VARIANCE ESTIMATORS FOR VARIABLES THAT HAVE BOTH OBSERVED AND IMPUTED VALUES

Sandra A.West, Diem-Tran Kratzke, and Kenneth W. Robertson, Bureau of Labor Statistics
Sandra A. West, 2 Massachusetts Ave. N.E., Washington, D.C. 20212

KEY WORDS: Imputation, Mean, Regression, Hot Deck, Multiple Imputation.

1. Introduction

We will present the results of theoretical and empirical investigations of different variance estimators in the presence of imputed and observed values in this paper. It is assumed that all the missing data are imputed by the same method. Imputation methods considered include mean, hot deck, regression, regression plus residual, and multiple imputation. Variance estimators considered include the standard, two versions of the jackknife, and random groups.

The data are employment from the Bureau of Labor Statistics' Universe Data Base (UDB). The UDB is a sampling frame of business establishments that is constructed from the State's Unemployment Insurance (UI) micro data file. The information used to maintain this file is obtained from quarterly UI reports which each employer is required to submit. Although the filing of the UI report is mandatory, there are always some late, incomplete, or missing reports. In previous studies, a single imputation procedure was developed that worked well for all industries within each State. For this study, the recommended imputation method and several alternatives will be considered. The actual data for non-repondents were never obtained. Thus non-response had to be simulated using the patterns of non-response observed on the files. For the most part, it was assumed that the non-respondents were missing at random. In addition, a fixed non-response rate was simulated in order to see the effect on the variance estimators when a large part of the sample is imputed.

In Section 2, we describe the data sets used and the design of the empirical investigations. The notation and evaluation criteria that are used to compare the various methods are presented in Section 3. Descriptions of the imputation methods and their properties are presented in Section 4 and Section 5, respectively. In Section 6, alternative variance estimators to the standard estimator are considered. The results of the empirical investigations are showed in Section 7, along with observations and conclusions. Future research is discussed in Section 8.

2. Data and Design of Empirical Investigation

Two months of UDB data were used for this study: December 1991 and January 1992. A unit (establishment) is classified as a non-respondent if it has not reported employment data for the current month.

Industries are classified on the UDB by a Standard Industrial Classification code (SIC). A 2-digit SIC represents a broad industry classification, with 3- and 4-digit SIC codes representing narrower industry definitions. As may be expected, many industry characteristics become more homogenous as we move from 2- to 3-digit SIC stratification.

We obtained data from Michigan in these industries (2-digit SIC code is in parenthesis): Agricultural Services (07), Lumber and Wood Products (24), Transportation Equipment (37), Trucking and Warehousing (42), Transportation Services (47), General Merchandise Stores (53), Apparel and Accessory Stores (56), Miscellaneous Retail (59), Nondepository Credit Institutions (61), Miscellaneous Repair Services (76), Membership Organizations (86), and Private Households (88).

Intuitively, an establishment's employment data are correlated with its own past employment and with the employment of similar establishments. If establishments are placed into strata based on characteristics related to employment, then the more homogenous the strata are, the higher the correlation will be. Within each 2-digit SIC chosen, we stratified the data further by (1) 3-digit SIC/county and (2) 3-digit SIC/size class.

Usually a measure of size is created for each establishment based on its most recent reported monthly employment. This was done in our study. Size classes were formed as follows:

Size Class 1 - Employment < 50
Size Class 2 - $50 \leq$ Employment < 250
Size Class 3 - Employment ≥ 250

After some initial results, we increased the number of size classes, as most units fell in the original Size Class 1. The original Size Class 1 was sub-divided as follows.

Size Class 1a - Employment < 5
Size Class 1b - $5 \leq$ Employment < 10
Size Class 1c - $10 \leq$ Employment < 20
Size Class 1d - $20 \leq$ Employment < 50

For our study we used two non-response patterns. In the first we simulated the pattern of non-response observed in the data as much as possible. If a particular industry had x% of imputed employment, then a non-response rate of x% was used. It was assumed that the missing data mechanism was ignorable, and a random set of units were chosen to represent the set of non-respondents. The second non-response pattern assumed that each industry had observed a 25% non-response rate.

For the empirical study, we allowed only continuous single units from private industries. Continuous units are units that existed on the file during the previous quarter. Single units are units that have only a single establishment. After discarding units which did not meet these requirements, we then determined the actual non-response rate within each industry. After which, all non-respondents were removed from the data set. Using this reduced data set, units were systematically placed, after a random start, into a Model set and a Test set based on the

chosen non-response pattern. All imputation methods used data from the Model set to determine parameters which were then applied to the units in the test set.

3. Notation and Evaluation Criteria

Notation

For a given 2-digit SIC let

$E_{j,t}$ denote the employment for unit j in month t,

$\hat{E}_{j,t}$ denote the predicted employment for unit j in month t,

B_t denote the set of units that have reported employment for months t and month t-1,

nr_t denote the percentage of units in month t that have imputed employment values,

NR_t denote the set of non-respondents that were obtained by randomly selecting the percentage nr_t of units from the set B_t (Test set.),

BR_t denote the set of units in B_t - NR_t (Model set.),

NNR_t denote the number of elements in NR_t

NBR_t denote the number of elements in BR_t.

Also let

V_t denote the variance of the employment variable for establishments in B_t; that is, the "true" variance,

$\hat{V}_{t,m,i}$ denote the estimator of V_t using variance method m and imputation method i, where i = 0 denotes no imputation and the variance estimator is based only on the respondents.

The following notation will be used for the different methods of computing the variance:

m = 1 - standard method, denoted by SD

m = 2 - jackknife A, denoted by JA,

m = 3 - jackknife B, denoted by JB,

m = 4 - random groups, denoted by RG.

The following notation will be used for the different methods of imputation:

i = 1 - stratum mean,

i = 2 - carry over,

i = 3 - hot deck nearest neighbor,

i = 4 - recommended regression,

i = 5 - as in i=4 plus residual,

i = 6 - as in i =4 plus multiple residuals.

Evaluation Criteria

Letting $\varepsilon_{m,i} = \hat{V}_{t,m,i} - V_t$ denote the error for variance method m and imputattion method i, then the Percent Relative Absolute Error will be used:

$$RAE_{m,i} = 100\ |\varepsilon_{m,i}|\ /V_t.$$

Note that the imputations were done by 3-digit SIC/county or 3-digit SIC/size class, but the variances were computed over the entire 2-digit SIC.

4. Imputation Methods

In a previous study by West, et al (1989), 32 methods of imputation with three sample designs were considered. The recommended method from the previous study and several commonly used methods will briefly be described.

Mean

The mean imputation method is a common method of imputation in many surveys, especially for those 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. For any fixed stratification, month t, employment is imputed as follows:

$$\hat{E}_{k,t} = \sum_{j \in BR_t} E_{j,t} \Big/ NBR_t \ , \ for\ all\ k \in NR_t \ .$$

Thus $\hat{E}_{k,t}$ is equal to the average of the monthly employment of all respondents in the stratum.

Carry-Over

Under the carry over method, each non-respondent's employment is imputed using its own history. The predicted value is therefore independent of size class and industry. It is computed as follows:

$$\hat{E}_{k,t} = E_{k,t-s},\ for\ all\ k \in NR_t \ .$$

where $s \geq 1$ and $t-s$ denotes the last time in which an employment value was reported for the establishment. (In the paper only s=1 is used.)

Hot Deck-Nearest Neighbor

For any fixed stratification, month t, let k denote a non respondent and c denote a respondent such that

$$\left|E_{c,t-1} - E_{k,t-1}\right| \leq \left|E_{j,t-1} - E_{k,t-1}\right| \ for\ all\ j \in BR_t.$$

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

For any particular non-respondent, this method selects the respondent that appears closest to the non-respondent in an ordered list, and substitutes the respondent's monthly employment value for the non-respondent's.

Regression Model

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 j in month t, given the vector of E-values (employment in month t-1 reported by units in set BR_t):

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

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

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

where β is some constant depending on t.

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

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

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

The model can be rewritten as:

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

where $\quad E\{\varepsilon_{j,t}\} = 0,$

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

In the previous studies, it was found that the model:

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

worked reasonably well for employment data. Thus the predicted employment value at time t is:

$$\hat{E}_{k,t} = \hat{\beta} E_{k,t-1} \text{ , for all } k \in NR_t \text{ .}$$

where $\quad \hat{\beta} = \sum_{j \in BR_t} E_{j,t} \Big/ \sum_{j \in BR_t} E_{j,t-1}$.

Adding Residuals to the Regression Model

The regression method could be thought of as imputing for missing employment by using the mean of the predicted E_t distribution, conditional on the predictors E_{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, drawing residuals r_k with mean zero to add to $\hat{E}_{k,t}$.

In the earlier studies, the residuals were chosen in three ways. For this study the residuals will be chosen from a normal distribution with mean zero and variance obtained from the model. Thus the predicted employment value at month t is imputed as:

$$\hat{E}_{k,t} = \hat{\beta} E_{k,t-1} + s\delta_k \text{ , for all } k \in NR_t \text{ .}$$

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

A slight modification of the previous method was obtained by drawing five random numbers and using the average value for the added residual. That is,

$$\hat{E}_{k,t} = \hat{\beta} E_{k,t-1} + s\bar{\delta} \quad \text{where} \quad \bar{\delta} = \sum_{k=1}^{5} \delta_k \Big/ 5.$$

5. Effects of Imputation on Standard Variance Estimator

Consider the population variance for a given 2-digit SIC at month t:

$$V_t = \sum_{j \in B_t} \left(E_{j,t} - \bar{E}\right)^2 \Big/ (NBR_t + NNR_t) \qquad (5.1)$$

where $\bar{E} = \sum_{j \in B_t} E_{j,t} \Big/ (NBR_t + NNR_t)$.

The variance can be rewritten as:

$$V_t = \left[\sum_{j \in BR_t} \left(E_{j,t} - \bar{E}\right)^2 + \sum_{j \in NR_t} \left(E_{j,t} - \bar{E}\right)^2 \right] \Big/ (NBR_t + NNR_t). \qquad (5.2)$$

Assuming that the missing data are missing at random, consider the effects of using imputation method i on V_t. First consider overall mean imputation, that is, i=1 with one stratum. In this situation, formula (5.2) become:

$$\hat{V}_{t,1,1} = \left[\sum_{j \in BR_t} \left(E_{j,t} - \hat{\bar{E}}\right)^2 + \sum_{k \in NR_t} \left(\hat{E}_{k,t} - \hat{\bar{E}}\right)^2 \right] \Big/ (NBR_t + NNR_t) \qquad (5.3)$$

where $\hat{\bar{E}} = \left(\sum_{j \in BR_t} E_{j,t} + \sum_{k \in NR_t} \hat{E}_{k,t} \right) \Big/ (NBR_t + NNR_t)$.

This method creates a spike in the employment distribution, since all the missing values are assigned the same value, the mean of the respondents, that is, $\hat{E}_{k,t} = \sum_{j \in BR_t} E_{j,t} / NBR_t$ *for all* $k \in NR_t$. The second term in (5.3) becomes zero since $\hat{E}_{k,t} = \hat{\bar{E}}$ resulting in the following variance estimator:

$$\hat{V}_{t,1,1} = \sum_{j \in BR_t} \left(E_{j,t} - \hat{\bar{E}}\right)^2 \Big/ (NBR_t + NNR_t) = \frac{(NBR_t - 1)}{(NBR_t + NNR_t)} S^2$$

where $S^2 = \sum_{j \in BR_t} \left(E_{j,t} - \hat{\bar{E}}\right)^2 \Big/ \left((NBR_t - 1)\right)$.

Since S^2, which is $\hat{V}_{t,0,1}$, is an unbiased estimator of V_t,

$$E(\hat{V}_{t,1,1}) = \frac{(NBR_t - 1)}{(NBR_t + NNR_t)} V_t$$

and hence,

$\dfrac{E(\hat{V}_{t,1,1})}{V_t} = \dfrac{(NBR_t - 1)}{(NBR_t + NNR_t)}$ is approximately equal to the expected response rate.

Note that the relative bias is approximately equal to minus the expected non-response rate:

$$\frac{E(\hat{V}_{t,1,1}) - V_t}{V_t} = -\frac{(NNR_t + 1)}{(NBR_t + NNR_t)} .$$

Next consider the case of mean imputation within strata; this method produces a series of spikes in the employment distribution at the means of the imputation strata. Let $\bar{E}_h$ denote the mean of the respondents in stratum h which has $NNR_{t,h}$ missing values, then the variance estimator can be written as:

$$\hat{V}_{t,1,1} = \left[\sum_{j \in BR_t} \left(E_{j,t} - \bar{E}_p\right)^2 + \sum_{h=1}^{H} NNR_{t,h} \left(\bar{E}_h - \bar{E}_p\right)^2 \right] \Big/ (NBR_t + NNR_t)$$

where H is the number of strata and,

$$\bar{E}_p = \left[\sum_{j \in BR_t} E_{j,t} + \sum_{h=1}^{H} NNR_{t,h} \bar{E}_h \right] \Big/ (NBR_t + NNR_t) .$$

which can be written as:

$$\bar{E}_p = \left[NBR_t\,\bar{E}_r + NNR_t\,\bar{E}_{w,h}\right]/(NBR_t + NNR_t), \text{ where}$$

$$\bar{E}_{w,h} = \sum_{h=1}^{H} NNR_{t,h}\bar{E}_h \Big/ NNR_t, \text{ since } NNR_t = \sum_{h=1}^{H} NNR_{t,h}.$$

And hence the variance estimator can be written as:

$$\hat{V}_{t,1,1} = \frac{(NBR_t - 1)}{(NBR_t + NNR_t)} S_p^2 + \frac{(NNR_t - 1)}{(NBR_t + NNR_t)} S_h^2$$

where $S_p^2 = \left[\sum_{j \in BR_t} (E_j - \bar{E}_p)^2\right] \Big/ (NBR_t - 1)$,

$$S_h^2 = \sum_{h=1}^{H} NNR_{t,h}(\bar{E}_h - \bar{E}_p)^2 \Big/ (NNR_t - 1).$$

Thus, the relative bias of $\hat{V}_{t,1,1}$ is approximately:

$$\frac{E(\hat{V}_{t,1,1}) - V_t}{V_t} \approx -\frac{(NNR_t)}{(NBR_t + NNR_t)}\left[1 - \frac{E(S_h^2)}{V_t}\right]$$

where $E(S_h^2)/V_t$ is the proportion of the variance explained by the imputation strata.

Similar results are obtained for imputation methods 2-4. For example, the formula for method 4 has the proportion of the variance explained by the regression. The predicted regression method curtails the spread of the employment distribution.

The random regression methods 5 and 6 for imputation adjust the employment distribution for the missing cases and retain the residual variability exhibited in the respondents' data. (In all these cases it is assumed that respondents always respond over conceptually repeated applications and non-respondents never do.)

In summary, the deterministic imputation methods (methods 1-4) distort the distribution and attenuate the variance, whereas the stochastic imputation methods (methods 5-6) yield approximately unbiased estimates of the distribution and the variance. In general for means, all the methods lead to at least approximately unbiased estimators.

6. Alternative Variance Estimators

In the empirical study three alternative estimators for the variance were considered: Two jackknife versions and a random groups method.

First consider the random groups method. Each unit was randomly assigned into a group g, where there are G random groups. (In this paper, G=20 was used). The random group estimator is defined as:

$$\hat{V}_{t,4,i} = \sum_{g=1}^{G} \hat{V}_{t,4,i,g} \Big/ G$$

where $\hat{V}_{t,4,i,g}$ is the standard variance estimator for group g.

The first jackknife estimator (jackknife A) was obtained by adding the jackknife estimator of bias to the standard estimator of the variance. First, the standard variance estimator is computed for all units except those in group g, which will be denoted by $\hat{V}_{t,2,i,(g)}$, for all G groups. The jackknife A estimator is defined as:

$$\hat{V}_{t,2,i} = G\hat{V}_{t,1,i} - (G-1)\hat{V}_{t,2,i(\cdot)}$$

where $\hat{V}_{t,1,i}$ is the standard estimator in (5.3), and

$$\hat{V}_{t,2,i,(\cdot)} = \sum_{g=1}^{G} \hat{V}_{t,2,i,(g)} \Big/ G.$$

To compute the jackknife B estimator, the jackknife A estimator of the variance of the mean was multiplied by the population size. Let $\hat{\bar{E}}_g$ denote the mean estimator of the population mean computed with only units in group g and $\hat{\bar{E}}_{(g)}$ denote the mean estimator of the population mean computed without units in group g, then the jackknife B estimator is defined as:

$$\hat{V}_{t,3,i} = NB_t \sum_{g=1}^{G} \left(\tilde{\bar{E}}_g - \tilde{\bar{E}}_{(\cdot)}\right)^2 \Big/ G(G-1)$$

where $\tilde{\bar{E}}_g = G\hat{\bar{E}}_g - (G-1)\hat{\bar{E}}_{(g)}$

and $\tilde{\bar{E}}_{(\cdot)} = \sum_{g=1}^{G} \tilde{\bar{E}}_g \Big/ G.$

7. Results / Conclusions

Tables 1 and 2 show the errors in computing variances using the standard variance estimator. Notation:

V_t = VAR, $NBR_t + NNR_t$ = N, $\hat{V}_{t,1,3+i}$ = REGi, i=1,2,3,

$\hat{V}_{t,1,1}$ = MEAN, $\hat{V}_{t,1,2}$ = CARRY, and $\hat{V}_{t,1,3}$ = NEAR.

Table 1.

Percent Relative Absolute Error incurred in Standard Variance Estimator due to Imputation

Stratified by 3 digit SIC/**county**. Non-response rates: as observed (OB) which is 3%-8% and fixed rate of 25%.

Nonresponse Rate: As observed on file=OB								
SIC	VAR	N	REG1	REG2	REG3	MEAN	CARRY	NEAR
7	256.15	1614	0.84	0.84	0.86	0.97	0.14	0.77
24	757.13	761	0.03	0.02	0.03	0.19	0.02	0.05
37	40954.39	503	0.10	0.10	0.10	0.45	0.16	0.63
42	1300.66	1836	2.64	2.65	2.66	1.56	3.53	1.95
47	1006.08	785	0.13	0.10	0.11	0.92	0.29	0.13
53	7711.62	262	0.01	0.01	0.01	0.44	0.01	0.01
56	3903.65	1622	1.10	1.09	1.11	2.42	0.79	2.37
59	2659.32	6099	3.39	3.38	3.39	65.24	1.33	64.94
61	15265.53	302	0.00	0.00	0.00	0.07	0.00	0.01
76	131.41	1459	0.97	1.03	0.96	2.95	1.37	0.36
86	921.87	2871	5.67	5.67	5.67	22.09	1.91	21.59
88	8.17	1495	1.47	1.59	1.59	4.53	1.59	1.35

Nonresponse Rate: 25%								
SIC	VAR	N	REG1	REG2	REG3	MEAN	CARRY	NEAR
7	255.67	1562	6.50	6.34	6.45	15.96	4.42	14.01
24	610.17	690	3.36	3.17	3.16	5.44	0.26	3.86
37	42829.52	470	0.12	0.12	0.12	1.44	0.77	1.70
42	1313.77	1816	0.62	0.55	0.59	19.52	1.90	9.13
47	1024.35	756	2.60	2.63	2.71	19.11	4.17	1.36
53	7964.20	223	2.27	2.28	2.28	5.84	3.88	1.17
56	4130.02	1530	3.23	3.27	3.29	30.20	11.14	25.18
59	2708.71	5975	0.81	0.81	0.81	4.82	0.13	3.09
61	16679.83	275	24.98	25.04	24.98	66.93	1.74	66.59
76	133.76	1428	2.86	3.22	2.98	16.60	5.24	9.52
86	951.31	2752	0.47	0.46	0.47	12.14	0.51	9.09
88	8.21	1483	2.80	4.14	2.92	16.32	1.46	1.71

Note that the stratification is used only in the imputation, not in calculating the variances. Also, the population size N changed in some SICs between the two response rates, because certain observations could not be used due to the requirements of certain imputation procedures.

Table 2.

Absolute Percent Errors incurred in Standard Variance Estimator due to Imputation

Stratified by 3 digit SIC/**size classe** (3 size classes). Non-response rates: as observed (OB) and 25%.

Nonresponse Rate: As observed on file=OB								
SIC	VAR	N	REG1	REG2	REG3	MEAN	CARRY	NEAR
7	252.43	1628	0.25	0.24	0.25	0.70	0.08	0.02
24	773.47	788	0.66	0.45	0.59	2.20	0.43	2.16
37	40728.98	506	0.05	0.05	0.05	0.35	0.16	0.24
42	1297.34	1841	2.78	2.77	2.79	3.63	3.53	3.13
47	1004.85	786	0.03	0.01	0.06	0.23	0.29	0.02
53	7495.46	270	0.01	0.01	0.00	0.04	0.01	0.01
56	3869.22	1637	0.74	0.79	0.75	2.06	0.79	0.62
59	927.41	6115	0.08	0.10	0.10	0.00	0.24	1.53
61	15265.53	302	0.00	0.00	0.00	0.02	0.00	0.00
76	130.67	1469	0.62	0.48	0.59	3.59	1.34	0.51
86	719.30	2879	0.01	0.03	0.02	0.13	0.05	0.04
88	8.16	1498	1.59	2.08	1.72	4.66	1.59	1.84
Nonresponse Rate: 25%								
SIC	VAR	N	REG1	REG2	REG3	MEAN	CARRY	NEAR
7	251.39	1626	6.79	5.85	7.06	8.93	3.29	17.42
24	607.01	787	2.39	1.92	2.22	3.57	1.07	1.65
37	40907.09	503	0.42	0.42	0.42	7.14	0.83	0.11
42	1297.34	1841	2.68	2.48	2.58	2.08	1.89	3.05
47	921.11	785	2.07	1.83	1.95	3.41	3.10	3.17
53	6701.98	267	2.47	2.47	2.47	5.13	3.84	1.15
56	3521.63	1634	1.88	1.83	1.89	3.78	13.85	1.11
59	2635.43	6116	1.52	1.49	1.50	2.02	0.20	1.32
61	15409.17	299	1.77	1.76	1.62	64.54	1.73	62.53
76	130.67	1469	1.52	1.84	1.67	12.79	5.32	1.61
86	895.19	2878	0.62	0.51	0.60	0.99	0.73	1.41
88	8.16	1498	1.35	2.82	1.72	17.40	1.35	0.37

Observations from Table 1

for OB%: REG1-3 and CARRY do well; both MEAN and NEAR can produce very large errors.

for 25%: REG1-3 and CARRY do well, however there is a large error for REG1-3 and for CARRY. Both MEAN and NEAR can produce very large errors.

Observations from Table 2

for OB%: REG1-3, for the most part, produce the smallest errors; however all the methods do fairly well. There are no large errors for MEAN and NEAR as in Table 1.

for 25%: REG1-3 do the best, there are no large error as in Table 1. CARRY, MEAN, and NEAR can produce large errors.

As one would expect, the errors, for the most part, are larger with 25% than with OB%.

County vs. Size Class Stratification

for OB%: Size class stratification produced smaller errors than county stratification, with the biggest improvements in the MEAN and NEAR methods. The maximum error of variance by REG methods became smaller by size class.

for 25%: Size class stratification did not produce the large error by REG methods as with county stratification. It also produced fewer number of large errors for MEAN and NEAR. For CARRY, there were no big differences between size class and county stratifications.

Note that outliers in an imputation cell formed by county are more likely to occur than in an imputation cell formed by size class. Thus, it is not surprising that larger errors were produced in the variances when the imputation was done by county.

In summary, if the standard variance formula is used, then the imputation method that least disturbs the population variance is one of the regression types. The simplest regression type which is the single model with no residual added should be used, and stratification should be by 3-digit SIC/size class. This method is robust for different response rates, and resulting error measures are relatively small.

Table 3 shows the errors in computing the variances using different variance methods. The stratification was done by 3-digit SIC/6 size classes, and only the 25% non-response rate was considered. Also, only the regression model with no residual added was considered for regression types. For m=1,2,3,4, the following notion is used: $\hat{V}_{t,1,0}$ = RespV, $\hat{V}_{t,m,4}$ = Rm (REG1), $\hat{V}_{t,m,1}$ = Mm (MEAN), $\hat{V}_{t,m,2}$ = Cm (CARRY), $\hat{V}_{t,m,3}$ = NNm (NEAR).

Table 3.

Absolute Percent Errors incurred in 4 Variance Estimators due to Imputation

Stratified by 3 digit SIC/**size classe** (6 size classes). Non-response rate: 25%.

SIC	VAR	N	R1	R2	R3	R4	M1	M2	M3	M4	RespV
7	251	1623	6.97	6.97	18.29	9.05	6.24	6.49	12.09	0.09	5.75
24	608	785	2.44	2.32	23.12	7.32	6.43	6.61	49.84	5.03	8.60
37	40907	503	0.41	0.28	29.74	1.74	7.23	8.93	23.19	14.64	25.00
42	1298	1840	2.75	2.76	51.74	1.79	3.06	2.96	30.69	4.14	5.83
47	921	785	2.09	1.95	0.37	6.39	2.10	2.11	0.29	0.48	12.98
53	6702	267	2.50	2.54	18.84	6.77	5.37	5.59	24.68	8.34	25.74
56	3522	1634	1.88	2.14	30.60	5.28	4.08	3.90	44.49	9.35	1.73
59	2636	6116	1.54	1.52	21.94	1.97	1.66	1.61	17.79	2.56	26.80
61	15360	300	1.77	1.60	15.67	7.16	64.41	64.29	71.73	67.04	56.49
76	131	1468	2.04	1.85	38.18	6.47	4.44	4.51	8.48	2.28	8.86
86	859	2877	0.08	0.33	42.50	6.59	0.97	0.54	50.49	7.77	22.26
88	8	1498	3.12	3.42	10.91	1.35	4.78	5.01	25.23	1.66	10.15
SIC	VAR	N	C1	C2	C3	C4	NN1	NN2	NN3	NN4	RespV
7	251	1623	3.18	2.87	23.13	10.50	17.52	17.40	0.11	22.54	5.75
24	608	785	1.07	1.23	39.19	0.04	1.56	1.74	51.07	0.40	8.60
37	40907	503	0.83	2.47	26.42	19.79	0.11	1.49	35.85	20.31	25.00
42	1298	1840	1.89	2.03	43.04	0.34	3.31	3.17	32.40	4.98	5.83
47	921	785	3.10	3.15	4.95	3.59	3.27	3.28	8.94	1.65	12.98
53	6702	267	3.84	4.05	24.54	6.76	1.18	1.40	26.45	4.29	25.74
56	3522	1634	13.85	13.62	62.77	19.97	1.09	0.73	28.39	10.29	1.73
59	2636	6116	0.20	0.25	21.51	0.64	1.32	1.27	23.16	2.29	26.80
61	15360	300	1.73	1.32	20.46	10.94	62.53	62.42	66.07	65.13	56.49
76	131	1468	5.33	5.26	33.72	7.32	4.18	4.03	52.24	7.87	8.86
86	859	2877	1.30	1.74	71.50	5.53	0.09	0.52	44.18	6.66	22.26
88	8	1498	1.38	1.60	28.78	1.41	0.87	0.58	32.96	5.04	10.15

Observations from Table 3

1. For the standard variance method, compare the imputation methods applied to three size classes (Table 2) to the methods applied to six size classes (Table 3). The only slight improvement in using six size classes was in

the mean method. The other three imputation methods performed about the same for both stratifications.

2. For each imputation method, the standard variance method and the jackknife A method produced the smallest errors of the four variance methods for most of the SICs. Occasionally, the random group method and less frequently the jackknife B method resulted in the smallest errors of the four variance methods, but it produced too many very large errors to be reliable. For the two promising variance methods, standard and jackknife A, the minimum and maximum errors across the SICs are listed in the following table for the four imputation methods.

	Standard		Jackknife A	
	Min. Error	Max. Error	Min. Error	Max. Error
REG1	.08	6.96	.28	6.97
MEAN	.97	64.40	.54	64.29
CARRY	.20	13.85	.25	13.62
NEAR	.09	62.53	.52	62.42

It is clear from the above table that REG1 imputation method with standard variance method has the smallest minimum errors, and the smallest maximum errors.

3. Consider the 16 possibilities from the four imputation methods and the four variance methods; the combination that resulted in the smallest and largest errors out of the 16 are given in the next table for each SIC.

SIC	Min. Error	Imputation / Variance Method	Max. Error	Imputation / Variance Method
7	.09	MEAN / RG	23.13	CARRY / JB
24	.04	CARRY / RG	51.06	NEAR / JB
37	.10	NEAR / SD	35.85	NEAR / JB
42	.34	CARRY / RG	51.74	REG1 / JB
47	.29	MEAN / JB	8.94	NEAR / JB
53	1.2	NEAR / SD	26.45	NEAR / JB
56	.7	NEAR / JA	62.77	CARRY / JB
59	.2	CARRY / SD	23.16	NEAR / JB
61	1.3	CARRY / JA	71.72	MEAN / JB
76	1.8	REG1 / JA	52.23	NEAR / JB
86	.08	REG1 / SD	71.50	CARRY / JB
88	.6	NEAR / JA	32.96	NEAR / JB

Clearly jackknife B is not a good method for computing variances, regardless of imputation methods. However, both REG1 and MEAN produced the largest error once, as opposed to CARRY and NEAR which produced the maximum error three and seven times respectively.

4. In Table 3, the last column indicates the error in the variance if only the respondents' values are used to compute the sample variance estimate, based on a sample of size NBR_i. The minimum and maximum values across the SICs are 1.73 and 56.49 respectively. Considering the best two variance methods, the minimum errors for the four imputation methods are all smaller than the minimum error, 1.73, resulting from no imputation. However, the resulting maximum error, 56.49, is in the range of the maximum errors. It is clear that imputing by REG1 or CARRY is better than no imputing, and from Table 3. it is clear that even MEAN and NEAR are better than no imputing. MEAN and NEAR have a slightly larger maximum value than no imputation, but they have fewer large errors.

Our recommendation for use in the Universe Data Base is the standard variance estimator along with the recommended REG1 method for imputation. For a data base where data are imputed by using either strata means, the carry over method or hot deck nearest neighbor, our results indicate that using the standard variance estimator is as good or better than using either of the jackknife methods or random groups. Although jackknife A method did well, the difference did not warrant its use over the simplicity of the standard estimator. In other more complex situations, other variance estimators might be considered, such as the jackknife variation suggested by Rao and Shao (1992).

8. Future Research

The next step will be to randomly select samples from the population, and consider variance estimators for various statistics, such as means, totals, and regression coefficients, when some of the data have been imputed. Imputation methods could include the popular methods, in particular the regression type methods. Robust variance estimators will be developed for variance estimators of total when the imputation is done by regression. In addition, the effect on the variance estimator of using two or more imputation methods on the same data set will be investigated.

References

Efron, B. (1982), *The Jackknife, the Bootstrap and other Resampling Plans*, SIAM, PA.

Rao, J. N. K., and Shao, J. (1992), "Jackknife variance estimation with survey data under hot deck imputation," *Biometrika* 79, 811-822.

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

Rubin, D., (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley and Sons Inc., NY.

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

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

West, S., Kratzke, D., and Robertson, K., (1993) "Alternative Imputation Procedures For Item Non-response from New Establishments in the Universe," *ASA Proceedings of the Section in Survey Research Methods.*

Wolter, K., (1985), *Introduction to Variance Estimation*, Springer-Verlag.