Alternative Imputation Procedures For Item Non-response from New Establishments in the Universe

S. West, D. Kratzke, and K. Robertson, Bureau of Labor Statistics
Sandra A. West, 2 Massachusetts Ave. N.E., Washington, D.C. 20212

KEY WORDS: Regression, Bayesian, Multiple Imputation, Births, Hot Deck

1. Introduction

In this paper the results of an empirical investigation of different imputation methods for item non-response from new establishments are presented. The imputation is for employment data given that wage data are known. This investigation began in connection with a revision project for the Bureau of Labor Statistics (BLS) program that maintains the Universe Data Base (UDB). The data base stores information received from the state Quarterly Unemployment Insurance Address (QUI) files. The QUI files represent a comprehensive list of all business establishments that are covered under the unemployment insurance system in the States. Each employer is required to submit a QUI report which contains, among other things, information on monthly employment for the quarter, total quarterly wages, a standard industrial classification (SIC) code and a county code for the establishment. Although the filing of the report is mandatory, there are always some reports that are filed late, delinquent, or with partial data in each quarter. In the case of partial data it is usually the employment data that are missing. In previous papers imputation methods for employment and wage data were considered for continuous units which are units that were present in the previous quarter. The situation of missing data from new establishments was not considered. This paper deals with the latter situation.

The goal of this project was to develop a single imputation procedure for new establishments that have reported total quarterly wages but not employment that would work reasonably well for all SIC groups within each state. The methods tested included regression modeling and distribution modeling with maximum likelihood estimators for the parameters, multiple imputation, as well as standard procedures such as hot deck, mean, and median.

The data used in this study are discussed in Section 2. 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 standard procedures such as mean, median, and several hot deck procedures. In Section 5, eight regression models for imputing employment given 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 describes the current method. Section 10 compares the results from the various imputation methods and summarizes the findings of this study.

2. Data

Six quarters of UDB data were available for this study, from quarter 1 of 1990 to quarter 2 of 1991. A unit (establishment) is classified as a birth unit if it can not be matched to any other unit in previous quarters by a number of criteria. To assure that we did not mistakenly label a unit as a birth, when perhaps it was inactive for a few quarters, we decided to use units in quarter 1 of 1991. The units in this quarter that are classified as birth units are not matched to units in any quarter of 1990.

Data from Michigan and California were obtained for the following industries: Special Trade Contractors, Chemical and Allied Products, Transportation Equipment, Trucking and Warehousing, Apparel and Accessory Stores, Miscellaneous Retail, Nondepository Institution, Personal Services, Membership Organizations, and Private Households. Additional industries from Michigan include: Agricultural Services, Lumber and Wood Products, Industrial Machinery and Equipment, Real Estate, and Miscellaneous Repair Services.

Intuitively, an establishment's total wages are highly correlated with its total employment at any given point in time. The more homogenous the strata, the higher the correlation will be. Several stratifications were tried. Within each 2-digit SIC chosen, we stratified the data further by:
(1) 3-digit SIC; (2) 3-digit SIC/size class; (3) 4-digit SIC/size class; (4) 4-digit SIC/county

Usually a measure of size is created for each establishment based on its most recent, nonmissing monthly employment. But since the target of this study is to impute employment for new units, we can not create a measure of size for these units based on employment. For our imputation procedures, size classes were formed by breaking units into wage classes at the 25th, 50th, 75th, and 95th percentile points of quarterly wages.

In order to validate our procedures, we only selected birth units that reported both non-zero employment and wages. Thus, the minimum quarterly employment (sum of monthly employment in the quarter) an establishment

could have is 3; so that in all the imputation procedures, 3 was set as the lower bound for quarterly employment. We simulated the pattern of non-response observed in the data as much as possible. If a particular industry has x% of imputed employment for birth units, then a response rate of (1-x)% was used. It was assumed that the missing data mechanism is ignorable, and random sets of units were chosen to represent the set of nonrespondents. Most of the imputation procedures imputed for total quarterly employment based on total quarterly wages.

3. Notation and Evaluation Criteria

Notation

$E_{i,t}$=quarterly employment for establishment i in quarter t,

$\hat{E}_{i,t}$=predicted quarterly employment for establishment i in quarter t,

$W_{i,t}$=quarterly wages for establishment i in quarter t.

The problem is to impute for a new establishment k, that has $W_{k,t}$, but is missing $E_{k,t}$. For a given stratified cell, let,

B_t=Set of birth units that have both reported wages and employment for quarter t,

A_t=Set of continuous units that have both reported wages and employment for quarter t,

nr_t=percentage of birth units in the t th quarter that have reported wages but no employment,

NR_t=Set of units that were obtained by randomly selecting the percentage, nr_t, from the set B_t,

BR_t=remaining set of units in B_t,

M_t = the set BR_t or set $BR_t \cup A_t$,

NNR_t=number of units in NR_t,

NM_t=number of units in BR_t or in $BR_t \cup A_t$.

The imputation methods will be applied to units of the set NR_t, The units in set M_t are used to fit different modelling methods or to obtain imputed values from standard procedures. The set NR_t is called the set of nonrespondents, or test set, and the set M_t is called the set of respondents, or the model set.

Evaluation Criteria

Let $\varepsilon_{k,t} = \hat{E}_{k,t} - E_{k,t}$ denote the error in the imputed value for establishment k. The following error measures for each stratum will be used.

Percent Relative Error:

$$RE = 100 \sum_{k \in NR_t} \varepsilon_{k,t} \Big/ \sum_{k \in NR_t} E_{k,t}$$

Percent Relative Absolute Error:

$$RAE = 100 \sum_{k \in NR_t} |\varepsilon_{k,t}| \Big/ \sum_{k \in NR_t} E_{k,t}.$$

The corresponding mean errors were also computed.

Errors were computed for each imputed value and then error measures were computed for each stratum, and then added across strata for total error for each 2-digit SIC. Note that RE represents a macro level statistic that indicates the effect that the imputation procedure has on total quarterly employment for each 2-digit SIC, while

RAE is a micro level statistic that indicates the effect of imputation on each unit's quarterly employment.

4. Standard Methods

Mean and Median

The mean imputation method is a common method of imputation in many surveys, especially for those surveys with a high response rate. If the response rate is low, 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, size class, and quarter t :

$$\hat{E}_{k,t} \quad = \quad \sum_{i \in M_t} E_{i,t} \Big/ NM_t$$

Thus $\hat{E}_{k,t}$ is equal to the average of the total quarterly employment of all respondents in the stratum. In the Tables, these estimates are referred to as Mean-3 or Mean-4, depending on whether the imputation was done at the 3- or 4-digit SIC level. $\hat{E}_{k,t}$ equaling the median of the total quarterly employment of all respondents in the stratum was also tried. These estimates are referred to as Med-3 or Med-4, again depending on whether the imputation was done at the 3- or 4-digit SIC level.

Mean and Median - Variations

The Mean Ratio method, denoted by MeanR, was calculated in the following manner. For any fixed SIC group, size class, and quarter t, the mean for total wages, $\overline{W}_t$, and the mean for total employment, $\overline{E}_t$, was calculated over M_t. The imputed employment is then:

$$\hat{E}_{k,t} = (\overline{E}_t / \overline{W}_t) \ W_{k,t}$$

(It will be seen that this is the same basic procedure as using Regression Model 2, which will be discussed in the next section.)

The Median Ratio method, denoted by MedR, is similar to the preceding one with median replacing mean.

Hot Deck - Nearest Neighbor

For any fixed SIC group, size class, and quarter t, let k denote a nonrespondent and c denote a respondent such that

$$\left| W_{c,t} - W_{k,t} \right| \le \left| W_{i,t} - W_{k,t} \right|, \quad \text{for all } i \in M_t.$$

then, $\qquad \hat{E}_{k,t} = E_{c,t}.$

The Nearest Neighbor hot deck method, denoted by NN, 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 employment value for the nonrespondent's.

Hot Deck - Nearest Neighbor Variations

Two variations were tried. One, denoted by NNI, used a linear interpolation in the ordered list. The second, denoted by NNIR, is identical to NNI, except in the border cases, when a ratio adjustment was made.

Hot Deck - Random Selection

For any fixed SIC group, size class, and quarter t,

$$\hat{E}_{k,t} = E_{r,t}^1$$

where $E_{r,t}^1$ is the employment value of an establishment randomly selected from M_t. In the tables the estimates are referred to as RAND.

5. Modeling Employment by Regression

Regression Models

A common method for imputing missing values is via least squares regression (Afifi and Elaskoff, 1969). In several 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 tth quarter, given the following vector of E - values for quarter t-1:

$$\overline{\mathbf{E}}_{t-1} = \left[E_{t-1,1}, E_{t-1,2}, E_{t-1,3}, \ldots, E_{t-1,n} \right]$$

is proportional to the establishment previous quarter's employment, $E_{t-1,i}$. That is,

$$E(E_{t,i} | \overline{\mathbf{E}}_{t-1} = \overline{\mathbf{e}}_{t-1}) = \beta \, E_{t-1,i}$$

where β is some constant depending on t.

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

$$\text{cov}(E_{t,i}, E_{t,j} | \overline{\mathbf{E}}_{t-1} = \overline{\mathbf{e}}_{t-1}) = \begin{cases} v_{t,i} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

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

The model can be rewritten as:

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

where

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

and

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

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

$$E_{t,i} = \beta \, E_{t-1,i} + \varepsilon_{t,i} \quad \text{with} \quad v_{t,i} = \sigma^2 E_{t-1,i}$$

worked reasonably well for employment data.

A similar model worked well for wages except the data were first transformed by applying the natural logarithm to each wage value. Since this model with the above assumptions worked well with employment and wage data, it was decided to apply variations of the same model with employment versus wage data. For the current data set, the following eight models were considered for total quarterly employment versus total quarterly wages.

$$E_{j,t} = \beta_1 \, W_{j,t} + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2) \qquad (1)$$

$$E_{j,t} = \beta_2 \, W_{j,t} + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2 \, w_{j,t}) \qquad (2)$$

$$\ln E_{j,t} = \beta_3 \left(\ln W_{j,t} \right) + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2) \qquad (3)$$

$$\ln E_{j,t} = \beta_4 \left(\ln W_{j,t} \right) + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2 \ln w_{j,t}) \quad (4)$$

$$E_{j,t} = \alpha_5 + \beta_5 \, W_{j,t} + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2) \qquad (5)$$

$$E_{j,t} = \alpha_6 + \beta_6 \, W_{j,t} + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2 \, w_{j,t}) \quad (6)$$

$$\ln E_{j,t} = \alpha_7 + \beta_7 \left(\ln W_{j,t} \right) + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2) \quad (7)$$

$$\ln E_{j,t} = \alpha_8 + \beta_8 \left(\ln W_{j,t} \right) + \varepsilon_{j,t} \, \text{with} \, \varepsilon_{j,t}.\sim N(0, \sigma^2 \ln w_{j,t}) (8)$$

The models will be fit over the set M_t by stratum. The models were fit for each 3-digit SIC and 4-digit SIC/county. The more homogenous the strata the better the models will be.

Example of fitting model 4:

$$\ln E_{j,t} = \beta_4 \left(\ln W_{j,t} \right) + \varepsilon_{j,t} \quad \text{with} \quad \varepsilon_{j,t}.\sim N(0, \sigma^2 \ln w_{j,t})$$

and β_4 is estimated as:

$$\hat{\beta}_4 = \sum_{i \in M_t} \ln E_{i,t} \Big/ \sum_{i \in M_t} \ln W_{i,t}$$

For establishment j in NR_t, the establishment's predicted total employment is:

$$\hat{E}_{j,t} = \exp\{ \hat{\beta}_4 \ln W_{j,t} \}.$$

The regression models 1-8 are denoted by REG1-REG8, respectively.

Adjustments for Log Models

Consider models r, for r = 3, 4, 7, 8. If it is assumed that $\varepsilon_{j,t}$ is normally distributed, then $E_{j,t}$ has a lognormal distribution with

Mean: $\quad \exp\{ \beta_r \ln(W_{j,t}) + .5Var(\varepsilon_{j,t}) \}$

Var: $\{\exp [Var(\varepsilon_{j,t})]^{-1}\} \exp\{ 2\beta_r \ln(W_{j,t}) + Var(\varepsilon_{j,t}) \}$.

Therefore, an unbiased estimator of $E_{i,t}$ is:

$$\exp\{ \beta_r \ln(W_{i,t}) + .5Var(\varepsilon_{i,t}) \}.$$

As an estimate of $Var(\varepsilon_{i,t})$, the residual mean square error, MSE, from the regression was used. The predicted total employment for r = 3 and 4 were computed as:

$$\hat{E}_{i,t} = \exp\{ \hat{\beta}_r \ln(W_{i,t}) + .5MSE\}$$

The log regression models with adjustment are denoted by REG3ADJ, REG4ADJ, REG7ADJ, and REG8ADJ, corresponding to the regression models REG3, REG4, REG7, and REG8 without the adjustment.

6. Adding Residuals to the Regression Models

The methods discussed in the previous section could be thought of as imputing for missing total quarterly employment by using the mean of the predicted E_t (or $\ln(E_t)$) distribution, conditional on the predictors, W_t (or $\ln(W_t)$). 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 $res_{j,t}$, with mean zero, to add to $\hat{E}_{k,t}$ (or the predicted $\ln(E_{k,t})$)

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

    1. A randomly selected respondent's residual, using each of the eight models. These models are denoted by REG1RES-REG8RES, corresponding to REG1-REG8.

    2. A random normal deviate, from the distribution with mean 0 and variance MSE. These models are denoted by REG1NOR-REG8NOR, corresponding to REG1-REG8.

    For example, using model 7 and the first method described above, a prediction of $E_{k,t}$ is:

$$\hat{E}_{k,t} = \exp\{\hat{\alpha}_7 + \hat{\beta}_7(\ln W_{k,t}) + res_{j,t}\},$$

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

$$res_{j,t} = [\ln E_{j,t} - \hat{\alpha}_7 - \hat{\beta}_7(\ln W_{j,t})].$$

Using model 6 and the second method described above:

$$\hat{E}_{k,t} = \hat{\alpha}_6 + \hat{\beta}_6 W_{k,t} + s\,\delta_k,$$

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

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 E_t, for $E_t, \in NR_t$, denoted by $E_{t,BAY}$. This is done by first drawing a parameter from the posterior distribution obtained in the estimation task and then drawing $E_{t,BAY}$ from its conditional posterior distribution given the drawn value of the parameter.

    For the modeling task, consider <u>model 1</u> and $E_{j,t}$ having a N($\beta_1 W_{j,t}$,s^2) distribution. This is the specification for the conditional density f($E_{j,t}$ x $W_{j,t}$, q) where q = (β_1,s). In order to complete the modeling task, the conventional improper prior for q, Prob(q) proportional to a constant, is assumed.

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

$$f(s^2 \text{ x } E_{j,t}) = \hat{\sigma}_0^2(n-1)\big/\chi_{n-1}^2$$

$$f(\beta_1, \text{ x } s^2) = N(\hat{\beta}_0, s^2 n)$$

where

$$\hat{\sigma}_0^2 = \sum_j \left(E_{j,t} - \hat{\beta}_0 W_{j,t}\right)^2 \big/(n-1) = MSE$$

$$\hat{\beta}_0 = \sum_j E_{j,t} W_{j,t} \big/ \sum_j W_{j,t}^2$$

$$v = 1 \big/ \sum_j W_{j,t}^2$$

where n = number of respondents.

    Since the posterior distribution of q is in terms of standard distributions, random draws can easily be computed. The imputation task for this model is as follows:

1. Estimate s^2 by a χ_{n-1}^2 random variable, say h, and let

$$\sigma_1^2 = \hat{\sigma}_0^2(n-1)(h)^{-1}$$

2. Estimate β_1 by drawing one independent N(0,1) variate, say Z_O, and let

$$\hat{\beta}_{00} = \hat{\beta}_0 + \sigma_1 v^{.5} Z_0$$

3. Let n_O be the number of values that are missing. Draw n_O values of $E_{t,BAY}$ as

$$\hat{E}_{k,t,BAY} = \beta_0 W_{k,t} + \sigma_1 Z_k$$

where the n_O normal deviates, Z_k are drawn independently.

    The above equation can be rewritten as:

$$\hat{E}_{k,t,BAY} = \hat{\beta}_0 W_{k,t} + MSE^{.5}(n-1)^{.5} h^{-.5}[v^{.5} Z_0 W_{k,t} + Z_k].$$

These Bayesian models are denoted by REG1BAY-REG8BAY, corresponding to REG1-REG8.

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 independent imputations were obtained by repeating the three steps. The average of these five values was taken as the imputed value. These methods are denoted by REG1BAYM-REG8BAYM, corresponding to REG1-REG8.

    Multiple imputation was also obtained for the regression model with randomly selected residuals and regression model with randomly generated residuals, N(0, MSE). For all of the multiple imputation methods, error measures were computed by using the average of five such repeated imputations. For randomly selected residuals, the models are denoted by REG1RESM-REG8RESM; and for randomly generated residuals, the models are denoted by REG1NORM-REG8NORM; corresponding to REG1-REG8.

9. The Current Method

    The current method is described in Appendix D of the Exportable ES-202 System. This method will be referred to as the EXPO method. The EXPO method is stratified by 4-digit SIC/county/ownership and it uses data from a year ago to form the ratio for imputing. In our paper, however, the ownership code will be excluded since only private ownership was considered in this study.

    A ratio of total quarterly employment to total quarterly wages of a macro record for the same quarter a year ago is computed. This ratio is multiplied by the unit's total quarterly wages to impute for quarterly employment. The

monthly employment is computed by dividing the quarterly employment by three times a prorate factor which indicates how many months the establishment is active in the quarter. For this research, only total quarterly employment is imputed. Note that this method is similar to using regression model 2, except with regression model 2 the ratio is computed at the current time period. That is, using REG2,

$$\hat{E}_{k,t} = \hat{\beta}_2 W_{k,t} \quad where \quad \hat{\beta}_2 = \sum_{i \in J_t} E_{i,t} \Big/ \sum_{i \in J_t} W_{i,t}$$

using EXPO,

$$\hat{E}_{k,t} = \hat{\beta} W_{k,t} \quad where \quad \hat{\beta} = \sum_{i \in J_{(t-4)}} E_{i,(t-4)} \Big/ \sum_{i \in J_{(t-4)}} W_{i,(t-4)}$$

where the subscript (t-4) denotes the quarter a year ago.

10. Comparison of Imputation Methods/Conclusions

At the beginning of the research, it was not clear whether to use establishments in the set BR_t or the set $BR_t \cup A_t$ to obtain information for imputing employment for establishments in the set NR_t.

In the first part of the research, model sets with only birth units were used, excluding those establishments that had total quarterly wages less than or equal to $110,500 (this figure was based on 50 employees making minimum wage of $4.25/hour each). For the States of California and Michigan, 18 imputation methods were applied to each of seven SICs (with an additional SIC in California) by various partitions and, accordingly, the error measures were computed for each combination. The 18 methods are: Mean-3, Mean-4, Med-3, Med-4, MeanR, MedR, NN, NNI, NNIR, RAND, and REG1-REG8. The regression models were done on 3-digt SIC. Within each state, the methods were ranked according to the error measures |RE| and RAE.

Selecting the best imputation method from the set of 18 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 birth total employment for all the SICs and the two states, the models were ranked according to several criteria. These criteria were as follows:

(1) The number of times a method yielded small errors, i.e., |RE| $\leq$ 15 and RAE $\leq$ 55
(2) The number of times a method yielded large errors, i.e, |RE| $\geq$ 30 or RAE $\geq$ 80
(3) The number of times a method ranked in the top 5 (or the top 10) according to |RE|
(4) The number of times a method ranked in the top 5 (or the top 10) according to RAE
(5) Total |RE| across all SIC's
(6) Total RAE across all SIC's

Note that the errors are relative and are summed only over the nonrespondents. After comparing the scores of the eighteen methods on the six criteria, eight methods were eliminated. When the ten remaining methods were reranked according to |RE| and RAE, Mean-3 and Mean-4 came to the top of the list.

Next, we included continuous units as well as birth units in the model set, that is, all establishments in the set $A_t \cup BR_t$. In this preliminary study on all units, the 18 methods metioned above and the EXPO procedure were done on the same seven SIC's from Michigan as with the birth units alone. After the 18 methods were ranked according to |RE| and RAE and the scores for the six criteria were compared, the promising methods were: MeanR, REG4-REG8.

In order to be able to directly compare our procedures with the current procedure, we decided to try the same stratification as the current procedure (which is 4-digit/county), using both continuous and birth units in the model sets, and including units making $110,500 or more in the study. Since the standard procedures did not do well in the preliminary phase using both birth and continuous units in the model sets, only the distribution modelling was done in this phase. The following methods were done: regression models, including adjustment to log models, regression models with residuals, Bayesian, and multiple imputation methods. However, due to time limitations, we only did the Bayesian for regression models 1, 7, and 8. The multiple imputation was done on the Bayesian method and on the regressions with residuals. A total of 51 procedures were done on the 12 SIC's from Michigan.

Based on the six criteria mentioned before, the ten best methods were REG2, REG2NOR, REG2NORM, REG6NOR, REG6NORM, REG7, REG7BAY, REG8, REG8ADJ, and REG8NOR. After comparing these methods, and noting the variances of |RE| and RAE across all 12 SIC's, the list was narrowed down to the following five mehtods: REG6NOR, REG6NORM, REG2, REG8ADJ, and REG8NOR.

Since these models did not differ markedly in their effectiveness, and because of cost considerations and the simplicity of the calculations under the model, we chose the REG2 model to be implemented in the ES-202 program. In practice this model can be implemented as a simple ratio adjustment. Also, this procedure is similar to the current procedure, except that more recent information is utilized.

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.
3. Royall, R. M. and Cumberland, W. G., (1978), "Variance Estimation in Finite Population Sampling", *Journal of the American Statistical Association*, vol. 73.
4. Rubin, D., (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons Inc., NY.
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", *ASA Proceedings of the Section in Survey Research Methods*.

8. West, S., Butani, S., Witt, M., (1990), "Alternate Imputation Methods for Wage Data", *ASA Proceedings of the Section in Survey Research Methods*.