Statistica Sinica **20** (2010), 263-280

EMPIRICAL LIKELIHOOD ESTIMATION FOR SAMPLES WITH NONIGNORABLE NONRESPONSE

Fang Fang[1], Quan Hong[2] and Jun Shao[3,4]

[1]*GE Consumer Finance,* [2]*Eli Lilly and Company,*
[3]*University of Wisconsin-Madison and* [4]*East China Normal University*

Abstract: Nonresponse is very common in survey sampling. Nonignorable nonresponse, a response mechanism in which the response probability of a survey variable Y depends directly on the value of Y regardless of whether Y is observed or not, is the most difficult type of nonresponse to handle. The population mean estimators ignoring the nonrespondents typically have heavy biases. This paper studies an empirical likelihood-based estimation method, with samples under nonignorable nonresponse, when an observed auxiliary categorical variable Z is available. The likelihood is semiparametric: we assume a parametric model on the response mechanism and the conditional probability of Z given Y, and a nonparametric model on the distribution of Y. When the number of Z categories is not small, a pseudo empirical likelihood method is applied to reduce the computational intensity. Asymptotic distributions of the proposed population mean estimators are derived. For variance estimation, we consider a bootstrap procedure and its consistency is established. Some simulation results are provided to assess the finite sample performance of the proposed estimators.

Key words and phrases: Empirical likelihood, nonignorable nonresponse, pseudo likelihood, sample survey, semiparametric likelihood, stratified samples.

1. Introduction

Nonresponse is a common phenomenon in sample surveys. Let Y be a variable of interest having nonrespondents and Z be a covariate with no nonresponse. If the propensity $P(\delta = 1 \mid Y, Z)$, where δ is the response indicator for Y, depends not only on Z and observed Y, but also on unobserved Y, then the nonresponse mechanism is nonignorable. Nonignorable nonresponse creates a great challenge in the estimation of the mean of Y based on incomplete survey data. Ignoring the dependence of nonresponse probability on unobserved Y typically leads to heavy bias.

Greenlees, Reece, and Zieschang (1982) studied maximum likelihood estimators for survey data with nonignorable nonresponse, based on a parametric model on the propensity $P(\delta = 1 \mid Y, Z)$ and a parametric (normal) model on $L(Y \mid Z)$, the distribution of Y conditional on Z. However, parametric models (especially normal models) on $L(Y \mid Z)$ for survey data are often not valid. In fact, Greenlees, Reece, and Zieschang (1982) admitted that the normality assumption on $L(Y \mid Z)$ was not valid for the data in their example, even though their method was better than the method of ignoring the fact that nonresponse was nonignorable.

On the other hand, it is impossible to develop a pure nonparametric method that produces a consistent estimator of the mean of Y in the presence of nonignorable nonresponse. Thus, some semiparametric methods assuming a parametric model on one of $P(\delta = 1 \mid Y, Z)$ and $L(Y \mid Z)$ have been proposed in the literature. Tang, Little, and Raghunathan (2003) developed a likelihood method by assuming a parametric model on $L(Y \mid Z)$; they assumed that $P(\delta = 1 \mid Y, Z) = P(\delta = 1 \mid Y)$ but otherwise is nonparametric. Qin, Leung and Shao (2002) proposed an empirical likelihood method by assuming a parametric model on $P(\delta = 1 \mid Y, Z)$ and a nonparametric model on $L(Y \mid Z)$; the resulting estimator of the mean of Y is similar to the estimator obtained by weighting each respondent by the inverse of an estimated propensity $P(\delta = 1 \mid Y, Z)$ (Robins, Rotnitzky and Zhao (1994)). For survey data, finding a suitable parametric model for $P(\delta = 1 \mid Y, Z)$ is much easier than finding an appropriate parametric model for $L(Y \mid Z)$. However, the estimation of $P(\delta = 1 \mid Y, Z)$ is still difficult under a parametric assumption on $P(\delta = 1 \mid Y, Z)$ because of the presence of unobserved Y values.

In many survey problems the covariate Z is categorical, e.g., age group, sex, race, education level, type of industry etc., while the main variable Y is continuous. If there is an appropriate parametric model on the conditional distribution $L(Z \mid Y)$ given Y (e.g., the logistic model), then we can improve the approach in Qin, Leung and Shao (2002). The purpose of this paper is to study an empirical likelihood approach under parametric models on $P(\delta = 1 \mid Y, Z)$ and $L(Z \mid Y)$ with a discrete Z, and under a nonparametric model on the distribution of Y. Our approach works for a stratified sampling design with a superpopulation within each stratum, which is commonly used in practice. Furthermore, we study a pseudo empirical likelihood to reduce the amount of computation when the number of Z categories is not small. Although losing some efficiency, the estimators based on the pseudo empirical likelihood are still consistent and asymptotically normal. Note that the same technique has been applied to the case of ignorable nonresponse (Fang, Hong and Shao (2009)).

This paper is organized as follows. Section 2 presents details on the sampling design and model, and gives results for estimation without imputation. In addition to the derivation of empirical likelihood estimators, their consistency and asymptotic normality are established. Section 3 discusses the pseudo empirical likelihood estimators. Section 4 considers variance estimation by bootstrapping.

In Section 5, we consider two imputation methods related to the results in Sections 2 and 3. Section 6 examines by simulation the finite sample performance of the proposed estimators, under some response patterns and models. The Appendix (available online at `http://www.stat.sinica.edu.tw/statistica`) contains proofs or sketched proofs.

2. Empirical Likelihood Approach

We consider the following sampling design commonly used in such business surveys as the Current Employment Survey conducted by the U.S. Bureau of Labor Statistics (Wolter, Shao and Huff (1998)), the Transportation Annual Survey conducted by the U.S. Census Bureau (Census Bureau (1987)), and the Financial Farm Survey conducted by Statistics Canada (Rancourt (1999)). The finite population $\mathcal{P}$ is stratified into H (a fixed positive integer) strata and samples are taken independently across the strata. Within each stratum, a large number of units are either independently sampled with replacement according to a probability sampling plan, or selected as a simple random sample without replacement with a negligible sampling fraction. According to the sampling plan, survey weights $\{\omega_i\}$ are constructed so that for any set of values $\{x_i\}$,

$$E_{\mathcal{S}}\left(\sum_{i\in\mathcal{S}}\omega_i x_i\right)=\sum_{i\in\mathcal{P}}x_i,$$

where $\mathcal{S}$ is the sample and $E_{\mathcal{S}}$ is the expectation with respect to sampling.

Let Y be the variable of interest in the survey and Z be a categorical covariate taking values in $\{z_1,\ldots,z_s\}$. We assume that values of (Y,Z) are i.i.d. from a superpopulation within each stratum, and are independent across strata. To present the main idea, we first consider the special case of one stratum so that the subscript for stratum is omitted.

Under the superpopulation model (within each stratum), Y has an unknown nonparametric distribution F, and we assume a parametric probability function

$$P(Z=z\mid Y=y)=f(y,z,\beta), \tag{2.1}$$

where β is an unknown parameter vector and f is a known function. For each sampled unit, the Z value is always observed, but the Y value may be a nonrespondent. We assume that the probability that an individual responds on Y can depend on both Y and Z according to

$$\phi(Y,Z,\gamma)=P(\delta=1\mid Y,Z), \tag{2.2}$$

where δ is the response indicator for Y, ϕ is a known function, and γ is an unknown parameter vector.

Without loss of generality, we assume that the first r sampled units are respondents and the rest of $n-r$ sampled units are nonrespondents. Thus, the observed data set is

$$\{(Y_i, Z_i), \quad i = 1, \ldots, r\} \cup \{Z_i, \quad i = r+1, \ldots, n\}.$$

Let $p_i = dF(Y_i)$ be the point mass that F places on Y_i. For observed Y_i, the likelihood is

$$\phi(Y_i, Z_i, \gamma) f(Y_i, Z_i, \beta) p_i.$$

For a nonrespondent Y_i, the likelihood is

$$\int \Big[1 - \phi(y, Z_i, \gamma)\Big] f(y, Z_i, \beta) dF(y).$$

Together with the survey weights (see, e.g., Chen and Qin (1993)), we obtain the following log-likelihood for the entire sample

$$\sum_{i=1}^{r} w_i \log\Big(\phi(Y_i, Z_i, \gamma) f(Y_i, Z_i, \beta) p_i\Big) + \sum_{i=r+1}^{n} w_i \log\bigg(\int [1 - \phi(y, Z_i, \gamma)] f(y, Z_i, \beta) dF(y)\bigg),$$

where $w_i = \omega_i / N$ and N is the finite population size. The use of w_i, instead of ω_i, does not change the maximization of the log-likelihood over the parameters. Since Z takes values $z_1, \ldots, z_s$, this log-likelihood can be written as

$$\sum_{i=1}^{r} w_i \log\Big(\phi(Y_i, Z_i, \gamma) f(Y_i, Z_i, \beta) p_i\Big) + \sum_{j=1}^{s} a_j \log(\pi_j), \tag{2.3}$$

where $a_j = \sum_{i=r+1}^{n} w_i I_{\{Z_i = z_j\}}$, I_A is the indicator function of the event A, and

$$\pi_j = P(\delta = 0, Z = z_j) = \int \Big[1 - \phi(y, z_j, \gamma)\Big] f(y, z_j, \beta) dF(y).$$

Note that π_j is a function of γ, β, and F. Maximizing (2.3) over γ, β, and F is equivalent to maximizing (2.3) over γ, β, p_i's, and π_j's subject to

$$p_i \geq 0, \quad \sum_{i=1}^{r} p_i = 1, \quad \pi_j = \sum_{i=1}^{r} p_i \Big[1 - \phi(Y_i, z_j, \gamma)\Big] f(Y_i, z_j, \beta), \quad j = 1, \ldots, s. \tag{2.4}$$

By introducing Lagrange multipliers, we can derive that

$$p_i = \frac{w_i}{\hat{N}_r + \sum_{j=1}^{s} \lambda_j [(1 - \phi(Y_i, z_j, \gamma)) f(Y_i, z_j, \beta) - \pi_j]}, \quad i = 1, \ldots, r, \tag{2.5}$$

where $\hat{N}_r = \sum_{k=1}^r w_k$ and λ_j's are Lagrange multipliers satisfying

$$\sum_{i=1}^r \frac{w_i[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]} = 0, \quad j = 1, \ldots, s. \tag{2.6}$$

Treating p_i in (2.5) as a function of β, γ, $\pi = (\pi_1, \ldots, \pi_s)$, and $\lambda = (\lambda_1, \ldots, \lambda_s)$, and substituting p_i into (2.3), the profile log-likelihood with Lagrange multipliers is

$$\begin{aligned} l(\beta, \gamma, \pi, \lambda) = & \sum_{i=1}^r w_i \log\Big(\phi(Y_i, Z_i, \gamma) f(Y_i, Z_i, \beta)\Big) + \sum_{j=1}^s a_j \log(\pi_j) \\ & + \sum_{i=1}^r w_i \log\left(\frac{w_i}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]}\right), \end{aligned}$$

Differentiating $l(\beta, \gamma, \pi, \lambda)$ with respect to π, λ, β, and γ, and setting the partial derivatives to 0, we have

$$\frac{a_j}{\pi_j} + \sum_{i=1}^r \frac{w_i \lambda_j}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]} = 0, \quad j = 1, \ldots, s, \tag{2.7}$$

$$\sum_{i=1}^r \frac{w_i[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]} = 0, \quad j = 1, \ldots, s, \tag{2.8}$$

$$\sum_{i=1}^r \left\{ \frac{w_i \partial \log f(Y_i, Z_i, \beta)}{\partial \beta} - \frac{w_i \sum_{j=1}^s \lambda_j (1-\phi(Y_i, z_j, \gamma)) \partial f(Y_i, z_j, \beta)/\partial \beta}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]} \right\} = 0, \tag{2.9}$$

$$\sum_{i=1}^r \left\{ \frac{w_i \partial \log \phi(Y_i, Z_i, \gamma)}{\partial \gamma} + \frac{w_i \sum_{j=1}^s \lambda_j \partial \phi(Y_i, z_j, \gamma)/\partial \gamma f(Y_i, z_j, \beta)}{\hat{N}_r + \sum_{j=1}^s \lambda_j[(1-\phi(Y_i, z_j, \gamma))f(Y_i, z_j, \beta) - \pi_j]} \right\} = 0. \tag{2.10}$$

From (2.5), (2.7), and the fact that $\sum_{i=1}^r p_i = 1$, we have

$$\lambda_j = -\frac{a_j}{\pi_j}, \quad j = 1, \ldots, s. \tag{2.11}$$

Let $(\hat{\beta}, \hat{\gamma}, \hat{\pi}, \hat{\lambda})$ be a solution to equations (2.8)−(2.11). The maximum empirical likelihood estimator (MELE) of (β, γ) is $(\hat{\beta}, \hat{\gamma})$, and the MELE of F is the empirical distribution $\hat{F}$ putting mass $\hat{p}_i$ at Y_i, $i = 1, \ldots, r$, where $\hat{p}_i$ is given by (2.5) with $(\beta, \gamma, \pi, \lambda)$ replaced by $(\hat{\beta}, \hat{\gamma}, \hat{\pi}, \hat{\lambda})$. If the parameter of interest is the finite population mean $\bar{Y} = \sum_{i \in \mathcal{P}} Y_i / N$, its MELE is

$$\hat{\bar{Y}} = \sum_{i=1}^r \hat{p}_i Y_i. \tag{2.12}$$

If the parameter of interest is the cell mean $\bar{Y}_j$, the finite population mean of Y given $Z = z_j$, the MELE is

$$\hat{\bar{Y}}_j = \frac{\sum_{i=1}^r \hat{p}_i f(Y_i, z_j, \hat{\beta}) Y_i}{\sum_{i=1}^r \hat{p}_i f(Y_i, z_j, \hat{\beta})}. \tag{2.13}$$

Let $\theta = (\beta, \gamma, \pi)$, $\hat{\theta} = (\hat{\beta}, \hat{\gamma}, \hat{\pi})$, and $\hat{\nu} = \hat{\lambda}/\hat{N}_r + J/(1 - \sum_{j=1}^s \hat{\pi}_j)$, where J is the s-vector of ones. The following result shows that $(\hat{\theta}, \hat{\nu})$ converges to $(\theta_0, 0)$, where $\theta_0 = (\beta_0, \gamma_0, \pi_0)$ is the true value of (β, γ, π). Also, $\hat{\bar{Y}}$ and $\hat{\bar{Y}}_j$ are consistent for $\bar{Y}$ and $\bar{Y}_j$, respectively. Furthermore, $(\hat{\theta}, \hat{\nu})$, $\hat{\bar{Y}}$, and $\hat{\bar{Y}}_j$ are asymptotically normal. The proof is given in the Appendix (available online at `http://www.stat.sinica.edu.tw/statistica`).

Theorem 1. *Assume the following.*

(i) *The sample from the finite population is selected with replacement according to a probability sampling plan or selected as a simple random sample without replacement. The values of (Y, Z) in the population is i.i.d. from a superpopulation according to* (2.1)−(2.2) *with a nonparametric Y-marginal F.*

(ii) *As $n \to \infty$, $N \to \infty$, $n/N \to 0$, $\max_{i \le N} w_i = O(1/n)$, and $n \sum_{i=1}^N w_i/N \to d$ for some constant d.*

(iii) *$f(y, z, \beta)$ and $\phi(y, z, \gamma)$ are twice continuously differentiable in β and γ for any y and z, and functions $\|\partial \log f(y, z, \beta)/\partial\beta\|^2$, $\|\partial \log \phi(y, z, \gamma)/\partial\gamma\|^2$, $\|\partial^2 f(y, z, \beta)/\partial\beta\partial\beta^\tau\|^2$, $\|\partial^2 \phi(y, z, \gamma)/\partial\gamma\partial\gamma^\tau\|^2$, $\|\partial f(y, z_j, \beta)/\partial\beta\|^3$, $\|\partial\phi(y, z_j, \gamma)/\partial\gamma\|^3$, $\|[\partial f(y, z_j, \beta)/\partial\beta][\partial f(y, z_k, \beta)/\partial\beta]^\tau\|^2$, $\|[\partial\phi(y, z_j, \gamma)/\partial\gamma][\partial\phi(y, z_k, \gamma)/\partial\gamma]^\tau\|^2$, and $\|[\partial f(y, z_j, \beta)/\partial\beta][\partial\phi(y, z_k, \gamma)/\partial\gamma]^\tau\|^2$ are bounded by some integrable functions in a neighborhood of β_0 and γ_0, $j, k = 1, \ldots, s$.*

(iv) *For any nonzero vector $c \in \mathcal{R}^{p+q}$, the value of $c^\tau \begin{pmatrix} \partial \log f(y, z_j, \beta_0)/\partial\beta \\ \partial \log \phi(y, z_j, \gamma_0)/\partial\gamma \end{pmatrix}$ depends on j, where p and q are the dimensions of β and γ.*

(v) *θ_0 is a unique root of $E[g(Y, \theta) \mid \delta = 1] = 0$ and $E[g(Y, \theta_0)g(Y, \theta_0)^\tau \mid \delta = 1]$ is positive definite, where $g = (g_1, \ldots, g_s)^\tau$ and*

$$g_j(y, \theta) = \frac{(1 - \sum_{k=1}^s \pi_k)[(1 - \phi(y, z_j, \gamma))f(y, z_j, \beta) - \pi_j]}{\sum_{k=1}^s \phi(y, z_k, \gamma) f(y, z_k, \beta)}. \tag{2.14}$$

(vi) *$\phi(y, z, \gamma)$ has a positive lower bound.*

Then, there exists a sequence $\{\hat{\theta}, \hat{\nu}, n = 1, 2, \ldots\}$ such that as $n \to \infty$,

$$P(\hat{\theta} \text{ is a solution to (2.8)−(2.10) }) \to 1, \tag{2.15}$$

$$\sqrt{n}\begin{pmatrix}\hat{\nu}-0\\ \hat{\theta}-\theta_0\end{pmatrix} \to_d N(0,\Sigma), \tag{2.16}$$

where the probability P and $\to_d$ (convergence in distribution) are with respect to the sampling and the superpopulation, and Σ is a positive definite matrix. Furthermore, if functions $\mid y(\partial f(y,z_j,\beta))/(\partial\beta)\|^2$ and midy$(\partial\phi(y,z_j,\gamma))/\ (\partial\gamma)\|^2$ are bounded by some integrable functions in a neighborhood of β_0 and γ_0 for each j, then

$$\sqrt{n}(\hat{\bar{Y}}-\bar{Y}) \to_d N(0,\sigma^2) \ \text{ and } \ \sqrt{n}(\hat{\bar{Y}}_j-\bar{Y}_j) \to_d N(0,\sigma_j^2), \quad j=1,\ldots,s, \tag{2.17}$$

where σ^2 and σ_j^2 are some constants.

In condition (v), $E[g(Y,\theta) \mid \delta=1]=0$ has a unique root θ_0 is equivalent to $\pi_j = \int[1-\phi(y,z_j,\gamma)]f(y,z_j,\beta)dF(y)$ is uniquely defined by (β,γ), i.e., a condition of identifiability of the π by (β,γ).

We now consider the stratified sample described in the beginning of this section. If (β,γ) in conditions (2.1) and (2.2) has different values in different strata, then we can solve (2.7)−(2.10) within each stratum to obtain an estimator of (β,γ) for each stratum. If (β,γ) is common for all strata, then constraint (2.4) is within each stratum, the sums in (2.7)−(2.8) are over each stratum, and the sums in (2.9)−(2.10) are over all strata. In any case, the marginal distribution of Y for stratum h is the empirical distribution putting mass $\hat{p}_i$ at Y_i with i in stratum h; the estimator of $\bar{Y}$ is the weighted average of estimators given by (2.12) over all strata with the weights $W_h = N_h/N$, where N_h is the population size for stratum h and $N=\sum_h N_h$; the estimator of $\bar{Y}_j$ is the ratio of the averages of the numerators and denominators in (2.13) with the weights W_h. Theorem 1 still holds if all conditions are given within each stratum and n_h/n coverges to a positive constant, where n_h is the sample size in stratum h and $n=\sum_h n_h$.

3. Pseudo Empirical Likelihood

When s (the number of Z categories) is not small, numerical solutions to (2.8)−(2.11) may be computationally intensive. Hence, we apply the idea of pseudo likelihood (Gong and Samaniego (1981)). That is, we substitute each π_j in (2.8)−(2.11) by a consistent estimator $\tilde{\pi}_j$. Note that consistent estimators of π_j's are easy to construct. For example, we may estimate π_j by

$$\tilde{\pi}_j = \frac{\sum_{i=1}^n w_i I_{\{\delta_i=0,Z_i=z_j\}}}{\sum_{i=1}^n w_i}. \tag{3.1}$$

Let $\tilde{\pi} = (\tilde{\pi}_j, j = 1, \ldots, s)$, $\tilde{\lambda}_j = -a_j/\tilde{\pi}_j$, and $\tilde{\lambda} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_s)$. Maximizing the pseudo empirical likelihood $l(\beta, \gamma, \tilde{\pi}, \tilde{\lambda})$ over (β, γ) results in the maximum pseudo empirical likelihood estimator (MPELE) $(\tilde{\beta}, \tilde{\gamma})$. Note that the MPELE is different from MELE since $\tilde{\pi}$ is not $\hat{\pi}$. However, we can directly establish the consistency and asymptotic normality of the MPELE.

Let $\tilde{p}_i$ be the estimator of p_i obtained by using (2.5) with β, γ, π_j, and λ_j replaced by $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{\pi}_j$, and $\tilde{\lambda}_j$, respectively. Because the MPELE is used, $\sum_{i=1}^r \tilde{p}_i \neq 1$, although $\sum_{i=1}^r \tilde{p}_i \to_p 1$. The MPELE of $\bar{Y}$ is

$$\tilde{\bar{Y}} = \frac{\sum_{i=1}^r \tilde{p}_i Y_i}{\sum_{i=1}^r \tilde{p}_i}, \tag{3.2}$$

and the MPELE of $\bar{Y}_j$ is

$$\tilde{\bar{Y}}_j = \frac{\sum_{i=1}^r \tilde{p}_i f(Y_i, z_j, \tilde{\beta}) Y_i}{\sum_{i=1}^r \tilde{p}_i f(Y_i, z_j, \tilde{\beta})}. \tag{3.3}$$

Estimators under stratified sampling can be obtained as described in the end of Section 2, with the sums in (3.1) within each stratum.

The following result shows that the MPELE is consistent and asymptotically normal.

Theorem 2. *Assume the conditions in Theorem* 1. *There exists a sequence* $\{\tilde{\beta}, \tilde{\gamma}, n = 1, 2, \ldots\}$ *such that, as* $n \to \infty$,

$$P\left(\frac{\partial l(\tilde{\beta}, \tilde{\gamma}, \tilde{\pi}, \tilde{\lambda})}{\partial(\beta, \gamma)} = 0\right) \to 1 \quad \text{and} \quad \sqrt{n}\begin{pmatrix} \tilde{\beta} - \beta_0 \\ \tilde{\gamma} - \gamma_0 \end{pmatrix} \to_d N(0, \Sigma_p), \tag{3.4}$$

where Σ_p *is a positive definite matrix. Furthermore,*

$$\sqrt{n}(\tilde{\bar{Y}} - \bar{Y}) \to_d N(0, \sigma_p^2) \quad \text{and} \quad \sqrt{n}(\tilde{\bar{Y}}_j - \bar{Y}_j) \to_d N(0, \sigma_{pj}^2), \quad j = 1, \ldots, s, \tag{3.5}$$

where σ_p^2 *and* σ_{pj}^2 *are some constants.*

4. Variance Estimation by Bootstrapping

It is a common practice in sample surveys to report a variance estimate for each estimate of the parameter of interest. We focus on the most commonly used estimators, the mean estimators $\hat{\bar{Y}}$, $\tilde{\bar{Y}}$ in (2.12) and (3.2), and the cell mean estimators $\hat{\bar{Y}}_j$, $\tilde{\bar{Y}}_j$ in (2.13) and (3.3). Because the formulation of these estimators is complicated, it is difficult to derive an analytic form of their asymptotic variances, σ^2, σ_j^2 in (2.17), and σ_p^2, σ_{pj}^2 in (3.5) . Thus, we apply the bootstrap method that consists of the following steps. In the following, $\hat{\eta}$ denotes any of $\hat{\beta}$, $\hat{\gamma}$, $\hat{\pi}$, $\hat{\nu}$, $\hat{\bar{Y}}$, $\hat{\bar{Y}}_j$, $\tilde{\beta}$, $\tilde{\gamma}$, $\tilde{\pi}$, $\tilde{\bar{Y}}$, and $\tilde{\bar{Y}}_j$.

1. Within stratum h, draw a simple random sample of size n_h with replacement from the set of sampled units (respondents or nonrespondents). Carry out this procedure independently across strata. For each unit in the bootstrap sample, the bootstrap data are the Z and Y values (if the Y is missing, the bootstrap datum is treated as missing) and their survey weights.
2. Compute $\hat{\eta}^*$, which is the same as $\hat{\eta}$ but with the original data replaced by the bootstrap data generated in Step 1.
3. Repeat the previous steps independently B times and obtain $\hat{\eta}^{*1}, \ldots, \hat{\eta}^{*B}$. Estimate the variance of $\hat{\eta}$ by the sample variance of $\hat{\eta}^{*1}, \ldots, \hat{\eta}^{*B}$.

The following result establishes the asymptotic validity of the bootstrap.

Theorem 3. *Assume the conditions in Theorem* 1.

(i) *Let* (2.8*)−(2.11*) *be the bootstrap analog of* (2.8)−(2.11). *Then there exists a sequence* $\{\hat{\theta}^*, \hat{\nu}^*, n = 1, 2, \ldots\}$ *such that, as* $n \to \infty$,

$$P_*(\hat{\theta}^* \text{ is a solution to } (2.8^*)-(2.10^*)) \to_p 1, \tag{4.1}$$

$$\sqrt{n}\begin{pmatrix} \hat{\nu}^* - \hat{\nu} \\ \hat{\theta}^* - \hat{\theta} \end{pmatrix} \to_{d^*} N(0, \Sigma), \tag{4.2}$$

where Σ *is given in* (2.16), P_* *denotes the bootstrap probability conditional on the data, and* $\vartheta_n^* \to_{d^*} \vartheta$ *means* $P_*(\vartheta_n^* \in B) - P(\vartheta \in B) \to_p 0$ *for any Borel set* B. *Furthermore,*

$$\sqrt{n}(\hat{\hat{Y}}^* - \hat{\hat{Y}}) \to_{d^*} N(0, \sigma^2) \;\text{ and }\; \sqrt{n}(\hat{\hat{Y}}_j^* - \hat{\hat{Y}}_j) \to_{d^*} N(0, \sigma_j^2), \tag{4.3}$$

where σ^2 *and* σ_j^2 *are defined in* (2.17).

(ii) *Let* $\tilde{\pi}^* = (\tilde{\pi}_1^*, \ldots, \tilde{\pi}_s^*)$, *with* $\tilde{\pi}_j^*$ *being the bootstrap analog of* $\tilde{\pi}_j$ *in* (3.1). *Then there exists a sequence* $\{\tilde{\beta}^*, \tilde{\gamma}^*, n = 1, 2, \ldots\}$ *such that, as* $n \to \infty$,

$$P_*\left(\frac{\partial l^*(\tilde{\beta}^*, \tilde{\gamma}^*, \tilde{\pi}^*, \tilde{\lambda}^*)}{\partial(\beta, \gamma)} = 0\right) \to_p 1 \;\text{ and }\; \sqrt{n}\begin{pmatrix} \tilde{\beta}^* - \tilde{\beta} \\ \tilde{\gamma}^* - \tilde{\gamma} \end{pmatrix} \to_{d^*} N(0, \Sigma_p), \tag{4.4}$$

where Σ_p *is given in* (3.4). *Further,*

$$\sqrt{n}(\tilde{\tilde{Y}}^* - \tilde{\tilde{Y}}) \to_{d^*} N(0, \sigma_p^2) \;\text{ and }\; \sqrt{n}(\tilde{\tilde{Y}}_j^* - \tilde{\tilde{Y}}_j) \to_{d^*} N(0, \sigma_{pj}^2), \tag{4.5}$$

where σ_p^2 *and* σ_{pj}^2 *are defined in* (3.5).

5. Imputation

Imputation is often carried out for practical reasons (Kalton and Kasprzyk (1986)). After imputation, estimates of parameters are computed by treating imputed values as observed data and using the standard formulas for the case of no nonresponse. In this section we consider imputation for the estimation of the population mean $\bar{Y}$ and the population cell mean $\bar{Y}_j$. Let $\hat{Y}_i = Y_i$ if Y_i is a respondent and $\hat{Y}_i$ be an imputed value if Y_i is a nonrespondent. After imputation, the population mean $\bar{Y}$ and cell mean $\bar{Y}_j$ are estimated by

$$\hat{\bar{Y}}_I = \sum_{i=1}^{n} w_i \hat{Y}_i, \tag{5.1}$$

$$\hat{\bar{Y}}_{jI} = \frac{\sum_{i=1}^{n} w_i \hat{Y}_i I_{\{Z_i=z_j\}}}{\sum_{i=1}^{n} w_i I_{\{Z_i=z_j\}}}, \tag{5.2}$$

respectively. Under stratified sampling, (5.1)$-$(5.2) should be modifed as described at the end of Section 2.

The naive mean imputation method imputes each nonrespondent with $Z = z_j$ by the cell sample mean $\sum_{i=1}^{r} w_i Y_i I_{\{Z_i=z_j\}} / \sum_{i=1}^{r} w_i I_{\{Z_i=z_j\}}$. The naive random imputation method imputes each nonrespondent with $Z = z_j$ by a random sample with replacement from respondents with $Z = z_j$, where each Y_i with $Z_i = z_j$ has probability $w_i I_{\{Z_i=z_j\}} / \sum_{i=1}^{r} w_i I_{\{Z_i=z_j\}}$ to be selected, $i = 1, \ldots, r$. The population mean estimators based on the naive imputation methods are inconsistent since they do not consider the difference between the respondents and the nonrespondents.

Using the MELE estimators developed in Section 2, we consider the following two imputation procedures.

1. Empirical Likelihood Mean Imputation. For each nonrespondent with $Z = z_j$, the imputed Y value is

$$\frac{\sum_{i=1}^{r} \hat{p}_i [1 - \phi(Y_i, z_j, \hat{\gamma})] f(Y_i, z_j, \hat{\beta}) Y_i}{\sum_{i=1}^{r} \hat{p}_i [1 - \phi(Y_i, z_j, \hat{\gamma})] f(Y_i, z_j, \hat{\beta})}.$$

2. Empirical Likelihood Random Imputation. Each nonrespondent with $Z = z_j$ is imputed by a random sample with replacement from all respondents, where the probability of each Y_i to be selected is

$$\frac{\hat{p}_i [1 - \phi(Y_i, z_j, \hat{\gamma})] f(Y_i, z_j, \hat{\beta})}{\sum_{i=1}^{r} \hat{p}_i [1 - \phi(Y_i, z_j, \hat{\gamma})] f(Y_i, z_j, \hat{\beta})}.$$

For stratified sampling, imputation should be carried out within each stratum.

Similarly, using the MPELE estimators developed in Section 3, we can develop Pseudo Empirical Likelihood Mean Imputation and Random Imputation. They are similar to the Empirical Likelihood Mean Imputation and Random Imputation that we described above. We just need to replace $\hat{\beta}$, $\hat{\gamma}$, and $\hat{p}_i$ by $\tilde{\beta}$, $\tilde{\gamma}$, and $\tilde{p}_i$, respectively.

The following result shows that the estimators of $\bar{Y}$ and $\bar{Y}_j$ based on these four imputation procedures are consistent and asymptotically normal.

Theorem 4. *Under the conditions of Theorem 1, for empirical likelihood mean imputation, empirical likelihood random imputation, pseudo empirical likelihood mean imputation, or pseudo empirical likelihood random imputation,*

$$\sqrt{n}(\hat{\bar{Y}}_I - \bar{Y}) \to_d N(0, \sigma_I^2), \quad \text{and} \quad \sqrt{n}(\hat{\bar{Y}}_{jI} - \bar{Y}_j) \to_d N(0, \sigma_{jI}^2), \quad j = 1, \ldots, s,$$

where σ_I^2 and σ_{jI}^2 are some constants.

The asymptotic variances σ_I^2 and σ_{jI}^2 do not have simple analytic forms. Variance estimation can be carried out using the bootstrap procedure described in Section 4. It should be emphasized that, to address the variability caused by imputation, nonrespondents in each bootstrap data set must be imputed using the bootstrap data and the same imputation method as that used to impute the original data set, as suggested by Shao and Sitter (1996).

6. Simulation Results

In this section, we report on simulation of the finite-sample properties of the MELE, MPELE, the empirical likelihood imputation, and the pseudo empirical likelihood imputation. We created a finite population similar to the Current Establishment Survey conducted by the U.S. Bureau of Labor Statistics. We chose four different industries as four strata with sizes $N_1 = 3,370$, $N_2 = 2,910$, $N_3 = 5,430$, and $N_4 = 4,110$. The variable Y is the total pay for each establishment and values of Y in stratum h were generated from a superpopulation F_h. The form of F_h was chosen to be the gamma distribution and $F_1 = \Gamma(43, 0.20)$, $F_2 = \Gamma(42, 0.19)$, $F_3 = \Gamma(38, 0.20)$, and $F_4 = \Gamma(50, 0.17)$, where $\Gamma(a, b)$ denotes the gamma distribution with shape parameter a and scale parameter b. The parameters in F_h's were chosen to match the mean and variance of a data set from the Current Establishment Survey.

The covariate $Z \in \{1, 2, 3, 4, 5\}$ was generated by the logistic model

$$P(Z = j \mid Y = y) = \frac{\exp\{\beta_j + \beta_5 y\}}{1 + \sum_{k=1}^{4} \exp\{\beta_k + \beta_5 y\}}, \quad j = 1, 2, 3, 4,$$

$$P(Z=5 \mid Y=y) = \frac{1}{1+\sum_{k=1}^{4} \exp\{\beta_k + \beta_5 y\}},$$

where β_k, $k = 1, 2, 3, 4, 5$, are unknown parameters whose values in the simulation are 0.25, 0.5, 0.75, 1, and -0.1, respectively.

The sampling plan was stratified simple random sampling. In each stratum, the sampling fraction was 0.05. For each sampled unit, the Y respondent was generated according to the response probability function

$$P(\delta = 1 \mid Y = y, Z = j) = \frac{\exp\{-10 - j + \gamma y\}}{1 + \exp\{-10 - j + \gamma y\}}$$

with a parameter $\gamma = 1.8$ or 2, or

$$P(\delta = 1 \mid Y = y, Z = j) = \frac{\exp\{10 + j + \gamma y\}}{1 + \exp\{10 + j + \gamma y\}}$$

with $\gamma = -1.4$. The following table lists the response rate for each Z and the mean response rate $E[P(\delta = 1 \mid Z)]$.

γ	1.8	2	-1.4
$P(\delta = 1 \mid Z = 1)$	0.888	0.951	0.457
$P(\delta = 1 \mid Z = 2)$	0.803	0.910	0.621
$P(\delta = 1 \mid Z = 3)$	0.697	0.842	0.751
$P(\delta = 1 \mid Z = 4)$	0.560	0.749	0.856
$P(\delta = 1 \mid Z = 5)$	0.469	0.675	0.908
$E[P(\delta = 1 \mid Z)]$	0.651	0.804	0.756

For each of the three γ, Table 1−3 respectively reports the relative bias (RB) and variance (VAR) of the MELE estimators in (2.12) and (2.13), the MPELE estimators in (3.2) and (3.3), the naive estimators that simply ignore nonrespondents, and the imputation estimators in (5.1) and (5.2) based on empirical, pseudo empirical, or naive mean imputation and random imputation. We also report their bootstrap variance estimators (Vboot) based on the bootstrap replication size $B = 200$, the coverage probabilities (CP) and the lengths (LEN) of the bootstrap confidence intervals of the form

$$\text{point estimate} \quad \pm 1.96\sqrt{\text{Vboot}}$$

that approximately have nominal coverage probability 95%.

Table 1. For $\gamma = 1.8$: Relative Bias (RB) in % and Variance (VAR) of the Estimators, Bootstrap Variance Estimates (Vboot), Coverage Probability (CP) in %, and Length (LEN) of 95% Confidence Interval.

Method		Naive					MELE					MPELE				
		RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN
Without Imputation	Y	5.92	0.0026	0.0025	0	0.19	0.23	0.0042	0.0041	91.5	0.24	0.16	0.0446	0.0522	96.5	0.84
	Y_1	1.95	0.0175	0.0160	74.0	0.49	0.26	0.0043	0.0044	93.5	0.25	0.19	0.0362	0.0414	96.5	0.75
	Y_2	3.45	0.0112	0.0127	30.0	0.44	0.26	0.0043	0.0044	92.8	0.25	0.17	0.0362	0.0414	96.9	0.75
	Y_3	5.75	0.0106	0.0107	0	0.40	0.24	0.0043	0.0044	91.8	0.25	0.18	0.0362	0.0414	96.5	0.75
	Y_4	8.63	0.0105	0.0097	0	0.38	0.23	0.0043	0.0044	90.8	0.25	0.20	0.0362	0.0414	96.9	0.75
	Y_5	10.59	0.0151	0.0141	0	0.46	0.16	0.0146	0.0150	94.1	0.47	0.41	0.1208	0.1333	94.2	1.37
Mean Imputation	Y	6.74	0.0026	0.0025	0	0.19	0.15	0.0031	0.0034	91.2	0.22	0.15	0.0055	0.0059	95.8	0.29
	Y_1	2.30	0.0173	0.0157	66.4	0.48	0.10	0.0151	0.0165	96.2	0.50	0.22	0.0173	0.0162	95.8	0.50
	Y_2	3.62	0.0137	0.0126	26.8	0.43	0.10	0.0138	0.0127	94.4	0.43	0.18	0.0127	0.0128	96.2	0.44
	Y_3	5.75	0.0105	0.0106	0	0.40	0.31	0.0101	0.0102	93.2	0.39	0.23	0.0122	0.0120	91.9	0.42
	Y_4	8.45	0.0103	0.0097	0	0.38	0.13	0.0070	0.0079	91.5	0.34	0.04	0.0141	0.0143	95.8	0.45
	Y_5	10.31	0.0149	0.0145	0	0.46	0.09	0.0131	0.0145	93.6	0.47	0.14	0.0183	0.0236	96.2	0.58
Random Imputation	Y	6.72	0.0032	0.0030	0	0.21	0.13	0.0033	0.0037	91.0	0.23	0.12	0.0058	0.0062	95.4	0.30
	Y_1	2.24	0.0194	0.0169	70.8	0.50	0.08	0.0154	0.0170	95.7	0.50	0.19	0.0176	0.0166	95.8	0.50
	Y_2	3.64	0.0124	0.0142	30.8	0.46	0.10	0.0147	0.0134	94.0	0.45	0.20	0.0138	0.0135	96.9	0.45
	Y_3	5.70	0.0124	0.0125	1.2	0.43	0.29	0.0109	0.0111	91.5	0.41	0.16	0.0135	0.0130	94.6	0.44
	Y_4	8.47	0.0117	0.0118	0	0.42	0.12	0.0081	0.0092	91.5	0.37	0.00	0.0155	0.0156	96.2	0.48
	Y_5	10.26	0.0185	0.0175	0	0.51	0.04	0.0163	0.0168	94.0	0.50	0.10	0.0213	0.0258	96.5	0.61

Table 2. For $\gamma = 2$: Relative Bias (RB) in % and Variance (VAR) of the Estimators, Bootstrap Variance Estimates (Vboot), Coverage Probability (CP) in %, and Length (LEN) of 95% Confidence Interval.

Method		Naive					MELE					MPELE				
		RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN
Without Imputation	Y	3.61	0.0021	0.0020	0	0.17	0.18	0.0026	0.0026	94.8	0.20	0.13	0.0243	0.0323	94.6	0.67
	Y_1	0.81	0.0158	0.0162	92.4	0.49	0.17	0.0032	0.0032	90.4	0.22	0.03	0.0196	0.0232	96.8	0.57
	Y_2	1.64	0.0120	0.0120	80.4	0.42	0.18	0.0032	0.0032	94.8	0.22	0.01	0.0196	0.0232	93.5	0.57
	Y_3	3.19	0.0105	0.0096	23.6	0.38	0.17	0.0032	0.0032	93.6	0.22	0.02	0.0196	0.0232	94.2	0.57
	Y_4	5.09	0.0079	0.0078	0.4	0.34	0.20	0.0032	0.0032	94.8	0.22	0.04	0.0196	0.0232	94.6	0.57
	Y_5	6.58	0.0108	0.0103	0	0.39	0.15	0.0118	0.0116	95.6	0.42	0.26	0.0752	0.0915	94.2	1.14
Mean Imputation	Y	3.96	0.0022	0.0021	0	0.17	0.14	0.0026	0.0024	94.4	0.19	0.06	0.0035	0.0035	96.2	0.23
	Y_1	1.11	0.0153	0.0161	88.8	0.49	0.12	0.0171	0.0168	94.8	0.50	0.18	0.0169	0.0165	94.6	0.50
	Y_2	1.79	0.0118	0.0119	76.8	0.42	0.17	0.0124	0.0126	95.2	0.43	0.05	0.0132	0.0126	93.5	0.44
	Y_3	3.21	0.0105	0.0095	22.0	0.38	0.13	0.0097	0.0098	94.4	0.38	0.27	0.0108	0.0104	95.0	0.40
	Y_4	4.97	0.0080	0.0078	0.4	0.34	0.15	0.0079	0.0076	94.8	0.34	0.03	0.0103	0.0107	96.2	0.40
	Y_5	6.61	0.0106	0.0104	0	0.39	0.12	0.0119	0.0116	95.6	0.42	0.15	0.0153	0.0162	96.5	0.49
Random Imputation	Y	3.94	0.0026	0.0023	0	0.18	0.12	0.0027	0.0025	94.4	0.19	0.08	0.0036	0.0037	95.8	0.23
	Y_1	1.15	0.0159	0.0167	87.6	0.50	0.12	0.0176	0.0169	95.2	0.50	0.18	0.0172	0.0167	94.6	0.50
	Y_2	1.78	0.0126	0.0128	79.6	0.44	0.18	0.0129	0.0129	95.2	0.44	0.08	0.0136	0.0129	94.6	0.44
	Y_3	3.20	0.0116	0.0106	26.8	0.40	0.12	0.0105	0.0103	94.8	0.39	0.26	0.0110	0.0108	94.2	0.41
	Y_4	4.92	0.0096	0.0091	0.4	0.37	0.15	0.0089	0.0083	94.8	0.35	0.07	0.0109	0.0112	96.2	0.41
	Y_5	6.58	0.0133	0.0123	0	0.43	0.06	0.0135	0.0127	96.0	0.44	0.15	0.0159	0.0172	95.0	0.50

Table 3. For $\gamma = -1.4$: Relative Bias (RB) in % and Variance (VAR) of the Estimators, Bootstrap Variance Estimates (Vboot), Coverage Probability (CP) in %, and Length (LEN) of 95% Confidence Interval.

Method		Naive					MELE					MPELE				
		RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN	RB	VAR	Vboot	CP	LEN
Without Imputation	Y	-3.96	0.0018	0.0019	0	0.17	0.20	0.0031	0.0031	94.0	0.21	0.57	0.1084	0.1195	95.4	1.18
	Y_1	-9.89	0.0201	0.0204	0	0.55	0.20	0.0043	0.0039	90.0	0.24	0.58	0.1224	0.1369	96.2	1.28
	Y_2	-7.11	0.0135	0.0131	0	0.44	0.25	0.0043	0.0039	92.5	0.24	0.63	0.1224	0.1369	95.4	1.28
	Y_3	-4.84	0.0085	0.0090	2.0	0.37	0.19	0.0043	0.0039	91.5	0.24	0.58	0.1224	0.1369	95.4	1.28
	Y_4	-2.86	0.0065	0.0068	19.2	0.32	0.22	0.0043	0.0039	94.5	0.24	0.60	0.1224	0.1369	95.8	1.28
	Y_5	-1.72	0.0085	0.0086	66.8	0.36	0.15	0.0112	0.0106	94.0	0.40	0.30	0.0823	0.0776	96.9	0.96
Mean Imputation	Y	-4.57	0.0018	0.0019	0	0.17	0.25	0.0027	0.0028	92.7	0.20	0.15	0.0054	0.0059	94.2	0.29
	Y_1	-9.46	0.0211	0.0208	0.4	0.55	0.26	0.0124	0.0128	95.4	0.44	0.08	0.0279	0.0269	97.3	0.60
	Y_2	-6.81	0.0132	0.0130	0	0.44	0.36	0.0112	0.0113	94.2	0.41	0.03	0.0180	0.0190	97.3	0.52
	Y_3	-4.80	0.0086	0.0089	2.4	0.36	0.14	0.0096	0.0097	93.1	0.38	0.20	0.0130	0.0134	96.5	0.44
	Y_4	-2.97	0.0067	0.0068	17.2	0.32	0.26	0.0070	0.0078	94.6	0.34	0.16	0.0098	0.0100	95.4	0.39
	Y_5	-1.93	0.0085	0.0086	60.4	0.36	0.27	0.0102	0.0099	94.2	0.38	0.35	0.0102	0.0099	93.5	0.39
Random Imputation	Y	-4.58	0.0022	0.0022	0	0.18	0.24	0.0031	0.0031	93.8	0.21	0.17	0.0059	0.0063	95.0	0.29
	Y_1	-9.47	0.0025	0.0024	0.8	0.60	0.25	0.0167	0.0186	96.5	0.53	0.01	0.0337	0.0327	97.3	0.67
	Y_2	-6.83	0.0152	0.0155	0	0.48	0.30	0.0132	0.0142	93.5	0.46	0.08	0.0218	0.0222	96.9	0.56
	Y_3	-4.80	0.0105	0.0103	3.6	0.39	0.18	0.0118	0.0112	93.5	0.41	0.18	0.0147	0.0149	95.8	0.47
	Y_4	-2.95	0.0076	0.0075	20.0	0.33	0.23	0.0086	0.0084	95.0	0.35	0.15	0.0106	0.0106	94.2	0.41
	Y_5	-1.94	0.0091	0.0092	60.8	0.37	0.25	0.0104	0.0103	95.0	0.39	0.34	0.0109	0.0104	92.7	0.40

Table 4. he Mean(Mean) and the Variance (VAR) of the Parameter Estimates. The true values are $\beta_1 = 0.25$, $\beta_2 = 0.5$, $\beta_3 = 0.75$, $\beta_4 = 1$, and $\beta_5 = -0.1$.

	MELE						MPELE					
	$\gamma = 1.8$		$\gamma = 2$		$\gamma = -1.4$		$\gamma = 1.8$		$\gamma = 2$		$\gamma = -1.4$	
	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR
β_1	0.2554	0.4074	0.2425	0.2987	0.2986	0.3863	0.3935	2.1449	0.1479	1.5564	0.2592	0.5873
β_2	0.5173	0.3873	0.5216	0.3025	0.5644	0.3883	0.6505	2.0736	0.4004	1.4972	0.4986	0.6183
β_3	0.7601	0.3990	0.7667	0.2984	0.8052	0.3792	0.9012	1.9614	0.6580	1.4382	0.7515	0.6906
β_4	1.0171	0.3878	1.0136	0.2945	1.0621	0.3713	1.1373	1.8363	0.9075	1.3843	0.9882	0.7577
β_5	-0.1006	0.0057	-0.1015	0.0041	-0.1075	0.0054	-0.1164	0.0214	-0.0896	0.0167	-0.0994	0.0127
γ	1.7952	0.0004	1.9960	0.0005	-1.4038	0.0002	1.8236	0.0597	2.0350	0.1310	-1.3912	0.0163

Table 5. The Ratio of MSE: mse(MPELE)/mse(MELE).

	Without Imputation						Mean Imputation						Random Imputation					
	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y	Y_1	Y_2	Y_3	Y_4	Y_5	Y	Y_1	Y_2	Y_3	Y_4	Y_5
$\gamma = 1.8$	9.85	7.38	7.41	7.51	7.54	7.78	1.23	0.97	0.84	1.11	1.39	1.27	1.16	0.95	0.84	1.06	1.28	1.15
$\gamma = 2$	8.01	5.38	5.34	5.35	5.24	7.39	1.01	0.85	1.05	1.21	1.16	1.17	1.00	0.85	1.04	1.15	1.09	1.28
$\gamma = -1.4$	23.91	24.90	24.13	25.16	24.74	4.34	1.46	1.69	1.04	1.13	1.33	1.00	1.52	1.47	1.03	1.04	1.40	1.04

Table 4 reports the mean and the variance (VAR) of the parameter estimates. Table 5 reports the ratios of the mean squared errors. Each MPELE is compared with its counterpart; that is, $\tilde{\bar{Y}}$ in (3.2) is compared with $\hat{\bar{Y}}$ in (2.12), $\tilde{\bar{Y}}_j$ in (3.3) is compared with $\hat{\bar{Y}}_j$ in (2.13), and $\hat{\bar{Y}}_I$ in (5.1) (or $\hat{\bar{Y}}_{jI}$ in (5.2)) with pseudo empirical likelihood mean (or random) imputation is compared with $\hat{\bar{Y}}_I$ in (5.1) (or $\hat{\bar{Y}}_{jI}$ in (5.2)) with empirical likelihood mean (or random) imputation described in Section 5.

The computation was done using MATLAB in a UNIX at the Department of Statistics, University of Wisconsin-Madison. For each γ and a single simulation, it took about 12 seconds to compute the MELE, MPELE, and imputed estimates for $\bar{Y}$ and $\bar{Y}_l$, $l = 1, \ldots, 5$. Because of the bootstrap, however, each simulation with a given γ took about 40 minutes. For each γ, we ran the simulation 250 times.

The simulation results can be summarized as follows.

1. In all cases, the proposed population mean and population cell mean estimators based on empirical likelihood or pseudo empirical likelihood (with imputation or not) performed well in terms of the relative bias (less than 1%) and variance, while the naive methods had heavy relative biases up to 10.31%.
2. The bootstrap variance estimate for our proposed estimators worked well in most cases in terms of its bias and the coverage probability of the bootstrap confidence interval. For the naive estimators, the coverage probability of the confidence interval was very low.
3. Although the MPELE estimators required less computational intensities, they were less efficient in terms of larger MSE compared with the MELE estimators. Most of the MSE ratios were greater than 1 (Table 5). For the estimators without imputation, the ratios were all greater than 5, and some of them were even greater than 20. The lengths of confidence intervals of the MPELE estimators were all greater than those of the MELE estimators, especially for the estimators without imputation.
4. Although the variances of the β and γ parameter estimates were a little bit large, the estimation of the population mean and population cell means, which is our major interest, was still good.

Acknowledgement

Jun Shao's work was partially supported by the NSF Grants DMS-0404535 and SES-0705033. The authors wish to thank the referees for their comments and suggestions.

References

Census Bureau (1987). Noncertainty sample specification. BSR-87 Action Memo D.06, the U.S. Census Bureau.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite population and the effective usage of auxiliary information. *Biometrika* **80**, 107-116.

Fang, F., Hong, Q. and Shao, J, (2009). A pseudo empirical likelihood approach for stratified samples with nonresponse. *Ann. Statist.* **37**, 371-393.

Gong, G. and Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and application. *Ann. Statist.* **9**, 861-869.

Greenlees, J. S., Reece, W. S. and Zieschang, K. Y. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. *J. Amer. Statist. Assoc.* **77**, 251-261.

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

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling. *J. Amer. Statist. Assoc.* **97**, 193-200.

Rancourt, E. (1999). Estimation with nearest-neighbor imputation at Statistics Canada. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 446-451.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-86.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *J. Amer. Statist. Assoc.* **91**, 1278-1288.

Tang, G., Little, R. J. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse. *Biometrika* **90**, 747-764.

Wolter, K., Shao, J. and Huff, L. (1998). Variance estimation for the current employment statistics program. *Proceedings of the Section on Survey Research Methods.* American Statistical Association, 775-780.

GE Consumer Finance, 1800 Cailun Road, Zhangjiang High-Tech Park, Shanghai 201203, P. R. China.

E-mail: fang.fang2@ge.com

Eli Lilly and Company, Lilly Corporate Center D/C 0734, Indianapolis, IN 46285, U.S.A.

E-mail: hong_quan@lilly.com

Department of Statistics, University of Wisconsin - Madison, Room 1220A, MSC 1300 University Ave. Madison, WI 53706-1685 U.S.A.

E-mail: shao@stat.wisc.edu

(Received December 2007; accepted October 2008)