JACKKNIFE VARIANCE ESTIMATION FOR NEAREST NEIGHBOR IMPUTATION

Jiahua Chen, University of Waterloo
Jun Shao, University of Wisconsin-Madison
Jun Shao, Department of Statistics, 1210 Dayton Street, Madison WI 53706 U.S.A.

Key Words: Adjusted jackknife; Hot deck; Reimputation; Sample mean; Stratified sampling; Unbiasedness.

Abstract:

Nearest neighbor imputation is a popular hot deck imputation method used to compensate for nonresponse in sample surveys. Although this method has a long history of application, the problem of variance estimation after nearest neighbor imputation has not been fully investigated. Since nearest neighbor imputation is a nonparametric method, a nonparametric variance estimation technique such as the jackknife is desired. We show that the naive jackknife that treats imputed values as observed data produces serious underestimation. We also show that Rao and Shao's (1992) adjusted jackknife or the jackknife with each pseudoreplicate re-imputed, which produces asymptotically unbiased and consistent jackknife variance estimators for other imputation methods (such as mean imputation, random hot deck imputation, ratio or regression imputation), produces serious overestimation in the case of nearest neighbor imputation. Two partially re-imputed and a partially adjusted jackknife variance estimators are proposed in this article and shown to be asymptotically unbiased and consistent. Some simulation results are provided to examine finite sample properties of these jackknife variance estimators.

1 Introduction

Imputation is commonly applied to compensate for item nonresponse in sample surveys (Kalton and Kasprzyk 1986, Sedransk 1985, Rubin 1987). The nearest neighbor imputation (NNI) method is used in many survey agencies such as the U.S. Bureau of Labor Statistics, the U.S. Census Bureau, and Statistics Canada. Consider a bivariate sample $(x_1, y_1), ..., (x_n, y_n)$ and suppose that $y_1, ..., y_r$ are observed (respondents), $y_{r+1}, ..., y_n$ are missing (nonrespondents), and all x-values are observed. The NNI method imputes a missing y_j by y_i, where $1 \leq i \leq r$ and i is the nearest neighbor of j measured by the x-variable, i.e., i satisfies

$$|x_i - x_j| = \min_{1 \leq l \leq r} |x_l - x_j|. \qquad (1)$$

If there are tied x-values, then there may be multiple nearest neighbors of j and i is randomly selected from them.

The NNI method has some nice features. First, it is a hot deck method in the sense that nonrespondents are substituted by respondents from the same variable; Second, it is shown in Chen and Shao (1999) that the NNI method provides asymptotically unbiased and consistent estimators for population means as well as quantiles. Third, the NNI method may be more efficient than other hot deck methods that do not make use of auxiliary information provided by the x-values (e.g., the mean imputation method). Finally, the NNI method uses a nonparametric model relating y and x (see Section 2) and, hence, it is expected to be more robust against model violations than methods based on parametric models, such as ratio imputation and regression imputation.

In this article we focus on jackknife variance estimation for the sample means (or weighted averages) based on data imputed by NNI. It is known that naive jackknife variance estimator often underestimates in the presence of imputation and some adjustments are necessary (Rao and Shao, 1992). Full implementation of the adjustment principle, however, overestimates the variance of the sample mean based on NNI. In view of these, we propose a partially re-imputed jackknife and a partially adjusted jackknife, and show that they produce asymptotically unbiased and consistent variance estimators for the sample means based on NNI, under some weak conditions for stratified samples.

The rest of the article is organized as follows. Some notation, assumptions, and details for NNI and the jackknife are given in Section 2. The asymptotic biases of the naive jackknife variance estimator

The research of Jiahua Chen was partially supported by Natural Sciences and Engineering Research Council of Canada; The research of Jun Shao was supported by National Science Foundation Grant DMS-9803112 and National Security Agency Grant MDA904-99-1-0032.

and Rao and Shao's adjusted jackknife variance estimator are derived in Section 3. Partially re-imputed and partially adjusted jackknife variance estimators are proposed in Section 4 and they are shown to be asymptotically unbiased and consistent. Section 5 contains some simulation results for these jackknife variance estimators, using a population that is close to a real data set from 1988 Current Population Survey (Valliant 1993).

2 Preliminaries

2.1 Sampling Design and Model

Let $\mathcal{P}$ be a finite population containing indices $1, ..., N$. Assume that $\mathcal{P}$ is stratified into H strata with N_h units in the hth stratum and that $n_h \geq 2$ units are selected without replacement from stratum h according to some probability sampling plan, independently across the strata. The overall sampling fraction n/N is assumed negligible, where $n = \sum_h n_h$, although n_h/N_h may be non-negligible for some h. Let $\mathcal{S}$ denote the sample. According to the sampling plan, survey weights w_i, $i \in \mathcal{S}$, are constructed so that for any set of values $\{z_i : i \in \mathcal{P}\}$,

$$E_s\left(\sum_{i\in\mathcal{S}} w_i z_i\right) = \frac{1}{N}\sum_{i=1}^{N} z_i,$$

where E_s is the expectation with respect to $\mathcal{S}$. This sampling design is commonly used in many business surveys conducted at the U.S. Bureau of Labor Statistics and the U.S. Census Bureau.

Let y be a variable of interest and x be an auxiliary variable. Let a be the response indicator for y (i.e., for the ith unit, $a_i = 1$ if y_i is a respondent and $a_i = 0$ otherwise). The validity of NNI is based on the following model assumption.

Assumption A. The finite population $\mathcal{P}$ is divided into K imputation classes such that within each imputation class, (x_i, y_i, a_i)'s are iid from a superpopulation and $P(a_i = 1|x_i, y_i) = P(a_i = 1|x_i)$. (x_i, y_i, a_i)'s from different imputation classes are independent. NNI is carried out within each imputation class.

We assume that K is fixed and the number of units in each imputation class is large. This is necessary for the validity of NNI, in fact, for any model-based and nonparametric imputation method (Valliant 1993)

2.2 Nearest Neighbor Imputation

Let $\mathcal{S}_k$ be the set of indices of sampled units in imputation class k, $\mathcal{R}_k$ be the set of indices of y-respondents in imputation class k, and $\mathcal{N}_k$ be the set of indices of y-nonrespondents in imputation class k ($\mathcal{S}_k = \mathcal{R}_k \cup \mathcal{N}_k$), $k = 1, ..., K$. Under assumption A, conditional on r_k, the number of respondents in $\mathcal{S}_k$, $\{(y_i, x_i), i \in \mathcal{R}_k\}$ and $\{(y_i, x_i), i \in \mathcal{N}_k\}$ are independent sets of iid random vectors from two possibly different distributions.

For $j \in \mathcal{N}_k$, let $\tilde{y}_j = y_i$ denote the value imputed by NNI, where i is selected according to

$$|x_i - x_j| = \min_{l\in\mathcal{R}_k} |x_l - x_j|. \qquad (2)$$

Note that (1) is a special case of (2). We focus on the case where the distribution of x related to the superpopulation is continuous so that there are no tied x-values. The NNI sample mean can be written as

$$\bar{y}_{\rm NNI}{=}\sum_{k=1}^{K}\sum_{i\in\mathcal{R}_k}\left(1 + d_i^{(k)}\right) w_i y_i, \qquad (3)$$

where

$$d_{ij} = \begin{cases} 1 & i \text{ is the nearest neighbor of } j \\ 0 & \text{otherwise} \end{cases}$$

and $d_i^{(k)} = \sum_{j\in\mathcal{N}_k} w_i^{-1} w_j d_{ij}$.

Chen and Shao (1999) showed that $\bar{y}_{\rm NNI}$ in (3) is asymptotically unbiased (for the population mean $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$) and has the following asymptotic variance

$$\begin{aligned} V &= \sum_{k=1}^{K} E\left[\sum_{i\in\mathcal{R}_k}\left(1 + d_i^{(k)}\right)^2 w_i^2 \mathrm{Var}(y_i|x_i)\right] \\ &\quad + \sum_{k=1}^{K} \mathrm{Var}\left[\sum_{i\in\mathcal{S}_k} w_i \psi_k(x_i)\right], \qquad (4) \end{aligned}$$

where $\psi_k(x) = E(y|x)$ in the kth imputation class. Their results were established under the stratified sampling design described in Section 2.1, assumption A, and the following technical assumption.

Assumption B. (i) The total number of sampled units $n \to \infty$ and $m_k^{-1} = O(n^{-1})$, $k = 1, ..., K$, where m_k is the number of sampled units in imputation class k.

(ii) The survey weights satisfy $\max_i w_i = O(n^{-1})$.

(iii) There exist constants $-\infty \leq M_1 < M_2 \leq \infty$ and C such that the function $\psi_k(x)$ is monotone when $x < M_1$ or $x > M_2$, and $|\psi_k(t) - \psi_k(s)| \leq C|t - s|$ when $M_1 \leq s, t \leq M_2$.

(iv) The marginal distribution of x has a density, $E|x|^3 < \infty$, $E|\psi_k(x)|^6 < \infty$, and $E|y_i|^6 < \infty$.

(v) The response probability $P(a = 1|x)$ satisfies $\inf_{x\in\mathcal{D}} P(a = 1|x) > 0$, where $\mathcal{D}$ is the support of the marginal distribution of x.

2.3 The Jackknife and the Adjusted Jackknife

If $\psi_k(x)$ and $\mathrm{Var}(y|x)$ in (4) have parametric forms, then we can estimate the variance in (4) by substitution (Chen and Shao 1999). Since NNI is nonparametric (i.e., the function forms of $\psi_k(x)$ and $\mathrm{Var}(y|x)$ are not known), a nonparametric variance estimation method is preferred.

The simplest nonparametric resampling method for variance estimation is the jackknife method, which estimates the variance of an estimator $\hat{\theta}$ by

$$v_{\mathrm{JACK}} = \sum_{k=1}^{K} \frac{m_k - 1}{m_k} \sum_{j \in \mathcal{S}_k} \left(\hat{\theta}^{(j)} - \hat{\theta}\right)^2, \qquad (5)$$

where m_k is the number of units in $\mathcal{S}_k$ and $\hat{\theta}^{(j)}$ is the same as $\hat{\theta}$ but is based on the jth jackknife pseudoreplicate, which is the dataset obtained by changing survey weight w_i to

$$w_i^{(j)} = \begin{cases} w_i & \text{if } j \in \mathcal{S}_k,\ i \notin \mathcal{S}_k \\ \frac{m_k}{m_k - 1} w_i & \text{if } i, j \in \mathcal{S}_k,\ i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Note that we adjust survey weights using m_k's, which are the imputation class sizes (or poststratum sizes), not the original stratum sizes n_h's. This is justified by the superpopulation model assumption in assumption A. In the case of no nonresponse, the jackknife estimator v_{JACK} is asymptotically unbiased and consistent when $\hat{\theta}$ is a differentiable function of sample means.

In the case of $\hat{\theta} = \bar{y}_{\mathrm{NNI}}$, however, directly applying formula (5) and treating imputed nonrespondents as observed data produces a v_{JACK} that underestimates the asymptotic variance of $\bar{y}_{\mathrm{NNI}}$. The precise form of the bias of v_{JACK} under NNI is given in Section 3.

For mean imputation and random hot deck imputation (which imputes y-nonrespondents in $\mathcal{S}_k$ by a random sample from y-respondents in $\mathcal{S}_k$), Rao and Shao (1992) proposed to first adjust the imputed values in each jackknife pseudoreplicate and then apply formula (5) with $\hat{\theta}^{(j)}$ computed based on the adjusted jth pseudoreplicate. Rao and Shao's adjustment can be described as follows. In the jth pseudoreplicate, each imputed value $\tilde{y}_i$ is adjusted to

$$\tilde{y}_{i,j}^{\mathrm{Adj}} = \tilde{y}_i + \tilde{E}(\tilde{y}_{i,j}) - \tilde{E}(\tilde{y}_i), \qquad (6)$$

where $\tilde{y}_{i,j}$ is the imputed value for y_i using the respondents in the jth pseudoreplicate and $\tilde{E}$ is the expectation for random imputation. For any nonrandom imputation (given the observed data), adjustment (6) reduces to

$$\tilde{y}_{i,j}^{\mathrm{Adj}} = \tilde{y}_{i,j}, \qquad (7)$$

i.e., Rao and Shao's adjustment amounts to reimputing nonrespondents in each pseudoreplicate using the observed data within the same pseudoreplicate.

Note that if y_j itself is a nonrespondent, then $\tilde{y}_{i,j}^{\mathrm{Adj}} = \tilde{y}_i$ under NNI. Hence, for NNI, adjustment (6) or re-imputation (7) needs to be applied for the jth pseudoreplicate only when y_j is a respondent.

Although Rao and Shao (1992) showed that the adjusted jackknife produces asymptotically unbiased and consistent variance estimators for sample means under mean imputation or random hot deck imputation, we show in Section 3 that the adjusted jackknife variance estimator overestimates the variance of $\bar{y}_{\mathrm{NNI}}$.

3 The Biases of Jackknife Estimators

Let $\bar{y}_{\mathrm{RI}}^{(j)}$ be the same as $\bar{y}_{\mathrm{NNI}}$ but based on the jth jackknife pseudoreplicate re-imputed according to (7) or, equivalently, adjusted according to (6). Then the re-imputed or adjusted jackknife variance estimator is

$$v_{\mathrm{JACK}}^{\mathrm{RI}} = \sum_{k=1}^{K} \frac{m_k - 1}{m_k} \sum_{j \in \mathcal{S}_k} \left(\bar{y}_{\mathrm{RI}}^{(j)} - \bar{y}_{\mathrm{NNI}}\right)^2.$$

Let us focus on imputation class k. Recall that $\mathcal{R}_k$ is the set of respondents and $\mathcal{N}_k$ is the set of nonrespondents. Consider the jth pseudoreplicate with a respondent y_j in $\mathcal{R}_k$. Let y_t be a nonrespondent in $\mathcal{N}_k$. If the original NNI value $\tilde{y}_t$ is not y_j, then the re-imputed value $\tilde{y}_{t,j}$ is still $\tilde{y}_t$. If $\tilde{y}_t = y_j$, i.e., j is the nearest neighbor of t, then the re-imputed value $\tilde{y}_{t,j}$ is the y-value of the second nearest neighbor of t. Let $x_{j_1} \in \mathcal{R}_k$ and $x_{j_2} \in \mathcal{R}_k$ be two nearest neighbor values of x_j ($x_{j_1} < x_j < x_{j_2}$ if x_j is not the smallest or largest x-value in $\mathcal{R}_k$; otherwise $x_{j_1} = x_{j_2}$). Then the second nearest neighbor of t must be either j_1 or j_2, i.e., $\tilde{y}_{t,j} = y_{j_1}$ or y_{j_2}. Let

$$c_{jt} = \begin{cases} 1 & j_1 \text{ is the second nearest neighbor of } t \\ 0 & j_2 \text{ is the second nearest neighbor of } t, \end{cases} \qquad (8)$$

given that j is the nearest neighbor of t,

$$c_j^{(k)} = \sum_{t \in \mathcal{N}_k} \frac{w_t}{w_j} c_{jt},$$

and let $d_j^{(k)}$ be the same as that in (3). Then

$$\begin{aligned}(m_k - 1)(\bar{y}_{\mathrm{NNI}} - \bar{y}_{\mathrm{RI}}^{(j)}) &= m_k w_j y_j - \bar{y}_k \\ &\quad + m_k w_j [d_j^{(k)} y_j - c_j^{(k)} y_{j_1} - (d_j^{(k)} - c_j^{(k)}) y_{j_2}], \qquad (9)\end{aligned}$$

if y_j is a respondent in $\mathcal{R}_k$, and

$$(m_k - 1)(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)}) = m_k w_j \tilde{y}_j - \bar{y}_k, \tag{10}$$

if y_j is a nonrespondent in $\mathcal{N}_k$, where $\bar{y}_k = \sum_{i \in \mathcal{R}_k} (1 + d_i^{(k)}) w_i y_i$.

For convenience, write $\epsilon_i = y_i - \psi_k(x_i)$. Then ϵ_i's are uncorrelated, conditional on x_i's, and $E(\epsilon_i | x_i) = 0$ and $\text{Var}(\epsilon_i | x_i) = \text{Var}(y_i | x_i)$. It follows from (9) that when y_j is a respondent in imputation class k,

$$\begin{aligned}
&(m_k - 1)(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)}) \\
&= m_k w_j [(1 + d_j^{(k)}) \epsilon_j - c_j^{(k)} \epsilon_{j_1} - (d_j^{(k)} - c_j^{(k)}) \epsilon_{j_2}] \\
&\quad + m_k w_j z_j + \eta_j^{(k)},
\end{aligned}$$

where $z_j = \psi_k(x_j) - \mu_k$, $\mu_k = E[\psi_k(x_i)]$,

$$\begin{aligned}
\eta_j^{(k)} = &-m_k w_j [c_j^{(k)} (z_{j_1} - z_j) + (d_j^{(k)} - c_j^{(k)})(z_{j_2} - z_j)] \\
&+ m_k w_j \mu_k - \bar{\psi}_k - \bar{\epsilon}_k,
\end{aligned}$$

and $\bar{\psi}_k$ (or $\bar{\epsilon}_k$) is the same as $\bar{y}_k$ but with y_i replaced by $\psi_k(x_i)$ (or ϵ_i).

It follows from Lemma 1 in the Appendix that

$$E\left[\sum_{j \in \mathcal{R}_k} \left(\frac{\eta_j^{(k)}}{m_k - 1}\right)^2\right] = o(n^{-1}), \tag{11}$$

where E is the joint expectation with respect to model and sampling. Let $\mu_j^{(k)} = (1 + d_j^{(k)})^2 + (c_j^{(k)})^2 + (d_j^{(k)} - c_j^{(k)})^2] w_j^2$. Using the fact that ϵ_j's are conditionally uncorrelated and assuming that the function $\sigma^2(x) = \text{Var}(y|x)$ is continuous in x, we obtain that

$$\begin{aligned}
E\left[\sum_{j \in \mathcal{R}_k} \left(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)}\right)^2\right] = &E\left[\sum_{j \in \mathcal{R}_k} \mu_j^{(k)} \text{Var}(y_j | x_j)\right] \\
&+ E\left(\sum_{j \in \mathcal{R}_k} w_j^2 z_j^2\right) + o(n^{-1}).
\end{aligned}$$

For nonrespondents in $\mathcal{N}_k$, a similar argument shows that

$$\begin{aligned}
E\left[\sum_{j \in \mathcal{N}_k} \left(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)}\right)^2\right] = &E\left[\sum_{i \in \mathcal{R}_k} d_i^{(k)} w_i^2 \text{Var}(y_i | x_i)\right] \\
&+ E\left(\sum_{i \in \mathcal{N}_k} w_i^2 z_i^2\right) + o(n^{-1}),
\end{aligned}$$

where the last equality follows from Lemma 2 in the Appendix.

The naive jackknife variance estimator is defined by (5) with $\hat{\theta} = \bar{y}_{\text{NNI}}$ and treating imputed values as observed data. By omitting the last term on the right hand side of (9) and using a similar argument, we obtain that

$$\begin{aligned}
E\left[\sum_{j \in \mathcal{R}_k} \left(\bar{y}_{\text{NNI}} - \bar{y}_{\text{NNI}}^{(j)}\right)^2\right] = &E\left[\sum_{j \in \mathcal{R}_k} w_j^2 \text{Var}(y_j | x_j)\right] \\
&+ E\left(\sum_{j \in \mathcal{R}_k} w_j^2 z_j^2\right) + o(n^{-1})
\end{aligned}$$

and

$$\begin{aligned}
&E\left[\sum_{j \in \mathcal{N}_k} \left(\bar{y}_{\text{NNI}} - \bar{y}_{\text{NNI}}^{(j)}\right)^2\right] \\
&= E\left[\sum_{j \in \mathcal{R}_k} d_j^{(k)} w_j^2 \text{Var}(y_j | x_j)\right] \\
&\quad + E\left(\sum_{j \in \mathcal{N}_k} w_j^2 z_j^2\right) + o(n^{-1}).
\end{aligned}$$

Combining these results and imputation classes, we obtain the following result on the biases of the naive jackknife variance estimator v_{JACK} and the re-imputed (or adjusted) jackknife variance estimator $v_{\text{JACK}}^{\text{RI}}$.

Theorem 1. Suppose that assumptions A and B hold and that the function $\sigma^2(x) = \text{Var}(y|x)$ is continuous in x. Then

$$\begin{aligned}
E(v_{\text{JACK}}) = &\sum_{k=1}^{K} E\left[\sum_{i \in \mathcal{R}_k} (1 + d_i^{(k)}) w_i^2 \text{Var}(y_i | x_i)\right] \\
&+ \sum_{k=1}^{K} \text{Var}\left[\sum_{i \in \mathcal{S}_k} w_i \psi_k(x_i)\right] + o(n^{-1})
\end{aligned}$$

and

$$\begin{aligned}
E(v_{\text{JACK}}^{\text{RI}}) = &\sum_{k=1}^{K} E\left[\sum_{i \in \mathcal{R}_k} \mu_j^{(k)} w_i^2 \text{Var}(y_i | x_i)\right] \\
&+ \sum_{k=1}^{K} \text{Var}\left[\sum_{i \in \mathcal{S}_k} w_i \psi_k(x_i)\right] + o(n^{-1}).
\end{aligned}$$

Compared with result (4), we conclude that the naive jackknife variance estimator v_{JACK} has a negative bias and the re-imputed jackknife variance estimator has a positive bias. Some numerical examples of these biases are given in the simulation study in Section 5.

4 Partially Re-imputed or Partially Adjusted Jackknife

In view of the fact that the naive jackknife underestimates and the re-imputed jackknife (which is the same as the adjusted jackknife) overestimates, we propose some jackknife methods with partial re-imputation or partial adjustment. The first partially re-imputed jackknife can be described as follows. For $j \in \mathcal{R}_k$, nonrespondents in pseudoreplicate j are re-imputed with probability $\rho_j^{(k)}$ and not re-imputed with probability $1-\rho_j^{(k)}$. Let δ be the indicator that equals 1 when re-imputation is conducted. Then a partially re-imputed jackknife variance estimator is

$$v_{\text{JACK}}^{\text{PRI}} = \sum_{k=1}^{K} \frac{m_k - 1}{m_k} \sum_{j \in \mathcal{S}_k} \left[\delta_j \left(\bar{y}_{\text{RI}}^{(j)} - \bar{y}_{\text{NNI}} \right)^2 + (1-\delta_j) \left(\bar{y}_{\text{NNI}}^{(j)} - \bar{y}_{\text{NNI}} \right)^2 \right]. \tag{12}$$

(Recall that for pseudoreplicate j with $j \in \mathcal{N}_k$, $\bar{y}_{\text{RI}}^{(j)} = \bar{y}_{\text{NNI}}^{(j)}$.) Using the same argument in establishing Theorem 1, we obtain that

$$E(v_{\text{JACK}}^{\text{PRI}}) = \sum_{k=1}^{K} E\left[\sum_{j \in \mathcal{R}_k} \{\rho_j^{(k)} \mu_j^{(k)} + 1 - \rho_j^{(k)} + d_j^{(k)}\} w_j^2 \text{Var}(y_j|x_j) \right] + \sum_{k=1}^{K} \text{Var}\left[\sum_{j \in \mathcal{S}_k} w_j \psi_k(x_j) \right] + o(n^{-1}).$$

By choosing $\rho_j^{(k)}$ so that the coefficient in front of $\text{Var}(y_j|x_j)$ equals $(1+d_j^{(k)})^2$, we can obtain an asymptotically unbiased $v_{\text{JACK}}^{\text{PRI}}$. This leads to

$$\rho_j^{(k)} = \frac{d_j^{(k)}(1+d_j^{(k)})}{(1+d_j^{(k)})^2 + (c_j^{(k)})^2 + (d_j^{(k)} - c_j^{(k)})^2 - 1}$$

($\rho_j^{(k)} = 0$ if $d_j^{(k)} = 0$).

To compute $v_{\text{JACK}}^{\text{PRI}}$ in (12), both $d_j^{(k)}$ and $c_j^{(k)}$ have to be computed in order to obtain the probability $\rho_j^{(k)}$. Note that the computation of $c_j^{(k)}$ is more complicated than that of $d_j^{(k)}$, since second nearest neighbors have to be located. The following slight modification of the previously described procedure avoids the computation of $c_j^{(k)}$. Instead of re-imputing nonrespondents in pseudoreplicate j by either y_{j_1} or y_{j_2}, where j_1 and j_2 are defined in (8), we re-impute nonrespondents by the average $(y_{j_1} + y_{j_2})/2$, whenever re-imputation is needed. The partially re-imputed jackknife variance estimator $\tilde{v}_{\text{JACK}}^{\text{PRI}}$ with this modification in the re-imputation procedure is then asymptotically unbiased if $\rho_j^{(k)} = 2(1+d_j^{(k)})/(3d_j^{(k)}+4)$ ($\rho_j^{(k)} = 0$ if $d_j^{(k)} = 0$).

Instead of randomizing re-imputation, we can use a partial adjustment, i.e., modify adjustment (7) to

$$\tilde{y}_{i,j}^{\text{Adj}} = \tilde{y}_i + g_j^{(k)}(\tilde{y}_{i,j} - \tilde{y}_i) \tag{13}$$

with a constant $g_j^{(k)} \in [0,1]$ to be specified later. If $g_j^{(k)} = 1$, then adjustment (13) is the same as adjustment (7); if $g_j^{(k)} = 0$, then there is no adjustment; if $0 < g_j^{(k)} < 1$, then there is a partial adjustment. For simplicity, we may use $\tilde{y}_{i,j} = (y_{j_1} + y_{j_2})/2$ when y_j is a respondent.

Our third proposed jackknife variance estimator is obtained by applying formula (5) with $\hat{\theta}^{(j)}$ being the sample mean computed based on the jth pseudoreplicate adjusted according to (13). This estimator is called the partially adjusted jackknife variance estimator and denoted by $v_{\text{JACK}}^{\text{PA}}$. Note that although re-imputed jackknife and adjusted jackknife are the same, partially re-imputed jackknife and partially adjusted jackknife are generally different.

Using the same argument in establishing Theorem 1, we can also show that $v_{\text{JACK}}^{\text{PA}}$ is asymptotically unbiased when

$$g_j^{(k)} = \frac{\sqrt{6(d_j^{(k)})^2 + 6d_j^{(k)} + 4} - 2}{3d_j^{(k)}}$$

($g_j^{(k)} = 0$ if $d_j^{(k)} = 0$). It is easy to verify that $0 \le g_j^{(k)} < 1$.

The following result summarizes the asymptotic performance of three proposed jackknife variance estimators.

Theorem 2. Assume the conditions in Theorem 1. Assume further that nV is bounded away from 0, where V is the asymptotic variance of $\bar{y}_{\text{NNI}}$ given by (4). Then

$$\frac{E(v_{\text{JACK}}^{\text{PRI}})}{V} \to 1 \tag{14}$$

and

$$\frac{v_{\text{JACK}}^{\text{PRI}}}{V} \to 1 \quad \text{in probability.} \tag{15}$$

Results (14) and (15) also hold with $v_{\text{JACK}}^{\text{PRI}}$ replaced by $\tilde{v}_{\text{JACK}}^{\text{PRI}}$ or $v_{\text{JACK}}^{\text{PA}}$.

The proof is given in the Appendix.

5 Simulation Results

As a complement to our theory, we present in this section some results from a simulation study. We examine five jackknife variance estimators for $\bar{y}_{\text{NNI}}$: the naive jackknife estimator v_{JACK}, the re-imputed jackknife estimator $v_{\text{JACK}}^{\text{RI}}$, two partially re-imputed jackknife estimators $v_{\text{JACK}}^{\text{PRI}}$ and $\tilde{v}_{\text{JACK}}^{\text{PRI}}$, and the partially adjusted jackknife estimator $v_{\text{JACK}}^{\text{PA}}$. The population distribution used to generate (y_i, x_i)'s is close to a real data set from 1988 Current Population Survey (Valliant 1993), where x is the hours worked per week and y is the weekly wage. Some descriptions of the data set can be found in Chen and Shao (1999).

We consider simple random sampling with $n = 100$ or 200 and a single imputation class. The y-respondents are generated according to the response probability function

$$P(a = 1|x) = \frac{\exp(\gamma_1 + \gamma_2 x)}{1 + \exp(\gamma_1 + \gamma_2 x)}$$

with various γ_1 and γ_2. When $\gamma_2 = 0$, respondents are generated with equal probability (uniform response); when $\gamma_2 \neq 0$, response rate depends on the value of x (non-uniform response). When uniform response is considered, the response rate is chosen to be between 0.5 and 0.88. Table 1 provides values of γ_1, γ_2, the ranges of $P(a = 1|x)$, and the average response rate $E[P(a = 1|x)]$.

Table 2 lists 10,000 Monte Carlo simulation estimates of the relative bias (in %) and the standard deviation for five jackknife variance estimators under 15 different models (different values of γ_1 and γ_2) described in Table 1. The values of the asymptotic variance V in (4) are also listed.

The results in Table 2 can be summarized as follows.

1. The naive jackknife variance estimator v_{JACK} has a serious negative bias. The relative bias of v_{JACK} can be as high as 50-70% and, as expected, is related to the average response rate $E[P(a = 1|x)]$. The re-imputed (or adjusted) jackknife variance estimator $v_{\text{JACK}}^{\text{RI}}$ has a serious positive bias, as indicated by Theorem 1. The relative biases of v_{JACK} and $v_{\text{JACK}}^{\text{RI}}$ are comparable in absolute value, but with different signs. Furthermore, the relative biases of v_{JACK} and $v_{\text{JACK}}^{\text{RI}}$ do not change much as n increases from 100 to 200.

2. The relative biases are small for the two partially re-imputed jackknife variance estimators, $v_{\text{JACK}}^{\text{PRI}}$ and $\tilde{v}_{\text{JACK}}^{\text{PRI}}$, and the partially adjusted jackknife variance estimator $v_{\text{JACK}}^{\text{PA}}$.

3. The standard deviations of $v_{\text{JACK}}^{\text{PRI}}$, $\tilde{v}_{\text{JACK}}^{\text{PRI}}$, and $v_{\text{JACK}}^{\text{PA}}$ are comparable and the standard deviation of $v_{\text{JACK}}^{\text{PA}}$ is slightly smaller. All standard deviations decrease substantially as n increases from 100 to 200, which supports the consistency result established in Theorem 2. The standard deviation of the naive jackknife variance estimator v_{JACK} is much smaller than those of other jackknife variance estimators, but this does not indicate a good performance of v_{JACK} since v_{JACK} has a large negative relative bias.

Appendix

Lemma 1. Under the conditions in Theorem 1, (11) holds.

Proof. From Theorem 2 in Chen and Shao (1999), $E(\bar{\psi}_k - \mu_k)^2 = O(n^{-1})$, under the conditions of Theorem 1. Hence,

$$E\left[\sum_{j\in\mathcal{R}_k}\left(\frac{m_k w_j \mu_k - \bar{\psi}_k}{m_k - 1}\right)^2\right] = o(n^{-1}).$$

Similarly,

$$E\left[\sum_{j\in\mathcal{R}_k}\left(\frac{\bar{\epsilon}_k}{m_k - 1}\right)^2\right] = o(n^{-1}).$$

Thus, the result follows from

$$E\left[\sum_{j\in\mathcal{R}_k} w_j^2[c_j^{(k)}(z_{j_1} - z_j) + (d_j^{(k)} - c_j^{(k)})(z_{j_2} - z_j)]^2\right] \\ = o(n^{-1}).$$

Let $b_j^{(k)} = \sum_{t\in\mathcal{N}_k} d_{jt}$. Since

$$w_j c_j^{(k)} \le w_j d_j^{(k)} = \sum_{t\in\mathcal{N}_k} w_t d_{jt} \le O(n^{-1}) b_j^{(k)},$$

the result follows from

$$E\left[\sum_{j\in\mathcal{R}_k}(b_j^{(k)})^2[\psi_k(x_{j_1}) - \psi_k(x_j)]^2\right] = o(n).$$

Let Var_* be the conditional variance, given x_j, $j \in \mathcal{R}_k$. Since $b_j^{(k)}$ is conditionally binomial, we have

$$E\left[\sum_{j\in\mathcal{R}_k}\text{Var}_*\{b_j^{(k)}[\psi_k(x_{j_1}) - \psi_k(x_j)]\}\right] \\ \le E\left[2m_k\sum_{j\in\mathcal{R}_k}[\psi_k(x_j) - \psi_k(x_{j_1})]^2[F_k(x_j) - F_k(x_{j_1})]\right] \\ = E\left[2m_k r_k \int_{x<y}[\psi_k(y) - \psi_k(x)]^2[F_k(y) - F_k(x)]\right. \\ \left.[1 + F_k(x) - F_k(y)]^{r_k} f_k(x) f_k(y) dx dy\right],$$

Table 1: Parameters in Response Models
$P(a=1|x) = \exp(\gamma_1 + \gamma_2 x)/[1+\exp(\gamma_1 + \gamma_2 x)]$

Model	γ_1	γ_2	$\min_x P(a=1\|x)$	$\max_x P(a=1\|x)$	$E[P(a=1\|x)]$
1	0	-0.02	0.12	0.50	0.32
2	0	-0.01	0.27	0.50	0.41
3	0	0.00	0.50	0.50	0.50
4	0	0.01	0.50	0.73	0.59
5	0	0.02	0.50	0.83	0.68
6	1	-0.03	0.12	0.73	0.46
7	1	-0.02	0.27	0.73	0.56
8	1	-0.01	0.50	0.73	0.65
9	1	0.00	0.73	0.73	0.73
10	1	0.01	0.73	0.88	0.80
11	2	-0.04	0.12	0.88	0.61
12	2	-0.03	0.27	0.88	0.70
13	2	-0.02	0.50	0.88	0.77
14	2	-0.01	0.73	0.88	0.83
15	2	0.00	0.88	0.88	0.88

where r_k is the size of $\mathcal{R}_k$, F_k is the conditional distribution of x given $a=0$ in imputation class k and f_k is the density function of F_k. Note that $r_k[F_k(y)-F_k(x)][1+F_k(x)-F_k(y)]^{r_k}$ has an upper bound independent of r_k. Also note that $x<y$. Hence, with the moment assumption,

$$r_k \int_{x<y} [\psi_k(y)-\psi_k(x)]^2 [F_k(y)-F_k(x)]$$
$$[1+F_k(x)-F_k(y)]^{r_k} f_k(x) f_k(y) dx dy < \infty.$$

Since the integrand converges to 0 almost surely when $r_k \to \infty$, by the dominate convergence theorem, the integral converges to 0 as $r_k \to \infty$. Therefore,

$$E\left[\sum_{j\in\mathcal{R}_k} \mathrm{Var}_*\{b_j^{(k)}[\psi_k(x_{j_1})-\psi_k(x_j)]\}\right] = o(n)$$

since $m_k = O_p(n)$. Let E_* be the conditional expectation, given x_j, $j \in \mathcal{R}_k$. The same argument can be used to establish

$$E\left[\sum_{j\in\mathcal{R}_k} \{E_*(b_j^{(k)})[\psi_k(x_{j_1})-\psi_k(x_j)]\}^2\right] = o(n).$$

This completes the proof.

Lemma 2. Under the conditions of Theorem 1,

$$E\left(\sum_{i\in\mathcal{R}_k} d_i^{(k)} w_i^2 z_i^2\right) = E\left(\sum_{i\in\mathcal{N}_k} w_i^2 z_i^2\right) + o(n^{-1}),$$

where $z_i = \psi_k(x_i) - \mu_k$, and

$$E\left[\sum_{i\in\mathcal{R}_k}\left(\sum_{j\in\mathcal{N}_k} w_j^2 d_{ij}\right) t_i\right] = E\left(\sum_{i\in\mathcal{R}_k} d_i^{(k)} w_i^2 t_i\right)$$
$$+o(n^{-1}),$$

where $t_i = z_i^2$ or $\mathrm{Var}(y_i|x_i)$.

Proof. The first result follows directly from Theorems 1 and 2 in Chen and Shao (1999). For the second result, note that

$$E\left[\sum_{i\in\mathcal{R}_k} w_i^2 t_i + \sum_{i\in\mathcal{R}_k}\left(\sum_{j\in\mathcal{N}_k} w_j^2 d_{ij}\right) t_i\right]$$
$$= E_s E_m \left[\sum_{i\in\mathcal{S}_k} a_i w_i^2 E_m(t_i|a_i=1)\right.$$
$$\left. + \sum_{j\in\mathcal{S}_k}(1-a_j) w_j^2 \sum_{i\in\mathcal{R}_k} d_{ij} t_i\right]$$
$$= E_s \left[\sum_{i\in\mathcal{S}_k} w_i^2 p_k E_m(t_i|a_i=1)\right.$$
$$\left. + \sum_{i\in\mathcal{S}_k} w_i^2(1-p_k) E_m(t_i|a_i=0)\right] + o(n^{-1})$$
$$= E_s\left[\sum_{i\in\mathcal{S}_k} w_i^2 E_m(t_i)\right] + o(n^{-1}),$$

where E_m is the expectation with respect to model under assumption A, E_s is the expectation with respect to sampling, $p_k = E(a_i)$ for i in imputation

Table 2: Relative Bias (RB) in % and Standard Deviation (SD) of Jackknife Variance Estimations

$n = 100$

		v_{JACK}		$v^{\text{RI}}_{\text{JACK}}$		$v^{\text{PRI}}_{\text{JACK}}$		$\tilde{v}^{\text{PRI}}_{\text{JACK}}$		$v^{\text{PA}}_{\text{JACK}}$	
Model	V	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD
1	2303.8	-74.9	194.1	63.7	2628.1	8.1	2100.7	3.1	1949.4	2.7	1602.0
2	1742.2	-66.1	168.5	59.2	1550.2	7.2	1222.8	2.1	1152.3	2.2	942.9
3	1360.6	-56.6	145.5	48.6	943.5	2.5	742.2	-0.6	677.5	-0.5	573.7
4	1070.2	-44.2	128.8	45.9	621.8	5.6	496.3	3.2	453.8	3.3	385.0
5	924.2	-35.0	117.4	35.5	443.8	2.9	342.1	1.8	327.4	1.8	282.5
6	1571.7	-62.1	164.8	61.5	1370.4	9.7	1091.0	6.2	1067.5	5.7	847.6
7	1240.0	-51.7	142.1	52.4	895.4	6.8	700.4	4.3	662.3	4.1	547.0
8	1025.2	-41.4	128.5	41.8	561.5	4.0	436.3	2.1	408.1	2.3	353.4
9	837.1	-28.3	115.1	40.1	401.5	8.0	313.4	7.2	294.6	7.0	256.7
10	782.3	-23.2	108.5	25.1	297.9	1.7	229.6	1.6	220.3	1.6	199.4
11	1163.3	-48.6	140.6	53.1	853.5	8.7	696.6	6.4	667.7	6.2	538.4
12	974.8	-38.4	125.7	40.1	528.8	4.6	415.5	3.1	393.1	3.2	337.7
13	850.7	-29.5	114.1	29.4	366.6	1.7	285.7	1.1	275.5	1.2	240.8
14	747.2	-19.5	105.6	24.6	269.9	3.5	211.4	3.2	207.0	3.1	183.9
15	689.6	-12.6	101.0	19.0	210.7	3.5	167.1	3.3	162.6	3.5	149.8

$n = 200$

		v_{JACK}		$v^{\text{RI}}_{\text{JACK}}$		$v^{\text{PRI}}_{\text{JACK}}$		$\tilde{v}^{\text{PRI}}_{\text{JACK}}$		$v^{\text{PA}}_{\text{JACK}}$	
Model	V	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD
1	1135.6	-74.1	70.1	64.4	901.8	8.7	715.2	2.3	676.3	2.2	545.9
2	849.9	-65.2	60.4	59.3	549.6	7.4	442.5	1.9	393.4	2.1	329.2
3	664.1	-55.4	51.4	50.4	330.0	4.2	259.3	0.7	237.1	0.8	200.0
4	527.3	-43.6	45.6	45.8	219.9	5.7	173.8	3.3	157.8	3.4	134.8
5	442.5	-32.7	41.2	39.7	155.5	6.2	118.8	4.8	111.4	4.8	96.9
6	770.0	-61.4	58.2	62.4	501.5	10.8	407.1	5.8	370.7	5.7	303.8
7	615.8	-51.7	50.3	50.8	301.6	5.1	233.9	2.9	221.4	2.6	181.8
8	498.5	-40.5	44.0	42.6	192.0	4.6	148.2	3.1	141.2	3.0	120.1
9	434.5	-31.5	40.0	32.3	136.2	2.3	105.5	1.8	101.4	1.6	87.8
10	384.5	-22.6	37.3	25.3	101.0	2.1	78.0	2.1	76.1	2.1	68.3
11	589.0	-49.5	49.6	48.0	293.9	5.1	236.8	2.7	220.5	2.7	181.9
12	483.2	-38.3	43.7	39.7	188.2	4.0	146.0	2.9	141.4	2.9	118.7
13	425.1	-29.9	39.9	28.8	130.5	1.2	103.2	0.3	96.0	0.4	84.3
14	367.6	-19.0	36.8	25.0	93.6	3.8	74.2	3.6	71.3	3.6	64.1
15	340.6	-12.6	34.4	18.4	71.9	3.3	57.4	3.2	55.9	3.2	51.5

class k, and the second equality follows from Theorem 1 in Chen and Shao (1999). The result then follows from the first result and $\sum_{i\in\mathcal{S}_k} w_i^2 E_m(t_i) = \sum_{i\in\mathcal{R}_k} w_i^2 E_m(t_i) + \sum_{i\in\mathcal{N}_k} w_i^2 E_m(t_i)$.

Proof of Theorem 2. Result (14) has already been established in the derivation of $v_{\text{JACK}}^{\text{PRI}}$ in Section 4. To show (15), it suffices to show that $\text{Var}(v_{\text{JACK}}^{\text{PRI}}) = O(n^{-3})$. From (9) and (10), a straightforward calculation shows that

$$\text{Var}[(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)})^2] = O(n^{-4})$$

uniformly in j and

$$\text{Cov}[(\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(j)})^2, (\bar{y}_{\text{NNI}} - \bar{y}_{\text{RI}}^{(l)})^2] = O(n^{-5})$$

uniformly in $j \neq l$. This proves the result.

6 References

Chen, J. and Shao, J. (1999). Nearest neighbor imputation for survey data. Manuscript submitted to *Journal of Official Statistics.*

Kalton, G. and Kasprzyk, D. (1986). The treatment of missing data. *Survey Methodology*, **12**, 1-16.

Rancourt, E., Särndal, C. E. and Lee, H. (1994). Estimation of the variance in the presence of nearest neighbor imputation. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 888-893.

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

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

Sedransk, J. (1985). The objective and practice of imputation, *Proceedings of the First Annual Research Conference*, 445–452, Bureau of the Census, Washington D.C.

Valliant, R. (1993). Poststratification and conditional variance estimation, *Journal of the American Statistical Association*, **88**, 89-96.