VARIANCE ESTIMATORS FOR VARIABLES THAT HAVE BOTH OBSERVED AND IMPUTED VAL

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 nvestigations of different variance estimators in the resence of imputed and observed values in this paper. It s assumed that all the missing data are imputed by the ame method. Imputation methods considered include nean, hot deck, regression, regression plus residual, and nultiple imputation. Variance estimators considered nclude the standard, two versions of the jackknife, and andom groups.

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

In Section 2, we describe the data sets used and the lesign of the empirical investigations. The notation and valuation criteria that are used to compare the various nethods are presented in Section 3. Descriptions of the mputation methods and their properties are presented in Section 4 and Section 5, respectively. In Section 6, lternative variance estimators to the standard estimator re considered. The results of the empirical investigations re showed in Section 7, along with observations and onclusions. 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) more homogenous as we move from 2- to 3- stratification.

We obtained data from Michigan in these indu digit SIC code is in parenthesis): Agricultural (07), Lumber and Wood Products (24), Trans Equipment (37), Trucking and Warehousi Transportation Services (47), General Merchandi (53), Apparel and Accessory Stores (56), Misc Retail (59), Nondepository Credit Institutio Miscellaneous Repair Services (76), Me Organizations (86), and Private Households (88).

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

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

Size Class 1 - Employment < 50
Size Class 2 - 50 ≤ Employment < 250
Size Class 3 - Employment ≥ 250

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

Size Class 1a - Employment < 5
Size Class 1b - 5 ≤ Employment < 10
Size Class 1c - 10 ≤ Employment < 20
Size Class 1d - 20 ≤ Employment < 50

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

For the empirical study, we allowed only c single units from private industries. Continuous units that existed on the file during the previou ısed data from the Model set to determine parameters vhich 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 nethods 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 nethods 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 nethod 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 vere computed over the entire 2-digit SIC.

4. **Imputation Methods**

method of imputation would not be desirable b adversely affects the distribution of the sample skewing the distribution toward the mean. For stratification, month t, employment is imputed as

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

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

Carry-Over

Under the carry over method, each non-res employment is imputed using its own histoı predicted value is therefore independent of size 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 employment value was reported for the establishı the paper only s=1 is used.)

Hot Deck-Nearest Neighbor

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

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

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

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

Regression Model

A common method for imputing missing valı least squares regression (Afifi and Elaskoff, 1! several papers on estimators for total employme 1982/1983, and West, et al, 1989), it was discov the most promising models for employment proportional regression models. These models sp the expected employment for establishment j in given the vector of E-values (employment in r reported by units in set BR_t):

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

is proportional to the establishment j's previous 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 con uncorrelated. That is,

$$\begin{cases} \nu_{j,t} \quad if \quad j = l \end{cases}$$

The model can be rewritten as:

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

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

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

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

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

vorked reasonably well for employment data. Thus the ›redicted employment value at time t is:

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

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

\dding Residuals to the Regression Model

The regression method could be thought of as imputing or missing employment by using the mean of the predicted Ξ_t distribution, conditional on the predictors E_{t-1}. As a esult, the distribution of the imputed values has a smaller ariance than the distribution of the true values, even if the ssumptions of the model are valid. A simple strategy of djusting for this problem is to add random errors to the ›redictive means, that is, drawing residuals r_k with mean ero to add to $\hat{E}_{k,t}$.

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

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

vhere δ_k is a random number from a $N(0,1)$ distribution nd s^2 is equal to the mean square error of the regression.

A slight modification of the previous method was ›btained by drawing five random numbers and using the verage 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 t month t:

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

Assuming that the missing data are missing a consider the effects of using imputation methoc First consider overall mean imputation, that is, one stratum. In this situation, formula (5.2) becon

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

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 em distribution, since all the missing values are ass same value, the mean of the respondents, $\hat{E}_{k,t} = \sum_{j \in BR_t} E_{j,t} \big/ NBR_t \quad for\ all\ k \in NR_t$. The secon (5.3) becomes zero since $\hat{E}_{k,t} = \hat{\bar{E}}$ resulting following variance estimator:

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

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

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

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

and hence,

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

Note that the relative bias is approximately 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 imputatic strata; this method produces a series of spik employment distribution at the means of the ir strata. Let $\bar{E}_h$ denote the mean of the respo stratum h which has $NNR_{t,h}$ missing values, variance estimator can be written as:

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

where H is the number of strata and,

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

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

And hence the variance estimator can be written as:

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

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

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

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

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

vhere $E(S_h^2)/V_t$ is the proportion of the variance explained ›y the imputation strata.

Similar results are obtained for imputation methods 2-4. 'or example, the formula for method 4 has the proportion ›f the variance explained by the regression. The predicted egression method curtails the spread of the employment listribution.

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

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

6. Alternative Variance Estimators

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

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

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

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

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

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

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

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

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

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

7. Results / Conclusions

Tables 1 and 2 show the errors in computing using the standard variance estimator. Notation: V_t = VAR, $NBR_t + NNR_t$ = N, $\hat{V}_{t,1,3+i}$ = REGi, i=1 $\hat{V}_{t,1,1}$ = MEAN, $\hat{V}_{t,1,2}$ = CARRY, and $\hat{V}_{t,1,3}$ = NE.

Table 1.

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

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

			Nonresponse Rate:	**As observed on file=OB**			
SIC	VAR	N	REG1	REG2	REG3	MEAN	CA
7	256.15	1614	0.84	0.84	0.86	0.97	
24	757.13	761	0.03	0.02	0.03	0.19	
37	40954.39	503	0.10	0.10	0.10	0.45	
42	1300.66	1836	2.64	2.65	2.66	1.56	
47	1006.08	785	0.13	0.10	0.11	0.92	
53	7711.62	262	0.01	0.01	0.01	0.44	
56	3903.65	1622	1.10	1.09	1.11	2.42	
59	2659.32	6099	3.39	3.38	3.39	65.24	
61	15265.53	302	0.00	0.00	0.00	0.07	
76	131.41	1459	0.97	1.03	0.96	2.95	
86	921.87	2871	5.67	5.67	5.67	22.09	
88	8.17	1495	1.47	1.59	1.59	4.53	
			Nonresponse Rate:	**25%**			
SIC	VAR	N	REG1	REG2	REG3	MEAN	CA
7	255.67	1562	6.50	6.34	6.45	15.96	
24	610.17	690	3.36	3.17	3.16	5.44	
37	42829.52	470	0.12	0.12	0.12	1.44	
42	1313.77	1816	0.62	0.55	0.59	19.52	
47	1024.35	756	2.60	2.63	2.71	19.11	
53	7964.20	223	2.27	2.28	2.28	5.84	
56	4130.02	1530	3.23	3.27	3.29	30.20	1

esponse rates, because certain observations could not be ısed due to the requirements of certain imputation ›rocedures.

Table 2.

\bsolute Percent Errors incurred in Standard /ariance Estimator due to Imputation

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

SIC	VAR	N	REG1	REG2	REG3	MEAN	CARRY	NEAR
Nonresponse Rate: As observed on file=OB								
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	40728.95	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

)bservations from Table 1

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

or 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.

)bservations from Table 2

or 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.

or 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 arger with 25% than with OB%.

:ounty vs. Size Class Stratification

or OB%: Size class stratification produced smaller errors than county stratification, with the biggest improvements in the MEAN and NEAR methods. The

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

In summary, if the standard variance formula then the imputation method that least dist population variance is one of the regression typ simplest regression type which is the single mode residual added should be used, and stratification : by 3-digit SIC/size class. This method is rc different response rates, and resulting error mea relatively small.

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

Table 3.

Absolute Percent Errors incurred in 4 Estimators due to Imputation

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

SIC	VAR	N	R1	R2	R3	R4	M1	M2	M3
7	251	1623	6.97	6.97	18.29	9.06	6.24	6.49	12.09
24	608	785	2.44	2.32	23.12	7.32	6.43	6.61	49.84
37	40907	503	0.41	0.28	29.74	1.74	7.23	8.93	23.19
42	1298	1840	2.75	2.76	51.74	1.79	3.06	2.96	30.69
47	921	785	2.09	1.95	0.37	6.39	2.10	2.11	0.29
53	6702	267	2.50	2.54	18.84	6.77	5.37	5.59	24.68
56	3522	1634	1.88	2.14	30.60	5.25	4.08	3.90	44.49
59	2635	6116	1.54	1.52	21.94	1.97	1.66	1.61	17.79
61	15360	300	1.77	1.60	15.67	7.16	64.41	64.29	71.73
76	131	1468	2.04	1.85	38.18	6.47	4.44	4.51	8.48
86	859	2877	0.08	0.33	42.50	6.59	0.97	0.54	50.49
88	8	1498	3.12	3.42	10.91	1.35	4.78	5.01	25.23

SIC	VAR	N	C1	C2	C3	C4	NN1	NN2	NN3
7	251	1623	3.18	2.87	23.13	10.50	17.52	17.40	0.11
24	608	785	1.07	1.23	39.19	0.04	1.56	1.74	51.07
37	40907	503	0.83	2.47	26.42	19.79	0.11	1.49	35.85
42	1298	1840	1.89	2.03	43.04	0.34	3.31	3.17	32.40
47	921	785	3.10	3.15	4.95	3.59	3.27	3.28	8.94
53	6702	267	3.84	4.05	24.54	6.76	1.18	1.40	26.45
56	3522	1634	13.85	13.62	62.77	19.97	1.09	0.73	28.39
59	2635	6116	0.20	0.25	21.51	0.64	1.32	1.27	23.16
61	15360	300	1.73	1.32	20.46	10.94	62.53	62.42	66.07
76	131	1468	5.33	5.26	33.72	7.32	4.18	4.03	52.24
86	859	2877	1.30	1.74	71.50	5.53	0.09	0.52	44.18
88	8	1498	1.38	1.60	28.78	1.41	0.87	0.58	32.96

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 clear that even MEAN and NEAR are better imputing. MEAN and NEAR have a slight maximum value than no imputation, but they ha large errors.

Our recommendation for use in the Universe I is the standard variance estimator along recommended REG1 method for imputation. F base where data are imputed by using either stra the carry over method or hot deck nearest neig results indicate that using the standard variance is as good or better than using either of the methods or random groups. Although jackknife A did well, the difference did not warrant its use simplicity of the standard estimator. In oth complex situations, other variance estimators considered, such as the jackknife variation sugg Rao and Shao (1992).

8. Future Research

The next step will be to randomly select sam the population, and consider variance estim various statistics, such as means, totals, and r coefficients, when some of the data have been Imputation methods could include the popular m particular the regression type methods. Robust estimators will be developed for variance estir total when the imputation is done by regress addition, the effect on the variance estimator of u or more imputation methods on the same data s investigated.

References

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

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

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

Rubin, D., (1987), *Multiple Imputation for Nor in Surveys*, John Wiley and Sons Inc., NY.

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

West, S., Butani, S., Witt, M., Adkins, C., "Alternate Imputation Methods for Employme *ASA Proceedings of the Section in Survey Methods*.

West, S., Kratzke, D., and Robertson, K. "Alternative Imputation Procedures For Ite response from New Establishments in the U