ALTERNATIVE IMPUTATION METHODS FOR WAGE DATA

Sandra A. West, Shail Butani, and Michael Witt, Bureau of Labor Statistcs
441 G Street NW, GAO Bldg Room 2126, Washington DC 20212

1. Introduction

In this paper the results of an empirical investigation of different imputation methods for wage data and the ratio of wage to employment data are presented. This study is a sequel to the paper entitled "Alternative Imputation Methods For Employment Data" (1989). Both projects 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 ES-202 microdata files. 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, total quarterly wages, 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 for total quarterly wages of an establishment that would work reasonably well for all SIC groups within each state. The methods included regression modeling, distribution modeling with maximum likelihood estimators for the parameters, multiple imputation, and standard procedures such as hot decks, and mean value.

The wage data used in this study are discussed in Section 2. Section 3 presents the notation used and the evaluation criteria that are used to compare the various imputation methods. Section 4 provides a description of the standard procedures such as: carryover method of imputation, two mean imputation procedures, and two hot deck procedures. In Section 5, eight regression models for imputing wages 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 imputation procedure would be reflected. Section 9 compares the results from the various imputation methods and summarizes the findings of this study.

2. Data

The data used for the wage study were the ES-202 microdata files obtained from the State Employment Security Agency (SESA) of Wisconsin. Although results of this project are needed for all the states, due to various reasons it was not possible to obtain a sufficient amount of data from any other state.

Five consecutive calendar quarters of data, (January 1988 through March 1989), were selected and used to impute wages for the latter four quarters. Four quarters were considered so that fluctuations in total wages due to seasonality could be incorporated into the analyses. Because most of the imputation methods required a unit's total wages from the previous quarter, data for five quarters were needed.

All the procedures were tested for three different SIC groups of establishments. They are: SIC 16, Heavy Construction Other Than Building Construction--Contractors; SIC 37, Transportation Equipment; and SIC 50, Wholesale Trade--Durable Goods. Two of the three SICs were chosen so that the results of this study could be compared against the authors previous study (West, et al, 1989).

Intuitively, an establishment's total wages are highly correlated with its total employment at any given point in time. Consequently, a measure of size was created for each establishment based on the establishment's oldest, nonmissing monthly employment value beginning with January 1988. This size measure was used to stratify the data set by three different size partitions (Table I) in order to examine the size class effect, if any, on the imputation procedures.

At the time of this study, there were no data available that would indicate the establishments for which the wages or employment were imputed by the SESA of Wisconsin. Therefore, it was assumed that the missing data mechanism is ignorable, and random sets of units were chosen to represent the set of nonrespondents. To examine the effects of nonresponse rates, two sets consisting of 10% and 20% of the datafile were selected and designated as the nonrespondents and imputation procedures were examined using the remaining 90% and 80% , respectively.

All of the imputation procedures were constructed using an establishment's total quarterly wages, and in an analogous fashion, the ratio of total quarterly wages to total quarterly employment (i.e. mean quarterly wages). The ratio of wages to employment was considered because it was felt that the ratio would stabilize total wages, which usually fluctuate across quarters much more than employment. In both cases, however, the error measures, defined in the next section, were computed for total quarterly wages.

The error measures for 90 different methods (Table II) by three size class partitions (1, 3, and 8) were computed for each of the three SICs and two response rates. This was done separately for models based on total quarterly wages and ratio of total quarterly wages to total employment. Due to space limitations, the results using all different combinations of these factors could not be presented but will be briefly discussed in the conclusion section.

3. Notation and Evaluation Criteria

In this Section and in Sections 4 through 7, the imputation procedures are discussed using total wage data. The imputation procedures applied to mean wages analogously follows by letting "Y" represent mean wages as opposed to total wages.

Notation

Let the variables:

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

$S_{t,r,m}$ = Set of respondents in domain of procedure m

$S_{t,nr,m}$=Set of nonrespondents in domain of procedure m

$Y_{t,i}$ = Reported quarterly wages of $ES_{t,i}$

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

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

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

$E_{t,i,m}$ = 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. Relative Error (%):

$$RE_m = \{\sum_{size} \sum_{t} \sum_{i} E_{t,i,m} / \sum_{size} \sum_{t} \sum_{i} Y_{t,i}\} * 100.00$$

b. Relative Absolute Error (%):

$$RAE_m = \{\sum_{size} \sum_{t} \sum_{i} AE_{t,i,m} / \sum_{size} \sum_{t} \sum_{i} Y_{t,i}\} * 100.00$$

where i ε $S_{t,nr,m}$.

Note that RE_m represents a macro level statistic that indicates the effect that the imputation procedure has on total quarterly wages, while RAE_m is a micro level statistic that indicates the effect on the unit's quarterly wages. The corresponding mean unit errors per nonrespondent were also computed but are not presented in this paper due to space.

4. Standard Methods

CO: Carryover Method of Imputation

Under this method, total quarterly wages of each nonrespondent, $ES_{t,j}$, is imputed as follows:

$$Y^p_{t,j,CO} = Y_{t-1,j} .$$

MN: Mean Imputation Method

For any fixed SIC group, employment size class, and quarter t and for all $ES_{t,j}$ ε $S_{t,nr,MN}$:

$$Y^p_{t,j,MN} = \sum_i Y_{t,i} / N_{t,r,MN} .$$

Thus $Y^p_{t,j,MN}$ is equal to the average of the total quarterly wages of all respondents in the stratum.

MNL: Mean Of Log Wages

This method is the same as the mean imputation method stated above except log wages are substituted for wages.

HD1: Hot Deck Imputation Method - Random Selection

For any fixed SIC group, employment size class, and quarter t, the quarterly wages of $ES_{t,j}$ ε $S_{t,nr,HD1}$ is:

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

where $Y^*_{t,i}$ is the total quarterly wages of a randomly selected respondent from $S_{t,r,HD1}$. 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 total quarterly wages value for the nonrespondent's.

Within any fixed SIC group, employment size class, and for each quarter t, all $ES_{t,i}$ 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 $ES_{t,j}$ ε $S_{t,nr,HD2}$, let $Y^{(1)}_{t,i}$ be the total quarterly wages for the first establishment $ES^{(1)}_{t,i}$ ε $S_{t,r,HD2}$ that precedes $ES_{t,j}$ on the ordered list, and $Y^{(2)}_{t,k}$ be the total quarterly wages for the first establishment $ES^{(2)}_{t,k}$ ε $S_{t,r,HD2}$ that succeeds $ES_{t,j}$ on the ordered list. If

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

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 and Wages by Regression

Regression Models

A common method for imputing missing values is via least squares regression (Afifi and Elaskoff 1969). In three papers on estimators for total employment (West 1982, 1983) and West, et al. (1989), 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 the establishment's previous month's employment, $Y_{t-1,i}$. That is,

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

where β is some constant depending on t. In the remaining sections, for clarity, the subscripts t and m are suppressed in conjunction with the parameters σ, α and β.

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

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

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

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

In previous papers, $\nu_{t,i} = \sigma^2 Y_{t-1,i}$ and $\nu_{t,i} = \sigma^2$ were considered and it was found that the model:

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

with $\nu_{t,i} = \sigma^2 Y_{t-1,i}$ worked reasonably well for employment data.

Since this model with the above assumptions worked well with employment data, it was decided to apply variations of the same model with wage data. For the current data set, the following eight models were considered for total quarterly wages.

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

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

Models 1 - 4 assume $\nu_{t,i} = \sigma^2$. 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 $\nu_{t,i} = \sigma^2 \mathrm{Ln}(Y_{t-1,i})$ for models 7 and 8.

When the imputation procedure is based on a regression model, m will be prefixed by RM. The regression model parameters were estimated using the establishments in the set $S_{t,r,m}$ and an imputed value was calculated for those establishments in the set $S_{t,nr,m}$. Note that in the case when $Y_{t,i}$ denotes the ratio of wages to employment it is assumed that employment is known. The model is conditional on $Y_{t-1,i}$ and the employment at time t.

<u>Example Using Model 8</u>

$\mathrm{Ln}(Y_{t,i}) = \beta \mathrm{Ln}(Y_{t-1,i}) + \varepsilon_{t,i}$ with $\nu_{t,i} = \sigma^2 \mathrm{Ln}(Y_{t-1,i})$

and β is estimated as:

$$\beta^p = \sum_{i \,\varepsilon\, S_{t,r,RM8}} \mathrm{Ln}(Y_{t,i}) \Big/ \sum_{i \,\varepsilon\, S_{t,r,RM8}} \mathrm{Ln}(Y_{t-1,i}) \;.$$

For any nonrespondent, $ES_{t,j}$, in $S_{t,nr,RM8}$, the establishment's predicted total wages at time t is:

$$Y^p_{t,j,RM8} = \exp\{\beta^p \mathrm{Ln}(Y_{t-1,j})\} \;.$$

<u>Adjustments for Models 4 and 8</u>

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

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

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

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

$$\exp\{\beta \mathrm{Ln}(Y_{t-1,j}) + .5\mathrm{Var}(\varepsilon_{t,j})\} \;.$$

As an estimate of $\mathrm{Var}(\varepsilon_{t,j})$, the residual mean square error, MSE, from the regression was used. The predicted total wages for m = 4 and 8 are computed as:

$$Y^p_{t,j,RMmA1} = \exp\{\beta^p \mathrm{Ln}(Y_{t-1,j}) + .5MSE\}$$

An alternative adjustment to the logarithmic regression models was also tried. This adjustment was used by David (1986), and led to the following unbiased prediction of $Y_{t,j}$ for models 4 and 8:

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

where $Z_{t-1,i} = \mathrm{Ln}(Y_{t-1,i})$. Thus,

$$Y^{p}_{t,j,RMmA2} = \exp\{ \beta^{p} Z_{t-1,j} + .5(MSE)(WGS)\}$$

where $WGS = 1 - \{Z^{2}_{t-1,j} \, / \, \sum_{i} Z^{2}_{t-1,i}\}$ for m = 4

and $WGS = 1 - \{Z_{t-1,j} \, / \, \sum_{i} Z_{t-1,i}\}$ for m = 8.

6. Adding Residuals to the Regression Models

The methods discussed in the previous section could be thought of as imputing for missing total quarterly wages by using the mean of the predicted Y_t (or $\ln(Y_t)$) distribution, conditional on the predictors, Y_{t-1} (or $\ln(Y_{t-1})$). 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, select residuals $res_{t,k}$, with mean zero, to add to $Y^{p}_{t,j,RMm}$ (or the predicted $\ln(Y_{t,jRMm})$).

In this project, it was decided to consider this imputation procedure with the residuals, $res_{t,k}$, equalling:

1. A randomly selected respondent's residual using model RMm (procedure denoted by RMmRS).

2. A random normal deviate, from the distribution with mean 0 and variance MSE $*\tau_j$, where τ_j takes on one of three values defined below, using model RMm (procedure denoted by RMmRGλ).

$$\tau_j = 1$$

$$= E_j = \{ (N_{t,r,m})^{-1} + A^{2}_{j} / \sum_{i} A^{2}_{i} \}$$

$$= P_j = 1 + E_j,$$

where for models 1 and 5

$$A_j = Y_{t-1,j} - (\sum_{i} Y_{t-1,i}) / N_{t,r,m})$$

and for models 2 and 6

$$A_j = Y_{t-1,j}.$$

For the corresponding log models $Y_{t-1,j}$ is replaced by Ln $(Y_{t-1,j})$. Note that the estimated variances, MSE * τ_j for $\tau_j = E_j$ and P_j are estimates of the variances of the estimator of the mean of $Y_{t,j}$, and a single new observation $Y^{p}_{t,j,m}$, respectively (Neter and Wasserman, 1974).

For each of the eight models, residuals were added to the model predictions by the above methods. For example, using model 8 and the first method described above, a prediction of $Y_{t,j}$ is:

$$Y^{p}_{t,j,RM8RS} = \exp\{\beta^{p} \ln(Y_{t-1,j}) + res_{t,k}\},$$

where $res_{t,k}$ is the residual from a randomly selected respondent k; that is,

$$res_{t,k} = \ln(Y_{t,k}) - \beta^{p} \ln(Y_{t-1,k}).$$

Using model 6 and the second method described above:

$$Y^{p}_{t,j,RM6RGt} = \beta^{p} Y_{t-1,j} + s\delta_j,$$

where δ_j is a random number from a $N(0,1)$ distribution and s^2 is equal to the MSE * τ_j.

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 selection can be made from it. The imputation task takes one random selection 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 $N(\beta Y_{t-1,i}, \sigma^2)$ 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, Probability(θ) proportional to a constant, is assumed.

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

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

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

where

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

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

$$\nu = 1 / \sum_{i} Y^{2}_{t-1,i}$$

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 $N(0,1)$ variate, say Z_o, and let

$$\beta_2 = \beta^p{}_1 + \sigma_2(n)^{.5}(Z_o)$$

3. Let n_o be the number of values that are missing, that is, the size of $S_{t,nr,BAY}$. Draw n_o values of $Y_{t,BAY}$ as

$$Y^p{}_{t,j,BAY} = \beta_2 Y_{t-1,j} + \sigma_2 Z_j \qquad (7.1)$$

where the n_o normal deviates, Z_j are drawn independently.

Equation (7.1) can be rewritten as:

$$Y^p{}_{t,j,BAY}=\beta^p{}_1 Y_{t-1,j} + \frac{(MSE)^{.5}(n-1)^{.5}}{(h)^{.5}} [(\nu)^{.5} Z_o Y_{t-1,j}+Z_j].$$

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 advantage of multiple imputation is that the resultant imputed values will account for sampling variability associated with the particular nonresponse model.

Multiple imputation was obtained from the Bayesian method by repeating the above three steps five times. The average of the five values was taken as the imputed value.

Multiple imputation was also obtained for the following procedures: hot deck random selection; regression model with randomly selected residuals; and regression model with randomly generated residuals, $N(0, \text{MSE} * \tau_j)$. For all of the multiple imputation methods, error measures were computed by using the average of five such repeated imputations.

9. Comparison of Imputation Methods and Conclusions

Each imputation method was applied to an establishment's total quarterly wages and to the ratio of total quarterly wages to total quarterly employment (i.e., mean quarterly wages). In order to have comparability between the two data types, the error measures, Percentage Relative Error (%RE) and Percentage Relative Absolute Error (%RAE) were based on total quarterly wages. For both the data types, each imputation method was applied to each of the three SICs by three sizes class partitions and two response rates, and, accordingly, the %RE and %RAE were computed for each combination. Due to space limitations, the results are presented in Table II only for SICs 16 and 37, and for the 80% response rate. Data are presented only for data type total quarterly wages for reasons stated below.

For the three SICs considered, imputing total wages based upon ratio of wages to employment faired about the same as imputing wages based on total wages, in terms of the two error measures. The knowledge of the employment values did not yield smaller errors. Additionally, the ratio of wages to employment assumes that total employment is known for every establishment on the file, which is generally not the case. In fact, because of the nature of the U.I. reports, one of the following occurs: (1) both the employment and wage data are missing or (2) wages are provided and employment data are missing. It is an extremely rare case when employment data are provided but not the wages. Since the effect of using an imputed employment value on predicting total wages has not been analyzed, at this time it is recommended that only total wages data be used for imputation.

Selecting the best imputation method based only on total wage data type from the set of 90 methods considered was difficult because one method of imputation did not consistently and clearly yield the smallest error measures. Consequently, in order to determine the best method of imputing total wages for the three SICs by two nonresponse patterns, it was decided to consider only those methods that yielded less than 10 for the |%RE| and less than 50 for the %RAE for any SIC group by any size class partition. From this set of imputation procedures and size class partitions, the subset that overlapped across the three SICs and the two response rates was retained. The resulting methods and size class partitions that had the |%RE| less than 10 and the %RAE less than 50 across the three SICs and two response rates are listed in Table A.

Table A: Procedures with |%RE| < 10 & %RAE < 50

METHOD	**SIZE CLASS**		
	1	3	8
Carryover	x	x	x
Regression Model 4	x		
Regression Model 8		x	x
.5(MSE) Model 8		x	x
.5(MSE)(WGS) Model 8		x	x
Randomly Generated Normal Residual:			
S^2 = MSE Model 8		x	x
S^2 = MSE Model 8*		x	x
S^2 = MSE*P Model 8		x	x
S^2 = MSE*P Model 8*		x	x
S^2 = MSE*E Model 4	x		
S^2 = MSE*E Model 4*	x	x	
S^2 = MSE*E Model 8		x	x
S^2 = MSE*E Model 8*		x	x

[*] Indicates data presented are multiple imputation results.

Model 8 with three and eight size class partitions also dominated the list when the ratio of wages to employment data were used. Note that the two basic models that are among the contenders use the logarithm of wages as the dependent variable. Most wage models in the literature, such as David, et al. (1986) and Greenlees, et al. (1982) are based on household surveys and have different independent variables in the model, but the dependent variable is generally the logarithm of wages.

Because of the dominance of some form of model 8 and the consistency of three and eight size class partitions, some form of model 8 is preferred over model 4. Of all the procedures involving model 8, the one with no adjustment is preferred in the interest of simplicity. In the above list, the three size class partition for the regression model based procedures usually performed as well as or better than the eight size class partition. Also, since many State/SIC cells will have only a small number of observations, it is recommended that three size classes be employed if a regression model is selected.

The above discussion limits the selection to either the carryover method or to regression model 8 with three size class partitions. Regression model 8 with three size class partitions is recommended instead of the carryover method, because the data used for this study were for January 1988 through March 1989, a relatively stable period economically. It is expected that the carryover method will not perform as well during a period of large economic growth or decline. Also a similar study conducted last year for employment data recommended the use of model 6, which is similar to model 8, the only difference being model 6 uses raw data while model 8 uses the transformed data.

Future work will include testing of both the carryover method and model 8 with three size class partitions for different SICs, States, and response patterns.

References

1. Afifi, A. A., and Elaskoff, R. M., (1969), "Missing Observations in Multivariate Statistics III: Large Sample Analysis of Simple Linear Regression", Journal of the American Statistical Association, vol. 64, pp. 337-358.

2. 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.

3. Greenlees, W. S., Reece, J. S., and Zieschang, K. D. (1982), "Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed," Journal of the American Statistical Association, vol. 77, pp. 251-261.

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

5. Neter, J., and Wasserman, W.,(1974), Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, Illinois.

6. 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.

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

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

9. 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.

10. West, S., Butani, S., Witt, M., Adkins, C., (1989), "Alternate Imputation Methods for Employment Data", ASA Proceedings of the Section in Survey Research Methods.

Table I: Establishment Size Class Definitions

Size class is determined by the establishment's oldest, nonmissing employment during the time period: January 1988 to March 1989. 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 +	20 - 49; 500 - 999
		50 - 99; 1000 +