Proceedings of the Second International Conference on Establishment Surveys

ALTERNATIVE IMPUTATION MODELS FOR WAGE RELATED DATA COLLECTED FROM ESTABLISHMENT SURVEYS

**Carl Barsky, James Buszuwski, Lawrence Ernst, Michael Lettau, Mark Loewenstein, Brooks Pierce,
Chester Ponikowski, James Smith, and Sandra West, Bureau of Labor Statistics
James Buszuwski, Bureau of Labor Statistics, Room 4160, 2 Massachusetts Avenue, NE,
Washington, D.C., 20212, Buszuwski_J@BLS.GOV**

Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

ABSTRACT

This study was undertaken to determine imputation methods for data collected from the National Compensation Survey (NCS), conducted by the Bureau of Labor Statistics (BLS). In this paper alternative regression models are compared for item nonresponse of wage related data, which are collected from establishments by detailed occupation level. The surveys involved are of a longitudinal nature and two separate cases of item nonresponse are considered. The first case involves establishment nonresponse at initiation in the survey, and the second case involves an update time for the establishment. The empirical study tests various regression models on real survey data. The nonresponse patterns in our tests were simulated using observed patterns on current NCS data.

Key Words: Missing Data, Item Nonresponse, National Compensation Survey

1. INTRODUCTION

In this paper the results of empirical investigations of alternative imputation models for nonresponse of wage related data are presented. The investigations began in connection with a development project for the Bureau of Labor Statistics (BLS) National Compensation Survey (NCS). The NCS collects wage related data from establishments by detailed occupation level. A missing data team, made up of mathematical statisticians, economists, collection and review specialists, and computer specialists, was formed. The goals included the development of new procedures for managing a variety of missing data issues in the NCS and the comparison to existing procedures. The issues include unit and item nonresponse with the primary focus on developing imputation methods for missing wage and benefit data at both initiation and during future updates.

The National Compensation Survey is an integration of earlier surveys. The purpose of NCS is to build a broader base of data concerning salaries and benefits. NCS replaces the Occupational Compensation Survey Program (OSCP) with a revised data collection procedure geared toward a broader coverage of occupations. In addition, NCS incorporates the Employment Cost Index (ECI), which measures changes in salaries and benefits; the Employer Cost for Employee Compensation (ECEC), which measures average employer costs for wages and benefits, and the Employee Benefits Survey (EBS), which studies the incidence and detailed characteristics of employer-provided benefits. For further description see the BLS Handbook of Methods (1997).

The NCS sample design comprises 154 primary sampling units (PSUs), which are either metropolitan areas or non-metropolitan counties. Wage estimates are published both nationally and for as many of the PSUs for which the sample is sufficiently large to support a publication. No locality estimates are produced for ECI, ECEC or EBS.

In this paper, the investigations will be presented only for the missing wage values at initiation and update times. At update time, it is assumed that a wage value for the establishment exists for an earlier time period. In contrast, at initiation, no earlier wage value is captured.

In Section 2, the theoretical background is discussed along with a comparison to earlier studies. The discussion of the empirical investigations begins in Section 3 with the description of the data sets used. Although nonrespondents were noted on the files, the actual values for the variables were never obtained. Thus nonresponse had to be simulated using the patterns of nonresponse observed on the files. We are not aware of previous research that might suggest a model for explaining the pattern of missingness in wage data. Therefore it was assumed that, within a stratum, the nonrespondents were missing at random. Also in this section the various regression models considered are discussed, along with the criteria used to evaluate the various models. Summary tables are listed to give an indication of the results. In Section 4 conclusions and plans for future work are presented.

2. MODELING WAGES BY REGRESSION – THEORETICAL BACKGROUND

Imputation models for wages will be considered for the situation where an earlier time period value for the wage variable is available; that is imputation for wage values missing during an update period. The situation dealing with imputation at initiation will be considered at the end of this section. In order to set the stage for the models considered in this study, results from earlier studies will now be presented.

2.1. Results from Other Studies

A common method for imputing missing values is via least squares regression (Afifi and Elaskoff, 1969). Previous work West (1982,1983,1989), has analyzed this and other methods of imputing wages using wage and employment data that are part of the Universe Data Base (UDB). Imputation methods were considered for both new and continuing establishments. The methods included regression modeling and distribution modeling with maximum likelihood estimators for the parameters, multiple imputation, as well as standard procedures such as hot deck, and mean value. It was discovered that the most promising models for employment and wages were the proportional regression models. Thus the other imputation methods are not re-studied in this paper.

The proportional regression models specify that the expected wage for quote i in cell j in the tth period, given the values for the (t-1)th period, is proportional to the quote's previous wage. A quote is defined as the average wage for an occupation within an establishment. We have,

$$E\left(W_{ijt} \;\middle|\; W_{ij(t-1)} = w_{ij(t-1)}\right) = \beta_{jt} w_{ij(t-1)}$$

where β_{jt} is some constant depending on j and t. Cells are defined by such variables as size class, industry, etc. The model can be rewritten as

$$(1) \qquad W_{ijt} = \beta_{jt} W_{ij(t-1)} + \varepsilon_{ijt}$$

where ε_{ijt} is an error term with mean 0 and variance σ^2_{ijt}. It was further assumed that errors are uncorrelated or, equivalently, $E(\varepsilon_{ijt}\varepsilon_{klt}) = 0 \text{ if } i \neq k \text{ or } j \neq l$. Three alternative assumptions about variances were considered:

$$(2a)\ \sigma^2_{ijt} = \sigma^2_t \qquad (2b)\ \sigma^2_{ijt} = \sigma^2_t W_{ij(t-1)} \qquad (2c)\ \sigma^2_{ijt} = \sigma^2_t W^2_{ij(t-1)} \qquad \text{for all i and j .}$$

Note that (2a) is the common assumption of homoscedasticity, which a priori seems unlikely to hold in the present case. In contrast, (2b) and (2c) represent alternative forms of heteroscedasticity.

Under assumption (2a), the least squares estimator of β_{jt} is given by the following weighted mean of wage ratios

$$(3) \qquad \hat{\beta}_{jt} = \sum_i c_{ij} \frac{W_{ijt}}{W_{ij(t-1)}}, \qquad c_{ij} = \frac{W^2_{ij(t-1)}}{\sum_i W^2_{ij(t-1)}}$$

where the sum is over the establishments in cell j reporting in both time periods.

Under assumption (2b), the weighted (inverse of variance) least squares estimator of β_{jt} is the ratio of the means;

(4) $$\hat{\beta}_{jt} = \left(\frac{1}{n_j} \sum_i W_{ijt}\right) \Big/ \left(\frac{1}{n_j} \sum_i W_{ij(t-1)}\right)$$

Where n_j is the number of matched quotes in the cell. One can obtain this estimate by performing ordinary least squares regression on the transformed equation:

(5) $$W'_{ijt} = \beta_{ijt} W'_{ij(t-1)} + \varepsilon'_{ijt}$$

where: $W'_{ijt} = W_{ijt} \Big/ \sqrt{W_{ij(t-1)}}$, $W'_{ij(t-1)} = \sqrt{W_{ij(t-1)}}$ and $\varepsilon'_{ijt} = \varepsilon_{ijt} \Big/ \sqrt{W_{ij(t-1)}}$.

Finally, under assumption (2c), the weighted least squares estimator is a mean of the ratios:

(6) $$\hat{\beta}_{jt} = \left(\frac{1}{n_j}\right) \sum_i \frac{W_{ijt}}{W_{ij(t-1)}}$$

One can obtain this estimate by performing ordinary least squares regression on the transformed equation:

(7) $$W''_{ijt-1} = \beta_{jt} + \varepsilon''_{ijt}$$

where: $W''_{ijt} = W_{ijt} \Big/ W_{ij(t-1)}$ and $\varepsilon''_{ijt} = \varepsilon_{ijt} \Big/ W_{ij(t-1)}$.

For a current nonrespondent k in cell j with prior quarterly wage $W_{kj(t-1)}$, the imputed current wage is:

(8) $$\hat{W}_{kjt} = \hat{\beta}_{jt} W_{kj(t-1)}$$

In previous papers, it was found that the estimator in (4) yielded much better fitting imputations than the other estimators. The imputations were even a little better when wages were replaced by their logs, that is, when the model was given by

(9) $$\ln W_{ijt} = \beta_{jt} \ln W_{ij(t-1)} + \varepsilon_{ijt} \quad \textit{with } E\left(\varepsilon_{ijt}\right) = 0, \textit{ and } E\left(\varepsilon_{ijt}^2\right) = \sigma_t^2 \ln\left(W_{ij(t-1)}\right)$$

Under (9), the weighted least square estimator of β_{jt} is:

(10) $$\hat{\beta}_{jt} = \left(\frac{1}{n_j} \sum_i \ln W_{ijt}\right) \Big/ \left(\frac{1}{n_j} \sum_i \ln W_{ij(t-1)}\right)$$

For a current nonrespondent k , with prior quarterly wage $W_{kj(t-1)}$, the imputed current wage is:

(11) $$\hat{W}_{kjt} = \exp\left(\hat{\beta}_{jt} \ln W_{kj(t-1)}\right)$$

It follows from (9) that if ε_{ijt} is normally distributed, then W_{kjt} is distributed lognormally with mean $\exp(\beta_{jt} \ln W_{kj(t-1)} + .5\sigma^2_{kjt})$. This suggests the alternative imputation

$$(12) \qquad \hat{\hat{W}}_{kjt} = \exp(\hat{\beta}_{jt} \ln W_{kj(t-1)} + .5\hat{\sigma}^2_{kjt})$$

where $\hat{\sigma}^2_{kjt}$ denote the estimated variance of ε_{kjt}. Taking into account the variance in the estimator of β_{jt} yields yet another adjustment to the imputation:

$$(13) \qquad \hat{\hat{\hat{W}}}_{kjt} = \exp\{\hat{\beta}_{jt} \ln W_{kj(t-1)} + .5[\hat{\sigma}^2_{kjt} + [\ln(W_{kj(t-1)})]^2] \operatorname{var}(\hat{\beta}_{jt})\}$$

In actual practice, the corrections (12) and (13) made very little improvement in the imputations.

2.2. The Approach Taken in the Present Study

The earlier studies by West fit a separate regression for every distinct cell defined by the relevant variables (size, industry, etc). An alternative approach is to estimate a single model for the entire sample, but to include the relevant variables, and perhaps their interaction terms, as explanatory variables in the estimated equation. This is the approach that the missing data team adopted in the present study.

Let X_i denote the row vector of explanatory variables and let β_{jt} be the corresponding vector of coefficients in the regression equation. Then instead of (1), the model now takes the form:

$$(14) \qquad W_{ijt} = (X_i\beta_{jt}) W_{ij(t-1)} + \varepsilon_{ijt}$$

Recall that a ratio of means estimation is optimal when the variance is given by (2b), and that a mean of ratios estimation is optimal when the variance is given by (2c). Dividing equation (14) by $\sqrt{W_{ij(t-1)}}$ and $W_{ij(t-1)}$ yields equations (15) and (16) respectively:

$$(15) \quad W'_{ijt} = (X_i\beta_{jt}) W'_{ij(t-1)} + \varepsilon'_{ijt} \qquad\qquad (16) \quad W''_{ijt} = (X_i\beta_{jt}) + \varepsilon''_{ijt}$$

These are the current analogues to the transformed OLS equations (5) and (7) respectively. Imputations obtained from (15) will be referred to as ratio of means imputations, and those from (16) as mean of ratios imputations.

The team adopted a slightly different log specification than that adopted in the earlier studies cited above. The log specification adopted in the present study arises from the following model with multiplicative error term:

$$(17) \qquad W_{ijt} = \beta_{ijt} W_{ij(t-1)} \varepsilon_{ijt}$$

Taking the logarithm, rearranging terms, and assuming that $\ln(\beta_{ijt}) = X_i\beta_{jt}$, yields:

$$(18) \qquad \ln(W_{ijt}) - \ln(W_{ij(t-1)}) = \ln(X_i\beta_{jt}) + v_{ijt}$$

where $v_{ijt} = \ln(\varepsilon_{ijt})$. Imputations obtained from estimates of this equation will be referred to as log imputations.

Experimentation was also conducted with a log difference specification. Exponentiating both sides of equation (18) and taking the conditional expectation of W_{ijt} given $W_{ij(t-1)}$ and X_i yields:

$$E(W_{it}) = W_{ij(t-1)} \exp(X_i \beta_{jt}) E(\exp(v_{ijt}))$$

Letting $\hat{v}_{ijt}$ denote the actual residuals from the wage model, $E(\exp(v_{ijt}))$ can be estimated by $\bar{v}_t = \sum_{i,j} \exp(\hat{v}_{ijt})$, giving us the alternative imputation:

$$\hat{W}_{ijt} = W_{ij(t-1)} \exp(X_i \beta_{jt}) \bar{v}_t \tag{19}$$

The large sample approximation for $E(\exp(v_{ijt}))$ was not accurate enough to improve the imputations.

The following set of explanatory variables were considered: The establishment's detailed (or major) industry, the major occupation of the job, a 0-1 variable indicating whether the job is full-time or part-time indicator, a union indicator, two size indicators (small, and large), and the payroll reference date. The models also include area and ownership indicator variables. In the estimation of the regression coefficients, observations are weighted using the establishment-occupation sample weights. As a precaution against outliers, wage level values below the first and above the ninety-ninth percentile in each survey are dropped. As will be seen in the next section, imputation results were slightly better when the less detailed industry variable was used. It will also be seen in the next section that the alternative functional forms perform similarly. Note that an advantage of the log difference specification is that it tends to reduce the effect of outliers. Also, there is some evidence that the inclusion of interaction variables leads to overspecification of the model and poorer performance. In the next section results are reported for the non-interacted log specification. Letting $\hat{\beta}_{jt}$ denote the estimated coefficient vector in (18), the imputed value for a missing wage is:

$$\text{(20a)} \quad \hat{r}_{ijt} = \exp(X_i \hat{\beta}_{jt}) \qquad \text{(20b)} \quad \hat{W}_{ijt} = W_{ij(t-1)} \hat{r}_{ijt}$$

The NCS consists of a collection of distinct area surveys. Both pooled and separate models for the different areas are considered. The pooled specification includes a 0-1 indicator variable for every area, but does not allow for any interactions between the area indicator variables and the other explanatory variables in the regression model. The alternative approach where one estimates a separate regression for every area is equivalent to allowing interactions between the area variables and the other explanatory variables in the regression equation. Both pooled and separate models for private, local government, and state government jobs are also considered.

Until now we have been discussing only imputation for wages at updates. We also considered imputation at initiation. Since at initiation the most important variable for predicting the current wage, the prior time period wage, is no longer available, it is not immediately clear that regression modeling would do better than say, mean imputation, or a nonresponse adjustment factor for the entire unit. After exploring various alternatives, a regression procedure was chosen for imputing missing NCS wage data at initiation. The explanatory variables are similar to the ones used at update with payroll reference date replaced by a variable indicating the month the job is surveyed. The number of factor points the job received at initiation and its squared value are added to this set. (Each occupation is evaluated based on 10 factors, including complexity, work environment, etc. Factor points are assigned based on an aggregation of the occupation's rank within each factor.) In all of the equations, observations are weighted using the establishment-occupation sample weights from the initial survey. The alternative functional forms we considered perform similarly.

Both pooled and separate equations for the different areas are considered. Both pooled and separate equations for private, local government, and state government jobs are also considered. Again the results for the pooled and non-pooled imputations are very similar. Summary results are shown in the next section.

3. EMPIRICAL INVESTIGATIONS

As mentioned earlier, the NCS is an integration of three surveys. Establishments that are selected to provide data for the index will be reporting quarterly, whereas establishments used only in the locality publications will be reporting data yearly. Therefore, both a quarterly and an annual imputation model are required for use at update time. Note that quotes used in the quarterly index will also be used in annual locality publications. Imputations from the quarterly model will also be retained when the quote is used in a locality publication.

3.1. Data Description and Design

The study of missing wage data at update time is based on ECI private sector wage data from September 1987-March 1994. The study of missing data at initiation uses private sector data from twelve NCS area surveys.

A straightforward procedure for conducting the simulations is adopted. First, we need to select a subset of nonmissing wage observations to be treated as missing. The proportion of observations selected as missing is determined by estimating a probit model and using it to predict the probability that each observation is missing. (The same set of explanatory variables was used in both the probit model and the imputation model.) We then compare the estimated probability to a probability selected as a random draw from a uniform distribution. If the estimated probability is greater than the random draw probability then the observation is treated as missing in the simulation. A crucial assumption implicit in this procedure is that the observations that are missing in reality are truly random. After randomly designating part of the sample as missing, the remaining observations are treated as non-missing and used to estimate the various wage growth models. The resulting regression coefficients are then used to obtain imputations for the subsample that is treated as missing. To guard against an unrepresentative draw, this procedure is repeated 10 times.

For the ECI, wage level and wage growth imputations were obtained for the first quarter of 1994. The wage level imputations assume that when a quote is placed in the missing subsample, one only has nonmissing wage information in a quote's initiation period. The quote's imputed value in the first quarter of 1994 is obtained by chaining together the imputed growth rates between the quote's initiation period and the first quarter of 1994. That is, letting 0 denote a quote's initiation period and letting τ refer to the first quarter of 1994, the imputed wage in March 1994 is given by:

$$\hat{W}_{ij\tau} = W_{ijo}\hat{r}_{ij1}\hat{r}_{ij2}...\hat{r}_{ij\tau} \qquad \text{where } \hat{r}_{ijt} \text{ is defined in (20a)}$$

3.2. Evaluation Criteria

There are a several criteria that can be used to evaluate the various imputation models. One statistic of interest is mean error, which provides information on bias. A second useful statistic is mean absolute error, which provides information on the accuracy of the imputation. Letting $\hat{W}_{it}$ denote the ith quote's imputed value and letting ω_{it} denote the ith quote sample weight, the mean error and mean absolute error in imputed wages are given by:

$$(21a) \quad ME_{Wjt} = \sum_i \omega_{ijt}(W_{ijt} - \hat{W}_{ijt}) \qquad (21b) \quad MAE_{Wjt} = \sum_i \omega_{ijt}\left|W_{ijt} - \hat{W}_{ijt}\right|$$

For imputed wage growth, these measures can be written as:

$$(22a) \quad ME_{rjt} = \sum_i \omega_{ijt}(r_{ijt} - \hat{r}_{ijt}) \qquad (22b) \quad MAE_{rjt} = \sum_i \omega_{ijt}|r_{ijt} - \hat{r}_{ijt}|$$

Inaccurate wage level and growth imputations may not have much effect on the estimated ECI and ECEC, since a relatively small part of the sample is missing, and the errors in the individual imputations may tend to cancel out. However, if errors are correlated, the index imputations may be poor. In order to measure the effects of the imputations on the overall index, one can compare the true index with the imputed indices. Specifically, let $\hat{I}_s$ be the index obtained in the sth imputation. This index is calculated in the standard way, except that imputed wage levels (or growth rates) are substituted for their actual values. A measure of the bias in the imputed index is provided by (23a) below. A second statistic of interest is the average absolute percentage difference between the true index and the imputed index. This statistic provides information about how the imputations affect the precision of the index and is given by (23b). Measures similar to (23a) and (23b) can also be computed for the ECEC.

$$(23a) \quad ME_I = \left(\frac{1}{10}\right)\sum_{s=1}^{10}\left(\hat{I}_s - I_s\right) \qquad (23b) \quad MAE_I = \left(\frac{1}{10}\right)\sum_{s=1}^{10}\frac{\left|\hat{I}_s - I_s\right|}{I_s}$$

3.3. Imputation Results

The first set of results is for nonrespondents at update for the quarterly ECI. Then the results for the annual update in the NCS Locality estimates will be shown. Finally, the last set of results shown will be for imputing for nonrespondents at initiation. Since initiation is handled in the same manner for the two groups only one set of studies was needed for initiation. The following notation will be used in describing the selected procedures. Recall,

W_{ijt} = reported wage in period t for quote i in cell j
$r_{ijt} = \left(W_{ijt}/W_{ij(t-1)}\right)$

The independent variables considered are denoted as:

MID_i = major industry division for quote I	UNION_i= indicator variable, denoting whether job i is union
MOG_i = major occupation group for quote I	FPTP_i = indicator variable, denoting full or part-time job
SIZE_i= size indicator based on employment	REGION_I = region indicator

The equation governing wage growth is assumed to take the form:

$$(24) \qquad \ln\left(r_{ijt}\right) = \beta_{0t} + \beta_{1t}MID_i + \beta_{2t}MOG_i + \beta_{3t}FTPT_i + \beta_{4t}UNION_i + \beta_{5t}SIZE_i + \beta_{6t}REGION_i + \varepsilon_{ijt}$$

where ε_{it} denotes an error term that has mean 0, a homoscedastic variance, and is uncorrelated with the independent variables. The imputed value for a missing $\hat{r}_{it}$ is given by:

$$(25) \qquad \hat{r}_{ijt} = \exp\left(\beta_{0t} + \beta_{1t}MID_i + \beta_{2t}MOG_i + \beta_{3t}FTPT_i + \beta_{4t}UNION_i + \beta_{5t}SIZE_i + \beta_{6t}REGION_i\right)$$

where it has been assumed that $E\left(\exp\left(\varepsilon_i\right)\right) \approx 1$. The imputed value for a missing wage is given by:

$$(26) \qquad \hat{W}_{ijt} = W_{ij(t-1)}\hat{r}_{ijt}$$

3.3.1 Quarterly Updates for the ECI

Table 1 summarizes the private sector wage growth imputations for the first quarter of 1994. The data in Table 1 are for averages over 10 iterations. The "Average MAE" line presents results for percent change imputations (eq. (22b)). All wage data in this table and Table 2 are in terms of dollars per hour. Column 2 presents this statistic when the regression includes all main effects and column 3 for a regression that only includes a constant term. The finding that the full regression yields about the same accuracy of imputations as the regression with a constant term reflects the fact that the explanatory variables do not do a very good job of explaining wage growth. Finally, column 4 shows the results for a fully interacted regression model. Clearly, adding the interaction effects does not improve the accuracy of the imputations.

Inaccurate wage growth imputations do not necessarily imply high variance in the estimated ECI since the errors in imputed wage growth may tend to cancel out. The results in Table 1 indicate that this is indeed the case. Referring to the row marked "Average ECI", column 1 of Table 1 presents the actual change in the ECI during the first quarter of 1994. Column 2 of the same row presents the average estimated private sector ECI when wage growth imputations are obtained from the expression in equation (25). The average percent difference from the actual ECI is quite small. Columns 3 and 4 present the relevant data for cases where the regression only includes a constant, and for a fully interacted model, respectively.

Table 1. Growth and ECI Imputations

Log Specification, Averages for 10 Iterations

		Main Effects			**Constant Term Only**			**Fully Interacted Model**		
	Actual Values (1)	**Value (2)**	**Percent Diff.**	**Absolute Percent Diff.**	**Value (3)**	**Percent Diff.**	**Absolute Percent Diff.**	**Value (4)**	**Percent Diff.**	**Absolute Percent Diff.**
Average MAE[1]		3.127			3.124			3.414		
Average ECI	0.6375	0.6353	-0.354	7.131	0.632	-0.867	6.307	0.671	5.217	6.137

1) Units in this row are in terms of percentage change over the first three months of 1994.

Table 2 summarizes the wage level imputations. The data in Table 2 are for averages over 10 iterations. The "Average MAE" line presents results for level imputations (eq. (21b)) for our chosen model with main effects only. The average MAE figure of 1.034 is about 6 times higher when we construct imputations from a wage level regression where the only explanatory variable is a constant term. This result, taken together with the results of Table 1, indicate that the wage level imputations are more successful than the wage growth imputations. This reflects the fact that a quote's past wage is helpful in predicting its current wage.

Table 2 also compares the imputed private sector ECEC with the actual private sector ECEC in the row marked "Average ECEC". The first column presents the actual ECEC in the first quarter of 1994. Column 2 presents the estimated ECEC when wage is imputed using (26). As with the ECI, the error in the ECEC imputation is much smaller than the error in the individual wage imputations. Furthermore, the average absolute percent difference from the ECEC is much smaller than that for the ECI. This finding that the imputed ECEC is more accurate than the imputed ECI is consistent with our result above that we are able to impute wage levels more accurately than wage growth rates.

Table 2. Level and ECEC Imputations

Log Specification, Averages for 10 Iterations

		Main Effects		
	Actual Values (1)	**Value (2)**	**Percent Diff.**	**Absolute Percent Diff.**
Average MAE[1]		1.034		
Average ECEC	13.049	13.032	-0.130	0.130

Tables 1 and 2 present results for the log specification. The alternative functional forms performed similarly, although there is some evidence that the specification where the ratio of the current to the previous wage is the dependent variable may be more sensitive to outliers. An advantage of the log difference specification is that it tends to reduce the effect of outliers.

Two alternative approaches for handling public sector jobs were considered. The first approach involves simply pooling the public sector jobs with the private sector jobs. This approach is reflected in equation 25 by use of MID variable. The MID is a broad grouping of industries. Separate MID code was assigned to private, State, and local government industry groupings. For example, schools are all part of services industry, but private schools were assigned a different services industry code then State or local schools. The second approach involves estimating separate regression equations for private, State, and local government jobs using the industry groupings. The approaches performed about the same. For example, the mean absolute error in the wage level predictions is 1.066 for the pooled approach and 1.073 when separate equations are estimated for the different sectors.

3.3.2 Annual Updates for the NCS Locality Estimates

The results for imputation at update for the yearly NCS are similar to the results for the ECI. Again, the log wage growth equation was used to impute for missing wage levels for 10 iterations of the simulation. The average MAE when the only explanatory variables are the month of survey, interval between surveys, and dummy variables for each area, is 1.032 (units are in dollars). When the areas are pooled and the proposed set of explanatory variables is used this figure is 1.039. Finally, when separate equations are estimated for each area with the same set of explanatory variables the average MAE is 1.044. The wage imputations are able to explain some of the variation in wages, but a great deal clearly remains unexplained. This reflects the fact that it is very difficult to predict wage growth.

The results also indicate the regressions used for the imputations have little explanatory power: adding covariates to the specification that only includes a constant term does not reduce the mean absolute error. Further analysis suggests, however, that using wage data from the previous survey yields a substantially better imputation than does a procedure, such as the current NCS occupational nonresponse adjustment, that does not use this information. When the imputations for the update survey come from a log wage equation and do not utilize information on a quote’s wage in the prior survey, the mean absolute difference is above three dollars. This is not very close to the one dollar figure obtained when the prior wage is used. These results are consistent with the results obtained in Lettau and Loewenstein (1997). Our current study does not consider the case when some of the imputed quotes also had imputed wage data for the previous interview. However, the results in Lettau and Loewenstein (1997) indicate that, although the quality of wage level imputations decreases with the amount of time since the last wage data were collected, a wage level imputation that revises previous wages by imputed wage growth is still superior to a direct wage level imputation that does not use prior wage information.

3.3.3 Imputation for Missing Wages at Initiation

The results for imputing for missing wages at initiation are now presented. The equation governing wages is:

$$\begin{aligned} W_{ijt} &= \beta_{0t} + \beta_{1t} MID_i + \beta_{2t} MOG_i + \beta_{3t} FTPT_i + \beta_{4t} UNION_i + \beta_{5t} SIZE_i + \beta_{6t} REF_i \\ &\qquad + \beta_{7t} AREA_i + \beta_{8t} FACPTS_i + \beta_{9t} FACPTS_i^2 + \varepsilon_{ijt} \end{aligned} \tag{27}$$

where ε_{ijt} denotes an error term that has mean 0, a homoscedastic variance, and is uncorrelated with the independent variables. The new variables in this model are defined as:
REF$_i$ = payroll reference date for quote i
AREA$_i$ = indicator variable for area
FACPTS = the number of factor points associated with the job.

The effects of using the wage level equation to impute for “missing” wage levels are summarized by comparing averages of MAE data for 10 iterations of our simulation. The team's recommended model has area, major occupation, industry, size, reference date, union, factor points, and factor points squared are the explanatory variables, and it’s dependent variable is the wage rate in its original units. We note that the addition of factor points and factor points squared adds significantly to the explanatory power of all the regressions considered. The average MAE for the recommended model is 3.85640. (This figure is measured in dollars and compares to a mean level of wages that is a bit above $15 with a standard deviation a bit above $10.) The use of a log wage model yields an average MAE of 3.89792. Finally, the average MAE is 3.88663 when one uses a log wage imputation with the addition of a correction taking into account that $E(\exp(v_i)) \neq 1$. Note that all three imputations perform similarly and, as expected, the error is much larger than for imputation at updates.

The recommended model pools quotes over the public and private sectors. The team investigated estimating separate equations for state, local, and private sector quotes. The average MAE of 3.85640 for the recommended model is very close to the average MAE of 3.78094 when separate equations are estimated. Pooling by ownership does not lead to a loss in accuracy, as all of the imputations perform similarly.

4. CONCLUSIONS FOR IMPUTATION FOR THE WAGE VARIABLE

After exploring various alternatives, the missing data team has chosen the following procedure for imputing missing NCS wage data at initiation. First, a regression model, where observations are weighted using the establishment-occupation sample weights, is estimated in which the dependent variable is a quote's current quarter wage and the independent variables are the set listed in Section 3. In estimating the regression coefficients, wage level outliers below the first and above the ninety-ninth percentile in each survey are dropped. The estimated coefficients from the regression model are used to impute for a quote's wage level when it is missing.

The recommended wage imputation requires that there is information on all variables other than wages. The team recommends that observations where other variables in addition to the wage are missing be handled using a weight adjustment for nonresponse. This recommendation is based on the consideration that there are not sufficiently many cases to justify a more complicated procedure that estimates different regression equations using different sets of explanatory variables.

The NCS data consists of both data that are collected quarterly and data that are only collected annually. The team proposes that missing wages in the quarterly data be imputed using just the good quarterly data, while missing wages in the annual data be imputed using all of the valid wage observations – both quarterly and annual.

Also, the team compared the imputations obtained when the separate localities are pooled together with imputations obtained when separate regressions are estimated for each locality. The team found that the differences in the estimates obtained from the two approaches are negligible and consequently decided on the pooling approach on the grounds that it is simpler, even though there are potential disadvantages. (For discussion on this see [3].)

Similarly, a regression model was chosen for wage imputation at post initiation time. In this situation the functional form chosen is the log of the ratio of the current to prior wages. The independent variables in the model are the ones discussed in Section 3. Note that in this situation the independent variables were not that helpful, which is different than at initiation, where they were definitely useful in predicting wage levels.

The proposed procedure does not distinguish between temporary and permanent non-respondents. The decision to keep permanent non-respondents in the sample and impute for their missing wages is based on the finding by Lettau and Loewenstein (1997) that a quote's past wage is useful for predicting its current wage far into the future. It should also be noted that permanent non-respondents belong in the sample only if they represent refusals and not if they represent deaths. The latter represent jobs that no longer exist and thus, ideally, should be dropped from the sample. For this purpose, the team recommends that a check be made as often as possible to determine whether businesses coded as refusals are still in business.

References

(1) BLS Handbook of Methods, April 1997.

(2) OCWC Missing Data Team. 1999. "Post-Initiation Imputation for Missing ECI Wage Data."

(3) OCWC Missing Data Team. 1999. "Post-Initiation Imputation for Missing NCS Wage Data."

(4) OCWC Missing Data Team. 1999. "Initiation Imputation for Missing NCS Wage Data."

(5) Lettau, Michael K. and Mark A. Loewenstein. 1997. "Imputation in the ECI." Unpublished paper, Bureau of Labor Statistics.

(6) Ponikowski, Chester H. "ECI Wage Imputation Study." Unpublished paper, Bureau of Labor Statistics.

(7) West, Sandra A. 1983. "A Comparison of Different Ratio and Regression-type Estimators for the Total of a Finite Population," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 388-393.

(8) West, Sandra A., Butani, Shail, and Witt, Michael. (1991), "Alternative Imputation Methods for Employment, Wage and Ratio of Wage to Employment Data," *Proceedings of the 78th Indian Science Congress, India*.(Also *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 254-259)