Journal of Official Statistics, Vol. 25, No. 1, 2009, pp. 55–62

Nonparametric Variance Estimation for Nearest Neighbor Imputation

Jun Shao[1]

Nearest neighbor imputation is a popular nonparametric hot deck imputation method used to compensate for nonresponse in sample surveys. Although the nearest neighbor imputation method has a long history of application, no asymptotically consistent nonparametric variance estimator for a survey estimator (such as the sample mean) based on data with nonrespondents imputed by nearest neighbor was available until the proposal of the adjusted jackknife variance estimator by Chen and Shao (2001). However, the adjusted jackknife method involves a somewhat artificial adjustment and is computationally complex because every jackknife pseudo-replicate has to be adjusted. We propose a consistent nonparametric variance estimator that is much easier to compute than the jackknife estimator. Some simulation results are provided to examine finite sample properties of the proposed variance estimator.

Key words: Nonrespondents; variance estimators; nearest neighbor; nonparametric method; consistency.

1. Introduction

The nearest neighbor imputation (NNI) method is often applied to compensate for nonresponse in many surveys such as the U.S. Census 2000 and the Current Population Survey conducted by the U.S. Census Bureau (Farber and Griffin 1998; Fay 1999), the Job Openings and Labor Turnover Survey and the Employee Benefits Survey conducted by the U.S. Bureau of Labor Statistics (Montaquila and Ponikowski 1993), and the Unified Enterprise Survey, the Survey of Household Spending, and the Financial Farm Survey conducted by Statistics Canada (Rancourt 1999). This trend will continue because of the availability of computer software, the Generalized Edit and Imputation System, which provides a simple way of performing NNI (Rancourt, Särndal and Lee 1994). It is shown in Chen and Shao (2000) that the NNI method provides asymptotically unbiased and consistent estimators for population means as well as quantiles, although these estimators are not exactly unbiased. The NNI method may be more efficient than the mean imputation and random hot deck imputation methods that do not make use of auxiliary information provided by covariates. On the other hand, the NNI method uses covariate information through nonparametric regression, instead of parametric regression (used in ratio or

[1] Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A. Email: shao@stat.wisc.edu

Acknowledgments: This research was partially supported by the National Science Foundation Grants DMS-0404535 and SES-0705033. The author would like to thank referees for their helpful comments.

regression imputation) that is sensitive to the choice of parametric models (see the simulation results in Shao and Wang 2008).

In this article we focus on variance estimation for estimators based on data imputed by NNI. We cannot treat the imputed values as observed data and use the standard variance formula for the case of no nonresponse, because the resulting variance estimator underestimates the true variance of the estimator based on NNI. Some variance estimation methods that take nonresponse and NNI into account are studied in Rancourt, Särndal, and Lee (1994) and in Chen and Shao (2000). However, these methods are based on parametric models. Since the NNI method is nonparametric, it is desired to have a nonparametric variance estimation method as well. Chen and Shao (2001) proposed an adjusted jackknife method for variance estimation. However, the adjusted jackknife involves a somewhat artificial adjustment to every jackknife pseudo-replicate, which increases computation complexity.

After introducing some notation, assumptions, and preliminary results in Section 2, we derive a variance estimator for NNI in Section 3. Our variance estimator is nonparametric and is computationally simple. The consistency of the proposed variance estimator is also established in Section 3. Some simulation results are presented in Section 4. The last section contains some conclusions.

2. Preliminaries

Let $\mathcal{P}$ be a finite population containing indices $1, \ldots, N$. 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 from stratum h according to some probability sampling plan, independently across the strata. 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 (U.S. Census Bureau 1987).

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

Assumption A. (x_i, y_i, δ_i), $i = 1, \ldots, N$, are independently from a superpopulation. The finite population $\mathcal{P}$ is divided into K imputation classes such that, within each imputation class, (x_i, y_i, δ_i)'s are identically distributed (with respect to the superpopulation) and $P(\delta_i = 1|x_i, y_i) = P(\delta_i = 1|x_i)$.

NNI is carried out within each imputation class. We assume that the sample size in each imputation class is large. This is necessary for the asymptotic consistency of estimators based on NNI. The number of imputation classes, K, is fixed. Imputation classes are often formed according to the value of an auxiliary variable that is observed for all sampled units. For example, in many business surveys, imputation classes are the same as strata or unions of strata. When there are many strata of small n_h's, imputation classes are often obtained through poststratification (Valliant 1993) and/or combining small strata. In applications, we may determine sample sizes and K by conducting a pilot or simulation study.

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 ($\mathcal{S}_k = \mathcal{R}_k \cup \mathcal{N}_k$), $k = 1, \ldots, K$. Under Assumption A, conditional on $\mathcal{S}_k$, $\{(y_i, x_i), i \in \mathcal{R}_k\}$ 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 y_{j_*} denote imputed y-value by NNI, where j_* is selected according to

$$|x_{j_*} - x_i| = \min_{i \in \mathcal{R}_k} |x_j - x_i|$$

We focus on the case where the superpopulation distribution of x is continuous so that there are no tied x-values. The NNI sample mean is

$$\bar{y}_{\mathrm{NNI}} = \sum_{k=1}^{K} \left(\sum_{i \in \mathcal{R}_k} w_i y_i + \sum_{j \in \mathcal{N}_k} w_j y_{j_*} \right) = \sum_{k=1}^{K} \left(\sum_{i \in \mathcal{R}_k} w_i y_i + \sum_{j \in \mathcal{N}_k} w_j \sum_{i \in \mathcal{R}_k} d_{ij} y_i \right)$$

$$= \sum_{k=1}^{K} \sum_{i \in \mathcal{R}_k} \left(1 + d_i^{(k)}\right) w_i y_i$$

where

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

and

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

When $w_i = w_t$ for all i and t in a given $\mathcal{S}_k$, $d_i^{(k)}$ is the number of times $i \in \mathcal{R}_k$ is used as the nearest neighbor for nonrespondents $j \in \mathcal{N}_k$.

The NNI sample mean $\bar{y}_{\mathrm{NNI}}$ is not exactly unbiased as an estimator of the population mean of y values. Under some regularity conditions (Assumption B stated in the Appendix), it is established in Chen and Shao (2000) and Shao and Wang (2008) that $\bar{y}_{\mathrm{NNI}}$ is asymptotically unbiased and asymptotically normal with the following asymptotic variance:

$$V_n = E\left[\sum_{k=1}^{K} \sum_{i \in \mathcal{R}_k} \left(1 + d_i^{(k)}\right)^2 w_i^2 \sigma_k^2(x_i) \right] + V\left[\sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} w_i \psi_k(x_i) \right] \qquad (1)$$

where $\psi_k(x) = E(y|x)$ and $\sigma_k^2(x) = V(y|x)$ in the kth imputation class. Throughout the article, $E(y|x)$ and $V(y|x)$ denote the conditional expectation and conditional variance, respectively, under the superpopulation model in Assumption A, and E and V denote the unconditional expectation and variance, respectively, with respect to the superpopulation and sampling.

3. Variance Estimation

If $\psi_k(x)$ and $\sigma_k^2(x)$ in (1) have parametric forms, then we can estimate the variance in (1) by substitution. However, we may not need NNI if $\psi_k(x)$ and $\sigma_k^2(x)$ have parametric forms. Since NNI is nonparametric, a nonparametric variance estimation method without parametric models on $\psi_k(x)$ and $\sigma_k^2(x)$ is preferred.

We propose a nonparametric estimator of V_n that is simple to compute. It follows from Lemma 1 in Shao and Wang (2008) that, for each k,

$$E\left[\sum_{i\in\mathcal{R}_k}(1+d_i^{(k)})w_i^2\sigma_k^2(x_i)\right]=E\left[\sum_{i\in\mathcal{S}_k}w_i^2\sigma_k^2(x_i)\right]+o\left(\frac{1}{n}\right) \tag{2}$$

as the sample size $n=\sum_{h=1}^{H}n_h\rightarrow\infty$. By conditioning, we have

$$V\left(\sum_{i\in\mathcal{S}}w_iy_i\right)=E\left[\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}w_i^2\sigma_k^2(x_i)\right]+V\left[\sum_{k=1}^{K}\sum_{i\in\mathcal{S}_k}w_i\psi_k(x_i)\right] \tag{3}$$

It follows from (1)–(3) that

$$V_n=V\left(\sum_{i\in\mathcal{S}}w_iy_i\right)+E\left[\sum_{k=1}^{K}\sum_{i\in\mathcal{R}_k}d_i^{(k)}(1+d_i^{(k)})w_i^2\sigma_k^2(x_i)\right]+o\left(\frac{1}{n}\right) \tag{4}$$

The first term on the right-hand side of (4) is the variance of the unbiased estimator of the population mean when there is no nonresponse. It can be estimated as follows: Let v_n be the standard variance estimator for $\sum_{i\in\mathcal{S}}w_iy_i$ when there is no nonresponse. For example, if the sampling design is stratified sampling, then

$$v_n=\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i\in\mathcal{S}_h}\left(w_iy_i-\frac{1}{n_h}\sum_{j\in\mathcal{S}_h}w_jy_j\right)^2$$

where $\mathcal{S}_h$ is the sample in the hth stratum, $h=1,\ldots,H$. Let $\tilde{v}_n$ be the same as v_n but with y_j replaced by the imputed value y_{j*} for any $j\in\mathcal{N}_k$. That is, $\tilde{v}_n$ is the naive variance estimator obtained by treating imputed values as observed data. It was shown by Chen and Shao (2001) that $\tilde{v}_n$ is a consistent estimator of the first term on the right-hand side of (4).

It remains to find a consistent estimator of the second term on the right-hand side of (4). For $i\in\mathcal{R}_k$, let i_* be the nearest neighbor of i in the set $\mathcal{R}_k-\{i\}$, i.e.,

$$|x_{i_*}-x_i|=\min_{l\in\mathcal{R}_k,l\neq i}|x_l-x_i|$$

Note that y_{i*} is the NNI value of y_i if y_i is treated as a nonrespondent. Then $y_i-y_{i_*}$ can be viewed as an imputation residual and a simple estimator of $\sigma_k^2(x_i)$ is $\hat{\sigma}_{ik}^2=(y_i-y_{i_*})^2/2$,

Table 1. Simulation estimates (10,000 runs)

Parameters			Simulation estimates								
γ_1	γ_2	p	RB(%)	$V/10^4$	$\tilde{v}_n/10^4$	$\hat{V}_n/10^4$	CV(%)	CP(%)	CP$_l$(%)	CP$_u$(%)	L/10^2
0.5	-1	0.607	0.45	24.65	14.83	23.95	20.44	91.82	99.19	92.63	19.10
0.5	1	0.610	0.47	26.28	13.81	25.52	26.19	90.62	99.34	91.28	19.66
0.5	0	0.628	0.13	21.44	14.75	22.01	14.19	94.64	98.49	96.15	18.35
1	-1	0.700	0.35	20.60	14.88	21.00	16.59	93.10	99.10	94.40	17.91
1	1	0.703	0.26	21.82	14.28	21.68	20.06	92.92	98.79	94.23	18.17
1	0	0.735	0.09	18.80	14.85	19.52	11.72	95.03	98.35	96.68	17.29
2	-1	0.846	0.17	17.64	14.93	17.73	10.81	94.09	98.62	95.47	16.48
2	1	0.848	0.08	17.39	14.80	17.90	12.59	94.99	98.43	96.56	16.55
2	0	0.883	0.02	16.69	14.98	17.07	9.73	95.02	98.02	97.00	16.18
∞	0	1.000	-0.01	14.84	15.59	15.05	8.65	95.46	98.05	97.41	15.46

RB: relative bias of $\bar{y}_{\text{NNI}}$, V: variance of $\bar{y}_{\text{NNI}}$.

$\tilde{v}_n$: the naive estimator of V.

$\hat{V}_n$: the proposed estimator of V.

CV: standard error of $\hat{V}_n$/the mean of $\hat{V}_n$.

CP: coverage probability of confidence interval $\bar{y}_{\text{NNI}} \pm 1.96\sqrt{\hat{V}_n}$.

CP$_l$: coverage probability of confidence bound $\bar{y}_{\text{NNI}} - 1.96\sqrt{\hat{V}_n}$.

CP$_u$: coverage probability of confidence bound $\bar{y}_{\text{NNI}} + 1.96\sqrt{\hat{V}_n}$.

L: Length of confidence interval $\bar{y}_{\text{NNI}} \pm 1.96\sqrt{\hat{V}_n}$.

$i \in \mathcal{R}_k$. This leads to the following estimator of V_n in (1):

$$\hat{V}_n = \tilde{v}_n + \sum_{k=1}^{K} \sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 \hat{\sigma}_{ik}^2 \tag{5}$$

Let $\mathcal{X}_k = \{x_i, i \in \mathcal{R}_k\}$ and $E(\cdot|\mathcal{R}_k, \mathcal{X}_k)$ be the conditional expectation with respect to the superpopulation, given $(\mathcal{R}_k, \mathcal{X}_k)$. Then

$$E\left(\hat{\sigma}_{ik}^2 | \mathcal{R}_k, \mathcal{X}_k\right) = E\left[\frac{(y_i - y_{i_*})^2}{2} | \mathcal{R}_k, \mathcal{X}_k\right] = \frac{[\psi_k(x_i) - \psi_k(x_{i_*})]^2}{2} + \frac{\sigma_k^2(x_i) + \sigma_k^2(x_{i_*})}{2}$$

and

$$\begin{aligned} E\left[\sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 \hat{\sigma}_{ik}^2 | \mathcal{R}_k, \mathcal{X}_k\right] = & \sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 \sigma_k^2(x_i) \\ & + \sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 \frac{[\psi_k(x_i) - \psi_k(x_{i_*})]^2}{2} \\ & + \sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 \frac{\sigma_k^2(x_{i_*}) - \sigma_k^2(x_i)}{2} \end{aligned}$$

Under Assumption B (in the Appendix), $\psi_k(x)$ and $\sigma_k^2(x)$ are Lipschitz continuous. Hence, the last two terms in the previous expression are bounded in absolute value by

$$A_n = C \sum_{i \in \mathcal{R}_k} d_i^{(k)}(1 + d_i^{(k)}) w_i^2 |x_i - x_{i_*}|$$

where $C > 0$ is a constant. It follows from the result in Shao and Wang (2008) that nA_n converges to 0 in probability. Thus, under Assumptions A-B, we have established the consistency of $\hat{V}_n$ as an estimator of V_n.

4. Simulation Results

A simulation study was carried out using a population similar to that in Chen and Shao (2001), which matches the main characteristics of an aggregated data set from the Financial Farm Survey (FFS) of 1998 published by Statistics Canada. The FFS is a biannual survey collecting information on agriculture operations in Canada. The survey collects information on revenues, expenses, assets, investments, and liabilities for the reference year. The results are mainly used by the Canadian System of National Accounts and by Agriculture and Agri-Food Canada. The FFS is based on stratified simple random sampling and NNI is used to impute nonrespondents for some variables. More details about FFS can be found in Rancourt (1999).

We considered dairy farms only. Strata were created using farm size (3 classes) and province (5 provinces and one group of small provinces, ALT). These 18 strata were also used as imputation classes. Two variables, the net assets (x) and the cash income (y), were considered. Information about population size, sample size, mean, and standard deviation of the variables under consideration is given in Table 1 in Chen and Shao (2001). In particular, the total sample size is 16,989 and the true population mean of y is 52,351.47.

To study the effect of response pattern, we generated the y-respondents according to the response probability function similar to that in Chen and Shao (2001). For each pair (x, y), a y-respondent was generated according to

$$P(y\,\text{is a respondent}|x) = \frac{\exp(\gamma_1 + \gamma_2(x - \mu_x)\sigma_x^{-1})}{1 + \exp(\gamma_1 + \gamma_2(x - \mu_x)\sigma_x^{-1})}$$

with some γ_1 and γ_2, where μ_x and σ_x are the mean and variance of x within an imputation class. When $\gamma_2 = 0$, the response mechanism is uniform; when $\gamma_2 \neq 0$, the response mechanism depends on the value of x. Values of γ_1 and γ_2 and the average response probability $p = E[P(a = 1|x)]$ are given in Table 1.

Table 1 shows 10,000 Monte Carlo simulation estimates of the relative bias of the sample mean based on NNI, $\bar{y}_{\text{NNI}}$, the variance of $\bar{y}_{\text{NNI}}$, the naive variance estimator $\tilde{v}_n$, the proposed variance estimator $\hat{V}_n$ defined in (5), the CV of $\hat{V}_n$ defined as the standard error of $\hat{V}_n$ divided by $\hat{V}_n$, the coverage probabilities of the lower confidence bound, of the upper confidence bound, and of the two-sided confidence interval for the true population mean of the cash income (y) based on $\bar{y}_{\text{NNI}} \pm 1.96\sqrt{\hat{V}_n}$, and the length of the two-sided confidence interval. The confidence bounds and interval are based on the asymptotic normality of $\bar{y}_{\text{NNI}}$ established by Shao and Wang (2008) and the consistency of $\hat{V}_n$.

The results in Table 1 can be summarized as follows. The relative bias of $\bar{y}_{\text{NNI}}$ is below 0.5% in all cases under consideration, although the bias is larger when the response rate is lower. The naive variance estimator $\tilde{v}_n$, which is obtained by treating imputed values as observed data and applying the standard variance formula, has a serious negative bias. The proposed variance estimator $\hat{V}_n$ has a negligible bias in all cases under consideration. The CV of $\hat{V}_n$ depends on the response mechanism. The coverage probability of the two-sided confidence interval is close to the nominal level 95% except for cases where the response rate is low and nonresponse depends on x ($\gamma_2 \neq 0$). The performance of the one-sided confidence bounds is not as good as that of the two-sided confidence interval (asymptotically, the two-sided confidence interval is second-order accurate but the one-sided confidence bounds are only first-order accurate). The coverage probability of the lower confidence bound is always larger than the nominal level 97.5% whereas the coverage probability of the upper confidence bound is always smaller than 97.5%, which indicates the skewness of the population under consideration.

5. Conclusion

We focus on variance estimation for the sample mean based on survey data with nonrespondents imputed by the nearest neighbor imputation. The proposed variance estimator does not require any parametric assumption on the conditional expectation $E(y|x)$ and the conditional variance $V(y|x)$. It is asymptotically consistent when $E(y|x)$ and $V(y|x)$ are smooth functions of x, under some regularity conditions. The finite sample performance of the proposed variance estimator is adequate. The advantage of our proposed variance estimator over the adjusted jackknife variance estimator in Chen and Shao (2001) is its simplicity: the jackknife method involves a somewhat artificial adjustment and requires a large amount of computation, since every jackknife pseudo-replicate has to be adjusted.

Appendix

Assumption B

(i) The total number of sampled units $n \rightarrow \infty$ and $m_k^{-1} = O(n^{-1})$, $k = 1, \ldots, K$, where m_k is the number of sampled units in imputation class k.
(ii) The survey weights satisfy $\max_i w_i = O(n^{-1})$ and $V_s(\sum_{i \in \mathcal{S}} w_i) = O(n^{-1})$.
(iii) The marginal distribution of the covariate x has a density, $E|x|^3 < \infty$, $E|\psi_k(x)|^6 < \infty$, and $E|y_i|^8 < \infty$.
(iv) 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.
(v) $\psi_k(x)$ and $\sigma_k^2(x)$ are Lipschitz continuous on $\mathcal{D}$.

6. References

Chen, J. and Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. Journal of Official Statistics, 16, 113–132.

Chen, J. and Shao, J. (2001). Jackknife Variance Estimation for Nearest Neighbor Imputation. Journal of the American Statistical Association, 96, 260–269.

Farber, J.E. and Griffin, R. (1998). A Comparison of Alternative Estimation Methodologies for Census 2000. Proceedings of the American Statistical Association, Section on Survey Research Methods, 629–634.

Fay, R.E. (1999). Theory and Application of Nearest Neighbor Imputation in Census 2000. Proceedings of the American Statistical Association, Section on Survey Research Methods, 112–121.

Montaquila, J.M. and Ponikowski, C.H. (1993). Comparison of Methods for Imputing Missing Responses in an Establishment Survey. Proceedings of the American Statistical Association, Section on Survey Research Methods, 446–451.

Rancourt, E. (1999). Estimation with Nearest Neighbour Imputation at Statistics Canada. Proceedings of the American Statistical Association, Section on Survey Research Methods, 131–138.

Rancourt, E., Särndal, C.E., and Lee, H. (1994). Estimation of the Variance in the Presence of Nearest Neighbour Imputation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 888–893.

Shao, J. and Wang, H. (2008). Confidence Intervals Based on Survey Data with Nearest Neighbor Imputation. Statistica Sinica, 18, 281–297.

U.S. Bureau of the Census (1987). Noncertainty Sample Specification. BSR-87 Action Memo D.06, the U.S. Census Bureau.

Valliant, R. (1993). Poststratification and Conditional Variance Estimation. Journal of the American Statistical Association, 88, 89–96.

Received March 2007
Revised March 2008