VARIANCE ESTIMATION IN THE PRESENCE OF IMPUTATION FOR MISSING DATA

J.N.K. Rao, Carleton University, Ottawa, Ontario, K1S 5B6, Canada
jrao@math.carleton.ca

ABSTRACT

Item nonresponse is usually treated by some form of deterministic or random imputation. We focus on deterministic imputation; in particular, ratio and nearest neighbour imputations commonly used in establishment surveys. Frequentist inference from imputed data is based on a repeated sampling framework and assumed response mechanism. On the other hand, use of imputation models requires only that the assumed model holds for the respondents. Treating the imputed values as true values and computing standard errors using standard formulae applicable to complete samples can lead to serious underestimation of true standard errors, especially when the item nonresponse is appreciable. This paper reviews some recent work on variance estimation under single imputation that takes proper account of the additional variability due to the unknown missing value; in particular, work on jackknife, jackknife linearization and modified balanced repeated replication.

Key Words: Item Nonresponse, Jackknife, Modified BRR, Ratio Imputation

1. INTRODUCTION

Unit (or total) nonresponse and item nonresponse both occur frequently in surveys. Unit nonresponse is customarily handled by forming nonresponse adjustment cells or weighting classes using auxiliary variables observed on all the sampled elements and then adjusting the survey weights of all respondents within a weighting class by a common nonresponse adjustment factor, with different adjustment factors in different classes (see e.g., Kalton and Kasprzyk, 1986). Little (1986) proposed a method for constructing weighting classes (or adjustment cells) using estimated response probabilities (propensities) obtained from logistic (or probit) regression of the response indicator on the auxiliary variables. Eltinge and Yansaneh (1997) developed useful diagnostics for formation of weighting classes.

Item nonresponse is usually handled by some form of imputation to fill in missing item values. Commonly used methods of imputation of business survey data include mean imputation; ratio or regression imputation, using auxiliary variables observed on all the sampled units and nearest neighbour imputation in which a nonrespondent item is assigned the item value of the "nearest" neighbour, where "nearest" is usually defined in terms of a distance function for the auxiliary variables or predicted values. The above methods are deterministic, and some of them may not preserve the distribution of item values. To avoid the latter problem, random donor imputation within imputation classes formed on the basis of auxiliary variables is often used, especially in socio-economic surveys. Random donor imputation and nearest neighbour imputation belong to the class of "hot-deck" imputation methods in which the value assigned for a missing response is taken from a respondent's item value. Imputation classes for random imputation are formed by using either estimated response probabilities or predicted items, obtained by fitting a regression equation to responses and associated auxiliary variables and then producing predicted item values for both respondents and nonrespondents. Imputation methods like the foregoing ensure that the results obtained from different analyses are consistent with one another, unlike the results of analyses from an incomplete data set. This is achieved through the use of the same survey weight for all items.

The foregoing methods belong to the class of marginal imputation methods, designed for making inference on the marginal parameters such as the population mean or total, cumulative distribution function and quantiles. These methods are "improper" in the sense of Rubin (1996), but all the same can lead to valid frequentist (or design-based) inferences under an assumed response mechanism or "design-model" inferences under an assumed "imputation model" requiring only that the model holds for the respondents. Moreover, a single completed data set is used for inference, unlike the use of multiple completed data sets in the method of Rubin (1996). However, the methods readily extend to multiple completed data sets (Fay, 1996), if desired.

In practice, it is also important to make inference on parameters measuring relationships (e.g. subclass means, correlation and regression coefficients). Typically, marginal imputation methods, without adjustment for bias of estimators, are not adequate for making frequentist references on relationships between items under an assumed response mechanism. Random "common donor" imputation is often used for relationships, but it also requires bias adjustment except in special cases (Skinner and Rao, 2000). Judkins (1997) proposed a simple random imputation procedure aimed at maintaining all the associations between the items, but its theoretical properties are not fully studied. Sophisticated methods of imputation, based on multivariate models, have also been proposed in the context of multiple imputation (Schaffer et al, 1993), but it is not clear how to make inference under such models with only

single imputation. Shao and Wang (1999) proposed a "joint random regression imputation" method that preserves asymptotic unbiasedness of marginal estimators as well as estimators of correlation coefficients, under an assumed model. This method is an extension of an earlier method proposed by Srivastava and Carter (1986).

It is a common practice to treat the imputed values as if they were observed and then compute estimates and variance estimates using standard formulas for a specified sampling design. The point estimates are generally unbiased (or approximately unbiased) under the assumed response mechanism or the imputation model, at least for the marginal parameters. But the variance estimates can lead to erroneous inferences even for large samples; in particular, serious underestimation of the variance of the imputed estimator because the additional variability due to estimating the unknown missing values is not taken into account.

This paper provides a brief account of some recent developments in variance estimation under single imputation that takes proper account of the variability due to estimating the missing values. In particular, work on jackknife and other resampling methods will be reviewed as well as jackknife linearization. Most of this work pertains to marginal parameters, but recent research on parameters measuring relationships looks promising. Comparisons with multiple imputation will also be presented. We study both stratified simple random sampling (commonly used in business surveys based on list frames) and stratified multistage sampling (commonly used in socio-economic surveys). To implement these methods, the completed data sets must include information on response status for each item as well as on the imputation class. Existing software (e.g., WESVAR for the jackknife) can be modified to implement the variance estimators using a single completed data set.

2. SIMPLE RANDOM SAMPLING

Establishment surveys based on list frames often use stratified simple random sampling. If the number of strata is small and within strata sample sizes are relatively large, then the strata are often used as imputation classes and imputation is done independently in each stratum. Due to space limitations, we focus here on ratio and nearest neighbour imputations that are commonly used in business surveys, assuming the above set-up. We need only study inference for the case of a single stratum (or simple random sampling) because estimates and variance estimates can be readily combined over strata. We consider the case of a large number of strata and relatively small within strata PSU sample sizes in section 3. In this case, imputation classes typically cut across sample clusters or strata. The methods in section 3 can be adapted to handle stratified simple random sampling with a large number of strata and relatively small within strata sample sizes.

2.1 Ratio Imputation

Let n be the number of units sampled from the N population units (in a stratum). In the case of complete response on an item y, a design-unbiased (or p-unbiased) estimator of the population mean $\bar{Y}$ is given by the sample mean $\bar{y} = \Sigma_s \, y_i / n$, where s denotes a sample. A p-unbiased estimator of variance of $\bar{y}$ is given by

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N} \right) s_y^2,$$

where $s_y^2 = (n-1)^{-1} \Sigma_s \, (y_i - \bar{y})^2$. A jackknife variance estimator of $\bar{y}$ is

$$v_J(\bar{y}) = \left(1 - \frac{n}{N} \right) \frac{n-1}{n} \sum_{j \in s} [\bar{y}(j) - \bar{y}]^2 \qquad (1)$$

where $\bar{y}(j)$ is the sample mean obtained by deleting $j \in s$. It extends readily to general statistics of the form $\hat{\theta} = g(\bar{y})$ by simply replacing $\bar{y}(j)$ and $\bar{y}$ by $\hat{\theta}(j) = g[\bar{y}(j)]$ and $\hat{\theta}$ respectively, in (1). In the linear case, $\hat{\theta} = \bar{y}$, we have $v_J(\bar{y}) = v(\bar{y})$.

In the presence of nonresponse on item y, suppose that r units respond to item y and that m units do not respond. We assume that an auxiliary variable, x, closely related to y, is observed on all sample units, s. Let $\bar{y}_r$, $\bar{x}_r$ be the means

for the respondents, s_r, and let y_i^* be the imputed value for unit $i \in s_m$, the sample of nonrespondents. Ratio imputation uses $y_i^* = (\bar{y}_r / \bar{x}_r) x_i$ for $i \in s_m$. The imputed estimator of the population $\bar{Y}$ is given by

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s_m} y_i^* \right), \tag{2}$$

and that of the population total Y by $\hat{Y}_I = N \bar{y}_I$. Under ratio imputation, $\bar{y}_I$ reduces to the ratio estimator

$$\bar{y}_I = (\bar{y}_r / \bar{x}_r)\bar{x}, \tag{3}$$

where $\bar{x}$ is the x-mean for the full sample s. Under uniform response (i.e., independent response across sample units and equal response probabilities) $\bar{y}_I$ is approximately unbiased (for large n); that is, $E_p E_r (\bar{y}_I) \approx \bar{Y}$, where E_p and E_r denote expectation with respect to the design and expectation with respect to response mechanism, respectively. It is also valid under the weaker assumption of missing at random (MAR), which permits a response probability depending on x but not on y, provided the following ratio model holds for the population units:

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \quad Cov_m(y_i, y_j) = 0, \quad i \neq j. \tag{4}$$

We have $E_m E_p E_r = E_m(\bar{Y})$ or $\bar{y}_I$ is *pm*-unbiased under MAR, where E_r now denotes expectation under MAR and the subscript m in (4) denotes the model. Note that y_i^* for ratio imputation is the best predictor of $y_i, i \in s_m$, under the ratio model. Särndal (1992), therefore, named the model as the imputation model, but it is no different from the assumption made in the model-dependent approach of Royall (1970). In any case, $\bar{y}_I$ is robust in the sense that it can be justified under both frequentist and model-dependent approaches. However, $\bar{y}_I$ is not justified under MAR if the assumed model (4) is not valid.

Treating the imputed values y_i^* as if they were observed and then using the jackknife variance estimator (1) with $\bar{y}(j)$ and $\bar{y}$ changed to $\bar{y}_I(j)$ and $\bar{y}_I$ could lead to serious underestimation if the item nonresponse is substantial. This problem is particularly serious in two-phase sampling when "mass" imputation is used; that is when the y values of units not sampled at the second phase are imputed using the first phase x-information. In this case, the proportion not sampled at the second phase (nonresponse rate) is larger than the proportion sampled (response rate).

A correct jackknife variance estimator is obtained by adjusting the imputed value y_i^* by the amount $y_i^*(j) - y_i^*$ only when a respondent $j \in s_r$ is deleted in the jackknife calculations. Here $y_i^*(j)$ is the value one would impute for the i-th nonrespondent if the j-th respondent is deleted from the sample s. Thus the adjusted imputed value equals $y_i^*(j) = [\bar{y}_r(j) / \bar{x}_r(j)] x_i$ under ratio imputation when $j \in s_r$ is deleted, where $\bar{y}_r(j)$ and $\bar{x}_r(j)$ are the respondent y- and x- means with j-th unit deleted. This method is equivalent to reimputing from the reduced respondent set; that is, it recognizes the fact that the donor set is changed when a respondent is deleted from the sample.

Denote the imputed estimator based on the respondent values and adjusted imputed values as $\bar{y}_I^a(j)$ when $j \in s_r$ is deleted. Then a valid jackknife variance estimator, $\tilde{v}_J(\bar{y}_I)$, under uniform response is given by

$$\tilde{v}_J(\bar{y}_I) = \frac{n-1}{n}\left\{ \sum_{j \in s_r} [\bar{y}_I^a(j) - \bar{y}_I]^2 + \sum_{j \in s_m} [\bar{y}_I(j) - \bar{y}_I]^2 \right\}, \tag{5}$$

provided the sampling fraction *n/N* is negligible. If *n/N* is not negligible, one should not use $(1-n/N)\widetilde{v}_J(\bar{y}_I)$ since it leads to underestimation, while $\widetilde{v}_J(\bar{y}_I)$ leads to overestimation. Lee et al. (1995) proposed a compromise jackknife variance estimator

$$v_{Jc}(\bar{y}_I) = \widetilde{v}_J(\bar{y}_I) - \frac{1}{N}s_{yr}^2 \tag{6}$$

where $s_{yr}^2 = (r-1)^{-1}\Sigma_{s_r}(y_i - \bar{y}_r)^2$.

By linearizing $\widetilde{v}_J(\bar{y}_I)$, Rao (1996) obtained a jackknife linearization variance estimator

$$\widetilde{v}_{JL}(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \frac{A}{r} + 2\left(\frac{\bar{x}}{\bar{x}_r}\right)\frac{B}{n} + \frac{C}{n}, \tag{7}$$

where $A = (r-1)^{-1}\Sigma_{s_r}e_i^2 = s_{er}^2$ with $e_i = y_i - (\bar{y}_r/\bar{x}_r)x_i$, $B = (\bar{y}_r/\bar{x}_r)\Sigma_{s_r}e_i x_i/(r-1)$ and $C = (\bar{y}_r/\bar{x}_r)^2\Sigma_s(x_i-\bar{x})^2/(n-1)$. He showed that $\widetilde{v}_{JL}(\bar{y}_I)$, and hence $\widetilde{v}_J(\bar{y}_I)$, are robust in the sense that they can be justified both under frequentist and model-dependent approaches, provided the sampling fraction is negligible.

Following Fay (1991), Shao and Steel (1999) reversed the customary sample-response path by assuming a census with nonrespondents from which a sample is taken. Under this set-up, we have

$$V(\bar{y}_I) = E_r V_p(\bar{y}_I) + V_r E_p(\bar{y}_I), \tag{8}$$

where the inner expectation and variance are with respect to sampling, conditional on the response indicators a_i: $a_i = 1$ if y_i is observed in the census and $a_i = 0$ otherwise, *i=1, ..., N*. The imputed estimator $\bar{y}_I$ can be expressed as a nonlinear function of the sample means of $z_{1i} = a_i y_i, z_{2i} = a_i x_i$ and $z_{3i} = (1-a_i)x_i : \bar{y}_I = \bar{z}_1(1+\bar{z}_3/\bar{z}_2) = g(\bar{z})$. It readily follows from (8) that the estimation of $E_r V_p(\bar{y}_I)$ is the same as the estimation of $V_p(\bar{y}_I)$, conditional on a set of respondents regardless of response mechanism. We can use either the standard jackknife variance estimator of $g(\bar{z})$ or a Taylor linearization variance estimator of $g(\bar{z})$ to estimate $V_p(\bar{y}_I)$. On the other hand, $E_p(\bar{y}_I) \approx g(\bar{\boldsymbol{Z}})$ where $\bar{\boldsymbol{Z}}$ is the vector of population means of z_1, z_2 and z_3, and it is necessary to specify the response mechanism to evaluate $V_r[g(\bar{\boldsymbol{Z}})]$ and in turn its estimator. However, it is not necessary to estimate $V_r E_p(\bar{y}_I)$ if *n/N* is negligible because this term is of lower order then the first term $E_r V_p(\bar{y}_I)$ which is of order $O(n^{-1})$. Shao and Steel (1999) showed that the Taylor linearization variance estimator of $g(\bar{z})$, in fact, equals $\widetilde{v}_{JL}$ given by (7) when *n/N* is negligible.

Under uniform response and ratio imputation, Shao and Steel (1999) showed that an estimator of $V_r[g(\bar{\boldsymbol{Z}})]$ is given by

$$v[g(\bar{\boldsymbol{Z}})] = \left(\frac{\bar{x}}{\bar{x}_r}\right)^2 \left[\frac{1}{r}\left(\frac{n}{N}\right) - \frac{1}{N}\right]s_{er}^2 \tag{9}$$

which is of lower order than $O(n^{-1})$ if *n/N=o(1)*.

The formula for the estimator of $V_r\left[g\left(\overline{\boldsymbol{Z}}\right)\right]$ under MAR and the ratio model is given by

$$v_m\left[g\left(\overline{\boldsymbol{Z}}\right)\right]=\left(\frac{\bar{x}}{\bar{x}_r}\right)^2\left[\frac{1}{r}\left(\frac{n}{N}\right)-\frac{1}{N}\left(\frac{\bar{x}_r}{\bar{x}}\right)\right]s_{er}^2. \tag{10}$$

Note that (10) is different from (9) but it remains valid under uniform response because $\bar{x}/\bar{x}_r \approx 1$ in this case. Therefore, a robust variance estimator, valid under both approaches, is given by combining $(1-n/N)\widetilde{v}_J(\bar{y}_I)$ or $(1-n/N)\widetilde{v}_{JL}(\bar{y}_I)$ with (10).

Combining (1 - *n/N*) $\widetilde{v}_{JL}$ and (9), we get a variance estimator of $\bar{y}_I$ valid for nonnegligible sampling fraction *n/N* and uniform response:

$$v_L\left(\bar{y}_I\right)=\left(\frac{\bar{x}}{\bar{x}_r}\right)^2\left(\frac{1}{r}-\frac{1}{N}\right)A+2\left(\frac{\bar{x}}{\bar{x}_r}\right)\left(\frac{1}{n}-\frac{1}{N}\right)B+\left(\frac{1}{n}-\frac{1}{N}\right)C \tag{11}$$

which agrees with the variance estimator of Rao and Sitter (1995) derived under a two-phase sampling approach. Note that conditionally given *r*, s_r is a simple random sample from *s* under uniform response (Oh and Scheuren, 1983). The compromise variance estimator (6) does not agree with the correct variance estimator (11) when $\widetilde{v}_J(\bar{y}_I)$ is replaced by $\widetilde{v}_{JL}(\bar{y}_I)$. The variance estimator under MAR and the ratio model (4) is obtained by combining (1 - *n/N*) $\widetilde{v}_{JL}$ and (10).

It is interesting to note that (7) is a valid estimator of variance of $\bar{y}_I$ under MAR without assuming an imputation model, but the estimator $\bar{y}_I$ is biased under MAR without the model assumption. Therefore, we need either uniform response without the model or the weaker assumption of MAR together with the model.

Ratio imputation does not preserve the distribution of *y*-values because it is nonrandom. For example, under the ratio model (4), the sample mean of $\widetilde{y}_i^2$ is asymptotically biased for the population mean of y_i^2, where $\widetilde{y}_i = y_i$ if $i \in s_r$; $= y_i^*$ if $i \in s_m$ (Shao and Wang, 1999). On the other hand, random ratio imputation leads to asymptotically unbiased estimates of both the first and second moments, (Shao and Wang, 1999). We refer the reader to Rao (1996) for details on random ratio or regression imputation and associated jackknife variance estimation based on adjusted imputed values.

Random ratio imputation introduces "imputation variance," unlike the (nonrandom) ratio imputation, which can be a significant component of the total variance if the item response rate is not high. Chen, Rao and Sitter (2000) studied random donor imputation and proposed a simple random imputation method that eliminates the imputation variance of the estimator of $\bar{Y}$ or *Y*, and at the same time preserves the distribution of item values. This method simply uses $\bar{y}_r + e_i^* - \bar{e}_m^*$ as imputed values in the data file instead of y_i^* for $i \in s_m$, where $e_i^* = y_i^* - \bar{y}_r$ and $\bar{e}_m^*$ is the mean of e_i^* for $i \in s_m$. Note that $\bar{y}_I$ reduces to $\bar{y}_r$. Chen, Rao and Sitter (2000) also proposed jackknife and bootstrap variance estimators that depend only on the reported values in the data file. Extension of this method to random ratio imputation is currently under investigation.

The imputed estimator of a domain total, Y_d, is given by

$$\hat{Y}_{dI} = \frac{N}{n}\left(\sum_{i\in s}\delta_i\widetilde{y}_i\right),$$

where $\delta_i = 1$ if the *i*-th sample unit belongs to the domain and $\delta_i = 0$ otherwise. Under ratio imputation, $\hat{Y}_{dI}$ reduces to

$$\hat{Y}_{dI} = \frac{N}{n}\left[r\left(\bar{y}_{rd} - \frac{\bar{y}}{\bar{x}} \bar{x}_{rd} \right) + n\left(\frac{\bar{y}}{\bar{x}} \bar{x}_d \right) \right],$$

where $(\bar{y}_{rd}, \bar{x}_{rd})$ are the domain respondent means and $\bar{x}_d$ is the domain sample mean. Under uniform response, $\hat{Y}_{dI}$ is asymptotically biased unless the domain population ratio Y_d / X_d equals the overall population ratio Y / X. On the other hand, $\hat{Y}_{dI}$ is *pm*-unbiased for Y_d under the ratio model (4) because $E_m(Y_d / X_d - Y / X) = 0$. The above results also hold under random ratio imputation. Note that the ratio model implies homogeneity of domain ratios.

We have assumed so far that the auxiliary variable, *x*, is observed on all the units in the sample, *s*. But, *x* may not be observable on all the sampled units. For example, in establishment surveys conducted at Statistics Canada, *x* is the previous period value; ratio imputation is used when the *x*-value is available and respondent mean imputation for the remaining sampled units with missing *y*-values (Rancourt, Lee and Särndal, 1994). Sitter and Rao (1997) studied the general case where the responses on either the variable of interest *y* or the auxiliary variable *x* or both may be missing. They used ratio imputation when the associated *x* is observed and different imputations when *x* is not observed. They obtained design consistent jackknife and linearization variance estimators under uniform response; the latter variance estimator incorporates the finite population corrections. Shao and Steel (1999) used the reverse approach (response – sample path) to handle complicated situations where a composite of some deterministic and /or random imputation methods is used, including the use of imputed data in subsequent imputations. They applied the method to imputed data from the Transportation Annual Survey conducted at the U.S. Census Bureau. This survey uses a composite of "cold deck" and ratio type imputation methods.

2.2 Nearest Neighbour Imputation

We confine ourselves to the case of scalar *x* observed on all the sample units. Nearest neighbour imputation (NNI) imputes a missing y_j, $j \in s_m$, by y_i, where $i \in s_r$ and *i* is the nearest neighbour of *j* in the sense that *i* satisfies $| x_i - x_j | = \min_{\ell \in s_r} | x_\ell - x_j |$. NNI is quite popular, but until recently its theoretical properties have not been studied.

Chen and Shao (1999) showed that NNI provides asymptotically valid estimators of population means, distribution functions and quantiles, under some regularity conditions, assuming MAR. These results are valid under almost no assumption on the model relating *y* and *x*, excepting that *x* is also assumed to be random. For variance estimation, the jackknife method of Rao and Shao (1992) based on adjusted imputed values is not applicable because the imputed estimator $\bar{y}_I$ is non-smooth. Chen and Shao (1999) proposed a linearization variance estimator as well as a modified jackknife variance estimator based on "partially" adjusted imputed values. Rancourt, Sarndal and Lee (1994) assumed the ratio model (4) and reported some simulation results. Rancourt (1999) studied some theoretical properties of nearest neighbour imputation, assuming the ratio model.

2.3 Post-stratification

In practice, post-stratification is commonly used to ensure consistency with known auxiliary totals. The basic design weights $w_i = 1/n$ are changed to $\tilde{w}_i = (1/n) g_i$, where g_i is the post-stratification adjustment factor. In the case of complete post-stratification, $g_i = (N_j / N) / (n_j / n)$ if *i*-th sample unit belongs to *j*-th post-stratum, where $n_j (N_j)$ is the number of sample (population) units in *j*-th post-stratum. Under complete response, the estimator of $\bar{Y}$ is $\bar{y}_{pst} = n^{-1} \Sigma_s g_i y_i$. Under item nonresponse, we use the post-stratified weights to perform ratio imputation: $y_i^* = \left[\Sigma_{s_r} \tilde{w}_j y_j / \Sigma_{s_r} \tilde{w}_j x_j \right] x_i$ and the imputed estimator $\bar{y}_I$ is given by (2). This estimator retains consistency with the known counts N_j. Yung and Rao (2000) studied jackknife variance estimation with post-stratified weights, in the context of stratified multistage designs.

3. STRATIFIED MULTISTAGE SAMPLING

3.1 Complete Response

Many large-scale surveys employ stratified multistage sampling designs with large number of strata, *H*, and relatively small numbers, $n_h \geq 2$, of primary sampling units or clusters, sampled without replacement with probabilities proportional to sizes within each stratum *h*. We assume that subsampling within sampled clusters, *hi*, is performed to ensure unbiased estimation of cluster totals. An unbiased estimator of the total *Y* for an item *y* is then of the form $\hat{Y} = \Sigma_s \; w_{hik} y_{hik}$, where *s* denotes the sample of elements, and $\left(w_{hik}, y_{hik} \right)$ respectively denote the basic design weight and the value for item *y* associated with (*hik*). Delete-cluster jackknife and balanced repeated replication (BRR) are commonly used to estimate the variance of $\hat{Y}$ as well as the variances of nonlinear statistics, but the jackknife may run into problems in estimating variances of nonsmooth estimators such as sample quantiles, unless the number of sample elements in each deleted cluster is large. BRR can handle nonsmooth estimators, but readily applicable only for the important special case of $n_h = 2$ clusters per stratum.

BRR works by forming balanced half-samples, re-computing the survey estimate on each replicate and taking the mean squared deviation of the replicate estimates as the variance estimate. A minimal set of *R* balanced half-samples can be constructed, for the case $n_h = 2$, from an *R* x *R* Hadamard matrix by choosing any *H* columns excluding the column of all + 1's, where $H + 1 \leq R \leq H + 4$. For the case of complete response, let $\hat{\theta}$ be the estimator of a parameter of interest, θ, based on the full sample, and $\hat{\theta}^{(r)}$ be the corresponding estimator computed from the *r*th half-sample. The estimator $\hat{\theta}^{(r)}$ is obtained using the same formula as for $\hat{\theta}$ with w_{hik} changed to $w_{hik}^{(r)}$, which equals $2w_{hik}$ or 0 according to whether or not the (*hi)*-th cluster is in the *r*th half-sample. A variance estimator for $\hat{\theta}$ is given by

$$v_{BRR}\left(\hat{\theta}\right) = \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\theta}^{(r)} - \hat{\theta}\right)^2 \quad . \tag{12}$$

This variance estimator is consistent and asymptotically unbiased when the first-stage sampling fraction, *n/N*, is negligible and $n \rightarrow \infty$, where $n = \Sigma \, n_h$ and $N = \Sigma \, N_h$ is the number of primary clusters in the population.

A draw-back of BRR is that the half-samples can result in domain sample sizes smaller than would have been tolerated in the full sample. For example, when the survey estimator is a ratio of two domain totals, some replicate ratios can be extremely large because of near-zero denominators or even undefined (Judkins, 1990). To overcome this problem, Fay (see Dippo, Fay and Morganstein, 1984) proposed a modified BRR method obtained by perturbing the weights by ± 100 ε% for the half-sample and its compliment, where $0 < \varepsilon < 1$, instead of the sharp perturbation (±100%) of weights used in the standard BRR. We denote this method as BRR (ε) and the standard method as BRR(1). Since BRR(ε) actually does not delete any sampled element, it does not have the same problem that the standard method has. Judkins (1990) gave a striking example of a ratio with large denominator coefficient of variation (c.v.) at the stratum level for which the C.V. of the BRR (1) variance estimator is about 900 compared to 1.4 for BRR (½).

Let $w_{hik}^{(r)}(\varepsilon) = (1+\varepsilon) w_{hik}$ or $(1-\varepsilon) w_{hik}$ be the replicate weights for the modified BRR, and $\hat{\theta}^{(r)}(\varepsilon)$ be the associated estimator of θ. Then, the variance estimator for $\hat{\theta}$ is given by

$$v_{BRR(\varepsilon)}\left(\hat{\theta}\right) = \frac{1}{\varepsilon^2 R} \sum_{r=1}^{R} \left[\hat{\theta}^{(r)}(\varepsilon) - \hat{\theta}\right]^2 \tag{13}$$

which reduces to (12) if ε=1. For any ε, $v_{BRR(\varepsilon)}$ is exactly the same as the standard variance estimator in the linear case $\hat{\theta} = \hat{Y}$ (Judkins, 1990). Rao and Shao (1999) established the asymptotic validity of (13) for both smooth and

nonsmooth estimators. Judkins (1990) studied by simulation the empirical performance of BRR(ε). He found that BRR(½) performs well and is a compromise between the standard BRR and the jackknife.

3.2 Item nonresponse

For stratified multistage sampling, imputation for item nonresponse is often carried out within imputation cells in which uniform response is assumed; imputation cells, υ, may cut across design strata, *h*. Rao and Shao (1992) studied jackknife variance estimation under imputation for item nonresponse. They proposed a consistent jackknife variance estimator, using adjusted imputed values. Rao (1996) linearized the jackknife to obtain jackknife linearization variance estimators that are computationally simpler. Yung and Rao (1999) extended these results to post-stratified weights. Shao, Chen and Chen (1998) extended the Rao–Shao approach to BRR. In this method, every imputed value y^*_{hik} in the υ-th imputation cell and *r*-th replicate is adjusted to $\tilde{y}^{(r)}_{hik} = y^*_{hik} + E^{(r)}_{*\upsilon}\left(y^*_{hik}\right) - E_{*\upsilon}\left(y^*_{hik}\right)$, $E_{*\upsilon}$ is the expectation with respect to the original imputation procedure in the υ-th cell and $E^{(r)}_{*\upsilon}$ is the same as $E_{*\upsilon}$ except that imputation uses respondents in the *r*-th replicate as donors. For any deterministic imputation method $E_{*\upsilon}\left(y^*_{hik}\right) = y_{hik}$ and $E^{(r)}_{*\upsilon}\left(y^*_{hik}\right)$ is the imputed value based on the data in the *r*-th replicate only, so that $\tilde{y}^{(r)}_{hik}$ is the same as re-imputing the missing y_{hik} in the *r*-th replicate using the data in the *r*-th replicate. In particular, for ratio imputation $y^*_{hik} = \left[\left(\Sigma_{s_{r\upsilon}} w_{hik} y_{hik}\right)/\left(\Sigma_{s_{r\upsilon}} w_{hik} x_{hik}\right)\right] x_{hik}$, and $\tilde{y}^{(r)}_{hik} = \left[\left(\Sigma_{s_{r\upsilon}} w^{(r)}_{hik} y_{hik}\right)/\left(\Sigma_{s_{r\upsilon}} w^{(r)}_{hik} x_{hik}\right)\right] x_{hik}$, where $s_{r\upsilon}$ is the sample of respondents to item *y* in the υ-th cell. The resulting estimators of total *Y* are given by

$$\hat{Y}^{(r)}_I = \Sigma_{s_{r\upsilon}} w^{(r)}_{hik}\, y_{hik} + \Sigma_{s_{m\upsilon}} w^{(r)}_{hik}\, \tilde{y}^{(r)}_{hik} \tag{14}$$

and

$$\hat{Y}_I = \Sigma_{s_{r\upsilon}} w_{hik}\, y_{hik} + \Sigma_{s_{m\upsilon}} w_{hik}\, y^*_{hik}, \tag{15}$$

where $s_{m\upsilon}$ is the sample of nonrespondents to item *y* in the υ-th cell . The adjusted BRR variance estimator is given by

$$v_{ABRR}\left(\hat{Y}\right) = \frac{1}{R}\sum_{r=1}^{R}\left[\hat{Y}^{(r)}_I - \hat{Y}_I\right]^2. \tag{16}$$

When the number of donors or respondents in a particular imputation cell, υ, is not large, the adjusted imputed value $\tilde{y}^{(r)}_{hik}$ under BRR may not be well-defined, in that the denominator term, $\Sigma_{s_{r\upsilon}} w^{(r)}_{hik} x_{hik}$ is 0 nearly zero. In other words, we may not be able to re-impute in some replicates since there are no or not enough donors in some cells. For ratio imputation, it is a common practice in survey agencies to collapse cells whenever a cell respondent sample size in a replicate is less than or equal to 2, but this could lead to significant bias in the variance estimator (16). We can overcome the difficulties by using BRR(ε) with adjusted imputed values $\tilde{y}^{(r)}_{hik}(\varepsilon)$ which are obtained from $\tilde{y}^{(r)}_{hik}$ by changing the BRR weights $w^{(r)}_{hik}$ to BRR(ε) weights $w^{(r)}_{hik}(\varepsilon)$. The resulting adjusted BRR(ε) variance estimator is given by

$$v_{ABRR(\varepsilon)}\left(\hat{Y}\right) = \frac{1}{\varepsilon^2 R}\sum_{r=1}^{R}\left[\hat{Y}^{(r)}_I(\varepsilon) - \hat{Y}_I\right]^2, \tag{17}$$

where $\hat{Y}^{(r)}_I(\varepsilon)$ is obtained from (14) by changing $w^{(r)}_{hik}$ to $w^{(r)}_{hik}(\varepsilon)$. Rao and Shao (1999) established the asymptotic validity of $v_{ABRR(\varepsilon)}\left(\hat{Y}\right)$. They also conducted a simulation study with stratified random sampling, treating the strata as imputation cells, and demonstrated that v_{ABRR} can perform poorly relative to $v_{ABRR(1/2)}$ when the cells are collapsed to insure that the number of cell donors in a replicate is greater than or equal to 2. The

relative bias and c.v. of v_{ABRR} are significantly larger than these of $v_{ABRR(1/2)}$, when the difference between the y-values of the cells that are collapsed increases. Rao and Shao (1999) also studied distribution functions and quantiles using a weighted random imputation method; ratio imputation can distort the distribution of items values, as noted in Section 2.

4. CONCLUDING REMARKS

We focussed on single imputation for item nonresponse and estimation of totals or means. In particular, we studied ratio and nearest neighbour imputation methods which are deterministic and "improper" in the sense of Rubin (1996), but all the same lead to asymptotically valid frequentist inferences under an assumed response mechanism or "design-model" inferences under an assumed imputation model requiring only missing at random (MAR) assumption. It should be noted that multiple imputation is not applicable to any nonrandom imputation method because the estimator remains unchanged over imputations.

Random imputation is necessary for handling marginal second moments, distribution functions and quantiles. We briefly considered random ratio imputation. Shao and Wang's (1999) "joint regression imputation" method looks very promising because it can handle general "Swiss cheese" patterns of missing data and can preserve asymptotic unbiasedness of marginal estimators as well as estimators of correlation coefficients, under an assumed model.

Random imputation methods readily extend to multiple completed data sets, if desired (Fay, 1996). But these methods are different from Rubin's proper multiple imputation. As noted by Rao (1996) and Judkins (1996), it may not be possible to develop an imputation scheme that is proper (i.e., satisfies Rubin's three conditions), for every statistic that the user may be interested in and for general sampling designs, Judkins (1996) noted that even the first condition (asymptotic unbiasedness of the imputed estimator) may not be satisfied under the assumed response mechanism. Binder and Sun (1996) and Kott (1995) showed that the conditions required for proper imputation are generally complex and difficult to satisfy in practice under an assumed response mechanism, unless the analyst's imputation model is valid.

Even if the imputation is proper and the data are i.i.d., Rubin's method can be very inefficient when we turn our attention from point to interval estimation. Wang and Robins (1998) derived a consistent variance estimator of the "improper" estimator (similar to Rao–Shao's approach) and showed that the median length of the resulting normal theory intervals can be much smaller than the median length of Rubin's t-intervals, when the number of imputed data sets, M, is small, say M=3. This poor performance arises because Rubin's variance estimator, for fixed M, although unbiased or asymptotically unbiased, is not consistent unlike the Rao-Shao type variance estimators.

5. ACKNOWLEDGEMENTS

My thanks are due to Mr. Eric Rancourt of Statistics Canada for useful comments. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

6. REFERENCES

Binder, D.A., and Sun, W. (1996), "Frequency valid multiple imputation for surveys with a complex design", *In Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 281-286.

Chen, J., and Shao, J. (1999), "Jackknife variance estimation for nearest neighbor imputation", unpublished manuscript.

Chen, J., Rao, J.N.K., and Sitter, R.R. (2000), "Efficient random imputation for missing data in complex surveys", *Statistica Sinica, 10*, in press.

Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984), "Comparing variances from complex samples with replicate weights", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 489-494.

Eltinge, J. and Yansaneh, I.S. (1997), "Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey," *Survey Methodology, 23,* 33-40.

Fay, R.E. (1991), "A design-based perspective on missing data variance", in *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, pp. 429-440.

Fay, R.E. (1996), "Alternative paradigms for the analysis of imputed survey data", *Journal of the American Statistical Association, 91*, 490-498.

Judkins, D.R. (1990), "Fay's method of variance estimation", *Journal of Official Statistics*, *6*, pp.223-239.

Judkins, D.R. (1990), "Comment", *Journal of the American Statistical Association, 91*, 507-510.

Judkins, D.R. (1997), "Imputing for Swiss cheese patterns of missing data", in *Proceedings of Statistics Canada Symposium*, 97, pp. 143-148.
Kalton, G., and Kasprzyk, D. (1986), "The treatment of missing survey data", *Survey Methodology*, 12, 1-16.
Kott, P.S. (1995), "A paradox of multiple imputation", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 380-383.
Lee, H., Rancourt, E., and Särndal, C.E. (1995), "Variance estimation in the presence of imputed data for the Generalized Estimation System", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 384-389.
Little, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means", *International Statistical Review*, *54*, pp. 139-157.
Oh, H.L., and Scheuren, F. (1983), "Weighting adjustments for unit nonresponse", in *Incomplete Data in Sample Surveys, Vol. 2* (Madow, W.G., Nisselson, H., and Olkin, I. Eds.), Academic Press, New York, pp. 143-184.
Rancourt, E. (1999), "Estimation with nearest neighbour imputation at Statistics Canada," in *Proceedings of the Section on Survey Research Methods,* American Statistical Association, in press.
Rancourt, E., Särndal, C.E., and Lee, H. (1994), "Estimation of the variance in the presence of nearest – neighbor imputation", in *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 888-893.
Rao, J.N.K. (1996), "On variance estimation with imputed survey data", *Journal of the American Statistical Association, 91*, 499-506.
Rao, J.N.K., and Shao, J. (1992), "Jackknife variance estimation with survey data under hot deck imputation", *Biometrika*, *79*, 811-822.
Rao, J.N.K., and Shao, J. (1999), "Modified balanced repeated replication for complex survey data", *Biometrika*, *86*, 403-415.
Rao, J.N.K., and Sitter, R. (1995), "Variance estimation under two-phase sampling with application to imputation for missing data," *Biometrika, 82,* 453-460.
Royall, R.M. (1970), "The linear least squares prediction approach to two-stage sampling", *Journal of the American Statistical Association*, *71*, 657-664.
Rubin, D.B. (1996), "Multiple imputation after 18 + years", *Journal of the American Statistical Association, 91*, 473-489.
Särndal, C.E. (1992), "Methods for estimating the precision of survey estimates when imputation has been used", *Survey Methodology, 18*, 242-252.
Schaeffer, J.L., Khare, M., and Ezzati-Rice, T.M. (1993), "Multiple imputation of missing data in NHANES III", in *Proceedings of the 1993 Annual Research Conference*, U.S. Bureau of the Census, 459-487.
Shao, J., and Steel, P. (1999), "Variance estimation for survey data with composite imputation and nonnegligible sampling fractions", *Journal of the American Statistical Association*, *94*, 254-265.
Shao, J., Chen, Y., and Chen, Y. (1998), "Balanced repeated replication for stratified multistage survey data under imputation", *Journal of the American Statistical Association*, *93*, 819-831.
Shao, J., and Wang, H. (1999), "Sample correlation coefficients based on survey data under regression imputation", unpublished manuscript.
Sitter, R., and Rao, J.N.K. (1997), "Imputation for missing values and corresponding variance estimation", *Canadian Journal of Statistics*, *25*, 61-75.
Skinner, C.J., and Rao, J.N.K. (2000), "Jackknife variance estimation for multivariate statistics under hot deck imputation from common donors", *Journal of Statistical Planning and Inference", 79,* in press.
Srivastava, M.S., and Carter, E.M. (1986), "The maximum likelihood method for non-response in sample surveys", *Survey Methodology*, *12*, 61-72.
Yung, W., and Rao, J.N.K. (2000), "Jackknife variance estimation under imputation for estimators using poststratification information", *Journal of the American Statistical Association, 95,* in press.

ACCOUNTING FOR IMPUTATION ERROR VARIANCE FOR ESTABLISHMENT DATA: AN EMPIRICAL EVALUATION

Thomas Krenzke, Jill Montaquila, and Leyla Mohadjer, Westat
Thomas Krenzke, Westat, 1650 Research Boulevard, Rockville, MD 20850, KrenzkT1@Westat.com

ABSTRACT

Several methods have been developed to account for the effects of imputation error in variance estimation. When performing variance estimation for establishment data containing imputed values, special considerations include the necessity of providing public file users with the information needed to incorporate the effects of imputation error into variance estimates, interpreting the effects of outliers, and estimating variance in situations involving mixed imputation methods, sequential imputation, or hierarchical response patterns. We use an establishment data file to explore ways of addressing these special considerations. This paper also contains empirical results from various approaches to variance estimation for establishment data.

Key Words: imputation error variance, empirical study, establishment data

1. INTRODUCTION

Imputation is often used in establishment surveys to compensate for item nonresponse. Imputation involves replacing a missing value with a nonmissing value, typically generated from a statistical model. Imputation error occurs when the imputed value is different from the true value. If imputed values are treated as if they had actually been observed or reported, the variance of the estimate may be substantially underreported, since the variance does not account for the imputation error in the data. There is a great deal of literature addressing variance estimation for data containing imputed values. Variance methods include multiple imputation (Rubin 1987), model-assisted variance estimation (Särndal 1990), the adjusted jackknife (Rao and Shao 1992), fractionally weighted imputations (Fay 1996), the bootstrap method (Shao and Sitter 1996), and the all-cases imputation (ACI) method (Montaquila and Jernigan 1997). Variance methods are discussed in Section 2.

When data contain imputed values, special considerations include the necessity of providing public file users with the information they need to incorporate the effects of imputation error in variance estimates, interpreting the effects of outliers, and performing variance estimation in situations involving mixed imputation methods, sequential imputation, or hierarchical response patterns. The aforementioned issues are common in establishment surveys. The objective of this paper is to discuss the relative performance and applicability of the imputation methods and variance estimation methods as each are applied to our establishment data. Several evaluation conditions and criteria are used. We discuss each of the criteria and provide results of the empirical study in Section 3.

2. VARIANCE METHODS THAT ACCOUNT FOR IMPUTATION ERROR VARIANCE

Because standard variance formulas do not account for the fact that imputed values are not true values, several methods have been developed to account for the effects of imputation error in variance estimation. Särndal (1990) decomposed the total variance as follows:

$$V_{TOT} = V_{SAM} + V_{IMP} + V_{MIX}, \qquad (1)$$

where, $V_{SAM} = V_{ORD} + V_{DIF}$ is the sampling error component; V_{DIF} is an adjustment needed for deterministic imputation because, without it, V_{SAM} would underestimate the true sampling error; V_{IMP} is the imputation error variance component; and V_{MIX} is the covariance between sampling error and imputation error.

In this section, we describe five variance estimation methods which account for imputation error and which we applied to the establishment data.[1] We describe how WesVar,[2] a software package for analyzing data from complex

[1] Finite population correction (fpc) factors were not used in this study because, under pps sampling, the factors overadjust the variance and produce estimates that are negatively biased. Sitter and Rao (1997) and Shao and Steel (1999) have studied the finite population correction problem in the presence of imputed values. It should be noted that Adjusted Jackknife and Bootstrap do not explicitely split the calculations into variance components, and therefore, if the fpc factors are applied, they would be applied to all variance components, when they should only be applied to the sampling error variance component.
[2] WesVar is developed and distributed by Westat (e-mail address: wesvar@westat.com).

surveys using replication methods, can be used to generate the variance estimates or components of the estimates of total variance.

The stratified jackknife (Wolter 1985) is introduced because of its use in many situations. We applied this method to generate 200 replicates for use in estimating variances for the establishment data. The ordinary variance component, $\hat{V}_{ORD}$, can be computed as follows:

$$\hat{V}_{ORD} = \sum_g k_g (\hat{\theta}_g - \hat{\theta}_0)^2 , \qquad (2)$$

where, $k_g = (n_{h'} - 1)/n_{h'}$ is the stratified jackknife factor for replicate g, h' identifies the stratum that is aligned with replicate g, and $n_{h'}$ is the number of variance units within stratum h'; $\hat{\theta}_0$ is the parameter estimate for θ using the full-sample weight; and $\hat{\theta}_g$ is the parameter estimate for replicate g. In the computations of $\hat{\theta}_0$ and $\hat{\theta}_g$, both observed and imputed values are used. In our analysis, the parameter θ is the population mean.

2.1. Model-Assisted Approach

The model-assisted approach to variance estimation for imputed data was introduced by Särndal (1990). A different variance formula is required for each sample design and each imputation model. Lee et al. (1995) give formulas for estimating each component in (1) for the mean within-cells hotdeck (MHD, refer to Section 3.4.1) and random within-cells hotdeck (RHD, refer to Section 3.4.2) methods. The ordinary variance component, $\hat{V}_{ORD}$, is computed as in (2), using observed and imputed values. The components $\hat{V}_{DIF}$ and $\hat{V}_{IMP}$ are computed using only the respondents, and each computation uses the expression $\sum_g f_g k_g (\hat{\theta}_g - \hat{\theta}_0)^2$. For the MHD method, the term $\hat{V}_{DIF}$ uses the factors $f_g = \frac{m_{h'}}{n_{h'}^2}(r_{h'} - 1)$, where $r_{h'}$ = the number of item respondents in stratum h'; $m_{h'}$ = the number of item nonrespondents in stratum h'; and $n_{h'}$ = the number of unit respondents in stratum h'. For the RHD method, $\hat{V}_{DIF} = 0$. For the second component, $\hat{V}_{IMP}$, the factors are $f_g = \frac{m_{h'}}{n_{h'}^2}(\frac{m_{h'}}{r_{h'}} + c)(r_{h'} - 1)$, where $c = 1$ for MHD and $c = 2$ for RHD. For domain estimation, a set of factors may be computed for each domain, by replacing $m_{h'}$ with $m_{h'd}$ (i.e., by replacing the number missing in stratum h' with the number missing in stratum h' for domain d). If the sampling and response mechanisms are unconfounded, then $\hat{V}_{MIX} = 0$, as we assumed for this study.

2.2. All-Cases Imputation

The all-cases imputation (ACI) method (Montaquila and Jernigan 1997) is a model-assisted approach where imputations are generated for all cases, respondents and nonrespondents, for variance estimation purposes only. The procedure is generally applicable to many sample design and imputation methods; however, the variance estimator must be rederived for each situation. The ACI concept is to apply the imputation scheme once to respondents as it was applied to nonrespondents. Then the imputation error, $\tau_i = y_i^* - y_i$, where y_i is the observed value of the characteristic y for case i and y_i^* is the imputed value of y for case i, is computed for respondents and used to estimate the imputation error variance among nonrespondents. Montaquila and Jernigan (1999) provide guidelines for constructing variance estimators for complex sample designs and/or various imputation models. The ordinary variance estimate, $\hat{V}_{ORD}$, is computed as in (2), using observed values for respondents and imputed values for nonrespondents. Using the stratified jackknife for the set of respondents, R, the second component, $\hat{V}_{IMP}$, is computed as $\hat{V}_{IMP} = \sum_g f_g k_g (\hat{\theta}_g - \hat{\theta}_0)^2$, where $f_g = \frac{r_{h'}}{n_{h'}} \times \frac{m_{h'}}{n_{h'}}$, $\hat{\theta}_g = \frac{\sum_{i \in R} w_{gi} \tau_i}{\sum_{i \in R} w_{gi}}$, and $\hat{\theta}_0 = \frac{\sum_{i \in R} w_{0i} \tau_i}{\sum_{i \in R} w_{0i}}$. More research is

needed for domain estimation and for the development of appropriate factors (f_g) for the ACI method. In the ACI method, a formula for the variance component $\hat{V}_{DIF}$ is not developed when deterministic imputation is applied; however, we recommend using respondents only in (2), which provided a good approximation for establishment data as shown in Section 3.5.1.

2.3. Adjusted Jackknife Method

The adjusted jackknife approach for hot-deck imputation was introduced by Rao and Shao (1992). Using the full sample, one imputed value (y_{ij}^*) is obtained for each nonrespondent. The adjusted jackknife variance estimator is used and separate adjustments are applied to the imputed values in each jackknife replicate, because each replicate (or reduced sample) includes only a subset of the donors available in the full sample. For the RHD method, the adjustment for item j is applied to each unit (i) and replicate (g), within each imputation class (k), as $a_{gjk} = \bar{y}_{gjki}(-g) - \bar{y}_{gjki}$, where $\bar{y}_{gjki}(-g)$ is the mean for the reduced sample (i.e., replicate g). Note that the terms in the adjustment are computed using item respondents only. One variance formula applies to most sample designs, although the specific adjustment for imputed values depends on the imputation method. Rao (1996) provides an overview of the jackknife technique and describes the theory underlying the adjusted jackknife for ratio imputation, regression imputation, and stochastic regression. The Rao (1996) paper also offers a methodology for stratified multistage sampling using a weighted hot deck within cells and for domain estimation. The adjusted jackknife approach was applied to the establishment data imputed using the RHD method. We considered a weighted adjustment of the imputed values, similar to the one specified by Nixon (1996). An unweighted adjustment could also be considered. For item nonrespondents, the imputed values were adjusted as $y_{gij}^* = a_{gjk} + y_{ij}^*$. Therefore, for each facility, for each replicate, there is an adjusted imputed value for each missing item (for observed values, the value is unadjusted for each replicate). However, these separate imputed values for each replicate do not need to be stored since they can be calculated using the full-sample imputed and observed values along with imputation class identifiers.

2.4. Bootstrap Method

A bootstrap method for variance estimation with imputed data was proposed by Shao and Sitter (1996). This method will be referred to henceforth as 'the bootstrap method', and should not be confused with the approximate bayesian bootstrap, which will be discussed in Section 2.5. The approach involves independently drawing B bootstrap samples[3] of size n such that n_h units are selected with replacement from stratum h. Shao and Sitter (1996) apply the same imputation procedure used on the full-sample data set to impute for nonrespondents in each bootstrap sample and compute the estimate based on the imputed bootstrap data set. The following Monte Carlo approximation is then used to obtain the bootstrap variance estimate:

$$v_B(\hat{\theta}) = \frac{1}{B}\sum\left(\hat{\theta}_b^* - \bar{\theta}^*\right)^2 , \qquad (3)$$

where $\hat{\theta}_b^*$ is the value of the estimate for the b^{th} bootstrap sample and $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*$. Shao and Sitter (1996) extend this approach to stratified, multistage sample designs; ratio and regression imputation; and multivariate estimators. The authors show that this bootstrap variance estimator is consistent under a variety of conditions (i.e., different sample designs, types of estimators, and imputation methods). The bootstrap is a flexible method of accounting for imputation error variance, in that it is applicable to virtually any imputation method, sample design, and type of estimator. The primary drawbacks of the bootstrap method are the computational burden associated with obtaining separate imputations within each bootstrap sample and the data storage issues associated with delivering imputed values for each bootstrap sample.

[3] For the evaluation presented in Section 3.5, we used the naive bootstrap procedure, selection of n_h units, described by Shao and Tu (1995) to select the bootstrap samples.

2.5. Multiple Imputation with the Approximate Bayesian Bootstrap

When the approximate bayesian bootstrap (ABB) is used to generate multiple imputations, separate analyses are performed for each of $M > 1$ completed data sets and the results are combined at the end, using the variation in the M estimates to account for variance due to imputation error (Rubin 1987). Using software developed by Frank Harrell,[4] the approximate bayesian bootstrap was applied to our establishment data. The ABB has practical and operational drawbacks. The M imputed values must be made available to the user (i.e., multiple imputations must be included in the data file). Furthermore, the validity of the ABB hinges on the creation of "proper" imputations; this term has a very specific meaning in the context of multiple imputation, and its interpretation is often ambiguous. One problem is that the definition of a proper imputation depends on the specific analysis of interest; what is proper for one analysis may not be proper for another. Fay (1992), Kott (1995), Judkins (1996), and Binder and Sun (1996) provide discussions about multiple imputation from the frequentist point of view.

3. EVALUATION CRITERIA AND RESULTS

The objective of this paper is to discuss the relative performance and applicability of the variance methods as applied to establishment data. Due to using one datafile and the fact that the true values are unknown for item nonrespondents, we are in no position to make conclusions about which method yields the most accurate and stable variance estimates. A simulation study is appropriate for that type of evaluation. However, using the establishment data proved valuable as several special considerations arose. We will discuss our observations as several criteria were used to evaluate the five variance estimation approaches described in Section 2.

3.1 Complex sample design. In this study, we used establishment data from a survey of facilities providing treatment in outpatient or inpatient settings. This survey used a stratified, systematic, probability proportional to size (pps) sample of approximately 2,400 responding treatment facilities from seven strata. The strata were based on the type of treatment and other characteristics of the facility's client population. Facilities were restratified based on their responses to the facility questionnaire. Stratum migration may have occurred for facilities that reported changes in their types of treatment and/or client types. For this empirical study, the variables imputed were limited to the highly skewed, highly correlated variables total annual revenues (denoted y) and total annual costs (denoted z). The auxiliary data, x, used in this study included total number of clients, type of ownership (private for profit, private nonprofit, and public), and Census region (Northeast, Midwest, South, and West).

We explore the use of replication methods when estimating variance components for this complex sample design. The adjusted jackknife method and the bootstrap method are the most adaptable to various sample designs because they were developed from their complete-data replication method counterparts. We found that replication methods (implemented using WesVar, for example) have the potential to relieve the ACI and model-assisted approaches, in part, of their limitation to certain sample designs. In addition, for the multiple imputation variance calculations, the stratified jackknife variance estimator was used to estimate the 'within' (or analogously, V_{SAM}) variances.

3.2 Types of Estimators. In this study we focus on the weighted mean. Domain estimation (subgroup analysis) is also investigated. The bootstrap and adjusted jackknife methods seem to have the most flexibility in terms of different estimators because they use replication methods. In our study, we observed that the bootstrap and adjusted Jackknife methods were the most attractive for domain estimation. Formulas have been developed for the model-assisted approach for domain estimation under certain sample designs, imputation methods, and estimation methods. The ACI method requires more development in this area.

3.3 Public file users' ability to capture imputation error. Variance methods may have different data requirements that allow for the calculations to be done. If the data must be provided to other users, it is always a good idea to include flags that identify imputed values. With just imputation flags added to the files, one can conceivably apply any of the above variance estimation methods. Because most data users are probably not familiar with the application of variance estimation methods in the presence of imputed values, it may be necessary to supply more information. Consider the following data file representations:

[4] Frank Harrell's software for multiple imputation may be obtained from the website http://hesweb1.med.virginia.edu/biostat/s/win/. The methodology used is described in Rubin (1987).

Model-Assisted:	$W_{n*(G+1)} \mid Y_{n*J} \mid F_{n*J}$
Adjusted Jackknife:	$W_{n*(G+1)} \mid Y_{n*J} \mid F_{n*J} \mid C_{n*k}$, where 1 <= k <= J
ACI:	$W_{n*(G+1)} \mid Y_{n*2*J} \mid F_{n*J}$
Approximate Bayesian Bootstrap:	$W_{n*(G+1)} \mid Y_{n*(M*J)} \mid F_{n*J}$
Bootstrap:	$W_{n*(G+1)} \mid Y_{n*((G+1)*J)} \mid F_{n*J}$

where, n= number of unit respondents, G= number of replicates; J= number of imputation variables, M = number of multiple imputations, F = matrix of imputation flags, T = matrix of imputation error; Y = matrix of imputation variables; W=matrix of weights; C = matrix of imputation cells.

The model-assisted approach requires the least amount of data for the analyst: the full sample and replicate weights, set of imputation variables, and their associated imputation flags. In addition to what is required for the model-assisted approach, the adjusted jackknife requires that the imputation cells be delivered with the file. The ACI method also requires more than the model-assisted approach since the ACI variance calculations need the observed and imputed values, as well as the imputed values for the item nonrespondents. The amount of information for the ABB is more than the ACI method if the number of multiple imputation is greater than 2. The bootstrap requires the most data since imputation are needed for each replicate and for each imputation item. To simplify the analyst calculations even further, follow the suggestion of Madow et al. (1983) in chapter 4 of volume 1. These authors recommend that if exact methods are not used to estimate variance, approximate "multipliers" (e.g., variance inflation factors (VIFs)) can be derived from more exact methods in order to inflate the sampling error variance. The VIFs may be used to account for imputation error variance for imputed items through the product of the appropriate VIF and the analyst's resulting variance estimate $\hat{V}_{SAM}$. It may be appropriate to provide the VIFs in documentation for public use files that contain imputed values. VIFs are usually computed as, $VIF = \hat{V}_{TOT} / \hat{V}_{SAM}$. In general, a tool to evaluate VIFs in practice is to check them to see if they are correlated with the item nonresponse rates.

3.4 Practicality of computations. The computations involved in applying the variance methods depend on the role of the statistician. The statistician could play the role of the file creator and/or the file user (analyst). The intensity of the calculations for the analyst will depend on the amount of work done by the file creator. In our study, we found that although the bootstrap approach is the most computationally intensive, it was straightforward to program. A special-purpose software program was written for the adjusted jackknife approach.

3.5 Imputation methods. Imputation is used to reduce nonresponse bias in survey estimates, simplify analyses, and improve the consistency of results across analyses. Imputations should also preserve multivariate distributions. The reduction of nonresponse bias and the preservation of multivariate distributions are particularly challenging if the item to be imputed is not missing at random. Most imputation methods assume that missingness occurs at random (Little and Rubin 1987). In this study, we considered five imputation methods, assuming missingness at random, as they are applied to establishment data. The first approach is a deterministic method, and the other four are stochastic methods.

3.5.1 Mean within-cells hot deck (MHD). The unweighted mean among the item respondents was imputed for each facility missing the particular item within each imputation cell. The imputation cells for the hot-deck methods were constructed using auxiliary data within the strata. The cells were decided upon by modeling the response propensity of the imputation items *y* and *z* on the auxiliary variables (x). The cells were formed within strata using the software CHAID. CHAID is the name given to one version of the Automatic Interaction Detector (AID) that has been developed for categorical variables. We applied the ACI and model-assisted variance estimation approaches to the deterministic MHD imputation model and computed standard error ratios relating the two approaches (with the ACI standard error in the denominator). When we used just item respondents to compute $\hat{V}_{SAM}$ under the ACI method (as discussed in Section 2.2), the standard error ratios were close to 1.00 (0.98 and 0.99 for total revenues and total costs, respectively); thus, the two variance estimation approaches yielded similar results.

3.5.2 Random within-cells hot deck (RHD) – mixed methods. In our study, mixed imputation methods were used to impute the same item. A single donor was randomly selected within each imputation cell, as formed for the MHD approach, for each case with at least one missing value for the set of imputation variables $\{y, z\}$. If both

imputation variables were missing, the recipient received the values from the donor. Donor ratios (either *y/z* or *z/y* were applied whenever there was one nonmissing imputation variable.

For the RHD imputation scheme, we used the ACI, adjusted jackknife, bootstrap, and model-assisted approaches to compute total variance estimates. We also produced repeated imputations by generating M=3 RHD imputations, then applied the multiple imputation variance formula to estimate the variance. The adjusted jackknife method and model-assisted approach were applied to the imputations under our "single donor" RHD approach by ignoring that imputations came from mixed methods. The issue of mixed imputation methods has not been resolved in the current adjusted jackknife approach. Skinner and Rao (1993) discussed this issue. For the bootstrap method, we selected 200 bootstrap samples from the establishment survey's sample. Table 3-1 provides standard error ratios, which estimate the increase in the standard error due to item nonresponse and imputation. The standard error ratios were calculated with $\sqrt{\hat{V}_{SAM}}$ as the base, where $\hat{V}_{SAM}$ was computed using the stratified jackknife replicates, treating imputed values as if they were observed[5]. One might consider the square of the standard error ratio to be the VIF. For this establishment survey datafile, the bootstrap standard error ratios vary the most. However, a simulation study is needed to make conclusions about the stability of the variance estimators. Another point is that the standard error ratios are attractively correlated (moderately) with the item nonresponse rate for the ACI, adjusted jackknife, and the model assisted approach.

Table 3-1. Comparing variance methods under the RHD imputation scheme (standard error ratios)

		Standard Error Ratios					Nonresponse
Item	Stratum	ACI	Jackknife	Bootstrap	M-A	Repeated	Rate (%)
Total revenues	1	1.04	1.08	0.87	1.10	0.99	9.7
	2	1.04	1.06	1.00	1.08	0.97	10.6
	3	1.06	1.02	1.14	1.09	1.00	4.0
	4	1.12	1.19	1.02	1.12	1.06	20.2
	5	1.09	1.02	1.33	1.06	1.64	4.0
	6	1.06	1.09	1.25	1.09	1.03	9.3
	Overall	1.03	1.05	0.95	1.08	1.03	9.4
Total costs	1	1.07	1.08	0.90	1.12	1.03	10.7
	2	1.04	1.05	0.69	1.04	0.86	14.2
	3	1.08	1.02	1.17	1.11	1.00	4.0
	4	1.11	1.25	1.00	1.14	1.08	21.7
	5	1.09	1.02	1.15	1.06	1.07	4.0
	6	1.08	1.06	1.24	1.10	1.07	10.5
	Overall	1.05	1.05	0.81	1.06	0.90	10.6

ACI, all-cases imputation; Jackknife, adjusted jackknife method; M-A, model-assisted method; Bootstrap, bootstrap method; Repeated: repeated RHD imputations three times and used the multiple imputation variance formulas to compute the variances.

With regard to mixed imputation methods, Rancourt et al. (1994) and Sitter and Rao (1997), describe using ratio imputation when auxiliary variable *x* is available and mean imputation when *x* is not available. Sitter and Rao (1997) obtained linearization variance estimators as well as jackknife variance estimators when ratio imputation and mean imputation were used to impute the same item. We found that the ACI can be adapted to this problem since when imputing respondents in mixed-methods situations under the ACI model, one simply applies the appropriate imputation methods to respondents under the same rules and at the same rate as when the imputation scheme is applied to nonrespondents.

3.5.3 Random regression – nonsequential approach. In this study, within each stratum, regression models were constructed for log(*y*) using auxiliary data[6], and imputations were created by adding random error (drawn from the

[5] When computing VIFs or standard error ratios to provide to data users, the denominator should be computed treating imputed values as if they were observed, using the replicate weights that are on the file. For instance, if bootstrap replicate weights are being provided to the analyst, then the VIFs or standard error ratios should use the bootstrap replicates. Otherwise, as seen in Table 3-1 (and Table 3-3), the VIFs or standard error ratios may be less than 1. For our study, we used the same denominator for all methods, in order to facilitate comparisons among methods.

[6] A log transformation was done on continuous independent variables.

appropriate normal distribution) to predicted values. The resulting values were transformed back to the original scale to arrive at the imputed values. Our nonsequential random regression imputation approach entailed imputing *z* without *y* in the imputation model, because of nonresponse in *y*. For the nonsequential random regression imputation scheme, we used the ACI, bootstrap, and the approximate bayesian bootstrap (ABB) approaches to compute total variance estimates.

3.5.4 Random regression – sequential imputation. Often, two items that are both subject to missingness are highly correlated, suggesting that one item should be used in the imputation of the other item. In such cases, it may be desirable to perform sequential imputation, where the former item is imputed first. After imputing for y, as described in Section 3.5.3, regression models were fit for log(*z*), given *x* and the observed and imputed values for log(*y*). A generalization of this approach is described in Judkins (1997), and involves cycling back and reimputing for *y* using the observed and imputed values of *z*. Because of its iterative nature, it is likely that the cyclical imputation approach does a better job of maintaining the joint distribution between *y* and *z*.

The ACI method can be applied to the sequential imputation scheme. Let *p* be the proportion of nonrespondents to item *z* that had imputed values for *y*. When imputing for respondents, the imputation model $z_{ri}^{*} = b_{r0} + \sum_{j=1}^{J-1} b_{rJ} x_{rij} + b_{rJ} y_{riJ}^{*} + \hat{e}_i$ is used 100·*p*% of the time and the imputation model $z_{ri}^{*} = b_{r0} + \sum_{j=1}^{J-1} b_{rJ} x_{rij} + b_{rJ} y_{riJ} + \hat{e}_i$ is used 100· (1-*p*)% of the time. For respondents to *z*, imputation error is computed as $\tau_{ri} = z_{ri}^{*} - z_{ri}$. Then the imputation error variance is computed as discussed in Section 2.2. We compared the sequential imputations to those obtained with the nonsequential approach, where imputations for *z* used auxiliary variables *x* but not observed and imputed values of *y*. We generated new regression models and used the respondents' auxiliary variables in the models to generate imputations. Table 3-2 provides the standard error ratios relating nonsequential imputation (the numerator) to sequential imputation (the denominator). The table also shows item nonresponse rates for *z*, observed correlations between *x* and *z* $\left(\rho_{obs}(x,z)\right)$, observed correlations between y and z $\left(\rho_{obs}(y,z)\right)$, and correlations when imputed values were present for sequential imputation $\left(\rho_{seq}^{(*)}(y,z)\right)$, and for nonsequential imputation $\left(\rho_{nonseq}^{(*)}(y,z)\right)$. We expected some mixed results because there is an added layer of imputation error for sequential imputation as a result of using imputed values in the prediction model; but at the same time stronger models are used. For our application, sequential imputation does a better job of preserving the correlations between *y* and *z*, especially where the nonresponse rate is high. Five of the six strata have standard error ratios greater than 1, demonstrating general reductions in variances when the sequential approach is used.

Table 3-2. Comparison of sequential and nonsequential imputation

Stratum	Standard Error Ratios	$\left(\rho_{obs}(x,z)\right)$	$\left(\rho_{obs}(y,z)\right)$	$\left(\rho_{seq}^{(*)}(y,z)\right)$	$\left(\rho_{nonseq}^{(*)}(y,z)\right)$	Nonresponse Rate (%)
1	0.975	0.267	0.929	0.929	0.902	10.7
2	1.119	0.216	0.340	0.412	0.363	14.2
3	1.137	0.457	0.991	0.990	0.984	4.0
4	1.104	0.537	0.915	0.882	0.790	21.7
5	1.060	0.523	0.842	0.846	0.825	4.0
6	1.078	0.748	0.985	0.984	0.945	10.5

3.5.5 Approximate Bayesian Bootstrap. Using the approximate bayesian bootstrap, we generated multiple imputations by generating bootstrap samples and then sampling from the bootstrap samples of respondents to obtain imputed values for nonrespondents. *M=5* imputations were generated for each nonrespondent, resulting in five completed data sets. Rubin (1987) states that modest values of *M* are sufficient when the fraction of missing information is modest. Table 3-3 provides the standard error ratios, using the stratified jackknife standard errors as the denominator, which treats imputed values as if they were observed. For all three approaches, the nonsequential imputation approach was used. As seen in Table 3-1, for the establishment datafile, Table 3-3 shows the standard error ratios varying across the variance methods. We note a couple of reasons for this. One factor is the model

building process. When processing the calculations for the approximate bayesian bootstrap, the results differed dramatically across different models depending on how the model was fit. The amount of care taken in the imputation modeling affects variance estimates from all three estimators shown in Table 3-3. Another reason, as will be discussed in Section 3.7, is the effect that outliers have on the variance estimates, which can occur from several ways including from the way the random error is added to the regression predicted values.

Table 3-3. Comparing variance methods under the random regression scheme (standard error ratios)

Item	Stratum	Standard Error Ratios			Nonresponse Rate (%)
		ACI	Bootstrap	ABB	
Total revenues	1	1.06	0.87	1.19	9.7
	2	1.04	0.96	1.01	10.6
	3	1.06	1.13	1.00	4.0
	4	1.35	1.46	1.48	20.2
	5	1.04	1.23	1.22	4.0
	6	1.05	1.07	1.46	9.3
	Overall	1.05	0.94	1.09	9.4
Total costs	1	1.09	0.88	1.41	10.7
	2	1.12	0.93	1.02	14.2
	3	1.28	1.20	1.28	4.0
	4	1.12	1.05	1.31	21.7
	5	1.02	1.14	1.13	4.0
	6	1.06	0.99	1.08	10.5
	Overall	1.08	0.91	1.21	10.6

To evaluate the effect that imputation methods had on the resulting variance estimates from the establishment file, the ACI variance estimate was computed under each imputation model. For total revenues, the ACI standard errors under RHD and random regression were 9% and 1% higher that of MHD, respectively. For total costs, the ACI standard errors under RHD, random regression sequential and nonsequential imputation were 24%, 3%, and 9% higher, respectively. The standard error ratios demonstrate the anticipated pattern, with the standard error being lowest under the MHD model [where $\hat{V}_{SAM}$ is estimated using respondents in (2)]. The random regression imputation model yielded lower ACI standard errors than did the RHD model. Because the continuous auxiliary variable was categorized for the hot-deck approach, we would expect to see a loss in predictability with the RHD method. Total variance was lower under the sequential imputation model than under the nonsequential model, due to the high correlation between revenues and costs and the use of that strong relationship in the sequential imputation model.

3.6 Hierarchical response patterns. In addition to mixed methods and sequential imputation, another fairly common special consideration results from skips in the questionnaire; once an item has been imputed, the applicability of items triggered from that item can be determined. Therefore, some items become missing (or skipped) only after their trigger items have been imputed. Thus, the imputation of the trigger item must also be considered as a component of imputation error for the ultimate item. The ACI method can be applied to a limited group of hierarchical nonresponse patterns. One set of questions in the establishment survey asked, "Did this facility admit clients of type Q? If so, how many were admitted?" The trigger item, y_1, must be imputed before the second item, y_2, can be imputed. Let p denote the proportion of nonrespondents to the second item, y_2, which had imputed values of the trigger item y_1. When imputing for respondents to the second item, y_2, we must first impute for the respondents' associated trigger item $100 \cdot p$% of the time when applying the RHD method, where donors and recipients are the same set of respondents to item y_1. For the respondent imputation of y_2, where $y_{r1}^{(*)}$ = yes, the donors are facilities with $y_{r2} > 0$. When $y_{r1}^{(*)}$=no, then $y_{r2}^{(*)}$ is set to 0. We conducted a small study to compare the following scenarios: 1) not imputing for the trigger item for respondents; that is, the trigger item value is assumed to be the true value, so no imputation error is accounted for in the trigger item; 2) imputing for the respondent's trigger item for $100 \cdot p$% (computed as 33%) of the respondents; that is, a proportion (i.e., $100 \cdot (1 - p\%)$) of the trigger item's values is assumed to be true; and 3) imputing the trigger item for all respondents, which would likely account for too much imputation error in the trigger item. The item nonresponse rate for item y_2 was 12.28%. The VIFs computed for y_2 for each of the three scenarios followed the expected pattern: 1.065, 1.070, and 1.101, respectively.

3.7 Establishment data. Establishment data are generally skewed, that is, many units are small in size, and few are large in size. Our study data is no different, as variables such as number of clients, revenues, and costs have skewed distributions. The skewed data from establishment surveys have an indirect effect on the variance estimates. If the imputation model is not appropriate for the data, then unreliable, highly variable and outlier imputations will result, affecting the variance estimates. We checked to see how these outliers affected the resulting variances from each variance method. When the adjusted jackknife estimate was investigated for total revenues for stratum 4 (standard error ratio = 1.19), shown in Table 3-1, we noticed that the largest adjustment factors for replicates and hot-deck cells corresponded to the replicates with the largest differences in mean total revenues between the adjusted and initial replicates. For these replicates, most of the dropped respondents were either much smaller or much larger than average (as a group) and usually had larger than average weights. We also noticed that most of these replicates fell into one of the three hot-deck cells. The nonresponse rate within this hot-deck cell was 37%. Therefore, if the reported values for a particular variance unit/hot-deck cell combination are larger (or smaller) than the average for the cell, the adjustment factor will be highly negative (or positive). In addition, if the imputed values in that hot-deck cell are small (or large), the revised imputed values will deviate further from average after adjustment, and thus the replicate's estimate will be further from the full-sample estimate. This is perhaps most likely to occur in establishment surveys (skewed distribution) where a domain has a large nonresponse rate.

In another example of how outlier imputed values affect imputation error variance, in Table 3-3, the standard error ratios for total revenues in stratum 4 (1.35) was investigated. It was noticed that the VIF (i.e., square of the standard error ratio) was out of line with the nonresponse rate (e.g., VIF = 1.81, Nonresponse Rate = 20.2%). With further investigation, it was determined that a large imputed value for an item respondent resulted in a large imputation error value, which made a large contribution to the imputation error variance component. The VIF value was reduced from 1.81 to 1.26 by discarding 1 of the 162 respondent imputed values calculated for the ACI method. This case demonstrates the sensitivity of the ACI approach to outliers and its ability to capture the effects of outliers in imputation for item nonrespondents. With the ACI method, imputation for respondents can be used as a diagnostic test, by examining how the imputation model performs on the respondents themselves. No other method has this feature built into its procedures; however, other methods (with the exception of the model-assisted approach can capture some of the effects of outlier imputed values in their computation of imputation error variance.

4. CONCLUDING REMARKS

In this paper we focused on applying five variance estimation methods for imputed data to a complex establishment survey's data. Our goal was to compare the performance and applicability of the variance methods under several conditions and criteria, giving file creators and analysts several factors to consider when choosing variance estimators. This paper examines several computational schemes for accounting for imputation error variance. The major conclusions are that the choice of variance estimators in the presence of imputed values largely depends on several factors. The sample design complexity, imputation methods, and estimators were discussed, as well as accounting for imputation error in the presence of special imputation situations, such as sequential imputation, mixed methods, and hierarchical response patterns. Another important consideration is to carefully consider the impact of outliers on the variance calculations. We saw that outliers can result from establishment data, inefficient variance models, and adding random error. Lastly, observations were discussed relating to the practicality of computations and the data delivery and analytical requirements for computing variances that account for imputation error in the data. Further research is needed to compare the stability of the variance estimators under establishment survey conditions.

5. REFERENCES

Binder, D.A., and Sun, W. (1996), "Frequency Valid Multiple Imputation for Surveys with a Complex Design," Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 281-286.

Fay, R. (1992), "When are Inferences from Multiple Imputation Valid," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 227-232.

Fay, R. (1996), "Alternative Paradigms for the Analysis of Imputed Survey Data," *Journal of the American Statistical Association*, Vol. 91, No. 434.

Judkins, D. (1996), "Comment," *Journal of the American Statistical Association*, Vol. 91, pp. 507-510.

Judkins, D. (1997), "Imputing for Swiss Cheese Patterns of Missing Data," in *Proceedings of Statistics Canada Symposium '97*, pp. 143-148.

Kott, P. (1995), "A Paradox of Multiple Imputation," *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 380-383.

Lee, H., Rancourt, E., and Särndal, C.E. (1995), "Variance Estimation in the Presence of Imputed Data for the Generalized Imputation System," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Lee, H., Rancourt, E., and Särndal, C.E. (1999), "Variance Estimation from Survey Data Under Single Imputation," To appear in the *Proceedings of the International Conference on Survey Nonresponse*.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with MissingData*. New York: Wiley.

Madow, W.G., Nisselson, H., and Olkin, I. (1983), *Incomplete Data in Sample Surveys*. Vol. 1, Chapter 4. New York: Academic Press.

Montaquila, J., and Jernigan, R.W. (1997), "Variance Estimation in the Presence of Imputed Data," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Montaquila, J., and Jernigan, R.W. (1999). "Variance Estimation. I: Accounting for Imputation." Seminar presented to the Washington Statistical Society.

Nixon, M., Kalton, G., and Brick, M. (1996), "Variance Estimation with Missing Best Values in the NIPRCS," *Proceedings of the Section on Survey Research Methods of the American Statistical Association*.

Rancourt, E., Lee, H., and Särndal, C.E. (1994), "Bias Corrections for Survey Estimates from Data with Ratio Imputed Values for Confounded Nonresponse," *Survey Methodology*, Vol. 20, pp. 137-147.

Rao, J.N.K. (1996), "On Variance Estimation with Imputed Survey Data," *Journal of the American Statistical Association*, Vol. 91, No. 434.

Rao, J.N.K., and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot-Deck Imputation," Biometrica, Vol. 79, pp. 811-822.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons.

Särndal, C.E. (1990), "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used," *Proceedings of Symposium '90: Measurement and Improvement of Data Quality*, pp. 337-347. Ottawa: Statistics Canada.

Shao, J., and Sitter, R. (1996), *Bootstrap for Imputed Survey Data*. Technical Report 227. Laboratory for Research in Statistics and Probability, Carleton University.

Shao, J., and Steel, P. (1999). Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions. Journal of the American Statistical Association, 94, 254-265.

Shao, J. and Tu, D. (1995), *The Jackknife and Bootstrap.* Springer

Sitter, R., and Rao, J.N.K. (1997), "Imputation for Missing Values and Corresponding Variance Estimation," *The Canadian Journal of Statistics*, Vol. 25, No. 1, pp. 61-73.

Skinner, C.J., and Rao, J.N.K (1993), "Jackknife Variance Estimation for Multivariate Statistics Under Hot Deck Imputation," Presented at the International Statistical Institute Meetings, Florence, Italy.

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

6. ACKNOWLEDGMENTS

The authors wish to thank Linda Libeg and Jane He for their computer programming assistance. We also thank David Judkins for his valuable review, comments, and formal discussion of this paper. We are also grateful for the helpful comments made by Hyunshik Lee, Mike Brick and Keith Rust.

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 t^{th} period, given the values for the $(t\text{-}1)^{th}$ 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) $$W_{ijt} = \beta_{jt} W_{ij(t-1)} + \varepsilon_{ijt}$$

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

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

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 $\exp\left(\beta_{jt} \ln W_{kj(t-1)} + .5\sigma^2_{kjt}\right)$. This suggests the alternative imputation

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

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

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

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

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

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

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

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:

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

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:

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

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:

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

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 i[th] quote's imputed value and letting ω_{it} denote the i[th] quote sample weight, the mean error and mean absolute error in imputed wages are given by:

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

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

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

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.

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

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

MOG_i = major occupation group for quote I

$SIZE_i$= size indicator based on employment

$UNION_i$= indicator variable, denoting whether job i is union

$FPTP_i$ = indicator variable, denoting full or part-time job

$REGION_i$ = region indicator

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

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

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:

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

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:

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

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:

$$W_{ijt} = \beta_{0t} + \beta_{1t} MID_i + \beta_{2t} MOG_i + \beta_{3t} FTPT_i + \beta_{4t} UNION_i + \beta_{5t} SIZE_i + \beta_{6t} REF_i + \beta_{7t} AREA_i + \beta_{8t} FACPTS_i + \beta_{9t} FACPTS_i^2 + \varepsilon_{ijt} \quad (27)$$

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)

DISCUSSION
Session 44: New Developments in Imputation of Business Survey Data

David R. Judkins, Westat
1650 Research Blvd., Rockville, MD 20850-3195, U.S.
judkind1@westat.com

One of the most difficult issues concerning imputation is to accurately discern customer needs. If they can be discerned, then it becomes possible to develop methodology to meet those needs. Dr. Rao's paper supplies the methodology for estimating post-imputation variance estimates when a single target variable has been imputed. The paper by Barsky, Buszuwski, Ernst, Lettau, Loewenstein, Pierce, Ponikowski, Smith, and West (a.k.a. the BLS paper) discusses alternative imputation strategies for preparing a specific set of statistics. The paper by Krenzke, Montaquila and Mohadjer (KMM) discusses several imputation strategies and post-imputation variance estimation strategies and includes example applications.

1. Rao Paper

While all three papers were interesting and useful, Dr. Rao's paper struck me as being a particularly useful survey of the state of the art for post-imputation variance estimation of marginal finite population means in the context of establishment surveys. In many such surveys, finite population correction factors can be non-negligible in some strata. I am thinking, for example, of surveys of refiners of petroleum. Dr. Rao in his equation (11) draws our attention to the marvelously simple and robust variance estimator for ratio imputation under uniform response with non-negligible sampling fractions recently published by Shao and Steel and links it to a variance estimator for double sampling by Rao and Sitter (1995). I found it particularly interesting to note that the FPC for the A term is $1-\frac{r}{N}$, instead of the $1-\frac{n}{N}$ for the B and C terms. Intuitively this makes sense: information on imputation errors comes only from the item respondents, but information about the predictor variable comes from the entire sample.

It was also nice to see the trick for reducing imputation variance for continuous targets from Chen, Rao and Sitter (1999). The idea of subtracting out average errors is very nice, provided of course, that one can accurately predict the analyses of interest. Other papers cited by Chen and Shao (1999) on consistency of nearest neighbor imputation, Shao and Wang (1999) on asymptotically unbiased estimation of correlation coefficients, Yung and Rao (1999) on post-imputation variance estimation with post-stratified weights, Rao and Shao (1999) on the asymptotic validity of Fay's modification to BRR, and others, all make it clear that the Rao school continues to be a prolific source of important work on the cutting edge of design-based survey inference techniques. The work with BRR raises my hopes that multivariate design-based methods can eventually be developed.

2. BLS Paper

I found it interesting that the authors found that adding interactions for localities does not improve the quality of the imputations. Since the purpose of the locality component of the National Compensation Survey is to set local pay raises for federal workers, the decision to pool models across localities cannot be made lightly. It does seem likely to me that a larger survey would have found local effects that would have been meaningful to federal workers in various industries and localities. I say this based on the intuition that workers in one industry do not easily migrate into another industry in the same locality nor do they easily migrate to other localities. So tightness in one industry in one location would not necessarily imply tightness in other industries at the same location nor tightness in the same industry at other localities. If this were not so, then there would be no attempt to measure local pay by industry in the survey. It seems to me that this would have been a perfect application for some of the generalized linear mixed-effect models. Such models that borrow strength across localities could also be used to improve the efficiency of the estimates based on reported data. See Robinson (1991) for a good introductory treatment of BLUPs (best linear unbiased predictors), Breslow and Clayton (1993) for a frequentist estimation approach, Folsom and Judkins (1997) for a frequentist approach that incorporates sample weights into Breslow and Clayton's approach, and Natarajan and Kass (2000) for the latest on Bayesian approaches to fitting this type of model.

3. KMM Paper

One of the things that I really liked about the KMM paper was the extension of Montaquila and Jernigan's ACI method to a simple sequential imputation. I am a strong supporter of sequential imputation. It is the best approach to extract information from partially complete vectors. The alternative is to only use information in naturally complete variables to impute the missing data. Their example illustrate some of the gains in variance reduction and covariance preservation that can be used with sequential imputation. However, it does pose very difficult problems for variance estimation. KMM presented a nice intuitive solution for the case of two variables with a hierarchical response pattern. I would like to see some theory or simulation studies to support their solution, but it is a step in the right direction.

I was also pleased to see them trying out the Shao-Sitter bootstrap. I think this method has good potential for design-based post-imputation variance estimation when the imputation procedure is complex. I note, however, that the procedures for selecting the stratum bootstrap sample sizes are more complex than they have hinted at. They drew bootstrap samples of the same size as the original sample in each stratum. Shao and Sitter (1996) pointed out that this is not a consistent approach. See sections 2.3 and 5.1 of their 1996 paper for full details.

The ratios of standard errors in their tables must be interpreted with caution. Since this is a case study rather than a simulation study, it is more instructive about the operation feasibility of the various procedures than about bias or stability. Also, the nonresponse rate may be a poor indicator of the true imputation variance due to outliers and the variation in size inherent in an establishment survey.

4. Unifying Themes

All three papers focus primarily on marginal means and means for sampling strata. There is some discussion in Dr. Rao's paper and in KMM on domain estimation, but only domains are treated where membership is not subject to missing data. While it seems sensible to solve the easy problems before tackling the hard ones, this approach is in stark contrast to multiple imputation which claims to apply to a full range of very complex imputation problems.

It appears to me that the objective of the BLS team could be met without any imputation. They used parametric models to impute missing values. In all cases, they imputed conditional means. The simulation of the imputation methods that they performed was unnecessary since the results could have been predicted from the fit of their models to the observed data. In this extreme case, where the analytic objectives are very narrow, imputation becomes uninteresting.

The KMM paper also focused considerable attention on the MHD procedure and noted that it is best for estimation of means. The MHD procedure is not really a hotdeck at all but a procedure for imputing conditional means. If stratum means are the only statistics of interest, then no imputation is required. If it is done despite the lack of need, then simple procedures work best for this narrow goal. All the post-imputation variance estimation procedures are likely to perform well. The superiority of MHD for stratum means compared to methods that add random residuals is hardly surprising since the only reason to add random residuals is to improve the estimation of statistics other than stratum means.

The methods in Dr. Rao's paper focus on functions of univariate marginal means. This is a broader focus that the BLS paper, but still fairly narrow. If there are k variables in a survey, then there are probably about k marginal univariate means of interest. If a good imputation procedure is developed strictly in terms of frame variables that are 100% complete and the procedures are well documented for each of these k variables, then the methods that he proposes can be used to estimate the post-imputation variances on those marginal means and functions of those marginal means. However, if the customers' needs are this narrowly focussed, then one could simply publish the k means along with their standard errors.

My contention is that customers usually want more complex analyses than simple means for domains defined in terms of naturally complete variables. Clients want to analyze the association of Y and Z, while controlling for the effects of X1 through XN, where each variable is complete on a different set of cases. Moreover, clients want secondary analysts to be able to choose Y and Z long after the imputation task has been completed. In the KMM example, the sequential procedure did a better job of preserving the correlation of Y with Z while simultaneously

reducing the variance on the mean of Z, an added bonus. KMM chose not to test data augmentation as an imputation method. If they had, I predict that they would have found that this combination is still better -- both at preserving the Y-Z association and at reducing the variance on the mean of Z. Moreover, they would have found a reduction in the variance on the mean of Y.

The magic allure of Bayesian methods is their promise to meet a much broader array of analytic objectives. This is more in line with the early promise of imputation methods. Which was to greatly simplify life for the secondary analyst. This promise has been difficult to fill.

In the fully parametric camp of statisticians (Bayesian and frequentist), imputation was never a priority since fitting the model was the primary objective. Once the model has been fit, it is more interesting to use it to predict future outcomes than to go back and fill in missing values in an old dataset. And to what end? Basically, so that some secondary analyst could come along later and recreate the model with less effort. But why would a secondary analyst have a burning desire to refit the same model? The statisticians of the parametric camps claim to get around this lack of purpose by emphasizing a thorough examination of the database to identify all important features of the dataset. If they succeed, then the model is correct, and other analyses of the dataset will be 'superefficient' since they will accelerate findings of no significance.

The design-based camp of statisticians have always had more modest goals. What happens if the user cross-tabulated two variables that I did not foresee as interesting in combination? We have tended to have fairly strict definitions of a complete record, with most seriously incomplete records receiving a zero weights, and many variables with high nonresponse rates going unanalyzed. This provided some protection against embarrassments but not much. First-order moments for simple cross-tabulations can go seriously astray with any imputation method. The Rao and KMM papers say, ok, that can happen, but at least I can calculate an appropriate variance estimate. Proponents of multiple imputation answer with the question, "Are these variance estimates useful?" If the cross-tabulation is biased, then having a consistent variance estimate will not lead to consistent confidence intervals. The Bayesians challenge us to focus on the inferential properties of confidence intervals produced by the imputation method and post-imputation variance estimation method. This is a fair but difficult challenge to meet.

Thus, I am thinking that we should give up thinking that we can ever create a truly multi-purpose imputation, at least an imputation that stays fixed in place and time. Whether we have one imputation per variable per case, or multiple imputed values is not very important. What we really need is general purpose software that will allow the secondary analyst to perform unanticipated primary analyses while extracting as much information as possible from partially complete records.

As we start to think about how to create such software, one can imagine a system that will automatically analyze the cluster of variables specified by the user and then use the automatic models to impute the missing data -- all repeated on bootstrap or BRR replications. The dataset with custom imputations could then be used for estimation of cross-tabulations or regressions by future versions of SUDAAN, WESVAR or VPLX. This is the direction that I believe can lead us to useful and robust design-based post-imputation variance estimation.

5. Additional References

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, **88**, pp 9-25.

Folsom, R. E., and Judkins, D. R. (1997), *Substance abuse in states and metropolitan areas: Model based estimates for the 1991-1993 National Household Surveys on Drug Abuse -- Methodology Report* (Methodological Series: M-1), Rockville, MD: Substance Abuse and Mental Health Services Administration.

Natarajan, R. and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, **95**, pp 227-237.

Robinson, G. K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, **6**, pp15-51.