A KERNEL SMOOTHING METHOD TO ADJUST FOR UNIT NONRESPONSE IN COMPLEX SURVEYS

Damião N. da Silva[*]
Universidade Federal do
Rio Grande do Norte (Brazil)

Jean D. Opsomer[**]
Iowa State University

1. Introduction

Nonresponse is a source of error in surveys that appears when part of the data to be collected are not observed. It has the potential to bias significantly the results of the survey and, consequently, to prevent valid inference. Typical causes for nonresponse involve refusal to participate in the survey, refusal to answer a question or unavailability of the interviewee. Nonresponse might also arise if the sampled unit can not be reached or located, or if they fail to have a measurement obtained.

Most methods to analyze data that contain nonresponse use adjustment procedures to compensate for the missing data. Kalton (1983) and Kalton and Kasprzyk (1986) classify such procedures as *weighting adjustments* and *imputation techniques.* Weighting adjustments are used to compensate for *unit nonresponse*, which occurs when no value for the characteristics of interest is recorded for the unit. These adjustments increase the weights of the units that respond to the survey in order to compensate for those who do not. Imputation techniques, on the other hand, are most often intended to handle *item nonresponse*, which occurs when there is partial data collection for some items of a given unit. We will only consider the case of unit nonresponse in this article.

In the context of unit nonresponse, when no value for the characteristics of interest is recorded for the unit, one possible method to compensate for the presence of nonrespondents weighs the remaining observations by incorporating estimates of the probabilities that the units are respondents. These probabilities are also known as response probabilities or *propensity scores*, following the Rosenbaum and Rubin (1983, 1985) theory for observational studies. This method was used by Nargundkar and Joshi (1975) to correct the Horvitz-Thompson estimator to account for the nonresponse. This same idea was applied by Cassel et al. (1983) to adjust the regression estimator. However, other types of weights based on response probabilities are possible as well. See, for example, the adjustments discussed in Chaubey and Crisalli (1995).

A critical step of the weighting with response probability procedure is the estimation of such probabilities. This step is usually processed under models relating the response occurrences and auxiliary variables. Using the terminology of Oh and Scheuren (1983), popular response models are the *uniform global response mechanism (UGRM)*, which assumes equal response probabilities for all units in the sample, and the *uniform response mechanism within subpopulations (URMWS)*, where the response probabilities are constant within each subpopulation or group. These models are popular in practice, because they only require simple ratio adjustments to the estimators.

A more sophisticated class of response models use an explicit parametric function to relate the response probabilities and the auxiliary variables. Typical choices for the parametric function are the logit and the probit models. Frequently, however, there is no a priori knowledge about the nonresponse process to substantiate the specification of a parametric function like the logit or probit. This represents a major disadvantage of the parametric approach to model the nonresponse process since, if the parametric model is misspecified, the resulting estimator is potentially biased.

It is therefore of interest to investigate the use of more flexible estimation techniques for the response probabilities. One particularly flexible technique is to estimate the response probabilities by nonparametric methods. Usually, these methods only assume that the response probabilities are related to the auxiliary variables by a "smooth" but unspecified function. The response probabilities may be estimated, for instance, by kernel-based smoothing techniques (see e.g. Wand and Jones 1995).

The use of kernel smoothing methods in the nonresponse context was first proposed by Giommi (1984). In that article, the author proposed to estimate the response probabilities at different values of an auxil-

[*] Departmento de Estatística, Campus Universitário, Natal RN 59072-970 Brazil, damiao@ccet.ufrn.br.

[**] Department of Statistics, Iowa State University, Ames IA 50011,jopsomer@iastate.edu.

iary variable by the response rates within neighborhoods centered at the those values. These resulting estimator of the response function is a running mean of the response indicators and can be also be viewed as a simple kernel smoother that uses a uniform kernel function. Giommi (1987) extends this previous estimator using a more general kernel function. In both articles, the author used the estimated response probabilities to construct estimators for the population mean according to the Cassel et al. (1983) adjusted regression estimator. The performance of the estimators was evaluated by Monte Carlo simulation experiments. The response probability estimators proposed by Giommi were also further considered by Niyonsenga (1994) and Niyonsenga (1997). These articles addressed the nonresponse problem following the method of selection in phases proposed by Särndal and Swensson (1985) and Särndal and Swensson (1987), where unit and item nonresponse correspond to subsequent phases in a multi-stage sampling framework.

In the current article, we extend the kernel-based approach of the previous authors in two ways. First, we will extend the estimation procedure from kernel regression to local polynomial regression, a method generally considered superior in the nonparametric literature (see Wand and Jones 1995, p.130). Second, we will develop an asymptotic framework in which the theoretical properties of the nonparametrically weighted estimator can be formally derived, and show that the estimator is consistent under the joint distribution of the sampling design and the response mechanism. Note that our results are directly applicable to the estimators in Giommi (1984) and Giommi (1987), because under equal-probability designs, these estimators can be viewed as local polynomial estimators with zero-degree polynomials.

The remainder of the article is organized as follows. Section 2 introduces the survey framework and the notation to account for nonresponse. Section 3 presents the construction of the proposed estimator for the population mean under nonresponse, and in Section 4, we establish the properties of the estimator. Finally, in Section 5, we summarize the main findings of the article. The derivations of the theoretical results in the article are found in Da Silva (2003).

2. Survey and Response Framework

We consider a finite population $U = \{1, 2, ..., N\}$, where N is known. Suppose that associated with U there are p characteristics of interest, $Y_1, Y_2, \cdots, Y_p$, and q auxiliary variables, $X_1, X_2, \cdots, X_q$. Let $\boldsymbol{Y}_i = (Y_{1,i}, Y_{2,i}, \cdots, Y_{p,i})'$ and $\boldsymbol{X}_i = (X_{1,i}, X_{2,i}, \cdots, X_{q,i})'$ denote the vectors of values of the characteristics of interest and the auxiliary variables corresponding to the i-th unit, $i \in U$, respectively. We denote $\boldsymbol{Y}_i$ by Y_i ($\boldsymbol{X}_i$ by X_i) when $p = 1$ ($q = 1$), and write $\boldsymbol{Y}, \boldsymbol{X}$ for the matrices (vectors) of the variables over the population U.

Let s be a sample selected from U ($s \subset U$) according to some probabilistic sampling design $p(\cdot)$. Using the information contained in s, the goal is to estimate quantities associated with the population U, such as means or totals of given characteristics. Sometimes, ratios of two means or totals are also of interest. In this article, we will consider the estimation of the population mean of $(Y_1, Y_2, \cdots, Y_p)'$, which is given by $\bar{\boldsymbol{Y}}_N = N^{-1} \sum_{i \in U} \boldsymbol{Y}_i$. Most popular estimators for $\bar{\boldsymbol{Y}}_N$ have the form $N^{-1} \sum_{i \in s} w_i \boldsymbol{Y}_i$ or equivalently, $N^{-1} \sum_{i \in U} w_i \boldsymbol{Y}_i \, I_i$, with w_i the sampling weight associated with the i-th unit and I_i an indicator variable for the event that the i-th unit is selected to the sample. We use $\boldsymbol{I} = (I_1, I_2, \cdots, I_N)'$ to represent the vector of sample inclusion indicators for all population units. For simplicity, we shall only consider here the Horvitz-Thompson estimator

$$\bar{\boldsymbol{y}}_\pi = \frac{1}{N} \sum_{i \in s} \pi_i^{-1} \boldsymbol{Y}_i, \qquad (1)$$

where $\pi_i = \Pr(i \in s)$ is the inclusion probability for the i-th unit.

In order to take into account nonresponse in the sample, we shall assume that each unit in the population is either a (potential) respondent or a nonrespondent. We introduce the response indicator R_i, assuming the value one if the i-th unit responds andzero otherwise, for all $i \in U$, and let $s_r = \{i \in s : R_i = 1\}$. The distribution of the vector $(R_i : i \in s)'$ is called the *response mechanism.* Unlike the vector $\boldsymbol{I}$, which has a known distribution once the sampling design is chosen, the survey sampler has no control over the response mechanism. However, since both the sampling design and the response mechanism are involved in the distribution of the survey estimator, it is necessary to specify a model for the nonresponse process.

In this article, we will assume that the R_i correspond to Poisson sampling, i.e. they are independent Bernoulli variables with

$$\Pr(R_i = 1 | \boldsymbol{Y}, \boldsymbol{X}, s) \equiv \phi_i \equiv \phi(\boldsymbol{X}_i) \qquad (2)$$

for all $i \in U$, where $\phi(\cdot)$ is a smooth but otherwise unspecified function of the $\boldsymbol{X}_i$ (with $0 < \phi(\cdot) \leq 1$). This response model implies that the nonresponse process does not depend on the units that are selected into the sample or the values of the vector $\boldsymbol{Y}$

corresponding to these units, but it allows for varying response probabilities, which depend on the auxiliary variables. These response probabilities correspond to *propensity scores*, whose theory was developed by Rosenbaum and Rubin(1983, 1985) in the context of observational studies. See David et al. (1983) and Little (1986) for the use of propensity scores in survey nonresponse problems.

Let $\mathbf{R} = (R_1, R_2, \cdots, R_N)'$. Combining the sampling design $p(\cdot)$ and the response model (2), we obtain a model for the joint distribution of $\mathbf{I}$ and $\mathbf{R}$, given $\boldsymbol{Y}$ and $\boldsymbol{X}$, namely,

$$\mathcal{L}\left(\mathbf{I},\mathbf{R}\big|\boldsymbol{Y},\boldsymbol{X}\right) = \mathcal{L}\left(\mathbf{I}\big|\boldsymbol{X}\right)\mathcal{L}\left(\mathbf{R}\big|\boldsymbol{X}\right), \qquad (3)$$

where $\mathcal{L}\left(\mathbf{I}\big|\boldsymbol{X}\right)$ denotes the distribution of the vector $\mathbf{I}$ generated by the sampling design $p(\cdot)$, which can depend on $\boldsymbol{X}$, and $\mathcal{L}\left(\mathbf{R}\big|\boldsymbol{X}\right)$ is the conditional distribution of $\mathbf{R}$ implied by the response mechanism (2). Evaluating inference under model (3) follows the *quasi-randomization* approach, in the terminology of Oh and Scheuren (1983). It treats the variability in the realized sample as dependent not only on the sampling design, but also (jointly) on the response model.

3. A Corrected Estimator for the Population Mean

In the presence of nonresponse, when classical estimators used in survey sampling are constructed by replacing the original sample s by the realized sample s_r, they no longer keep their usual statistical properties. For example, the nonresponse version of (1)

$$\bar{\boldsymbol{y}}_{\pi r} = \frac{1}{N}\sum_{i \in s_r} \pi_i^{-1}\boldsymbol{Y}_i \qquad (4)$$

is biased for the population mean. The bias, under the joint distribution of the sampling design and the response mechanism (2), is given by $\mathrm{B}(\bar{\boldsymbol{y}}_{\pi r}) = -N^{-1}\sum_{i \in U}(1-\phi_i)\boldsymbol{Y}_i$, (Cassel et al. 1983). Clearly, in the case of a nonnegative variable Y, (4) underestimates $\bar{Y}_N$, with the absolute bias increasing with the magnitude of the nonresponse probabilities for the units in the population.

One method to correct the estimator (4) for bias was discussed by Nargundkar and Joshi (1975), who argued that if the response probabilities were known,

$$\bar{\boldsymbol{y}}_{\pi\phi} = \frac{1}{N}\sum_{i \in s_r} w_i\phi_i^{-1}\boldsymbol{Y}_i, \qquad (5)$$

with $w_i = \pi_i^{-1}$, would be unbiased for $\bar{\boldsymbol{Y}}_N$. However, since in practice the response probabilities are unknown, estimator (5) is infeasible. But, it suggests the estimator

$$\bar{\boldsymbol{y}}_{\pi\widehat{\phi}} = \frac{1}{N}\sum_{i \in s_r} w_i\widehat{\phi}_i^{-1}\boldsymbol{Y}_i, \qquad (6)$$

where $\widehat{\phi}_i$ is an estimator of ϕ_i, $i \in s_r$. If it is required that the adjusted weights add up to unity, one should divide (6) by $\sum_{i \in s_r} w_i\widehat{\phi}_i^{-1}$. For this modification, see for example Little and Rubin (2002, p. 46).

One estimator of the form (6) is the weighting-class estimator (Oh and Scheuren 1983). It assumes that the population U can be divided into G disjoint, exhaustive and prespecified classes of elements $U_1, U_2, \cdots, U_G$. In the sample, there are n_g units from the g-th class, among which r_g are respondents ($0 \le r_g \le n_g$). The response probabilities are estimated within each class by $\widehat{\phi}_i \equiv r_g/n_g$, $i \in s_{r_g} = s_r \cap U_g$. One generalization of this method is the *fully efficient fractional imputed* (FEFI) estimator, proposed by Kim and Fuller (1999). The estimated response probabilities are $\widehat{\phi}_i \equiv \sum_{j \in s_{rg}} w_j \big/ \sum_{j \in s_g} w_j$, for all $i \in s_{rg}$, where $s_g = s \cap U_g$ and $g = 1, 2, \cdots, G$. Thus, under both the weighting–class and the FEFI methods, all response probabilities within a class have the same estimate. If the true corresponding probabilities are not homogeneous, however, these methods produce biased estimators. In the following section, we present a kernel smoothing method to estimate the response probabilities that does not rely on the specification of homogeneous cells.

In this article, we will consider (6) as the estimator for the population mean. Instead of using fixed classes U_g, the estimates $\widehat{\phi}_i$, $i \in s_r$, will be obtained by a kernel regression for each i. The idea behind the method is that if the function $\phi(\cdot)$ in (2) is smooth, then the estimation of $\phi_i = \phi(\boldsymbol{X}_i)$ should be possible by local averaging of the R_j in the sample for which $\boldsymbol{X}_j$ is "close" to $\boldsymbol{X}_i$. The observations that are used in the averaging process at i are identified by a window (kernel) centered around $\boldsymbol{X}_i$, which works somewhat like the classes in the FEFI method. The kernel method has the advantage that the response probabilities do not have to be assumed to be equal for all observations in a cell, and that the probabilities will not vary abruptly at cell boundaries.

We now describe the local polynomial estimator of the response probabilities, and for simplicity we will only consider the situation where the nonresponse process depends on one auxiliary variable X. Under (2), the indicator response variables are independent random variables, with $\mathrm{E}(R_i|\boldsymbol{X}) = \phi(X_i)$

and $\mathrm{Var}(R_i|\,\boldsymbol{X}) = \phi(X_i)(1-\phi(X_i))$, for all $i \in U$. The procedure to estimate $\phi_i = \phi(X_i)$ fits the centered polynomial

$$\beta_0 + \beta_1(\cdot - X_i) + \cdots + \beta_k(\cdot - X_i)^k$$

to $\{(X_j, R_j) \ : \ j \in s\}$ by weighted least squares. The observation weights are given by $w_j K_h(X_j - X_i)$, where k is the degree of the polynomial, $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a continuous positive kernel function and h is the smoothing parameter, also known as the *bandwidth.* Hence, the estimate of $\phi(X_i)$ is $\widehat{\phi}_i = \widehat{\beta}_0$, obtained by minimizing

$$\widehat{S}_i(\boldsymbol{\beta}) \equiv \sum_{j\in s} w_j K_h(X_j - X_i)\{R_j - \beta_0 - \beta_1(X_j - X_i) - \cdots - \beta_k(X_j - X_i)^k\}^2. \qquad (7)$$

with respect to $\beta_0, \beta_1, \ldots, \beta_k$.

This estimator differs from the "classical" local polynomial regression estimator described in Wand and Jones (1995) because of the inclusion of the sampling weights w_i. These weights are included to make design-based inference possible, as was also done in e.g. Breidt and Opsomer (2000). This procedure makes it possible to estimate ϕ_i for any $i \in U$ without having to specify a parametric form for the response function $\phi(\cdot)$. As long as $\phi(\cdot)$ is a continuous and smooth function, the local polynomial regression estimator $\widehat{\phi}_i$ can be used to properly adjust the design-based estimator for the effect of the nonresponse.

Two simple expressions for $\widehat{\phi}_i$ are obtained for the local constant ($k = 0$) and local linear ($k = 1$) fits. In the former case, the resulting estimator is

$$\widehat{\phi}(X_i, 0, h) \;=\; \frac{\sum_{j\in s} w_j K_h(X_j - X_i)R_j}{\sum_{j\in s} w_j K_h(X_j - X_i)}, \qquad (8)$$

while, for the latter, it is

$$\widehat{\phi}(X_i, 1, h) = \frac{\sum\limits_{j\in s}\{\widehat{s}_2(X_i,h) - \widehat{s}_1(X_i,h)(X_j - X_i)\}a_{ij}(h)}{n\left(\widehat{s}_2(X_i,h)\widehat{s}_0(X_i,h) - \widehat{s}_1(X_i,h)^2\right)},$$

with

$$a_{ij}(h) = w_j K_h(X_j - X_i)R_j$$

and

$$\widehat{s}_\ell(X_i,h) \;=\; \frac{1}{n}\sum_{j\in s}(X_j - X_i)^\ell w_j K_h(X_j - X_i).$$

Because the dependent variables R_i can only assume the values zero and one, $\widehat{\phi}(X_i, 0, h)$ and $\widehat{\phi}(X_i, 1, h)$ produce estimates restricted to the interval $[0, 1]$, as can readily be checked. That property does not hold for $k \geq 2$, however. When the sampling weights are constants, $\widehat{\phi}(X_i, 0, h)$ corresponds to the Nadaraya–Watson estimator used by Giommi (1987) in the response probability estimation context.

4. Theoretical Results

4.1 Asymptotic Framework and Assumptions

To study asymptotic properties under the quasi-randomization model (3), we assume that the population U is embedded in an increasing sequence of finite populations $\{U_\nu\}_{\nu=1}^\infty$, where the ν-th population has size the N_ν ($N_\nu > N_{\nu-1}$). Define $\boldsymbol{Y}_\nu = (Y_1, Y_2, \cdots, Y_{N_\nu})'$ to be the vector of values of one characteristic of interest, Y, associated with U_ν, and similarly, let $\boldsymbol{X}_\nu = (X_1, X_2, \cdots, X_{N_\nu})'$ be corresponding vector for the auxiliary variable X. For each ν, we assume that $\boldsymbol{X}_\nu$ is known and that a sample s_ν of size n_ν ($n_\nu \geq n_{\nu-1}$) is selected from U_ν, according to sampling design $p_\nu(\cdot)$. This increasing-population setup is commonly used for studying the asymptotic properties of survey estimators (see Isaki and Fuller (1982) for an early reference).

Let $\boldsymbol{I}_\nu = (I_1, I_2, \cdots, I_{N_\nu})'$ be the sample inclusion indicator vector for the ν-th population. Suppressing the ν for ease of notation, let $\pi_i = \Pr(I_i = 1)$, and let

$$\Delta_{j_1,\cdots,j_k} \equiv \mathrm{E_d}\left(\prod_{\ell=1}^k (I_{j_\ell} - \pi_{j_\ell})\right) \qquad (9)$$

denote higher moments for the sample inclusion indicators $I_{j_1}, I_{j_2}, \cdots, I_{j_k}$, where the subscript "d" indicates the expectation is taken with respect to the sampling design. Let $\mathbf{R}_\nu = (R_1, R_2, \cdots, R_{N_\nu})'$ denote the response indicator vector for the ν-th population.

We now state the assumptions needed to derive our main results. We shall assume that there are strictly positive constants $\lambda_1, \lambda_2, \ldots, \lambda_6$ such that:

(A1) $\lambda_1 < N_\nu n_\nu^{-1}\pi_i < \lambda_2 < \infty, \ \forall\, i \in U_\nu$;

(A2) $N_\nu^{-1}n_\nu \to \pi$, for some $0 < \pi < 1$, as $\nu \to \infty$;

(A3) For distinct $j_1, j_2, \cdots, j_k \in U_\nu$, where $k = 2, 3, \cdots, 8$,

$$\left|\Delta_{j_1,\cdots,j_k}\right| \leq \begin{cases} \left[\prod_{\ell=1}^k (N - \ell + 1)\right]^{-1} n_\nu^{\frac{k}{2}}\lambda_3 & , \\ \text{if } k \text{ is even,} & , \\ \left[\prod_{\ell=1}^k (N - \ell + 1)\right]^{-1} n_\nu^{\frac{k-1}{2}}\lambda_4, & \\ \text{if } k \text{ is odd} & , \end{cases}$$

(A4) $\lim_{\nu\to\infty} N_\nu^{-1}\sum_{i\in U_\nu} Y_i = \mu \in (-\infty,\infty)$ and $N_\nu^{-1}\sum_{i\in U_\nu} |Y_i|^4 \le \lambda_5$, for all $\nu \ge 1$.

In addition to these assumptions on the sampling design and the population distribution of the Y_i, we will also need the following assumptions on the response mechanism:

(B1) $R_1, R_2, \cdots, R_{N_\nu}$ are independent random variables;

(B2) $\Pr\{R_i = 1 \big| \boldsymbol{I}_\nu, \boldsymbol{Y}_\nu, \boldsymbol{X}_\nu\} = \Pr\{R_i = 1\big|\boldsymbol{X}_\nu\} \equiv \phi_i,\ \forall\ i \in U_\nu$;

(B3) $\phi_i = \phi(X_i),\ \forall\ i \in U_\nu$, where $\phi(\cdot)$ is a twice continuously differentiable function with $\lambda_6 < \phi(\cdot) \le 1$. The first derivative $\phi'(\cdot)$ has a finite number of sign changes.

Finally, we require assumptions on the distribution of the X_i and the kernel estimator:

(C1) For all $\nu \ge 1$, $X_1, X_2, \cdots, X_{N_\nu}$ are independent and identically distributed random variables with distribution $F_X(x) = \int_{-\infty}^x f_X(t)\,dt$, where $f_X(\cdot)$ is a continuous and positive probability density function on a compact set $[a_X, b_X]$. Without loss of generality, we shall take $[a_X, b_X] \equiv [0,1]$;

(C2) The kernel function $K(\cdot)$ is a bounded and continuous probability density, which is symmetric around zero and supported on [-1,1];

(C3) $\int_{-1}^1 z^4 K(z)dz < \infty$;

(C4) For all $\nu \ge 1$, $\{h_\nu\}$ is a sequence of bandwidths satisfying $0 < h_\nu \le 1$, $h_\nu \to 0$, $n_\nu h_\nu^2 \to \infty$ and $N_\nu h_\nu / \log N_\nu \to \infty$, as $\nu \to \infty$;

(C5) The first derivatives $f_X'(\cdot)$ and $K'(\cdot)$ have a finite number of sign changes on the intervals $[a_X, b_X]$ and [-1,1], respectively.

A detailed discussion of assumptions (A1)–(A4), (B1)–(B3) and (C1)–(C5) is provided in Da Silva (2003).

4.2 Estimation of the Population Mean under a Local Constant Fit

Consider the local constant fit $\widehat{\phi}(X_i, 0, h_\nu)$ of (8), for which we will formally describe the statistical properties in this section. The objective is to construct a consistent estimator to the reciprocal of the i-th response probability, ϕ_i^{-1}, and then use it to adjust the estimator for the population mean according to (6). Notice, however, that the random variable $\widehat{\phi}(X_i, 0, h_\nu)$ can be equal to zero for some $i \in s_\nu$, if there are no respondents in the interval $(X_i - h, X_i + h)$. In these situations, the reciprocal of the corresponding estimated response probability is undefined and, by consequence, so is $\bar{y}_{\pi\widehat{\phi}}$. More generally, local polynomial estimators of degree k will be undefined whenever there are less than $k+1$ respondents in the interval $(X_i - h, X_i + h)$, which can happen even if $i \in s_r$.

One way of handling this problem formally is by modifying the local polynomial estimator to always produce positive values with probability one. Define

$$\begin{aligned}\widehat{\boldsymbol{m}}_{i\nu} &\equiv (\widehat{m}_{1i\nu}, \widehat{m}_{2i\nu})' \\ &= \frac{1}{N_\nu h_\nu}\sum_{j\in s_\nu} w_j K\left(\frac{X_j - X_i}{h_\nu}\right)(R_j, 1)'. \quad (10)\end{aligned}$$

Observe that $\widehat{\phi}(X_i, 0, h_\nu) = \widehat{m}_{2i\nu}^{-1}\widehat{m}_{1i\nu}$, for all $i \in U_\nu$, and therefore is not invertible when $\widehat{m}_{1i\nu}$ is equal to zero. So, one possible correction for this estimator is to add a small (nonrandom) positive quantity to $\widehat{m}_{1i\nu}$. This type of adjustment was previously advocated for similar reasons by Fan (1993) and Breidt and Opsomer (2000), for instance. We shall bound $\widehat{\phi}(X_i, 0, h_\nu)$ away from zero, by replacing $\widehat{m}_{1i\nu}$ by

$$\widehat{m}_{1i\nu}^* = \max\{\widehat{m}_{1i\nu}, (N_\nu h_\nu)^{-1}\delta\}, \qquad (11)$$

where δ is a fixed positive constant, and the corresponding (adjusted) response probability estimator will be given by

$$\widehat{\phi}_{0i\nu} \equiv \widehat{m}_{2i\nu}^{-1}\widehat{m}_{1i\nu}^*, \quad i \in U_\nu. \qquad (12)$$

Note that (12) is strictly positive because of the adjustment $(N_\nu h_\nu)^{-1}\delta$. Also, when $\widehat{m}_{1i\nu} \ge (N_\nu h_\nu)^{-1}\delta$, which happens with high probability for ν sufficiently large, then both estimators $\widehat{\phi}_{0i\nu}$ and $\widehat{\phi}(X_i, 0, h_\nu)$ are positive and $\widehat{\phi}_{0i\nu}^{-1} = \widehat{\phi}(X_i, 0, h_\nu)^{-1}$; that is, the reciprocal of the adjusted estimator of (12) coincides with the reciprocal of the zero order kernel regression estimator of (8).

According to prescription (6), the estimator for the population mean is therefore given by

$$\bar{y}_{\pi 0\nu} \equiv \frac{1}{N_\nu}\sum_{i\in s_r} w_i \widehat{\phi}_{0i\nu}^{-1} Y_i. \qquad (13)$$

The following results present the theoretical properties of (12) and (13). In what follows, let $P_{\boldsymbol{X}}$ denote the joint probability distribution of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots$.

Theorem 1. *Consider a sequence of increasing populations $\{U_\nu : \nu \ge 1\}$, where U_ν has size N_ν. Assume that for each ν, a sample s_ν of fixed size n_ν $(n_\nu \ge n_{\nu-1})$ is selected from U_ν by a probabilistic sampling design $p_\nu(\cdot)$, satisfying (A1)–(A3).*

Suppose the response mechanism satisfies the conditions (B1)–(B3), $\boldsymbol{X}_\nu$ satisfies (C1) and assumptions (C2)–(C5) hold. Consider the estimator (12) for the response probability $\phi(X_i)$, where $i \in U_\nu$ is fixed. Then, with $P_{\boldsymbol{X}}$–probability one,

$$\begin{aligned}\mathrm{E}\left[|\widehat{\phi}_{0i\nu}^{-1} - \phi(X_i)^{-1}| \,\middle|\, \boldsymbol{X}_\nu\right] \le \\ O(h_\nu)\, I_{\{X_i \in [0,h_\nu] \cup (1-h_\nu,1]\}} + \\ O\big(h_\nu^2\big)\, I_{\{X_i \in (h_\nu,1-h_\nu]\}} + O\big((n_\nu h_\nu)^{-1/2}\big).\end{aligned} \tag{14}$$

Theorem 1 implies directly that $\widehat{\phi}_{0i\nu}^{-1}$ is consistent and asymptotically unbiased for $\phi(X_i)^{-1}$.

Theorem 2. *Suppose the conditions of Theorem 1 hold. Consider the estimator $\bar{y}_{\pi 0\nu}$ in (13) to estimate the population mean of any variable Y satisfying (A4). Then, with $P_{\boldsymbol{X}}$–probability one,*

$$\mathrm{E}\left[\big(\bar{y}_{\pi 0\nu} - \bar{y}_{\pi\psi\nu}\big)^2 \,\middle|\, \boldsymbol{X}_\nu\right] = O\big((n_\nu h_\nu)^{-2}\big), \tag{15}$$

where $\bar{y}_{\pi\psi\nu}$ is a random variable such that, with $P_{\boldsymbol{X}}$–probability one,

$$\begin{aligned}\mathrm{E}\big(\bar{y}_{\pi\psi\nu} - \bar{Y}_{N_\nu} \,\big|\, \boldsymbol{X}_\nu\big) = \\ O\left(h_\nu^{3/2}\right) + O\left((n_\nu h_\nu)^{-1}\right)\end{aligned} \tag{16}$$

and

$$\mathrm{Var}\big(\bar{y}_{\pi\psi\nu} \,\big|\, \boldsymbol{X}_\nu\big) = O\big((n_\nu h_\nu)^{-1}\big). \tag{17}$$

Corollary 1. *Under the conditions of Theorem 2, suppose that the sampling design is such that, conditioned on $\boldsymbol{X}_\nu$,*

$$\frac{\big(\bar{y}_{\pi\psi\nu} - \bar{Y}_{N_\nu} - B_\nu\big)}{\mathrm{Var}(\bar{y}_{\pi\psi\nu})} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{a.s. } P_{\boldsymbol{X}},$$

as ν tends to ∞. Suppose also that

$$V \equiv \lim_{\nu\to\infty} (n_\nu h_\nu)\mathrm{Var}\big(\bar{y}_{\pi\psi\nu} \,\big|\, \boldsymbol{X}_\nu\big) \in (0,\infty) \;\; \textit{a.s. } P_{\boldsymbol{X}}.$$

Then, conditioned on $\boldsymbol{X}_\nu$,

$$\frac{\big(\bar{y}_{\pi 0\nu} - \bar{Y}_{N_\nu} - B_\nu\big)}{\mathrm{Var}(\bar{y}_{\pi\psi\nu})} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \quad \textit{a.s. } P_{\boldsymbol{X}},$$

as ν tends to ∞.

Remark 1. The results in Theorems 1 and 2 should be interpreted as statements with probability one with respect to the distribution of the sequence of values for the auxiliary variable X in the population; that is, the results are valid for almost all such possible sequences, under the distribution $P_{\boldsymbol{X}}$. So even though the results depend on the regularity conditions on the distribution of X given in (C1), it is still a "finite population" result in the sense of being valid for (almost) any sequence of finite populations U_ν.

Remark 2. The statement (15) implies that

$$\bar{y}_{\pi 0\nu} = \bar{y}_{\pi\psi\nu} + O_p\left(\frac{1}{n_\nu h_\nu}\right). \tag{18}$$

and hence that $\bar{y}_{\pi 0\nu} - \bar{Y}_{N_\nu}$ and $\bar{y}_{\pi\psi\nu} - \bar{Y}_{N_\nu}$ have the same asymptotic distribution (up to terms of sufficiently high order). Under the additional assumption that $\bar{y}_{\pi\psi\nu}$ is asymptotically normal, we obtain the asymptotic normality result given in Corollary 1.

Remark 3. The bandwidth parameter plays an important role in the asymptotic properties of $\bar{y}_{\pi\psi\nu}$. By (16), the bias has two components that converge to zero, as a consequence of (C4). The first one is at most of order $h_\nu^{3/2}$. In this local polynomial regression context, this order can be seen as a compromise between $O(h_\nu)$ and $O\big(h_\nu^2\big)$, which represent the orders of the biases when the estimation of $\phi(X_i)$ takes place at the "boundary" and "bounded away from the boundary", respectively (see Theorem 1). Despite this, it is possible for this component to achieve order $O\big(h_\nu^2\big)$ by requiring additional assumptions; for example, if the assumption of finite fourth moment on the Y variable — assumption (A4) — is strengthened to uniformly boundedness. The second component in the bias of $\bar{y}_{\pi\psi\nu}$ is at most of order $1/n_\nu h_\nu$, and hence is a negligible component in the mean squared error compared to the order of the variance in (17).

Remark 4. The $1/n_\nu h_\nu$ convergence rate in the approximation (18) and in variance (17) plays the same role as $1/n_\nu$ in the parametric context. Speaking somewhat loosely, it is possible to think of $n_\nu h_\nu$ as the "equivalent sample size" in the nonparametric regression context. Hence, the price paid for not specifying a parametric shape for the nonresponse function is a slower rate of convergence and a larger asymptotic variance, compared to a those obtained under a parametric model specification.

Remark 5. Combining the results in (15), (16) and (17), it follows that

$$\mathrm{E}\Big[\big(\bar{y}_{\pi 0\nu} - \bar{Y}_{N_\nu}\big)^2 \,\Big|\, \boldsymbol{X}_\nu\Big] = a h_\nu^3 + \frac{b}{n_\nu h_\nu},$$

for some positive constants a, b. So, an "optimal" bandwidth choice, in the sense of minimizing the mean square error of $\bar{y}_{\pi 0\nu}$ for n_ν fixed, is given by $h_0 = [b/3an_\nu]^{1/4}$. Furthermore, as $\bar{Y}_{N_\nu}$ converges to a finite constant by (A4), then on a set of $P_{\boldsymbol{X}}$–

probability one, for all $\varepsilon > 0$,

$$P\left\{|\bar{y}_{\pi 0\nu} - \bar{Y}_{N_\nu}| \geq \varepsilon \big| \boldsymbol{X}_\nu\right\} \leq \frac{1}{\varepsilon^2}\mathrm{E}\left\{\left[\left(\bar{y}_{\pi 0\nu} - \bar{y}_{\pi\psi\nu}\right) + \left(\bar{y}_{\pi\psi\nu} - \bar{y}_{N_\nu}\right)\right]^2 \big| \boldsymbol{X}_\nu\right\} \to 0,$$

as ν tends to infinity. Therefore, $\bar{y}_{\pi 0\nu}$ is consistent for $\bar{Y}_{N_\nu}$.

Remark 6. Expression (17) in Theorem 2 gives an asymptotic rate for the variance of $\bar{y}_{\pi 0\nu}$. Although this variance can be explicitly expressed in a quadratic form, the structure of the variance weights results in a long and complicated expression that is unsuited for constructing direct variance estimators. We shall address the estimation of this variance in a further article.

5. Conclusions

In this article, we addressed the problem of unit nonresponse in sample surveys by considering a weighting procedure that adjusts the sampling weights by the reciprocal of estimates for the response probabilities. The estimated response probabilities are obtained by nonparametric regression, a procedure that makes it possible to avoid prespecifying a parametric form for the nonresponse model. The procedure is shown to estimate consistently the population mean of any characteristic of interest that has a finite fourth population moment. The asymptotic rates of convergence for the bias and variance of the asymptotic distribution are provided. The effect of the bandwidth parameter on the bias and variance is also examined.

A number of open questions still need to be addressed in the study of kernel-based nonresponse adjustments for survey estimators. In the methodological area, a useful variance estimator still needs to be derived. In addition, the results given in this article need to be generalized from zero-order local polynomials to those of higher order. Finally, the practical behavior of the estimators needs to be evaluated through simulation experiments and examples on real data.

References

Breidt, F. J. and J. D. Opsomer (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics 28*, 1026–1053.

Cassel, C.-M., C.-E. Särndal, and J. H. Wretman (1983). Some uses of statistical models in connection with the nonresponse problem. In W. G. Madow, I. Olkin, and D. B. Rubin (Eds.), *Incomplete data in sample surveys (Vol. 3): Theory and bibliographies*, pp. 143–160. Academic Press (New York; London).

Chaubey, Y. P. and A. N. Crisalli (1995). Adjustment of the inclusion probabilities in case of nonresponse. In *Statistical Society of Canada Proceedings of the Survey Methods Section*, pp. 75–79. Statistical Society of Canada (McGill University, Montreal).

Da Silva, D. N. (2003). *Adjustments for survey unit nonresponse under nonparametric response mechanisms.* Ph. D. thesis, Iowa State University.

David, M. H., R. Little, M. Samuhel, and R. Triest (1983). Imputation models based on the propensity to respond. In *ASA Proceedings of the Business and Economic Statistics Section*, pp. 168–173.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics 21*, 196–216.

Giommi, A. (1984). A simple method for estimating individual response probabilities in sampling from finite populations. *Metron 42*(4), 185–200.

Giommi, A. (1987). Nonparametric methods for estimating individual response probabilities. *Survey Methodology 13*, 127–134.

Isaki, C. and W. Fuller (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association 77*, 89–96.

Kalton, G. (1983). *Compensating for missing survey data.* Institute of Social Research.

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

Kim, J.-K. and W. A. Fuller (1999). Jackknife variance estimation after hot deck imputation. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 825–830. American Statistical Association (Alexandria, VA).

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*, 139–157.

Little, R. J. A. and D. B. Rubin (2002). *Statistical Analysis With Missing Data.* Wiley.

Nargundkar, M. and G. B. Joshi (1975). Nonresponse in sample surveys. In *40th session of* *the ISI, Warsaw 1975, Contributed papers*, pp. 626–628.

Niyonsenga, T. (1994). Nonparametric estimation of response probabilities in sampling theory. *Survey Methodology 20*, 177–184.

Niyonsenga, T. (1997). Response probability estimation. *Journal of Statistical Planning and Inference 59*, 111–126.

Oh, H. L. and F. J. Scheuren (1983). Weighting adjustments for unit non-response. In W. G. Madow, I. Olkin, and D. B. Rubin (Eds.), *Incomplete data in sample surveys (Vol. 2): Theory and bibliographies*, pp. 143–184. Academic Press (New York; London).

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rosenbaum, P. R. and D. B. Rubin (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician 39*, 33–38.

Särndal, C. E. and B. Swensson (1985). Incorporating nonresponse modelling in a general randomization theory approach. In *Proceedings of the 45th Session of the International Statistical Institute*, Volume 15.2, pp. 1–15. International Statistical Institute (Voorburg, The Netherlands).

Särndal, C. E. and B. Swensson (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review 55*, 279–294.

Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman and Hall.