1000 and above

SIC Group Definitions

1972 SIC Code	Industry
121	Bituminous Coal and Lignite Mining
373	Ship and Boat Building and Repairing
508	Machinery, Equipment and Supplies

TABLE II: Error Measures for SICs 121 and 373

Imputation Method	SIC 121						SIC 373					
	Number of Employment Sizes						Number of Employment Sizes					
	1		3		8		1		3		8	
	ME	MAE	ME	MAE	ME	MAE	ME	MAE	ME	MAE	ME	MAE
ES 202 Method	6.6	17.4	6.6	17.4	6.6	17.4	-4.0	24.6	-4.0	24.6	-4.0	24.6
Mean	-78	206	-6.1	119	12.3	55.7	49.7	679	91.2	588	-14	275
Hot Deck:												
Rand Selection	-60	266	-3.1	161	14.5	75.4	-94	685	93.5	684	39.2	318
Near Neighbor	-2.6	17.7	-2.6	17.7	-2.6	17.7	-48	75.8	-48	75.8	-48	75.8
Reg Method												
Model 2	-.0	8.6	.0	8.7	-.2	9.0	-3.3	15.8	-3.3	15.8	-4.0	17.2
Model 4	-3.3	10.5	-1.7	9.6	-.5	9.6	-5.0	19.2	-7.5	21.9	-3.3	17.4
Model 6	-.1	8.5	-.1	8.6	.2	9.0	-3.2	16.0	-3.2	16.2	-2.7	16.6
Model 8	-3.8	10.9	-1.7	9.7	-.4	9.7	-2.5	22.1	-7.3	23.1	-3.0	17.6
Adjust Equals (.5) (MSE)												
Model 4	1.4	8.6	.6	8.7	1.2	9.4	3.1	19.5	-2.0	20.7	-1.1	17.4
Model 8	-1.8	9.8	-1.2	9.4	-.0	9.6	.6	21.8	-6.2	22.9	-2.5	17.6
Adjust Equals (.5) (MSE) (EMP)												
Model 4	1.1	8.6	.4	8.7	1.0	9.3	.1	18.9	-4.2	21.2	-1.9	17.4
Model 8	-1.9	9.8	-1.3	9.4	-.1	9.6	-.6	22.0	-6.6	22.9	-2.7	17.6
Rand Generate Normal Resid												
Model 2	-.4	18.3	.4	16.9	-.4	16.0	-4.9	61.1	-5.7	39.2	-1.6	30.8
Model 4	-1.2	37.9	-3.0	25.2	.2	18.2	20.9	70.8	5.1	57.4	-.6	37.2
Model 6	-.1	8.8	-.2	8.9	.1	9.1	-3.2	16.4	-3.2	16.5	-2.6	16.6
Model 8	-3.4	25.8	-1.0	15.4	.4	12.4	15.6	56.6	-6.8	42.7	-1.3	21.7
Rand Sel Resid												
Model 2	.4	11.4	.4	12.7	.1	11.4	-7.2	26.8	-3.1	26.2	-.8	23.1
Model 4	1.9	19.3	1.8	14.9	1.5	15.0	16.6	63.5	-2.0	29.6	-4.4	27.9
Model 6	.3	12.1	1.6	11.3	1.1	10.3	-2.6	27.1	-4.3	28.1	-3.0	22.3
Model 8	2.3	21.2	-2.2	13.7	-.8	11.8	-7.1	37.7	1.1	33.4	-2.4	26.2
Rand Sel Resid After Restrat												
Model 2	.4	11.4	1.0	11.8	1.7	10.6	-7.2	26.8	-2.1	28.8	-.7	24.4
Model 4	1.9	19.3	.5	11.5	-2.5	13.5	16.6	63.5	4.8	32.1	-13	33.0
Model 6	.3	12.1	.4	11.0	.4	11.5	-2.6	27.1	-3.9	23.4	-.3	25.4
Model 8	2.3	21.2	-1.7	13.1	-.7	11.6	-7.1	37.7	-29	53.4	3.6	32.6
Bayes Model												
Model 2	.3	17.6	-.4	16.8	-1.4	17.5	-6.2	36.5	-1.2	34.9	-2.9	35.1
Model 6	-.3	8.8	-.6	9.4	-.3	9.6	-2.6	16.3	-2.7	17.5	-2.4	19.3
Mult Imputat Bayes Model												
Model 2	-1.2	13.9	-1.7	23.1	2.3	27.8	7.9	57.3	4.0	69.6	39.4	111
Model 6	-.2	8.7	.3	8.7	-.1	9.2	-2.5	16.3	-4.7	16.7	-2.1	17.6
Mult Imp Rand Gen Norm Resid												
Model 2	-.2	12.5	-.3	11.3	-1.0	10.8	-.0	34.6	-.1	24.6	-3.8	22.7
Model 4	-2.3	18.5	-.6	13.8	-.9	12.4	9.8	37.4	-3.7	40.9	.4	23.4
Model 6	-.1	8.6	-.1	8.7	.2	9.1	-3.2	16.1	-3.1	16.2	-2.7	16.7
Model 8	-2.1	16.5	-1.0	10.9	-.2	10.5	2.9	30.0	-12	31.4	-2.6	18.2

Note: ME = Mean Error, MAE = Mean Absolute Error
Monthly Average Nos. of (Respond.,Nonrespond.): SIC 121 (337,49); SIC 373 (318,40)