Journal of Official Statistics, Vol. 28, No. 2, 2012, pp. 239–277

Calibration Inspired by Semiparametric Regression as a Treatment for Nonresponse

Giorgio E. Montanari[1] *and M. Giovanna Ranalli*[2]

In the last decade, calibration has been used to reduce both sampling error and nonresponse bias in surveys. In the presence of auxiliary variables with known population totals or with known values on the originally sampled units, the calibration procedure generates final weights for observations that, when applied to those auxiliary variables, yield their population totals or unbiased estimates of these totals, respectively. A single set of variables and a single calibration step is employed to this end. In this article, we extend this approach to allow for more flexible implicit description of the relationship of the auxiliary variables with either the response probabilities or the survey variable(s). By using penalized splines the simplicity of the original proposal and the linearity of the estimator are preserved. The conditions under which the proposed estimator of the total is design consistent and its asymptotic properties are explored, and its finite sample behavior is investigated via simulations.

Key words: Auxiliary information; nonparametric regression; penalized splines; nonresponse bias; shrinkage; unit nonresponse.

1. Introduction

Nonresponse can harm the quality of the estimates of a survey. In particular, since we have to accept that those who respond are in general different from those who do not respond, bias is introduced. In this article we will not deal with imputation, but only with design weights modification to adjust for unit nonresponse bias. Note that techniques for handling nonresponse can be employed also for nonresponse adjustments in censuses. Commonly, a two-phase approach is used, with the response mechanism as the second phase; this is based on quasi-randomization theory, where the response distribution has corresponding response probabilities assumed to be independent of the realized sample (e.g., Särndal et al. 1992, Ch. 9). In practice, such response probabilities have to be estimated assuming a response model. The prefix "quasi" is added to emphasize that inference depends not only on the design, but also on the assumed response model.

[1] Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia. Email: giorgioeduardo.montanari@unipg.it
[2] Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia. Email: giovanna@stat.unipg.it
Acknowledgments: This study was first presented as an invited paper at the First Italian Conference on Survey Methodology, Siena, Italy, June 10-12, 2009. The work reported here has been developed under the support of the project PRIN 2007 "Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics" awarded by the Italian Government. The authors wish to thank the associate editor and three referees for their close reading of the manuscript and for many thoughtful comments, Wayne Fuller and Mingue Park for very useful suggestions on some theoretical issues.

One of the most common and simple techniques for handling nonresponse is given by constructing response homogeneity groups: the population (or the sample) can be partitioned into groups in such way that units belonging to the same group are assumed to have the same response probability. Sometimes, a more complex direct modeling of the response probabilities is conducted, for example through logistic regression models (Little 1986; Ekholm and Laaksonen 1991). Asymptotic properties in this situation are explored in Kim and Kim (2007). More generally, the probability of response can be assumed to be the inverse of a known *link* function of an unknown (but estimable) linear combination of model variables (Folsom 1991; Fuller et al. 1994; Kott 2006). Nonparametric regression that allows relaxing the assumption of a known functional relationship between response probabilities and model variables has been explored through kernel smoothing (Giommi 1987; Silva and Opsomer 2006).

Lundström and Särndal (1999) propose a simple approach for the treatment of nonresponse based on *calibration* (Deville and Särndal 1992). This is pursued by the construction of a single set of weights for all variables of interest that are as close as possible to specified initial weights (usually the design weights), while satisfying benchmark constraints on known auxiliary information. No discrimination is made within the set of auxiliary variables available to the researcher: a single set of variables is employed at the same time for nonresponse treatment, sampling error reduction and coherence among estimates. No explicit model is specified for the treatment of the nonresponse mechanism; it is implicitly given by the calibration procedure.

The relationship between regression estimation and calibration is well known (Deville and Särndal 1992; Särndal 2007): the efficiency of the calibration procedures relies on how well a linear model describes the relationship between the variable(s) of interest and the auxiliary ones. It therefore may be inefficient when the underlying relationship is indeed nonlinear (Wu and Sitter 2001; Montanari and Ranalli 2005). We argue that the approach in Lundström and Särndal (1999) can be usefully generalized to include more complex relationships through semiparametric regression (Rupper et al. 2003) without losing in simplicity. Semiparametric regression based on penalized splines (Eilers and Marx 1996) has been usefully employed for model-assisted inference in the case of complete response (Breidt et al. 2005). More easily than with kernel smoothing, it allows for the treatment, at the same time, of categorical and continuous auxiliary variables. Categorical variables can be inserted parametrically, while continuous variables can be accounted for nonparametrically. Recently, a kernel-based model-assisted estimator that can handle both continuous and categorical covariates has been proposed in Sànchez-Borrego et al. (2011).

The article proceeds as follows: in Section 2, calibration with particular regards to treatment of nonresponse is reviewed. In Section 3, semiparametric regression is employed to extend nonparametrically calibration to the treatment of nonresponse. Simulation studies that explore the finite sample behavior of the proposed estimator are reported in Section 4. Some concluding remarks and directions for future research are provided in Section 5.

2. Calibration as a Treatment for Nonresponse

Consider a finite population of N elements $U = \{1, \ldots, k, \ldots, N\}$; the aim is to estimate the total $Y = \sum_U y_k$, where y_k is the value of the variable of interest y for the kth unit.

We will use the shorthand $\sum_A$ for $\sum_{k\in A}$, with $A \subseteq U$ an arbitrary set. A sample s of size n is drawn from U through the sampling design $p(s)$ that induces positive first and second order inclusion probabilities $\pi_k = P(k \in s)$ and $\pi_{kj} = P(k,j \in s)$, respectively, with $\pi_{kk} = \pi_k$. Let I_k be the indicator variable for unit k selected in the sample, so that $E(I_k|\mathcal{F}) = \pi_k$, where $\mathcal{F} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, $\boldsymbol{u}_k = \left(y_k, \boldsymbol{x}_k^T\right)$, and $\boldsymbol{x}_k$ is the value of the p-vector of auxiliary variables $\boldsymbol{x}$ on unit k. Therefore, expectation is taken with respect to the sampling design and conditional of the realized finite population $\mathcal{F}$. We will denote with $d_k = 1/\pi_k$ the design weight of unit k. Since nonresponse occurs, the response set r of size m is obtained, with $r \subseteq$ s and $m \leq n$. Let $\delta_k = 1$ if unit k responds and 0 otherwise.

Lundström and Särndal (1999) consider auxiliary information on two separate levels. In particular, $\boldsymbol{x}$ is considered a vector of auxiliary variables assumed to contain information for reducing both the sampling error and the nonresponse bias, and the two following "information levels" are considered:

info-s: $\boldsymbol{x}_k$ is known for all $k \in$ s;
info-U: $\boldsymbol{x}_k$ is observed for all $k \in r$ and $\sum_U \boldsymbol{x}_k$ is known.

In the first case, information goes up to the sample level, while in the second case, it refers to the population. A combination of the two can of course be considered (Särndal and Lundström 2005), but we will consider them separately in order to keep this discussion simple.

Design weights d_k for responding units are on average too small to produce reasonable Horvitz-Thompson estimates of totals when there is nonresponse. They need to be adjusted by a factor v_k. Calibrated weights $w_k = d_k v_k$ used to compute the estimator $\hat{Y}_{c,r} = \sum_r w_k y_k$ of Y are obtained so that they satisfy calibration equations given by either

info-s: $\sum_r w_k \boldsymbol{x}_k = \sum_s d_k \boldsymbol{x}_k$, or
info-U: $\sum_r w_k \boldsymbol{x}_k = \sum_U \boldsymbol{x}_k$.

A simple choice for the factors v_k may be $v_k = 1 + \boldsymbol{\xi}^T \boldsymbol{x}_k$, which is linear and has considerable computational advantages. Note that this choice is equivalent to finding calibrated weights by minimizing a chi-squared distance measure from basic design weights (see e.g. Deville and Särndal 1992; Särndal and Lundström 2005, p. 58). Other forms for v_k are considered in Deville (2000) and Kott (2006). The vector $\boldsymbol{\xi}$ is determined after substitution in the calibration equations. The calibration estimator in these cases takes the following forms:

info-s: $\hat{Y}_{c,r} = \sum_r d_k v_{sk} y_k$, with

$$v_{sk} = 1 + \left(\sum_s d_k \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k\right)^T \left(\sum_r d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \boldsymbol{x}_k, \quad \text{for } k \in r;$$

info-U: $\hat{Y}_{c,r} = \sum_r d_k v_{Uk} y_k$, with

$$v_{Uk} = 1 + \left(\sum_U \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k\right)^T \left(\sum_r d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \boldsymbol{x}_k, \quad \text{for } k \in r. \tag{1}$$

Note that the estimator is constructed using calibration in a single phase and without explicit introduction of a model for the response mechanism. In addition, the weights do not depend on the study variable y, and therefore the estimator is said to be *linear.* Such a property is very valuable in a survey setting, because the calibrated weights can then be applied to all variables of interest. Poststratification is included as a special case of $\hat{Y}_{c,r}$ if the auxiliary vector $\boldsymbol{x}_k$ denotes membership to poststrata. In case of full response, when $r = s$, $\hat{Y}_{c,s} = \sum_s d_k y_k$, that is, the Horvitz-Thompson estimator for info-s, and $\hat{Y}_{c,s} = \sum_s d_k g_{c,k} y_k$, with $g_{c,k} = 1 + \left(\sum_U \boldsymbol{x}_k - \sum_s d_k \boldsymbol{x}_k\right)^T \left(\sum_s d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \boldsymbol{x}_k$, that is, the generalized regression estimator for info-U.

In the two-phase approach to handling nonresponse, an unbiased estimator is given by

$$\hat{Y}_{2p} = \sum_r \frac{d_k}{\theta_k} y_k \tag{2}$$

where $\theta_k = P(\delta_k = 1 | I_k = 1)$ is the conditional probability that unit k responds, given that the unit is selected in the sample (Särndal et al. 1992, Ch. 9). Of course such estimator cannot be computed, since nonresponse probabilities are not known. However, we can note that $\hat{Y}_{c,r}$ uses proxy values given by v_{sk} and v_{Uk} to approximate θ_k^{-1}, that is, the inverse of the response probability for unit k is implicitly approximated by a linear combination of the vector of auxiliary variables $\boldsymbol{x}_k$.

Weights in (1) provide a calibration estimator that is equivalent to the regression estimator in the presence of nonresponse. In fact, it can be written in the form

$$\hat{Y}_{c,r} = \sum_r d_k y_k + \left(\sum_U \boldsymbol{x}_k - \sum_r d_k \boldsymbol{x}_k\right)^T \hat{\boldsymbol{b}}_r, \tag{3}$$

where $\hat{\boldsymbol{b}}_r = \left(\sum_r d_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_r d_k \boldsymbol{x}_k y_k$, that is, the regression estimator when the regression coefficient is computed only on respondents. This estimator has been considered and studied in Fuller et al. (1994) and Fuller and An (1998).

To review the large sample properties of $\hat{Y}_{c,r}$, we will consider the traditional finite population asymptotic framework considered in Isaki and Fuller (1982), where the population U and the sampling design $p(\cdot)$ are embedded into a sequence of finite populations and associated probability samples. The set of indices of the elements in the Nth finite population is $U_N = \{1, 2, \ldots, N\}$ with $N = p + 1, p + 2, \ldots$, while the design is $p_N(\cdot)$ and the sample size n_N is assumed to grow with N. Let $\mathcal{F}_N = \left\{\left(y_{1N}, \boldsymbol{x}_{1N}^T\right), \left(y_{2N}, \boldsymbol{x}_{2N}^T\right), \ldots, \left(y_{NN}, \boldsymbol{x}_{NN}^T\right)\right\}$ be the set of vectors of both survey and auxiliary variables for the Nth finite population. In the following, the subscript N on the vectors will often be dropped for ease of notation.

Now, under regularity conditions such as those reported in Fuller (2002), $\hat{\boldsymbol{b}}_r$ is a design-consistent estimator of

$$\boldsymbol{c}_U = \left(\sum_U \theta_k \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_U \theta_k \boldsymbol{x}_k y_k, \tag{4}$$

in the sense that, given $\mathcal{F}_N$ for all $N > p + 2$, $\lim_{N\to\infty} P\{|\hat{\boldsymbol{b}}_r - \boldsymbol{c}_U| > \epsilon|\mathcal{F}_N\} = 0$ for all $\epsilon > 0$. The population total of y can be written as

$$Y = \sum_U a_k + \sum_U \boldsymbol{x}_k^T \boldsymbol{c}_U$$

with $a_k = y_k - \boldsymbol{x}_k^T \boldsymbol{c}_U$. Therefore, the regression estimator in (3) – and consequently $\hat{Y}_{c,r}$ for info-U – will be a design-consistent estimator for Y, in the sense that $\lim_{N\to\infty} P\{|\hat{Y}_{c,r} - Y| > N\epsilon|\mathcal{F}_N\} = 0$, if the probability limit of $\sum_U a_k$ is zero. There are several ways in which this occurs. Fuller et al. (1994) give the three following situations.

(*i*) The probability limit of $\sum_U a_k$ is zero when the sequence of finite populations is a sequence of random samples from an infinite population in which the linear model $y_k = \boldsymbol{x}_k^T \boldsymbol{b} + e_k$, with the e_k independent of the $\boldsymbol{x}_k$ and with zero expectation, holds for all k.

(*ii*) The total $\sum_U a_k$ is zero when θ_k is constant for all k, because in this case $\boldsymbol{c}_U = \boldsymbol{b}_U = \left(\sum_U \boldsymbol{x}_k \boldsymbol{x}_k^T\right)^{-1} \sum_U \boldsymbol{x}_k y_k$, and $\sum_U y_k - \boldsymbol{x}_k^T \boldsymbol{b}_U = 0$.

(*iii*) A sufficient condition for $\sum_U a_k$ to be zero is the existence of a vector $\boldsymbol{d}$ such that

$$\boldsymbol{x}_k^T \boldsymbol{d} = \theta_k^{-1} \tag{5}$$

for all $k \in U$. Therefore, if the inverse of the response probability is a linear function of the auxiliary variables, the regression estimator is consistent for Y.

Note that to have design consistency, it is sufficient that any one of these conditions holds. Apart from condition (*ii*), which is unlikely to hold in practice, it is enough that the auxiliary variables fulfill either the prediction model in (*i*) or the response model in (*iii*) to achieve a vanishing bias. This property has been called "double protection" against nonresponse bias.

The third condition sheds some light on the implicit modeling of the response mechanism done with the one-step calibration technique. A sufficient condition for consistency is that the inverse of the response probabilities belongs to the space spanned by the columns of the $N \times p$ matrix of population values of $\boldsymbol{x}$. In the following section, the approach in Lundström and Särndal (1999) is generalized to make condition (*i*) above valid for a wider range of models through semiparametric regression without loss of simplicity.

3. Calibration Inspired by Semiparametric Regression for the Treatment of Nonresponse

Semiparametric regression that relies on penalized splines has been usefully employed for model-assisted inference in the case of complete response (Breidt et al. 2005). Penalized splines are now often referred to as p-splines and have been brought to attention by Eilers and Marx (1996). P-splines provide an attractive smoothing method due to their simplicity of implementation, being a relatively straightforward extension of linear regression, and to their flexibility, as they are applicable in a wide range of modeling contexts. Ruppert et al. (2003) provide a thorough treatment of p-splines and their applications. In this section,

we first describe p-splines in the general context of model-assisted estimation regardless of nonresponse, and then move to semiparametric regression-based calibration for treatment of nonresponse.

3.1. Review of p-splines for Model-Assisted Regression Estimation

Let us first consider only smoothing with one covariate z. In Breidt et al. (2005), a nonparametric superpopulation regression model is written as

$$y_k = m(z_k) + \varepsilon_k, \tag{6}$$

where the errors ε_k are independent random variables with mean zero and variance $v(z_k)$. The p-spline estimator of the unknown function $m(\cdot)$ may be given by

$$m(z; \boldsymbol{\beta}) = \beta_0 + \beta_1 z + \sum_{l=1}^{L} \beta_{1+l}(z - \kappa_l)_+ \tag{7}$$

where the so-called plus functions $(t)_+$ are such that $(t)_+ = t$ if $t > 0$ and 0 otherwise (see Figure 1), κ_l for $l = 1, \ldots, L$ is a set of fixed knots, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{1+L})^T$ is the coefficient vector made of a parametric portion (the first two coefficients) and a spline part (the last L coefficients). The latter portion of the model allows for handling departures from a linear fit in the structure of the relationship. If the number of knots L is sufficiently large, the class of functions in (7) is very large and can approximate most smooth functions. In the p-splines context, a knot is placed every 4 or 5 observations; however, to avoid an excessive number of knots (and therefore parameters), a maximum number of allowable knots, say 35, is recommended. In addition, knots are usually placed at the quantiles of the distribution of z, making unequally spaced intervals so as to more properly account for the possible skewness of such a distribution. More details on knots choice can be found in Ruppert et al. (2003, Ch. 3 and 5). In the survey context, the choice of the

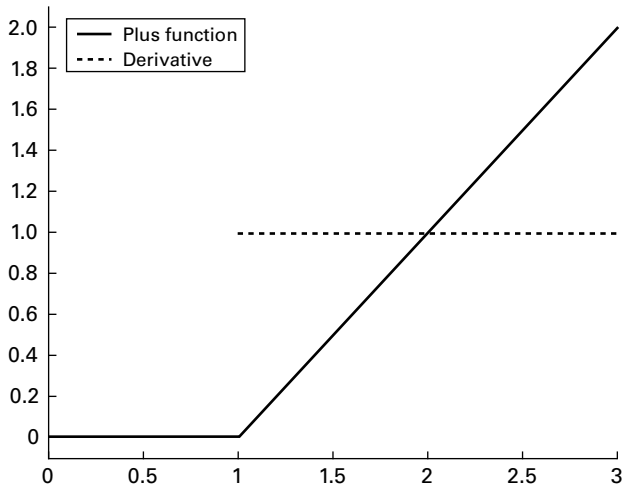

Fig. 1. Example of a truncated linear spline basis $(x - \kappa)_+$ with $\kappa = 1$ (solid line) and of its derivative (dashed line).

location of the knots may be also influenced by the available population level auxiliary information. More details on this aspect will be given in Section 3.2.

The spline model (7) uses a truncated linear spline basis to approximate the function $m(\cdot)$. In this case, the slope of the relationship between y and z is allowed to change when it hits a knot. The amount of the change is given by the coefficient of the corresponding base function, which represents the change in the first derivative of the approximating function at the knot (see Figure 1). Truncated polynomials of higher order, say 2 or 3, can be used, in which, similarly, coefficients for the base functions provide the change in the 2nd and 3rd, respectively, derivative. Different bases can be used, like thin plate splines or B-splines; more details on choice of base can be found in Ruppert et al. (2003, Ch. 3 and 5). In this article we will use p-splines with a truncated linear basis, not only for their simplicity of interpretation, but also because of their implication on the auxiliary information required. We will see this in more detail in the next section.

Given the large number of knots, model (7) can be too complex; the influence of the knots can be limited by putting a constraint on the size of the spline coefficients. Estimation can be accommodated by including this constraint in the least squares criterion, so that the census level estimator of the parameter vector is given by the minimizer of

$$\sum_U \{y_k - m(z_k; \boldsymbol{\beta})\}^2 + \lambda \sum_{l=1}^{L} \beta_{1+l}^2$$

for some fixed positive constant λ. The smoothness of the resulting fit depends on the value of λ, with larger values corresponding to smoother fits. Choice of λ will be discussed later. Let $\boldsymbol{z}_k = (1, z_k, (z_k - \kappa_1)_+, \ldots, (z_k - \kappa_L)_+)^T$ and $\boldsymbol{\Lambda} = \text{diag}\{0, 0, \lambda, \ldots, \lambda\}$ be an $L + 2$ diagonal matrix. Then, the census level penalized least squares estimator of the coefficient vector has the following ridge regression representation

$$\boldsymbol{\beta}_U = \left(\sum_U \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \sum_U \boldsymbol{z}_k y_k.$$

The role of the matrix $\boldsymbol{\Lambda}$ is to shrink the magnitude of the value of the coefficient $\boldsymbol{\beta}_U$ for the spline part of the function in (7). Under the conditions in Breidt et al. (2005), consistent design-based estimates of $\boldsymbol{\beta}_U$ can be obtained as $\hat{\boldsymbol{\beta}}_s = \left(\sum_s d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \sum_s d_k \boldsymbol{z}_k y_k$. Finally, sample-based fits $\hat{m}_k = m(\boldsymbol{z}_k; \hat{\boldsymbol{\beta}}_s)$ are used to define the model-assisted p-spline estimator

$$\hat{Y}_{p,s} = \sum_U \hat{m}_k + \sum_s d_k (y_k - \hat{m}_k) = \sum_s d_k g_{p,k} y_k \tag{8}$$

with $g_{p,k} = 1 + \left(\sum_U \boldsymbol{z}_k - \sum_s d_k \boldsymbol{z}_k \right)^T \left(\sum_s d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{z}_k$. Breidt et al. (2005, Sec. 2.2) discuss in detail the properties of this estimator. Here note that it can be seen as a calibration estimator in which calibration constraints are met for the first two variables $\{1, z\}$ and are relaxed for the other L. The amount of relaxation depends on the smoothing parameter λ. Relaxing the constraints on the L variables related to the knots has a shrinkage effect on the range of the final set of weights.

3.2. P-splines and Calibration for the Treatment of Nonresponse

Now, we want to exploit the enhanced flexibility provided by using p-splines in a model-assisted framework while retaining the simplicity of the proposal of Lundström and Särndal (1999) for handling nonresponse. To achieve this, we introduce the following calibration estimator based on p-splines:

info-s: $\hat{Y}_{p,r} = \sum_r d_k v_{sk} y_k$, with

$$v_{sk} = 1 + \left(\sum_s d_k \boldsymbol{z}_k - \sum_r d_k \boldsymbol{z}_k \right)^T \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{z}_k, \tag{9}$$

info-U: $\hat{Y}_{p,r} = \sum_r d_k v_{Uk} y_k$, with

$$v_{Uk} = 1 + \left(\sum_U \boldsymbol{z}_k - \sum_r d_k \boldsymbol{z}_k \right)^T \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \boldsymbol{z}_k. \tag{10}$$

It is easy to see that in the case of full response, that is, when $r = s$, $\hat{Y}_{p,r}$ reduces to the Horvitz-Thompson estimator for info-s and to $\hat{Y}_{p,s}$ in (8) for info-U. Recall that the auxiliary information for unit k used to compute these estimators is given by $\boldsymbol{z}_k = \left(1, z_k, \boldsymbol{z}_k^{*T}\right)^T$, with $\boldsymbol{z}_k^* = ((z_k - \kappa_1)_+, \ldots, (z_k - \kappa_L)_+)^T$. The first two entries of the vector are those usually employed for calibration, while $\boldsymbol{z}_k^*$ allows for handling departures from a linear fit in the structure of the relationship between y and z as illustrated in the previous section. Calibrating on the whole vector $\boldsymbol{z}_k$, however, may lead to a very erratic final set of weights. Therefore, the influence of the knots is limited by relaxing the binding constraint for that part of the auxiliary information. This is accomplished by minimizing a penalized version of the chi-square distance measure between final and initial weights. In particular, weights satisfy either of the two following conditions:

$$\textbf{info-s}: \ \min_{w_k} \sum_r \frac{(w_k - d_k)^2}{d_k} + \left(\sum_r w_k \boldsymbol{z}_k^* - \sum_s d_k \boldsymbol{z}_k^* \right)^T \boldsymbol{\Lambda}_*^{-1} \left(\sum_r w_k \boldsymbol{z}_k^* - \sum_s d_k \boldsymbol{z}_k^* \right)$$

$$\text{under the constraint } \sum_r w_k (1, z_k) = \sum_s d_k (1, z_k);$$

$$\textbf{info-U}: \ \min_{w_k} \sum_r \frac{(w_k - d_k)^2}{d_k} + \left(\sum_r w_k \boldsymbol{z}_k^* - \sum_U \boldsymbol{z}_k^* \right)^T \boldsymbol{\Lambda}_*^{-1} \left(\sum_r w_k \boldsymbol{z}_k^* - \sum_U \boldsymbol{z}_k^* \right)$$

$$\text{under the constraint } \sum_r w_k (1, z_k) = \sum_U (1, z_k),$$

where $\boldsymbol{\Lambda}_* = \text{diag}\{\lambda, \ldots, \lambda\}$ and λ here represents the inverse cost of relaxing those constraints. In general, smaller values of λ mean a large penalization and therefore that the calibration constraints are more stringent. Larger values imply increasingly relaxing the constraint for those variables and, therefore, a shrinkage effect on the range of the final set of weights. The results of those constrained minimization problems provide the estimators

considered above. See also Rao and Singh (1997) on relaxing the calibration constraints, and Fuller (2002, Sec. 9), Park and Fuller (2009) and Guggemos and Tillé (2010) on the link between the penalized minimum distance criterion and mixed effects models, and Beaumont and Bocci (2008) for a review on ridge calibration.

Let us consider again the estimator under the two-phase approach to handling nonresponse in (2). Here, for $\hat{Y}_{p,r}$ we can see that the inverse of the response probability is approximated by proxy values given by v_{Uk} and v_{sk} for info-U and info-s respectively, which depend on the whole vector of auxiliary variables $\boldsymbol{z}_k$ and not only on its linear part $(1, z_k)$ as it would be the case under classical calibration estimation. This allows for a more flexible implicit description of the nonresponse mechanism.

Let us now consider the auxiliary information required to compute these estimators. The vector $\boldsymbol{z}_k$ indeed contains information on only one variable z, so that the info-s needed to compute it reduces to $\boldsymbol{z}_k$ known for each $k \in s$, that is, the same information needed to compute $\hat{Y}_{p,r}$ with the auxiliary vector given by $\boldsymbol{x} = (1, z)^T$. As for info-U, on the other hand, the information required to compute $\hat{Y}_{p,r}$ is more than that needed to compute $\hat{Y}_{c,r}$ with $\boldsymbol{x} = (1, z)^T$. In particular, we need $\sum_U \boldsymbol{z}_k$ to be known. This means that, other than N and $\sum_U \boldsymbol{z}_k$, we need population counts and totals of z in subgroups defined by the knots, that is, $\sum_U I(z_k > \kappa_l)$ and $\sum_U z_k I(z_k > \kappa_l)$ for $l = 1, \ldots, L$. In fact, $\sum_U (z_k - \kappa_l)_+ = \sum_U (z_k - \kappa_l) I(z_k > \kappa_l) = \sum_U z_k I(z_k > \kappa_l) - \kappa_l \sum_U I(z_k > \kappa_l)$. Note that with other nonparametric techniques, like local polynomials or neural networks, the amount of auxiliary information required is much larger; in particular, z_k has to be known for all $k \in U$ (see e.g. Montanari and Ranalli 2005).

A particularly valuable property in the survey estimation contexts of $\hat{Y}_{p,r}$ – inherited by $\hat{Y}_{p,s}$ – is that of being a linear estimator. This result assumes that the number and placement of the knots and the value of the penalty constant λ are all determined and fixed before the model is fitted. The efficiency of the estimator will depend on the choice of these factors. However, for p-splines it is sufficient to focus on the choice of λ, since the choice of the other settings has been shown to have a limited effect on the final fit once the value of λ is allowed to vary (see e.g. Ruppert 2002; Ruppert et al. 2003, Ch. 5).

In addition, Breidt et al. (2005) note that, in the survey context, trying to find the optimal penalty is not as relevant as in the classical nonparametric regression context: the estimator is not constructed for a single variable, but for a large set of variables collected during the survey. A penalty that is optimal for a variable may well not be adequate for another one and using different sets of weights would not be feasible for practical purposes and for coherence issues. We will therefore consider a single fixed value for λ representing a compromise choice that may work reasonably well for many variables in a survey. In Section 3.4, we will give some guidelines to select such a value and in the simulation studies we will look at its effects on the final performance of the proposed estimator.

3.3. Asymptotic Properties

To study the asymptotic properties of $\hat{Y}_{p,r}$, we will follow closely the approach mentioned in Section 2 to discuss the properties of $\hat{Y}_{c,r}$. In particular, to discuss the large sample properties of $\hat{Y}_{p,r}$, we will again consider the asymptotic framework discussed in Section 2

in which $\mathcal{F}_N = [\boldsymbol{u}_{1N}, \boldsymbol{u}_{2N}, \ldots, \boldsymbol{u}_{NN}]$ is the set of vectors $\boldsymbol{u}_{kN} = (y_{kN}, \boldsymbol{z}_{kN}^T)$ for Nth finite population. In this regard, given that the regression coefficient in $\hat{Y}_{p,r}$ resembles a ridge type coefficient, conditions on the value of λ_N as the population and sample sizes increase should also be added. In particular, we can consider the following two cases.

Case A. In the first case one accepts that the shrinking effect vanishes asymptotically and, therefore, that the penalized coefficient vector converges to the unpenalized one. This can be reasonable, given that the number of knots is kept fixed asymptotically. For the shrinking effect to vanish, as $n_N \to \infty$, λ_N can remain constant or go to zero. More generally, λ_N can also grow as n_N grows, but at a slower rate, that is, $\lambda_N = O(n_N^{\alpha})$ with $\alpha < 1$, so that the ratio between λ_N and n_N goes to zero. In this case, the properties of $\hat{Y}_{p,r}$ coincide with those discussed for the classical calibration estimator $\hat{Y}_{c,r}$.

Case B. In the second case, one wants to ensure that the shrinking effect does not vanish asymptotically. This is reasonable when one wants to keep a smooth relationship between y and z also asymptotically. In this case, λ_N is allowed to grow as n_N grows. Theorem 3.1 proves the consistency of $\hat{Y}_{p,r}$ in this case. To this purpose, consider the following assumptions.

A1. *Assume L is fixed and the knots κ_l for $l = 1, \ldots, L$ are fixed and such that $\{\boldsymbol{u}_{kN}\}$ is a sequence of $(L+3)$-dimensional independent random vectors with bounded eighth moments.*

A2. *Assume $\{\mathcal{F}_N, p_N(\cdot)\}$ is a sequence of populations and designs such that for any $\boldsymbol{u}$ with bounded fourth moments the Horvitz-Thompson estimator of its mean for a complete sample satisfies a central limit theorem:*

$$\frac{\sqrt{n_N}}{N}\left(\sum_s d_k \boldsymbol{u}_k - \sum_U \boldsymbol{u}_k\right) \Big| \mathcal{F}_N \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{\Sigma}), \quad a.s.$$

where

$$\boldsymbol{\Sigma} = \lim_{N\to\infty} n_N \sum_U \sum_U \frac{\pi_{kj} - \pi_k \pi_j}{N^2} \frac{\boldsymbol{u}_k \boldsymbol{u}_j^T}{\pi_k \pi_j}$$

is positive definite.

A3. *Assume* $\lim_{N\to\infty} n_N^{-1} \lambda_N = \lambda^*$, *where λ^* is a positive constant.*

A4. *Assume the sampling rate is such that* $\lim_{N\to\infty} N^{-1} n_N = \pi$ *with* $0 < \pi < 1$ *and* $0 < l_1 \le \pi_k \le l_2 < 1$ *for all $k \in U$. In addition, assume for a sample with nonresponse that* $0 < l_3 \le \theta_k \le 1$ *for all $k \in U$.*

A5. *Assume for a sample with nonresponse that there exists a vector $\boldsymbol{d}$ such that* $\boldsymbol{z}_k^T \boldsymbol{d} = \tilde{\boldsymbol{z}}_k^T \boldsymbol{d}_1 + \boldsymbol{z}_k^{*T} \boldsymbol{d}_2 = \theta_k^{-1}$, *with* $\tilde{\boldsymbol{z}}_k = (1, z_k)^T$ *and* $\boldsymbol{z}_k^* = ((z_k - \kappa_1)_+, \ldots, (z_k - \kappa_L)_+)^T$ *and such that* $\boldsymbol{d}_2 = O(n_N^{-1})$.

A6. *Assume for a sample with nonresponse that responses are independent, i.e.,* $P(\delta_k = 1 \,\&\, \delta_j = 1 | I_k = I_j = 1) = \theta_{kj} = \theta_k \theta_j$.

A7. *Assume that the Horvitz-Thompson estimator of the variance of the mean of any variable with finite fourth moments for a complete sample has a variance that is* $O_p\left(n_N^{-3}\right)$ *almost surely*.

Assumption A1 requires that the number and placement of knots is kept fixed over repeated sampling and asymptotically, and together with A2 it allows that the variance of the Horvitz- Thompson estimator of a mean of certain variables for a complete sample has a variance that is $O_p(n^{-1})$. Assumption A3 allows λ_N to grow at the same rate as n_N and, therefore, that λ_N/n_N does not vanish asymptotically. Assumptions A4 and A6 concern the nonresponse mechanism, while Assumption A5 has a particular relevance. It plays a similar role to the sufficient condition (*iii*) considered at the end of Section 2 for the regression estimator. It is the main condition for consistency that makes the mean of population residuals, similarly to the $a_k s$ of Section 2, vanish asymptotically. Note that in this case the inverse of the response probabilities is allowed to be a linear combination of both the linear and the spline part of the auxiliary vector $\boldsymbol{z}$, where the coefficients for the spline part are required to decrease with n to make shrinkage reasonable also for large samples. This also implies that the part of population mean of the residuals associated with the spline part is decreasing by the assumption on $\boldsymbol{d}_2$. Finally, Assumption A7 ensures that the Horvitz-Thompson estimator of the variance of a mean is a consistent estimator.

Theorem 3.1

Assume A1–A6. Then, estimator $\hat{Y}_{p,r}$ *is design* $\sqrt{n_N}$*-consistent for Y, in the sense that* $\hat{Y}_{p,r} - Y = O_p\left(Nn_N^{-1/2}\right)$. *Furthermore,*

$$\hat{Y}_{p,r} - Y = \sum_r \frac{e_k}{\pi_{2k}} + O_p\left(Nn_N^{-1}\right) \quad a.s. for\ info-U \tag{11}$$

$$\hat{Y}_{p,r} - Y = \sum_r \frac{e_k}{\pi_{2k}} - \sum_s \frac{e_k}{\pi_k} + \sum_s \frac{y_k}{\pi_k} - \sum_U y_k + O_p\left(Nn_N^{-1}\right) \quad a.s.\ for\ info-s \tag{12}$$

where $\pi_{2k} = \pi_k\theta_k = \theta_k/d_k$, $e_k = y_k - \boldsymbol{z}_k^T\boldsymbol{\gamma}_U$ *and* $\boldsymbol{\gamma}_U = \left(\sum_U \theta_k \boldsymbol{z}_k\boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N\right)^{-1}\sum_U \theta_k\boldsymbol{z}_k y_k$, *with* $\boldsymbol{\Lambda}_N = diag\{0, 0, \lambda_N, \ldots, \lambda_N\}$. *In addition*,

$$\left\{V_\infty\left(\hat{Y}_{p,r}\right)\right\}^{-1/2}\left(\hat{Y}_{p,r} - Y\right)|\mathcal{F}_N \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad a.s. \tag{13}$$

where

$$V_\infty\left(\hat{Y}_{p,r}\right) = \sum_U\sum_U(\pi_{2kj} - \pi_{2k}\pi_{2j})\frac{e_k}{\pi_{2k}}\frac{e_j}{\pi_{2j}}, \quad for\ info-U \tag{14}$$

$$V_\infty\left(\hat{Y}_{p,r}\right) = \sum_U\sum_U(\pi_{2kj} - \pi_{2k}\pi_{2j})\frac{y_k}{\pi_{2k}}\frac{y_j}{\pi_{2j}} + \sum_U \frac{e_k^2 - y_k^2}{\pi_{2k}}(1 - \theta_k), \quad for\ info-s \tag{15}$$

with $\pi_{2kj} = \pi_{kj}\theta_k\theta_j$.

Proof. See the Appendix. ■

Theorem 3.2.
Assume A1–A7. Then,

$$\hat{V}(\hat{Y}_{p,r}) = \sum_r \sum_r \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} \hat{w}_k \hat{e}_k \hat{w}_j \hat{e}_j + \sum_r (\pi_k - 1/\hat{w}_k) \hat{w}_k^2 \hat{e}_k^2$$

$$= V_\infty(\hat{Y}_{p,r}) + O_p\left(N^2 n_N^{-3/2}\right), \quad a.s. for\ info-U$$

$$\hat{V}(\hat{Y}_{p,r}) = \sum_r \sum_r \frac{\pi_{kj} - \pi_k \pi_j}{\pi_{kj}} \hat{w}_k y_k \hat{w}_j y_j + \sum_r (1 - 1/\hat{w}_k \pi_k) \hat{w}_k^2 \hat{e}_k^2$$

$$- \sum_r (1 - 1/\hat{w}_k \pi_k)(1 - \pi_k) \hat{w}_k^2 y_k^2 = V_\infty(\hat{Y}_{p,r}) + O_p\left(N^2 n_N^{-3/2}\right),$$

$$a.s. for\ info-s$$

where

$$\hat{w}_k = \sum_U \boldsymbol{z}_k^T \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \boldsymbol{z}_k d_k \ \ for\ info-U, \tag{16}$$

$$\hat{w}_k = \sum_s d_k \boldsymbol{z}_k^T \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \boldsymbol{z}_k d_k \ \ for\ info-s, \tag{17}$$

$$\hat{e}_k = y_k - \boldsymbol{z}_k^T \hat{\boldsymbol{\beta}}_r, \text{ and } w_k = \sum_U \boldsymbol{z}_k^T \left(\sum_U \theta_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \boldsymbol{z}_k d_k.$$

Proof. See the Appendix. ■

Theorem 3.1 proves consistency of $\hat{Y}_{p,r}$ when λ is allowed to grow as n. On the other hand, as pointed out for Case A, if λ remains constant or goes to zero as $n_N \to \infty$, then the penalized coefficient vector converges to the unpenalized one and provides an estimator that is unbiased with respect to model (6) with mean function approximated by (7). Therefore, the model with respect to which $\hat{Y}_{p,r}$ is model-unbiased under assumption A3 is not model (6) with mean function given by (7), but a model in which it is reasonable to shrink the coefficients of the spline part, even in large samples. Then, exploiting the relationship between penalized splines and the mixed effects model (e.g., Ruppert et al. 2003, Sec. 4.9), we would restate condition (i) at the end of Section 2 as follows: the probability limit of $N^{-1} \sum_U e_k$ is zero when the sequence of finite populations is a sequence of random samples from an infinite population in which the linear mixed model

$$y_k = \tilde{\boldsymbol{z}}_k^T \boldsymbol{\beta}_1 + \boldsymbol{z}_k^{*T} \boldsymbol{\beta}_2 + \epsilon_k$$

holds, where the ϵ_ks and $\boldsymbol{\beta}_2$ are uncorrelated random variables such that $E(\epsilon_k | \boldsymbol{z}_k) = 0$, $E(\boldsymbol{\beta}_2 | \boldsymbol{z}) = \boldsymbol{0}$. In other words, an alternative sufficient condition for $\hat{Y}_{p,r}$ to be design-consistent is that the finite population is a random sample from an infinite superpopulation mixed-effects model in which $\tilde{\boldsymbol{z}} = (1, z)^{\mathrm{T}}$ is the fixed component of the model and $\boldsymbol{z}^* = ((z - \kappa_1), \ldots, (z - \kappa_L))^{\mathrm{T}}$ is the random component. Park and Fuller (2009) study the properties of the regression estimator based on a mixed effects model in the case of full response.

Note that the double protection property considered in Section 2 holds here too. In particular, then, if either the conditions in the Theorems stated above – and in particular Assumption A5 that is the counterpart of the corresponding condition (*iii*) of Section 2 – or the aforementioned mixed-effect model holds, then the proposed estimator is design-consistent. We have explicitly worked out the asymptotic properties under the former, because most large surveys involve many y-variables, and to achieve a low bias the mixed-effects model has to hold for all of them. We argue that, given its flexibility, this would be much more frequent than with a simple linear regression model, but it could also not be the case. We have then focused instead on the modeling for the response distribution. Condition A5, in fact, is sufficient to have the population mean of the residuals e_k vanish asymptotically. Note that this is similar to assuming mixed-effect model for the inverse of the response probabilities θ_k, and then having more flexible modeling for it. This comment builds a bridge to most nonresponse literature in which such a condition would be comparable with a case in which data is missing at random (MAR). This latter situation arises when the response probability θ_k depends on $\boldsymbol{z}_k$ but not on y_k; then nonresponse depends only on observed values and can be successfully modeled. Now, if probability θ_k depends on $\boldsymbol{z}_k$, then it cannot be independent of y_k given that usually $\boldsymbol{z}_k$ and y_k are related themselves, however, A5 tells us that if data is MAR conditional on $\boldsymbol{z}_k$, then the proposed estimator is design-consistent. Evidence of this implicit modeling for θ_k emerges from the simulation studies of Section 4.

Theorem 3.2 provides variance estimators for $\hat{Y}_{p,r}$ under both the info-s and the info-U settings. Such estimators follow closely the proposal in Särndal and Lundström (2005, Ch. 11) where variance estimation for the calibration approach is derived using the connection with a two phase design. See also Fuller (2009, Ch. 5) for the variance estimator under the info-U setting for the regression estimator. Note that also for variance estimation in (16) and (17) θ_k^{-1} is replaced by its proxy values v_{Uk} and v_{sk} respectively.

3.4. Selection of λ

The properties of the proposed estimator have been provided when λ is decided in advance and kept fixed over repeated sampling. As we saw in Section 3.2, in this context λ has a double interpretation. From a calibration perspective, it can be considered as the quantity that governs the amount of relaxing of the constraints on the L truncated linear variables and, therefore, the shrinking of the final set of weights (see Rao and Singh 1997; Fuller 2002; Beaumont and Bocci 2008, for different ways of selecting the amount of relaxing). From a smoothing perspective, as noted earlier, it provides the degree of smoothness of the final function fit. To determine the optimal value of λ for a particular variable of interest, Breidt et al. (2005) exploit the fact that penalized splines can be seen as mixed effect models, and use for λ the ratio between the estimates of the variances of the two random components (the spline and the error) obtained via restricted maximum likelihood. We will not look at this latter interpretation to select its value, but will look at an alternative way to try to find a compromise value for a set of different y-variables, instead of an optimal one for a single y-variable.

In particular, let $\hat{\boldsymbol{m}}_U = (\hat{m}_1, \ldots, \hat{m}_k, \ldots, \hat{m}_N)^T$ denote the vector of predictions that use $\boldsymbol{\beta}_U$ in Equation (7), that is, for which $\hat{m}_k = m(\boldsymbol{z}_k; \boldsymbol{\beta}_U) = \boldsymbol{z}_k^T \boldsymbol{\beta}_U$. Now

$$\hat{\boldsymbol{m}}_U = \boldsymbol{Z}_U \left(\boldsymbol{Z}_U^T \boldsymbol{Z}_U + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{Z}_U^T \boldsymbol{y}_U =: \boldsymbol{S}_U \boldsymbol{y}_U \tag{18}$$

where $\boldsymbol{Z}_U$ is the $N \times (L+2)$ matrix with $\boldsymbol{z}_k$ on its kth row and $\boldsymbol{y}_U$ is the vector of population y_k values. The degrees of freedom used to approximate the relationship between y and z can be computed as the trace of the smoother matrix $\boldsymbol{S}_U$. In particular

$$df(\lambda_U) = \text{trace}\{\boldsymbol{S}_U\} = \text{trace}\left\{ \left(\sum_U \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \sum_U \boldsymbol{z}_k \boldsymbol{z}_k^T \right\} \tag{19}$$

We can see that increasing values of λ provide a decreasing number of degrees of freedom. Therefore, a value for λ_U can be chosen by fixing in advance the number of degrees of freedom, i.e., λ_U defined through $df(\lambda_U) = d^*$, where the number of degrees of freedom d^* should not be either too few in order to be able to capture a complex relationship, nor too many so that overfitting may be an issue. This quantity does not depend on y and represents a compromise that accounts for the multipurpose aim of a survey. In the simulation studies in Section 4 we investigate the performance of the proposed estimator for a wide range of values of d^*. Note that, since it depends on population quantities, it can be computed only when the auxiliary information available is such that the population totals involved in (19) are known. In addition, once it is computed, it is a fixed quantity over repeated sampling and theoretical results in Section 3.3 apply.

When we are in an info-s setting, we can consider the vector $\hat{\boldsymbol{m}}_s = (\hat{m}_1, \ldots, \hat{m}_k, \ldots, \hat{m}_n)^{\mathrm{T}}$ of predictions based on $\boldsymbol{\beta}_s$. In particular, in this case

$$\hat{\boldsymbol{m}}_s = \boldsymbol{Z}_s \left(\boldsymbol{Z}_s^T \boldsymbol{D}_s \boldsymbol{Z}_s + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{Z}_s^T \boldsymbol{D}_s \boldsymbol{y}_s =: \boldsymbol{S}_s \boldsymbol{y}_s$$

where subscript s denotes sample versions of matrices and vectors used in (18) and $\boldsymbol{D}_s = \text{diag}\{d_k\}_{k \in s}$. In this case the aforementioned rule of thumb can be applied to

$$df(\lambda_s) = \text{trace}\{\boldsymbol{S}_s\} = \text{trace}\left\{ \left(\sum_s d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \sum_s d_k \boldsymbol{z}_k \boldsymbol{z}_k^T \right\}.$$

In this case, the value of λ_s defined through $df(\lambda_s) = d^*$ changes with the sample selected. However, consistency of the estimator still holds since, under the regularity conditions considered in Section 3.3, it is straightforward to show that for a given d^*, $\lambda_s = \lambda_U + O_p\left(n_N^{-1/2}\right)$ (see e.g. the technique proposed in Wu and Sitter 2001).

Finally, in both information settings, a value for λ may also be determined only looking at respondents. In particular, if subscript r denotes matrices and vectors that include only respondent information,

$$\hat{\boldsymbol{m}}_r = \boldsymbol{Z}_r \left(\boldsymbol{Z}_r^T \boldsymbol{D}_r \boldsymbol{Z}_r + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{Z}_r^T \boldsymbol{D}_r \boldsymbol{y}_r =: \boldsymbol{S}_r \boldsymbol{y}_r$$

and

$$df(\lambda_r) = \text{trace}\{\boldsymbol{S}_r\} = \text{trace}\left\{ \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda} \right)^{-1} \sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T \right\}.$$

In this case, for a given d^*, λ_r converges in probability to a different quantity than λ_U because of nonresponse. In particular, let $\lambda_{\theta U}$ be such that

$$df(\lambda_{\theta U}) = \text{trace}\left\{\left(\sum_U \theta_k z_k z_k^T + \mathbf{\Lambda}\right)^{-1} \sum_r \theta_k z_k z_k^T\right\} = d^*,$$

then $\lambda_r = \lambda_{\theta U} + O_p\left(n_N^{-1/2}\right)$. Simulation studies in Section 4 explore the behavior of $\hat{Y}_{p,r}$ for different values of d^* and choices of λ.

As far as variance estimation is concerned, while for λ_U the result in Theorem 3.2 holds, for λ_s and λ_r the variance estimator proposed does not account for the extra variability introduced with estimation of λ.

3.5. Moving to Multivariate Auxiliary Information: Semiparametric Modeling

Multivariate auxiliary information can be easily considered in $\hat{Y}_{p,r}$. In fact, additional auxiliary variables – both categorical and continuous – can be inserted parametrically by adding them to the binding part of the calibration procedure; namely, they will be part of the set of auxiliary variables for which the calibration constraints are met exactly. Additional continuous variables can be added nonparametrically by adding the linear part to the binding part of the calibration procedure, and another set of relaxed constraints with a different penalty on the nonbinding one. In particular, assume that we want to insert the vector $\boldsymbol{x}$ of p variables parametrically and the variables z_1 and z_2 nonparametrically. The v_k weights of the proposed estimator can be then written in these cases as

$$\textbf{info}-\textbf{s}: v_{sk} = 1 + \left(\sum_s d_k \tilde{\boldsymbol{x}}_k - \sum_r d_k \tilde{\boldsymbol{x}}_k\right)^T \left(\sum_r d_k \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^T + \tilde{\mathbf{\Lambda}}\right)^{-1} \tilde{\boldsymbol{x}}_k,$$

$$\textbf{info}-\textbf{U}: v_{Uk} = 1 + \left(\sum_U \tilde{\boldsymbol{x}}_k - \sum_r d_k \tilde{\boldsymbol{x}}_k\right)^T \left(\sum_r d_k \tilde{\boldsymbol{x}}_k \tilde{\boldsymbol{x}}_k^T + \tilde{\mathbf{\Lambda}}\right)^{-1} \tilde{\boldsymbol{x}}_k,$$

where $\tilde{\boldsymbol{x}}_k = \left(1, \boldsymbol{x}_k^T, \boldsymbol{z}_{1k}^T, \boldsymbol{z}_{2k}^T\right)^T$, $\boldsymbol{z}_{ik} = (z_{ik}, (z_{ik} - \kappa_{i1})_+, \ldots, (z_{ik} - \kappa_{iL_i})_+)^T$ for $i = 1, 2$ with L_1 and L_2 number of knots for z_1 and z_2, respectively. In addition, $\tilde{\mathbf{\Lambda}} = \{\mathbf{0}_{p+1}, 0, \lambda_1, \ldots, \lambda_1, 0, \lambda_2, \ldots, \lambda_2\}$, with $p + 1$ zeroes on the diagonal – intercept and $\boldsymbol{x}$ variables – followed by a zero and L_1 penalty constants λ_1, and by a zero and L_2 penalty constants λ_2.

Extension to bivariate smoothing is also possible – although not pursued here – by using a different set of basis functions than truncated linear, such as radial basis functions. Smoothing in two dimensions is particularly relevant when auxiliary information comes in the form of geographic coordinates. For more details see Ruppert et al. (2003, Ch. 11).

4. Simulation Studies

In this section, results from a simulation study that aims at investigating the finite sample behavior of the proposed estimator are presented. In particular, we wish to explore the double protection provided by the proposed estimator with respect to the description of the relationship between y and z and of that between θ and z. To this end we consider different

relationships and different combinations of such relationships. Firstly, values of a finite population of $N = 5{,}000$ units are generated for an auxiliary variable z from a uniform [0,1] distribution. Then, six survey variables are obtained by using the following three regression functions:

LIN : $m\{z) = 0.8 + 3z;$

SIN : $m(z) = 1.8 + 1.5z\sin[4\pi(z - 0.6)];$

DIS : $m(z) = (0.8 - 1.5z)I(z < 0.25) + (0.8 + 2z)I(0.25 < z < 0.50)$

$$+ (-1.7 + 5z)I(0.50 < z < 0.75) + (2.8 - 3z)I(z > 0.75).$$

Units are then randomly divided into two strata of equal dimension 2,500, to simulate stratification on a variable different from z. Then, a constant value of 0.3 is added to $m(z)$ only for units in the first of the two strata. Then, the survey variables are constructed by adding to $m(z_k)$ for $k = 1, \ldots, 5{,}000$ a heteroskedastic error component of the form $2\sqrt{z_k}\varepsilon_k$, where $\varepsilon_k \sim \mathcal{N}(0, \sigma)$ and σ is set to 0.15 for a first set of three survey variables, and to 0.50 for a second set.

Figure 2 shows the scatter plots of the six survey variables thus obtained. Grey crosses and black circles distinguish units belonging to different strata. The LIN populations (the first column) are considered as cases in which a calibration estimator that uses $\{1, z\}$ as auxiliary variables should provide a good protection against nonresponse bias. The SIN populations (the second column) provide a situation in which the aforementioned vector of auxiliary variables is not sufficiently adequate and for which gains in bias reduction are expected from the proposed splines estimator. Finally, the DIS populations (the third column) are generated under a discontinuous function of z, for which the spline estimator is also based on a misspecified model.

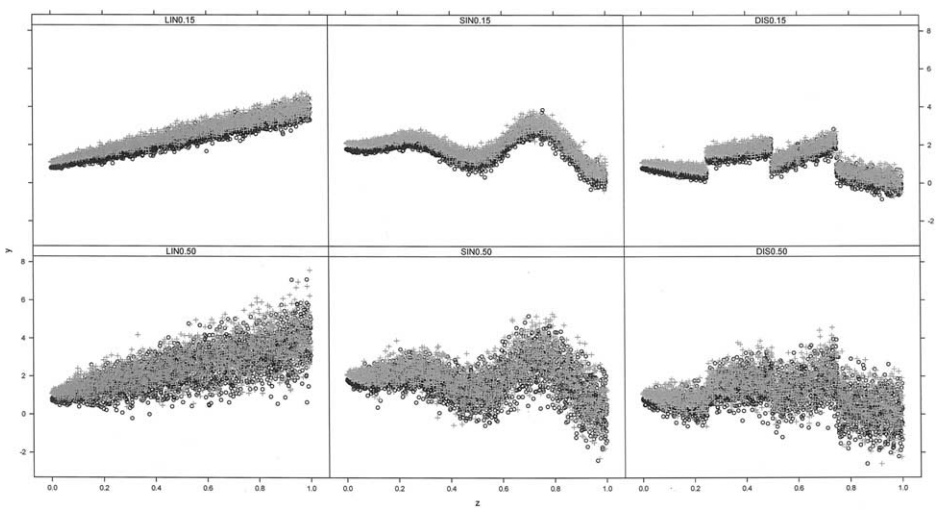

Fig. 2. Scatter plot of the six survey variables versus the auxiliary variable. Variables in the first row have errors with standard deviation 0.15, while those on the second row have errors with standard deviation 0.50. Grey crosses and black circles denote units belonging to the two different strata.

Each unit in the population has its own response probability attached. To study under which circumstances the proposed estimator provides more protection against nonresponse bias with respect to the classical calibration estimator, we will consider different relationships between z_k and θ_k. In particular, what is relevant here is the relationship between z_k and $1/\theta_k$ as considered in Section 2 and 3.3. For this reason we have considered the following four cases:

LIN : $\theta_k = 1/(1.2 + z_k)$;

LOG+ : $\theta_k = 0.3 + 0.5/[1 + \exp(6 - 15z_k)]$;

LOG− : $\theta_k = 0.3 + 0.5/[1 + \exp(-6 + 10z_k)]$;

GAU : $\theta_k = 0.5/\exp[-(z_k - 0.5)^2/0.4]$.

The response rate is approximately 60% in all cases. Figure 3 depicts these four sets of θ_ks, together with $1/\theta_k$. Different levels of complexity of the relationship between $1/\theta_k$ and the auxiliary variable allow to investigate in which situations the double protection property of the calibration estimators holds with respect to the proposed spline estimator. For example, the GAU case is inspired by the kernel of a Gaussian distribution and θ_k takes a U-shape.

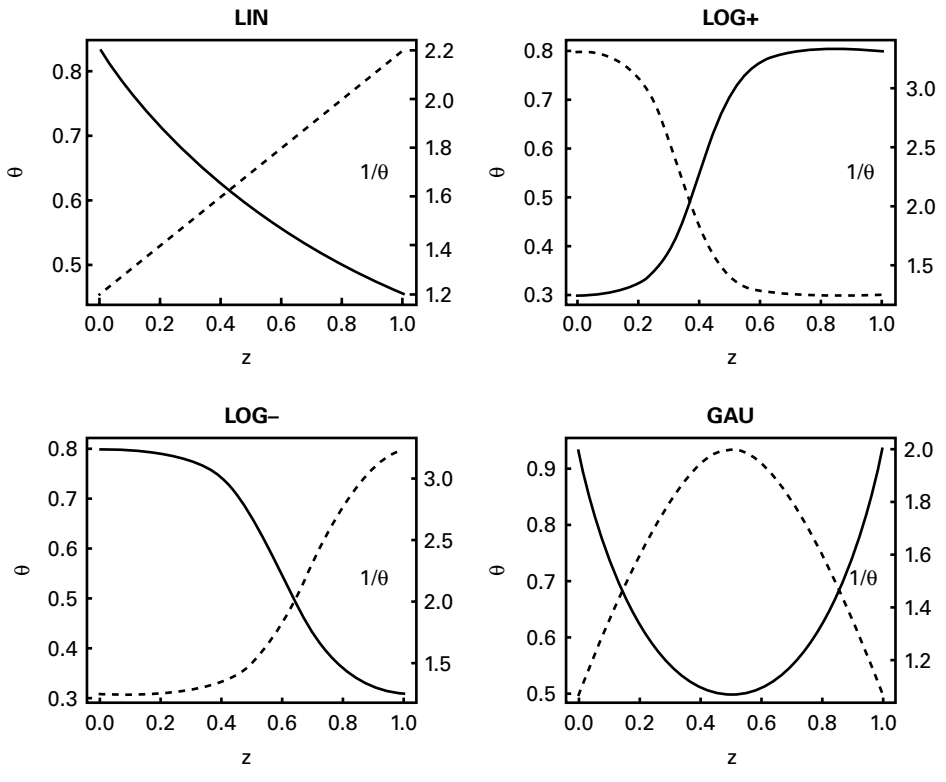

Fig. 3. Scatter plot of the four sets of response probabilities versus the auxiliary variable (black). The dashed line plots the inverse of the response probabilities versus z.

From each population, $J = 1{,}000$ stratified random samples of dimensions $n = 250$, $n = 500$ and $n = 800$ have been selected. Disproportionate allocation is considered so that 40% of the sample comes from the first stratum and the remaining 60% comes from the second. Recall that the first stratum is the one with the increased values. For all survey variables this makes a 3×4 design for the simulation – 3 sample sizes by 4 types of response probabilities. For each unit in the sample, a Bernoulli experiment with probability of success given by its response probability is conducted to simulate the response mechanism.

On the response set the following estimators of the total of each survey variable have been computed:

- $\text{exp} = N\bar{y}_r$, the expansion estimator where $\bar{y}_r = \sum_r d_k y_k / \sum_r d_k$, no auxiliary information used;
- $\text{pwa} = \sum_{p=1}^{P} N_p \bar{y}_{r_p}$ population weighting adjustment, poststratified estimator where $P = 3$ poststrata are defined using the 0.33 and 0.66 quantiles of z and $\bar{y}_{r_p} = \sum_{r_p} d_k y_k / \sum_{r_p} d_k$, with r_p the respondents set in poststratum p;
- $\text{wc} = \sum_{p=1}^{P} \hat{N}_p \bar{y}_{r_p}$ weighting class estimator, poststratified estimator with estimated population counts $\hat{N}_p = \sum_{s_p} d_k$, with s_p the sample set in poststratum p;
- $\text{ra} = \sum_U z_k \sum_r d_k y_k / \sum_r d_k z_k$, the ratio estimator;
- $\text{reg} = \text{exp} + \left(\sum_U z_k - \text{exp}_z\right) b$, the regression estimator with $\text{exp}_z = N \sum_r d_k z_k / \sum_r d_k$ and $b = \sum_r d_k (z_k - \bar{z}_r)(y_k - \bar{y}_r) / \sum_r d_k (z_k - \bar{z}_r)^2$;
- $\text{reg2} = \text{exp} + \left(\sum_U z_k - \text{exp}_z\right) b_1 + \left(\sum_U z_k^2 - \text{exp}_{z^2}\right) b_2$, the quadratic regression estimator;
- $\text{reg3} = \text{exp} + \left(\sum_U z_k - \text{exp}_z\right) b_1 + \left(\sum_U z_k^2 - \text{exp}_{z^2}\right) b_2 + \left(\sum_U z_k^3 - \text{exp}_{z^3}\right) b_3$, the cubic regression estimator;
- $\text{sepra} = \sum_{p=1}^{P} \sum_{U_p} z_k \bar{y}_{r_p} / \bar{z}_{r_p}$, the separate ratio estimator (the three poststrata in pwa are used);
- $\text{sepreg} = \sum_{p=1}^{P} N_p \left\{ \bar{y}_{r_p} + \left(\sum_{U_p} z_k - \bar{z}_{r_p} \right) b_p \right\}$, the separate regression estimator with $b_p = \sum_{r_p} d_k (z_k - \bar{z}_{r_p})(y_k - \bar{y}_{r_p}) / \sum_{r_p} d_k (z_k - \bar{z}_{r_p})^2$ (the three post-strata in pwa are used);
- spline*df*, eight different p-splines estimators according to the value of the degrees of freedom used to approximate all survey variables; in particular, λ is chosen so that $df = \{3,4,6,8,10,12,14,16\}$.

Estimators ra, reg, reg2, reg3, sepra, sepreg, and all the spline estimators are computed in the estimators info-U and info-s scenario. For the latter case, an extra 's' will be attached to the name of the estimator. In addition, for the spline estimators the value of λ has been determined in two different ways for info-U and for info-s. In particular, for info-U λ is determined (*i*) at the population level – using values of z_k for $k \in U$ – and kept fixed over repeated sampling and (*ii*) for each sample, at the response set level – using values of z_k for $k \in r$. Similarly, for info-s λ is determined for each sample (*i*) at the sample level – using values of z_k for $k \in s$ – and (*ii*) at the response set level – using values of z_k for $k \in r$. Estimators with λ determined as in *(ii)* for either info-s or info-U will be denoted with an extra 'r' in the name of the estimator. So, for example, spline4 denotes the estimator that also uses 4 degrees of freedom, auxiliary information of type info-U and λ determined at the population level and then kept fixed over repeated sampling; while

spline4rs denotes the estimator that also uses 4 degrees of freedom, but auxiliary information of type info-s and λ computed at each replication at a response set level. The spline-based estimators all use $L = 35$ knots, placed at the quantiles of population values of z and kept fixed over repeated sampling. Note that the choice of the position of the knots is not as crucial as the choice of the position of thresholds for poststrata, once penalization is included in the estimation procedure.

The performance of the estimators is evaluated for each survey variable using the following measures in which $\hat{Y}_j$ denotes the value taken by a generic estimator $\hat{Y}$ of Y at replication j, with $j = 1, \ldots, J$.

- % Relative Bias, given by

$$\%\mathrm{RB} = \frac{\hat{B}(\hat{Y})}{Y} 100$$

  where $\hat{B}(\hat{Y}) = \hat{E}(\hat{Y}) - Y$ is the Monte Carlo estimate of the bias with $\hat{E}(\hat{Y}) = J^{-1} \sum_{j=1}^{J} \hat{Y}_j$;

- % Coefficient of Variation, given by

$$\%\mathrm{CV} = \frac{\sqrt{\widehat{MSE}(\hat{Y})}}{Y} 100$$

where the Monte Carlo estimate of the mean squared error is given by $\widehat{MSE}(\hat{Y}) = J^{-1} \sum_{j=1}^{J} (\hat{Y}_j - Y)^2$.

In addition, the performance of the variance estimators for the proposed estimator illustrated in Theorem 3.2 has also been tested by means of the empirical coverage rate for a 95% nominal confidence interval based on the normal approximation. Note that estimators from exp to sepregs are "conventional" and also considered in Särndal and Lundström (2005). We will see that results are in line with those in for instance, Särndal and Lundström (2005, Sec. 10.3).

We will report results only for $n = 500$ and then discuss the differences occurring when considering a smaller or a larger sample size. Tables 1 and 2 report the % Relative Bias in the different settings. Estimators that use info-U are displayed in the first half of the tables. In general, it can be noted that for the same estimator, info-s shows the same performance as info-U in terms of bias. Under the columns with the heading θ LIN in Table 1 we report results when the reciprocal of the nonresponse probabilities is a linear function of the auxiliary variable. This is a situation in which condition (5) holds when the auxiliary vector contains an intercept and the values of z_k. This is the case for all reg estimators – reg, reg2, reg3 – that, in fact, show an almost zero bias also for any population of interest. This is also true for the sepreg and all the spline estimators even if they are using a more complicated set of auxiliary variables than needed. Poststratification corresponds to a piecewise constant approximation to the linear function that provides some reduction in bias compared to exp, but not as well as the others. Estimators ra and sepra use an auxiliary information vector which suffices approximate neither the nonresponse model nor the population model, and in most cases show a larger bias than does exp.

When the inverse of the response probability is a more complicated function of z, as for the case θ LOG+ in Table 1 and θ LOG− and θ GAU in Table 2, then the reg estimator

Table 1. Percent Relative Bias – %RB – for all estimators and survey variables. Response probabilities type LIN and LOG+ , n = 500

	θ LIN						θ LOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
exp	−5.99	1.04	1.42	−5.96	1.09	1.28	11.19	−1.54	−0.45	11.16	−1.50	−0.49
pwa	−0.70	0.53	0.77	−0.66	0.65	0.71	0.98	0.23	0.43	0.97	0.17	0.41
ra	4.27	12.17	12.60	4.31	12.23	12.45	−6.02	−16.74	−15.80	−6.05	−16.70	−15.83
reg	−0.02	0.05	0.15	0.03	0.13	0.11	−0.02	0.98	8.58	−0.18	1.01	8.35
reg2	−0.03	0.08	−0.01	0.02	0.17	−0.05	−0.01	−1.33	0.58	−0.01	−1.38	0.51
reg3	−0.03	0.06	−0.08	0.02	0.15	−0.13	−0.02	−1.82	0.33	0.00	−1.89	0.27
sepra	0.75	2.66	2.72	0.79	2.77	2.65	−1.03	−2.46	−2.51	−1.05	−2.52	−2.52
sepreg	−0.04	−0.04	−0.17	0.01	0.05	−0.19	−0.03	−0.52	−0.07	−0.02	−0.58	−0.07
spline3	−0.03	0.07	0.08	0.02	0.16	0.04	−0.02	−0.08	4.21	−0.09	−0.09	4.07
spline3r	−0.03	0.07	0.04	0.02	0.15	0.00	−0.02	−0.36	2.98	−0.07	−0.39	2.86
spline4	−0.03	0.06	0.00	0.02	0.15	−0.04	−0.02	−0.68	1.45	−0.05	−0.72	1.35
spline4r	−0.03	0.05	−0.02	0.02	0.14	−0.06	−0.02	−0.73	0.90	−0.04	−0.77	0.81
spline6	−0.04	0.01	−0.05	0.01	0.10	−0.08	−0.02	−0.51	0.27	−0.04	−0.55	0.20
spline6r	−0.04	−0.01	−0.05	0.01	0.08	−0.09	−0.02	−0.38	0.17	−0.05	−0.41	0.10
spline8	−0.05	−0.03	−0.05	0.00	0.06	−0.09	−0.03	−0.25	0.08	−0.05	−0.27	0.01
spline8r	−0.05	−0.04	−0.05	0.00	0.04	−0.09	−0.03	−0.17	0.03	−0.05	−0.19	−0.04
spline10	−0.05	−0.05	−0.06	0.00	0.03	−0.09	−0.03	−0.14	0.00	−0.05	−0.16	−0.07
spline10r	−0.06	−0.05	−0.06	−0.01	0.02	−0.09	−0.04	−0.12	−0.04	−0.06	−0.13	−0.12
spline12	−0.06	−0.06	−0.07	−0.01	0.02	−0.10	−0.04	−0.11	−0.05	−0.06	−0.13	−0.13
spline12r	−0.07	−0.06	−0.08	−0.01	0.01	−0.11	−0.05	−0.10	−0.09	−0.06	−0.12	−0.18
spline14	−0.07	−0.07	−0.08	−0.01	0.01	−0.11	−0.05	−0.10	−0.10	−0.06	−0.12	−0.18
spline14r	−0.07	−0.07	−0.10	−0.02	0.00	−0.14	−0.06	−0.10	−0.14	−0.07	−0.12	−0.22
spline16	−0.07	−0.07	−0.10	−0.02	−0.01	−0.14	−0.06	−0.10	−0.13	−0.07	−0.12	−0.22
spline16r	−0.08	−0.08	−0.12	−0.03	−0.02	−0.16	−0.07	−0.11	−0.18	−0.07	−0.13	−0.26

Table 1. Continued

	θ LIN						θLOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
wc	−0.75	0.54	0.81	−0.71	0.65	0.74	0.94	0.24	0.46	0.92	0.17	0.44
ras	4.16	12.00	12.43	4.19	12.06	12.27	−6.11	−16.84	−15.92	−6.14	−16.81	−15.95
regs	−0.07	0.05	0.10	−0.02	0.13	0.06	−0.07	0.97	8.55	−0.23	0.99	8.31
reg2s	−0.07	0.09	0.12	−0.02	0.17	0.08	−0.06	−1.30	0.73	−0.05	−1.36	0.66
reg3s	−0.07	0.06	0.08	−0.02	0.14	0.03	−0.06	−1.85	0.50	−0.04	−1.93	0.46
sepras	0.71	2.66	2.74	0.75	2.77	2.67	−1.08	−2.47	−2.49	−1.10	−2.53	−2.50
sepregs	−0.08	0.03	0.05	−0.02	0.11	0.02	−0.07	−0.46	0.16	−0.06	−0.52	0.16
spline3s	−0.07	0.07	0.12	−0.02	0.15	0.08	−0.06	−0.08	4.25	−0.14	−0.11	4.10
spline3rs	−0.07	0.07	0.12	−0.02	0.15	0.07	−0.06	−0.37	3.06	−0.12	−0.40	2.94
spline4s	−0.07	0.06	0.11	−0.02	0.15	0.07	−0.06	−0.69	1.55	−0.09	−0.73	1.46
spline4rs	−0.07	0.06	0.11	−0.02	0.14	0.07	−0.06	−0.73	1.02	−0.08	−0.77	0.95
spline6s	−0.08	0.03	0.10	−0.03	0.11	0.06	−0.06	−0.49	0.42	−0.08	−0.53	0.36
spline6rs	−0.08	0.02	0.10	−0.03	0.10	0.06	−0.06	−0.35	0.32	−0.08	−0.38	0.26
splinc8s	−0.08	0.01	0.10	−0.03	0.09	0.06	−0.06	−0.20	0.24	−0.08	−0.23	0.18
spline8rs	−0.08	0.01	0.10	−0.03	0.09	0.06	−0.06	−0.12	0.19	−0.09	−0.15	0.13
spline10s	−0.08	0.01	0.10	−0.03	0.08	0.07	−0.06	−0.09	0.16	−0.09	−0.11	0.10
spline10rs	−0.08	0.00	0.10	−0.04	0.08	0.07	−0.07	−0.06	0.13	−0.09	−0.08	0.06
spline12s	−0.08	0.00	0.11	−0.04	0.08	0.07	−0.07	−0.05	0.11	−0.09	−0.07	0.05
spline12rs	−0.09	0.00	0.10	−0.04	0.07	0.07	−0.07	−0.03	0.09	−0.09	−0.06	0.02
spline14s	−0.09	0.00	0.10	−0.04	0.07	0.07	−0.07	−0.03	0.08	−0.09	−0.06	0.02
spline14rs	−0.09	0.00	0.10	−0.04	0.06	0.06	−0.08	−0.03	0.06	−0.09	−0.06	0.00
spline16s	−0.09	0.00	0.10	−0.04	0.06	0.06	−0.08	−0.03	0.06	−0.09	−0.06	0.00
spline16rs	−0.09	−0.01	0.10	−0.05	0.06	0.06	−0.08	−0.03	0.04	−0.10	−0.06	−0.02

Table 2. Percent Relative Bias – %RB – for all estimators and survey variables. Response probabilities type LOG−, and GAU, n = 500

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
exp	−10.62	0.92	6.48	−10.61	0.89	6.30	0.17	−2.14	−7.97	0.34	−2.14	−7.96
pwa	−1.06	1.31	3.07	−1.04	1.39	3.07	0.01	−3.04	−3.80	0.12	−3.02	−3.88
ra	8.18	22.25	28.97	8.19	22.21	28.75	−0.03	−2.23	−8.05	0.15	−2.23	−8.03
reg	−0.06	−0.77	6.01	−0.05	−0.76	5.97	0.02	−2.07	−7.81	0.19	−2.06	−7.80
reg2	−0.04	−1.23	0.17	0.04	−1.22	0.17	−0.01	0.28	−0.20	0.00	0.36	−0.32
reg3	−0.05	−1.06	0.28	0.02	−1.05	0.25	−0.01	0.23	−0.25	0.00	0.32	−0.37
sepra	0.54	2.81	4.86	0.56	2.90	4.86	1.13	0.20	−1.11	1.24	0.20	−1.21
sepreg	−0.06	−0.41	−0.03	0.02	−0.41	−0.02	−0.02	−0.03	0.00	0.00	0.06	−0.10
spline3	−0.05	−0.81	2.81	−0.01	−0.80	2.78	0.00	−0.89	−3.31	0.08	−0.84	−3.37
spline3r	−0.05	−0.80	1.95	0.00	−0.79	1.92	0.00	−0.69	−2.52	0.07	−0.63	−2.59
spline4	−0.05	−0.73	0.86	0.01	−0.72	0.84	−0.01	−0.29	−1.01	0.03	−0.22	−1.11
spline4r	−0.05	−0.67	0.50	0.01	−0.65	0.47	−0.01	−0.21	−0.68	0.02	−0.13	−0.78
spline6	−0.05	−0.39	0.10	0.00	−0.38	0.07	−0.02	−0.09	−0.19	0.00	−0.01	−0.29
spline6r	−0.05	−0.29	0.06	−0.01	−0.28	0.02	−0.02	−0.09	−0.13	0.00	0.00	−0.23
spline8	−0.06	−0.20	0.03	−0.02	−0.18	−0.01	−0.03	−0.08	−0.08	−0.01	0.01	−0.18
spline8r	−0.06	−0.15	0.02	−0.02	−0.14	−0.02	−0.03	−0.08	−0.07	−0.01	0.01	−0.16
spline10	−0.07	−0.13	0.01	−0.03	−0.12	−0.03	−0.03	−0.08	−0.07	−0.01	0.01	−0.16
spline10r	−0.07	−0.12	0.00	−0.04	−0.10	−0.05	−0.04	−0.08	−0.07	−0.02	0.00	−0.16
spline12	−0.07	−0.11	−0.01	−0.04	−0.10	−0.05	−0.04	−0.09	−0.07	−0.02	0.00	−0.16
spline12r	−0.08	−0.11	−0.03	−0.06	−0.10	−0.07	−0.05	−0.09	−0.08	−0.02	−0.01	−0.18
spline14	−0.08	−0.11	−0.03	−0.06	−0.10	−0.08	−0.05	−0.09	−0.09	−0.02	−0.01	−0.18
spline14r	−0.09	−0.11	−0.05	−0.07	−0.11	−0.11	−0.05	−0.10	−0.10	−0.03	−0.02	−0.20
spline16	−0.09	−0.11	−0.05	−0.07	−0.11	−0.10	−0.05	−0.10	−0.10	−0.03	−0.02	−0.20
spline16r	−0.10	−0.12	−0.08	−0.08	−0.11	−0.14	−0.06	−0.11	−0.12	−0.03	−0.03	−0.22

Table 2. Continued

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
wc	−1.11	1.32	3.10	−1.08	1.41	3.11	−0.04	−3.02	−3.76	0.07	−3.02	−3.85
ras	8.06	22.07	28.78	8.07	22.03	28.57	−0.13	−2.39	−8.21	0.04	−2.39	−8.20
regs	−0.10	−0.76	5.96	−0.10	−0.75	5.91	−0.03	−2.07	−7.84	0.14	−2.07	−7.84
reg2s	−0.09	−1.23	0.30	−0.01	−1.23	0.30	−0.05	0.30	−0.06	−0.04	0.38	−0.17
reg3s	−0.09	−1.06	0.44	−0.02	−1.07	0.42	−0.05	0.24	−0.09	−0.04	0.32	−0.20
sepras	0.49	2.81	4.89	0.51	2.90	4.89	1.08	0.21	−1.08	1.19	0.20	−1.19
sepregs	−0.10	−0.34	0.17	−0.02	−0.33	0.20	−0.06	0.05	0.21	−0.04	0.13	0.12
spline3s	−0.09	−0.81	2.83	−0.05	−0.80	2.81	−0.04	−0.89	−3.24	0.04	−0.84	−3.30
spline3rs	−0.09	−0.81	2.01	−0.04	−0.80	1.99	−0.05	−0.68	−2.44	0.02	−0.63	−2.51
spline4s	−0.09	−0.74	0.96	−0.03	−0.73	0.94	−0.05	−0.28	−0.89	−0.01	−0.21	−0.97
spline4rs	−0.09	−0.66	0.61	−0.03	−0.65	0.59	−0.05	−0.20	−0.55	−0.02	−0.13	−0.64
spline6s	−0.09	−0.37	0.24	−0.04	−0.35	0.22	−0.06	−0.06	−0.04	−0.04	0.02	−0.13
spline6rs	−0.09	−0.26	0.20	−0.04	−0.24	0.17	−0.06	−0.04	0.02	−0.04	0.03	−0.08
spline8s	−0.09	−0.15	0.18	−0.05	−0.13	0.15	−0.06	−0.03	0.07	−0.04	0.05	−0.02
spline8rs	−0.09	−0.10	0.17	−0.06	−0.08	0.14	−0.06	−0.02	0.09	−0.04	0.05	0.00
spline10s	−0.10	−0.08	0.17	−0.06	−0.06	0.13	−0.06	−0.02	0.10	−0.04	0.06	0.01
spline10rs	−0.10	−0.06	0.17	−0.07	−0.04	0.13	−0.06	−0.02	0.10	−0.04	0.06	0.01
spline12s	−0.10	−0.05	0.16	−0.07	−0.03	0.12	−0.07	−0.02	0.10	−0.04	0.06	0.01
spline12rs	−0.10	−0.04	0.16	−0.08	−0.02	0.11	−0.07	−0.02	0.10	−0.04	0.05	0.01
spline14s	−0.10	−0.04	0.16	−0.08	−0.02	0.11	−0.07	−0.02	0.10	−0.04	0.05	0.01
spline14rs	−0.11	−0.04	0.15	−0.09	−0.02	0.10	−0.07	−0.03	0.10	−0.04	0.05	0.01
spline16s	−0.11	−0.04	0.15	−0.09	−0.02	0.10	−0.07	−0.03	0.10	−0.04	0.05	0.01
spline16rs	−0.11	−0.04	0.14	−0.09	−0.02	0.09	−0.07	−0.03	0.10	−0.05	0.05	0.00

successfully reduces bias to almost zero only with a LIN population. For the other populations, reg always suffers from a misspecified response or population model. By contrast, reg2 and reg3, that use, respectively, a quadratic and a cubic model for either the relationship between y and z or between $1/\theta$ and z, allow the reduction of nonresponse bias also in the case of the SIN or DIS populations, when the response probabilities are of type GAU. Note that for info-U reg2 requires the knowledge of the population total of z^2, and reg3 further requires also the population total of z^3. Estimator sepreg succeeds in decreasing bias every time a piecewise linear approximation in each poststratum provides a good description of the relationship between y and z – for instance the DIS cases – or between $1/\theta$ and z – for instance the GAU cases.

On the other hand, the spline estimators almost always succeed in taking the bias to zero because the inclusion of the basis functions allow handling departures from linearity in either the response model or the population model. Note, for instance, that the DIS population is based on a function of z that the spline estimators cannot handle because the function is discontinuous. In these cases also, though, bias is reduced because the implicit estimation of the inverse of the response probabilities allows to handle the LOG and the GAU functions.

The ability of the spline estimators to capture either the response model or the population model depends on the penalty λ and, therefore, on the number of degrees of freedom used. The simulation studies show that it is better to have a relatively larger value for the degrees of freedom: this allows the handling of even complicated structures, like the SIN population or the LOG response models, and does not provide significant losses when in the presence of simple linear structures. In addition, it is hard to detect differences in the performance of the alternative spline estimators, once at least 8 degrees of freedom are used.

Tables 3 and 4 report %CV for the simulations. In these tables, as expected, the difference between info-s and info-U versions of the same estimator are more clear, with the latter providing gains in efficiency over the former when the vector of auxiliary variables employed by the estimator provides a good approximation of the population model. It is the case of reg in the LIN populations, and of spline estimators for LIN and SIN populations. Estimator sepreg, that showed a good performance in decreasing bias, suffers from its coarse approximation of functions like the SIN or the LOG and GAU, by a relatively larger overall error.

As for the role of λ for the spline estimators, again here there is very little difference in performance among estimators with a number of degrees of freedom going from 8 to 16. In addition, virtually no difference can be detected for each spline estimator with a given number of degrees of freedom when λ is chosen at the population (sample) level on the one hand or at a response set level on the other. This provides evidence of little increase in variability due to the estimation of its value at a response set level (see Section 3.4).

In general, simulations with a larger (smaller) sample size show, other things being equal, an increase (decrease) in the role of bias as opposed to that of variance. The spline estimators, as all nonparametric regression techniques, suffer from a reduced number of observations and therefore provide better performances both in terms of %RB and %CV when $n = 800$.

As for the performance of the variance estimators for the proposed estimators, Tables 5 and 6 report coverage rates for 95% confidence intervals for all spline estimators.

Table 3. % Coefficient of variation for all estimators and survey variables. Response probabilities type LIN and LOG+, n = 500

	θ LIN						θ LOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
exp	6.34	2.17	3.52	6.49	2.96	4.60	11.35	2.70	3.63	11.43	3.53	5.02
pwa	1.15	2.20	3.22	2.03	3.15	4.73	1.33	1.71	2.95	1.99	2.67	4.28
ra	4.59	13.14	13.86	4.90	13.38	14.18	6.09	17.00	16.36	6.27	17.08	16.65
reg	0.57	2.19	3.49	1.78	3.13	4.87	0.56	2.06	9.30	1.60	2.90	9.57
reg2	0.57	2.12	2.35	1.79	3.08	4.23	0.56	2.32	2.51	1.56	3.08	4.00
reg3	0.58	1.75	2.32	1.79	2.83	4.24	0.56	2.62	2.27	1.57	3.35	3.88
sepra	1.24	4.19	4.57	2.09	4.80	5.75	1.52	3.97	4.07	2.13	4.45	5.20
sepreg	0.58	1.15	2.16	1.80	2.54	4.15	0.56	1.14	2.14	1.58	2.35	3.75
spline3	0.57	2.06	2.62	1.78	3.04	4.34	0.55	1.67	4.93	1.57	2.61	5.69
spline3r	0.57	1.99	2.47	1.78	2.99	4.27	0.56	1.63	3.80	1.57	2.58	4.82
spline4	0.57	1.75	2.29	1.79	2.83	4.19	0.56	1.55	2.63	1.57	2.54	4.03
spline4r	0.57	1.59	2.24	1.79	2.73	4.18	0.56	1.46	2.35	1.57	2.49	3.88
spline6	0.58	1.10	2.11	1.79	2.47	4.13	0.56	1.06	2.06	1.57	2.28	3.73
spline6r	0.58	0.98	2.06	1.79	2.42	4.10	0.56	0.91	1.96	1.57	2.21	3.68
spline8	0.58	0.86	1.95	1.80	2.38	4.05	0.56	0.79	1.84	1.58	2.17	3.61
spline8r	0.58	0.82	1.88	1.80	2.37	4.02	0.56	0.75	1.75	1.58	2.15	3.57
spline10	0.58	0.80	1.83	1.80	2.37	3.99	0.56	0.73	1.70	1.58	2.15	3.55
spline10r	0.59	0.78	1.77	1.81	2.37	3.96	0.57	0.72	1.64	1.59	2.15	3.52
spline12	0.59	0.78	1.75	1.81	2.37	3.95	0.57	0.72	1.63	1.59	2.15	3.52
spline12r	0.59	0.77	1.70	1.81	2.37	3.94	0.57	0.72	1.59	1.59	2.15	3.51
spline14	0.59	0.77	1.69	1.81	2.37	3.94	0.57	0.72	1.59	1.59	2.15	3.51
spline14r	0.59	0.77	1.66	1.82	2.38	3.93	0.57	0.72	1.56	1.60	2.15	3.50
spline16	0.59	0.77	1.66	1.82	2.38	3.93	0.57	0.72	1.56	1.60	2.15	3.51
spline16r	0.59	0.77	1.63	1.82	2.38	3.93	0.58	0.73	1.53	1.61	2.16	3.51

Table 3. Continued

	θ LIN						θ LOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
wc	1.83	2.25	3.45	2.48	3.17	4.85	1.91	1.78	3.21	2.42	2.71	4.41
ras	4.53	12.55	13.27	4.84	12.80	13.55	6.40	17.03	16.33	6.57	17.11	16.62
regs	1.61	2.20	3.52	2.33	3.14	4.90	1.60	2.05	9.27	2.19	2.90	9.56
reg2s	1.61	2.15	3.04	2.33	3.10	4.59	1.60	2.31	3.18	2.18	3.07	4.34
reg3s	1.61	1.96	3.06	2.33	2.96	4.62	1.60	2.76	3.00	2.17	3.47	4.24
sepras	1.77	3.76	4.52	2.45	4.42	5.66	2.12	3.64	4.10	2.58	4.16	5.16
sepregs	1.61	1.73	3.01	2.32	2.82	4.57	1.60	1.69	2.98	2.18	2.63	4.20
spline3s	1.61	2.13	3.15	2.33	3.08	4.65	1.60	1.77	5.25	2.18	2.68	5.94
spline3rs	1.61	2.09	3.10	2.33	3.05	4.62	1.60	1.80	4.29	2.17	2.69	5.15
spline4s	1.61	1.98	3.03	2.33	2.97	4.58	1.60	1.86	3.34	2.17	2.74	4.44
spline4rs	1.61	1.91	3.02	2.33	2.92	4.58	1.60	1.85	3.12	2.17	2.73	4.29
spline6s	1.61	1.72	2.99	2.33	2.79	4.56	1.60	1.70	2.94	2.18	2.62	4.18
spline6rs	1.61	1.68	2.97	2.33	2.77	4.55	1.60	1.64	2.90	2.18	2.58	4.15
spline8s	1.61	1.64	2.94	2.33	2.75	4.53	1.60	1.60	2.86	2.18	2.55	4.12
spline8rs	1.61	1.63	2.92	2.33	2.74	4.51	1.60	1.59	2.83	2.18	2.54	4.10
spline10s	1.61	1.63	2.90	2.33	2.74	4.49	1.60	1.59	2.81	2.18	2.54	4.08
spline10rs	1.61	1.62	2.89	2.33	2.74	4.48	1.60	1.58	2.80	2.18	2.54	4.07
spline12s	1.61	1.62	2.88	2.33	2.74	4.48	1.60	1.58	2.79	2.18	2.54	4.07
spline12rs	1.61	1.62	2.88	2.33	2.74	4.47	1.60	1.58	2.78	2.18	2.54	4.06
spline14s	1.61	1.62	2.87	2.33	2.74	4.47	1.60	1.58	2.78	2.18	2.54	4.06
spline14rs	1.61	1.62	2.87	2.33	2.74	4.46	1.60	1.58	2.77	2.18	2.54	4.05
spline16s	1.61	1.62	2.86	2.33	2.74	4.46	1.60	1.58	2.77	2.18	2.54	4.05
spline16rs	1.61	1.62	2.86	2.33	2.75	4.46	1.60	1.58	2.76	2.18	2.54	4.05

Table 4. % Coefficient of variation for all estimators and survey variables. Response probabilities type LOG− and GAU, n = 500

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
exp	10.78	1.98	7.20	10.87	2.77	7.56	2.21	2.97	8.66	2.71	3.61	9.17
pwa	1.44	2.74	4.76	2.33	3.78	5.98	0.90	3.59	4.62	1.80	4.12	5.69
ra	8.38	22.85	29.58	8.57	22.92	29.57	1.49	5.04	9.64	2.13	5.52	10.13
reg	0.62	2.48	7.15	1.92	3.50	7.90	0.55	2.86	8.45	1.63	3.49	8.96
reg2	0.65	2.71	2.56	1.98	3.70	4.63	0.56	2.06	2.25	1.66	2.92	3.97
reg3	0.65	2.27	2.53	1.98	3.36	4.67	0.56	1.77	2.20	1.66	2.72	3.97
sepra	1.10	4.45	6.46	2.16	5.24	7.45	1.55	3.19	3.47	2.27	3.83	4.78
sepreg	0.65	1.29	2.32	1.98	2.87	4.61	0.57	1.19	2.04	1.67	2.45	3.88
spline3	0.63	2.46	4.08	1.94	3.50	5.47	0.55	2.05	4.09	1.64	2.87	5.20
spline3r	0.64	2.41	3.37	1.95	3.47	5.02	0.55	1.94	3.40	1.64	2.79	4.70
spline4	0.64	2.19	2.67	1.97	3.32	4.66	0.56	1.64	2.37	1.65	2.61	4.05
spline4r	0.64	1.98	2.50	1.97	3.19	4.60	0.56	1.51	2.22	1.65	2.54	3.97
spline6	0.65	1.35	2.32	1.98	2.83	4.56	0.56	1.06	2.01	1.66	2.31	3.88
spline6r	0.65	1.14	2.26	1.98	2.74	4.54	0.56	0.95	1.95	1.66	2.27	3.85
spline8	0.65	0.98	2.15	1.98	2.68	4.50	0.56	0.82	1.84	1.66	2.23	3.80
spline8r	0.66	0.92	2.06	1.99	2.66	4.47	0.56	0.78	1.78	1.67	2.22	3.77
spline10	0.66	0.89	2.00	1.99	2.66	4.45	0.56	0.76	1.73	1.67	2.22	3.74
spline10r	0.66	0.86	1.93	2.00	2.65	4.43	0.56	0.75	1.68	1.67	2.22	3.72
spline12	0.66	0.86	1.90	2.00	2.65	4.43	0.57	0.75	1.65	1.67	2.22	3.71
spline12r	0.66	0.85	1.84	2.00	2.66	4.41	0.57	0.75	1.61	1.68	2.23	3.69
spline14	0.66	0.85	1.84	2.00	2.66	4.41	0.57	0.75	1.60	1.68	2.23	3.69
spline14r	0.67	0.85	1.80	2.01	2.66	4.41	0.57	0.75	1.57	1.68	2.23	3.68
spline16	0.67	0.85	1.80	2.01	2.66	4.41	0.57	0.75	1.57	1.68	2.23	3.68
spline10r	0.67	0.85	1.76	2.02	2.67	4.41	0.57	0.75	1.54	1.69	2.24	3.68

Table 4. Continued

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
wc	2.01	2.79	4.96	2.71	3.80	6.13	1.72	3.60	4.81	2.31	4.15	5.80
ras	8.30	22.44	29.22	8.48	22.51	29.20	1.57	3.90	9.18	2.17	4.42	9.67
regs	1.63	2.48	7.10	2.42	3.51	7.87	1.59	2.89	8.51	2.20	3.54	9.03
reg2s	1.64	2.72	3.18	2.46	3.72	4.95	1.59	2.10	2.96	2.21	2.95	4.32
reg3s	1.64	2.40	3.20	2.46	3.46	5.02	1.59	1.99	2.94	2.21	2.86	4.32
sepras	1.74	4.09	6.45	2.53	4.90	7.41	1.91	2.49	3.43	2.51	3.24	4.68
sepregs	1.64	1.82	3.09	2.46	3.11	5.00	1.59	1.75	2.90	2.21	2.71	4.28
spline3s	1.63	2.50	4.42	2.44	3.53	5.71	1.59	2.16	4.42	2.20	2.96	5.42
spline3rs	1.63	2.47	3.86	2.44	3.52	5.33	1.59	2.08	3.84	2.20	2.90	4.97
spline4s	1.63	2.34	3.33	2.45	3.43	5.02	1.59	1.92	3.06	2.21	2.79	4.39
spline4rs	1.64	2.21	3.22	2.45	3.34	4.98	1.59	1.86	2.97	2.21	2.76	4.33
spline6s	1.64	1.86	3.13	2.46	3.11	4.96	1.59	1.70	2.89	2.21	2.64	4.27
spline6rs	1.64	1.77	3.10	2.46	3.05	4.96	1.59	1.67	2.87	2.21	2.62	4.26
spline8s	1.64	1.71	3.06	2.46	3.01	4.94	1.59	1.63	2.85	2.21	2.60	4.23
spline8rs	1.64	1.68	3.03	2.46	3.00	4.93	1.59	1.62	2.83	2.21	2.60	4.22
spline10s	1.64	1.67	3.00	2.46	3.00	4.92	1.59	1.61	2.82	2.21	2.60	4.20
spline10rs	1.64	1.67	2.97	2.47	3.00	4.91	1.59	1.61	2.81	2.21	2.60	4.19
spline12s	1.64	1.67	2.96	2.47	3.00	4.91	1.59	1.61	2.81	2.21	2.60	4.19
spline12rs	1.64	1.66	2.94	2.47	3.00	4.90	1.59	1.61	2.80	2.21	2.60	4.18
spline14s	1.64	1.66	2.94	2.47	3.00	4.90	1.59	1.61	2.79	2.21	2.60	4.18
spline14rs	1.64	1.66	2.93	2.47	3.00	4.90	1.59	1.61	2.79	2.21	2.61	4.17
spline16s	1.64	1.66	2.93	2.47	3.00	4.90	1.59	1.61	2.79	2.21	2.61	4.17
spline16rs	1.64	1.67	2.92	2.47	3.00	4.91	1.59	1.61	2.78	2.21	2.61	4.17

Table 5. Coverage rate for 95% confidence intervals for all p-splines based estimators and survey variables. Response probabilities type LIN and LOG+, n = 500

	θ LIN						θ LOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
spline3	94.6	93.8	95.3	95.4	94.0	95.6	95.3	96.1	62.9	94.7	95.1	85.6
spline3r	94.6	93.9	95.0	95.4	93.8	95.6	95.1	95.4	74.6	94.5	94.6	90.1
spline4	94.6	94.6	94.8	95.1	93.9	95.5	94.8	94.2	89.7	94.3	93.8	93.6
spline4r	94.4	94.6	95.0	95.1	94.1	95.5	94.8	92.7	92.9	94.3	93.6	95.0
spline6	94.0	95.3	95.1	94.5	94.4	95.3	94.9	93.2	94.9	94.3	93.4	95.0
spline6r	93.9	95.2	95.1	94.4	94.3	95.1	94.8	94.7	94.5	94.3	93.5	95.2
spline8	93.9	95.5	94.0	94.0	94.1	95.1	94.5	95.3	94.9	94.3	93.4	95.2
spline8r	94.0	95.1	93.6	94.0	94.1	94.7	94.3	95.1	94.6	94.2	94.1	95.1
spline10	94.0	94.8	93.7	93.8	94.0	94.3	94.3	94.9	94.7	94.2	94.0	95.4
spline10r	93.5	95.1	93.9	93.6	93.9	94.3	94.2	95.0	94.5	94.2	93.8	95.8
spline12	93.4	94.6	93.7	93.6	93.6	94.4	94.2	95.1	94.5	94.2	93.9	95.7
spline12r	93.5	94.5	93.1	93.3	93.7	94.1	94.0	95.0	94.7	93.9	93.8	95.7
spline14	93.4	94.4	92.8	93.3	93.6	94.1	94.0	94.8	94.7	93.8	93.8	95.7
spline14r	93.3	93.8	92.4	93.0	93.6	94.1	93.6	95.0	94.8	93.4	93.6	95.5
spline16	93.3	93.8	92.4	93.0	93.6	94.1	93.7	95.0	94.8	93.4	93.6	95.5
spline16r	93.4	93.4	92.2	92.9	93.4	94. lj	93.5	94.6	94.2	93.4	93.1	95.2

Table 5. Continued

	θ LIN						θ LOG+					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
spline3s	94.9	94.8	95.4	94.6	94.3	95.5	94.8	96.1	70.8	94.0	95.1	86.6
spline3rs	95.0	95.2	95.3	94.6	94.3	95.6	95.0	95.8	81.4	93.8	94.7	91.0
spline4s	95.0	94.6	95.2	94.6	94.2	95.5	95.2	94.9	90.8	94.0	94.4	94.2
spline4rs	95.0	94.5	94.8	94.6	93.9	95.4	95.4	94.7	92.9	94.0	94.3	95.4
spline6s	95.0	94.7	94.7	94.7	94.1	95.7	95.4	94.8	94.7	94.1	94.5	95.7
spline6rs	95.0	94.7	94.7	94.8	94.3	95.7	95.4	95.1	94.9	94.1	94.8	95.8
spline8s	95.0	95.0	94.2	94.8	94.3	95.6	95.3	95.4	94.8	94.1	94.5	95.7
spline8rs	95.0	95.3	94.3	94.8	94.3	95.4	95.4	95.1	94.6	94.0	94.9	95.9
spline10s	95.0	95.1	94.5	94.8	94.2	95.2	95.4	94.9	94.6	94.0	94.8	95.9
spline10rs	95.0	95.2	94.7	94.7	94.3	94.8	95.5	94.9	94.6	94.0	95.0	96.0
spline12s	95.0	95.2	94.6	94.7	94.4	94.9	95.5	94.9	94.8	94.0	95.0	96.2
spline12rs	94.9	95.3	94.6	94.7	94.5	94.7	95.5	95.0	94.7	94.1	95.2	96.3
spline14s	94.9	95.4	94.7	94.7	94.5	94.8	95.5	95.0	94.8	94.2	95.2	96.3
spline14rs	94.9	95.5	94.7	94.7	94.5	94.8	95.5	95.0	94.8	94.0	95.3	96.4
spline16s	94.9	95.5	94.7	94.7	94.4	94.8	95.5	95.0	94.8	94.0	95.3	96.4
spline16rs	94.9	95.5	95.1	94.6	94.5	94.7	95.6	95.2	94.9	94.1	95.1	96.4

Table 6. Coverage rate for 95% confidence intervals for all p-splines based estimators and survey variables. Response probabilities type LOG− and GAU; n = 500

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
spline3	94.0	92.5	86.6	94.5	93.7	91.8	94.5	91.2	71.8	94.2	93.1	87.3
spline3r	94.0	92.2	90.3	94.4	93.5	93.5	94.5	92.3	81.9	94.1	94.0	90.8
spline4	93.8	92.2	93.4	94.3	93.8	94.9	94.4	94.4	93.2	94.1	94.1	95.0
spline4r	93.9	92.9	94.0	94.2	93.7	94.8	94.3	95.7	94.7	94.2	94.0	95.5
spline6	93.3	94.1	94.4	94.2	93.2	94.9	93.9	94.9	94.7	93.9	94.0	96.1
spline6r	93.3	94.9	94.4	94.2	93.5	94.7	93.9	94.5	94.6	94.0	94.4	96.3
spline8	93.2	94.4	94.2	93.9	93.5	94.5	94.0	93.8	94.3	93.9	94.3	95.9
spline8r	93.0	94.4	94.5	93.8	93.7	94.4	93.9	93.7	93.9	93.8	94.4	95.7
spline10	92.6	94.0	94.9	93.8	93.9	94.5	93.9	92.7	93.7	93.8	94.3	95.5
spline10r	92.8	94.3	94.5	93.7	94.0	94.7	93.7	92.7	93.2	93.8	94.1	95.3
spline12	92.6	94.2	94.2	93.2	93.9	94.6	93.5	92.8	93.3	93.8	94.0	95.4
spline12r	92.1	93.9	94.0	92.8	93.6	94.3	93.2	92.9	93.4	93.8	93.8	95.5
spline14	92.1	93.9	93.8	92.8	93.6	94.3	93.1	92.9	93.6	93.9	93.6	95.5
spline14r	91.8	93.7	93.5	92.6	93.0	94.2	93.0	92.7	94.0	93.7	93.3	95.4
spline16	91.9	93.7	93.6	92.6	93.0	94.2	93.0	92.5	94.0	93.7	93.3	95.4
spline16r	91.5	93.8	93.1	92.6	92.4	93.8	92.7	92.5	93.9	93.5	93.4	95.7

Table 6. Continued

	θ LOG−						θ GAU					
	$\sigma = 0.15$			$\sigma = 0.50$			$\sigma = 0.15$			$\sigma = 0.50$		
	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS	LIN	SIN	DIS
spline3s	94.9	93.0	87.4	95.3	93.7	91.6	95.8	92.2	82.4	96.0	93.3	89.5
spline3rs	95.1	93.0	91.9	95.3	93.3	94.1	95.8	93.2	88.0	96.0	93.9	91.9
spline4s	95.3	93.1	94.3	95.2	93.3	95.6	95.5	94.2	95.1	95.8	94.5	95.1
spline4rs	95.3	93.3	94.3	95.4	93.6	95.3	95.4	94.2	95.5	95.7	94.5	95.6
spline6s	95.5	94.7	94.3	95.4	94.1	95.4	95.2	94.7	95.3	95.9	95.2	95.9
spline6rs	95.5	95.3	94.4	95.3	94.3	95.4	95.1	94.8	95.2	95.9	95.0	95.9
spline8s	95.4	95.2	94.3	95.2	94.5	95.5	95.1	94.8	95.0	95.9	95.2	96.1
spline8rs	95.3	95.1	94.6	95.0	94.7	95.3	95.1	94.9	94.9	95.8	95.2	96.1
spline10s	95.3	94.9	94.8	95.0	95.0	95.3	95.1	95.0	94.6	95.8	95.1	96.2
spline10rs	95.3	94.6	94.7	95.0	95.0	95.1	95.1	95.2	94.4	95.7	95.0	96.3
spline12s	95.3	94.6	94.6	94.9	95.0	95.2	95.1	95.2	94.4	95.7	94.9	96.2
spline12rs	95.4	94.6	94.9	94.9	94.9	94.9	95.1	95.2	94.5	95.7	95.0	96.3
spline14s	95.4	94.6	94.8	94.9	94.9	94.9	95.2	95.2	94.5	95.7	95.1	96.2
spline14rs	95.4	94.6	95.1	94.8	94.8	94.7	95.2	95.2	94.2	95.7	95.1	96.1
spline16s	95.4	94.6	95.1	94.8	94.8	94.7	95.2	95.2	94.2	95.7	95.1	96.1
spline16rs	95.4	94.6	95.1	94.9	94.6	94.6	95.1	95.2	94.3	95.7	94.9	96.2

Coverage rates are satisfactory, with almost all rates between 93% and 96%. Serious undercoverage is displayed essentially only for the spline3 estimators when a large relative bias was also recorded, that is, in those cases in which 3 degrees of freedom are far too few to estimate complicated structures such as the GAU or the LOG+ response models in combination with the DIS or the SIN populations.

5. Conclusions

It is well known that nonresponse can harm the quality of the estimates from a survey by introducing bias. Put simply, this can happen in two ways: either the response probabilities depend on the variable of interest – direct effect – or on other variables that, in turn, influence the variable of interest – mediated effect. In both cases, it is possible to reduce the bias only if we have some auxiliary information that is able to describe either the variable of interest or the nonresponse probabilities, either at the level of the original sample (info-s) or for the whole population (info-U).

In this article we propose to use such auxiliary information to build a calibration type estimator following in the footsteps of those studied in Särndal and Lundström (2005). These latter estimators can reduce bias as long as the auxiliary information used in the calibration procedure provides a good proxy for the values of the variable of interest or, alternatively, for the inverse of the response probabilities. In classical calibration, such proxy values are constructed as linear combinations of the auxiliary information introduced in the calibration procedure. The estimator proposed here tries to bring such proxy values closer to the values of the variable of interest in a larger class of situations by using the results from model-assisted estimation based on nonparametric regression models. In particular, here we look at the penalized splines regression estimator proposed by Breidt et al. (2005) in the case of full response, since it has a close relationship with calibration.

The p-splines calibration estimator proposed here allows us to account for situation in which the effect of some auxiliary variables on the variable(s) of interest is more complicated than a linear function. In addition, it allows also for handling auxiliary information in the form of geographical coordinates and complicated spatial structures. Such flexibility grants a better description of the variable of interest for both respondents and nonrespondents and, therefore, more chances to reduce nonresponse bias. This comes at the price of extra auxiliary information required in the info-U setting, while it can be computed without extra auxiliary information in the info-s setting.

The asymptotic properties of the proposed estimator have been studied, conditions for consistency discussed and variance estimation proposed. The finite sample behavior has been explored via a limited simulation study on simulated data. Results show that the proposed estimator allows the reduction of bias, is not any less efficient than competing estimators that use the same auxiliary variables, and may be more efficient for complex survey variables. Of course, estimators that use auxiliary information at info-U level are more efficient than the corresponding estimators that use the info-s level. However, if bias is the main concern, then estimators that use info-s can provide the same reduction in bias as the info-U ones.

Like all nonparametric regression-based estimators, the performance of the proposed estimator depends on the selection of a smoothing parameter that governs its approximation ability. However, note that in the survey context, trying to find the optimal parameter is not as relevant as in the standard context: the estimator is not constructed for a single variable, but for a large set of variables collected during the survey. A penalty that is optimal for one variable may well not be adequate for another, but using different sets of weights would not be feasible for coherence issues. We have therefore considered a single fixed value for it and given some guidelines to selecting its value. In this regard, Särndal and Lundström (2008) have proposed an indicator that allows the ranking of different auxiliary vectors for their potential to reduce the bias for the calibration estimator. It will be interesting to investigate how this indicator can be modified to encompass penalized calibration (and hence p-splines) by comparing the use of a continuous auxiliary variable as it stands (linear), with dividing it into different groups or poststrata (piecewise linear), and with p-splines (nonparametrical).

A. Proofs

Proof of Theorem 3.1. First note that for info-U $\hat{Y}_{p,r} = \sum_U z_k^T \hat{\boldsymbol{\beta}}_r$, with

$$\hat{\boldsymbol{\beta}}_r = \left(\sum_r d_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \sum_r d_k \boldsymbol{z}_k y_k, \tag{A.1}$$

because $\sum_r d_k \left(y_k - z_k^T \hat{\boldsymbol{\beta}}_r \right) = 0$ for the properties of GLS estimators and noting that the first element of $\hat{\boldsymbol{\beta}}_r$ remains unpenalized. Therefore, $\hat{Y}_{p,r} - Y = \sum_U z_k^T (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) - \sum_U e_k$, where $e_k = y_k - z_k^T \boldsymbol{\gamma}_U$. For info-s $\hat{Y}_{p,r} = \sum_s d_k z_k^T \hat{\boldsymbol{\beta}}_r$ and

$$\begin{aligned} \hat{Y}_{p,r} - Y = & \sum_U z_k^T (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) + \left(\sum_s d_k \boldsymbol{z}_k - \sum_U \boldsymbol{z}_k \right)^T \boldsymbol{\gamma}_U \\ & + \left(\sum_s d_k \boldsymbol{z}_k - \sum_U \boldsymbol{z}_k \right)^T (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) - \sum_U e_k \end{aligned} \tag{A.2}$$

Note that $\boldsymbol{\gamma}_U$ is such that

$$\begin{aligned} \boldsymbol{\gamma}_U &= \left(\sum_U \theta_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \left[\sum_U \theta_k \boldsymbol{z}_k e_k + \sum_U \theta_k \boldsymbol{z}_k \boldsymbol{z}_k^T \boldsymbol{\gamma}_U + \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U - \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U \right] \\ &= \boldsymbol{\gamma}_U + \left(\sum_U \theta_k \boldsymbol{z}_k \boldsymbol{z}_k^T + \boldsymbol{\Lambda}_N \right)^{-1} \left[\sum_U \theta_k \boldsymbol{z}_k e_k - \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U \right]. \end{aligned}$$

This implies that $\sum_U \theta_k z_k e_k = \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U$. Now,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U &= \left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)^{-1} \sum_r d_k z_k y_k - \boldsymbol{\gamma}_U \\
&= \left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)^{-1} \left[\sum_r d_k z_k y_k - \left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)\boldsymbol{\gamma}_U\right] \\
&= \left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)^{-1} \left[\sum_r d_k z_k e_k + \sum_r d_k z_k z_k^T \boldsymbol{\gamma}_U - \sum_r d_k z_k z_k^T \boldsymbol{\gamma}_U - \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U\right] \\
&= \left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)^{-1} \left[\sum_r d_k z_k e_k - \sum_U \theta_k z_k e_k.\right]
\end{aligned}$$

The conditional expectation of the components of $\hat{\boldsymbol{\beta}}_r$ in (A.1) is given by

$$E\left\{\sum_r d_k z_k z_k^T | \mathcal{F}_N\right\} = \sum_U \theta_k z_k z_k^T \text{ and } E\left\{\sum_r d_k z_k y_k | \mathcal{F}_N\right\} = \sum_U \theta_k z_k y_k$$

Then, by Assumption A2 it follows that $\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U = O_p\left(n_N^{-1/2}\right)$ because $N^{-1}\left(\sum_r d_k z_k z_k^T + \boldsymbol{\Lambda}_N\right)$ is bounded by bounding arguments on $\boldsymbol{z}$ and Assumptions A3 and A4. In addition, given that $\sum_U \theta_k z_k e_k = \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U$ by the first part of assumption A5, for which there exists a vector $\boldsymbol{d}$ such that $z_k^T \boldsymbol{d} = \tilde{z}_k^T \boldsymbol{d}_1 + z_k^{*T} \boldsymbol{d}_2 = \theta_k^{-1}$, with $\tilde{z}_k = (1, z_k)^T$ and $z_k^* = ((z_k - \kappa_1)_+, \ldots, (z_k - \kappa_L)_+)^T$, then we can write $\sum_U e_k = \boldsymbol{d}^T \boldsymbol{\Lambda}_N \boldsymbol{\gamma}_U = \left(0, 0, \lambda_N \boldsymbol{d}_2^T\right)\boldsymbol{\gamma}_U = O\left(N n_N^{-1}\right)$ The last equality follows from the second part of Assumption A5.

To obtain representation (11) we follow Fuller (2009, Ch. 5). Assume, without loss of generality, that the first element of $\boldsymbol{z}$ is θ_k^{-1}. In fact, because of A5, θ_k^{-1} is in the space spanned by the columns of $\boldsymbol{Z}_r$ and we can transform the matrix of values of z_k so that the first element is the inverse of θ_k. In particular, consider the following transformation of the vector $\boldsymbol{z}$ by $\boldsymbol{\zeta}_k = \hat{\boldsymbol{Q}} \boldsymbol{z}_k$ with

$$\begin{aligned}
\zeta_{1k} &= \theta_k^{-1} \\
\zeta_{lk} &= z_{lk} + \zeta_{1k} \hat{q}_{1l}
\end{aligned}$$

where

$$\hat{q}_{1l} = -\left(\sum_r d_k \zeta_{1k}^2\right)^{-1} \sum_r d_k \zeta_{1k} z_{lk},$$

for $l = 2, 3, \ldots, L+2$ and $\hat{\boldsymbol{Q}} = diag\{1, \hat{q}_{1l}\}_{l=2,\ldots,L+2}$. Now, for info-U, $\hat{Y}_{p,r} - Y = \sum_U \boldsymbol{z}_k^T (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) + \sum_U e_k = \sum_U \boldsymbol{\zeta}_k^T \hat{\boldsymbol{Q}}^{-1} (\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) + O\left(N n_N^{-1}\right)$. Note that

$$\sum_U \boldsymbol{\zeta}_k^T = \left(\sum_U \boldsymbol{\zeta}_{1k}, 0, \ldots, 0\right) + O_p\left(N n_N^{-1/2}\right)$$

because of A1, A2 and A4. Now

$$\hat{\boldsymbol{Q}}^{-1}(\hat{\boldsymbol{\beta}}_r - \boldsymbol{\gamma}_U) = \left(\sum_r d_k \boldsymbol{\zeta}_k \boldsymbol{\zeta}_k^T + \hat{\boldsymbol{Q}}\boldsymbol{\Lambda}_N\hat{\boldsymbol{Q}})^{-1}\right)\left(\sum_r d_k \boldsymbol{\zeta}_k e_k + \hat{\boldsymbol{Q}}\boldsymbol{\Lambda}_N\boldsymbol{\gamma}_U\right),$$

whose first element is such that

$$\left(\sum_r d_k \zeta_{1k}^2\right)^{-1}\sum_r d_k \zeta_{1k} e_k = \left(\sum_U \zeta_{1k}\right)^{-1}\sum_r d_k \zeta_{1k} e_k + O_p\left(Nn_N^{-1}\right),$$

so that (11) is obtained. Finally, (12) can be obtained for info-s from (A.2). The e_k (and the $\boldsymbol{z}_k^T\boldsymbol{\gamma}_U$ for info-s) have bounded fourth moments by the moment assumptions so that, by assumption A2, $N^{-1}\sqrt{n_N}(\hat{Y}_{p,r} - Y)|\mathcal{F}_N$ has a normal distribution in the limit. The form of the variance in (14) follows from (11), while (15) can be obtained from

$$V\left(\sum_r \frac{e_k}{\pi_{2k}} - \sum_s \frac{e_k}{\pi_k} + \sum_s \frac{y_k}{\pi_k}|\mathcal{F}_N\right) = V_{p(s)}\left[E\left(\sum_r \frac{e_k}{\pi_{2k}} - \sum_s \frac{e_k}{\pi_k} + \sum_s \frac{y_k}{\pi_k}|s\right)\right]$$

$$+ E_{p(s)}\left[V\left(\sum_s \frac{e_k}{\pi_{2k}}|s\right)\right] = V_{p(s)}\left[\sum_s \frac{y_k}{\pi_k}\right] + E_{p(s)}\left[\sum_s \frac{e_k^2}{\pi_{2k}^2}\theta_k(1-\theta_k)\right], \qquad \text{(A.3)}$$

where subscript $p(s)$ denotes expectation and variance taken with respect to the sampling design. ■

Proof of Theorem 3.2. First note that by using conditional arguments as in (A.3), variances in (14) and in (15) can be rewritten as

$$V_\infty(\hat{Y}_{p,r}) = V_{p(s)}\left[\sum_s \frac{e_k}{\pi_k}\right] + E_{p(s)}\left[V\left(\sum_r \frac{e_k}{\pi_{2k}}|s\right)\right], \quad \text{for info-U}$$

$$V_\infty(\hat{Y}_{p,r}) = V_{p(s)}\left[\sum_s \frac{y_k}{\pi_k}\right] + E_{p(s)}\left[V\left(\sum_r \frac{e_k}{\pi_{2k}}|s\right)\right], \quad \text{for info-s.}$$

The expectation of the following Horvitz-Thompson variance estimator for $\sum_r w_k e_k$ is

given by

$$
\begin{aligned}
&E\left\{\sum_r\sum_r \frac{\pi_{kj}-\pi_k\pi_{kj}}{\pi_{kj}} w_k e_k w_j e_j | \mathcal{F}_N\right\} \\
&\quad = \sum_U \left(\pi_k - \pi_k^2\right)\theta_k w_k^2 e_k^2 + \sum_U\sum_{U_{k\neq j}} (\pi_{kj}-\pi_k\pi_{kj})\theta_k\theta_j w_k e_k w_j e_j \\
&\quad = \sum_U \left(\pi_k - \pi_k^2\right)\theta_k w_k^2 e_k^2 + \sum_U\sum_U (\pi_{kj}-\pi_k\pi_{kj})\theta_k\theta_j w_k e_k w_j e_j \\
&\qquad - \sum_U \left(\pi_k\theta_k - \pi_k^2\theta_k^2\right) w_k^2 e_k^2 \\
&\quad = \sum_U\sum_U (\pi_{2kj}-\pi_{2k}\pi_{2j}) w_k e_k w_j e_j - \sum_U (\pi_k - \pi_{2k})\pi_{2k} w_k^2 e_k^2.
\end{aligned}
$$

By A7 and given that $\boldsymbol{\beta}_r - \boldsymbol{\gamma}_U = O_p\left(n_N^{-1/2}\right)$ the variance estimator constructed using $\hat{e}_k$ is asymptotically equivalent to the one that uses e_k, and the result is proven. ■

6. References

Beaumont, J.F. and Bocci, C. (2008). Another Look at Ridge Calibration. Metron, 66, 5–20.

Breidt, F.J., Claeskens, G., and Opsomer, J.D. (2005). Model-Assisted Estimation for Complex Surveys Using Penalised Splines. Biometrika, 92, 831–846.

Deville, J.-C. (2000). Generalized Calibration and Application to Weighting for Non-Response. J. G. a. Bethlehem and P. G. M. a. van der Heijden (Eds), COMPSTAT – Proceedings in Computational Statistics, 14th Symposium, 65–76. Physica-Verlag Ges.m.b.H.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible Smoothing with B-splines and Penalties. Statistical Science, 11, 89–121.

Ekholm, A. and Laaksonen, S. (1991). Weighting via Response Modeling in the Finnish Household Budget Survey. Journal of Official Statistics, 7, 325–337.

Folsom, R.E. (1991). Exponential and Logistic Weight Adjustments for Sampling and Nonresponse Error Reduction. ASA Proceedings of the Social Statistics Section, 197–202.

Fuller, W.A. (2002). Regression Estimation for Survey Samples. Survey Methodology, 28, 5–25.

Fuller, W.A. (2009). Sampling Statistics. Hoboken, NJ: Wiley, Wiley Series in Survey Methodology.

Fuller, W.A. and An, A.B. (1998). Regression Adjustments for Nonresponse. Journal of the Indian Society of Agricultural Statistics, 51, 331–342.

Fuller, W.A., Loughin, M.M., and Baker, H.D. (1994). Regression Weighting in the Presence of Nonresponse with Application to the 1987–1988 Nationwide Food Consumption Survey. Survey Methodology, 20, 75–85.

Giommi, A. (1987). Nonparametric Methods for Estimating Individual Response Probabilities. Survey Methodology, 13, 127–134.

Guggemos, F. and Tillé, Y. (2010). Penalized Calibration in Survey Sampling: Design-based Estimation Assisted by Mixed Models. Journal of Statistical Planning and Inference, 140, 3199–3212.

Isaki, C.T. and Fuller, W.A. (1982). Survey Design under the Regression Superpopulation Model. Journal of the American Statistical Association, 77, 89–96.

Kim, J.K. and Kim, J.J. (2007). Weighting Adjustment using Estimated Response Probability. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 35, 501–514.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. Survey Methodology, 32, 133–142.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

Lundström, S. and Särndal, C.E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. Journal of Official Statistics, 15, 305–327.

Montanari, G.E. and Rananlli, M.G. (2005). Nonparametric Model Calibration Estimation in Survey Sampling. Journal of the American Statistical Association, 100, 1429–1442.

Park, M. and Fuller, W.A. (2009). The Mixed Model for Survey Regression Estimation. Journal of Statistical Planning and Inference, 139, 1320–1331.

Rao, J.N.K. and Singh, A.C. (1997). A Ridge-Shrinkage Method for Range Restricted Weight Calibration in Survey Sampling. In Proceedings of the American Statistical Association, Section on Survey Research Methods, 57–65.

Ruppert, D. (2002). Selecting the Number of Knots for Penalized Splines. Journal of Computational and Graphical Statistics, 11, 735–757.

Ruppert, D., Wand, M.P., and Carrol, R. (2003). Semiparametric Regression. Cambridge, New York: Cambridge University Press.

Sànchez-Borrego, I., Opsomer, J. D., Rueda, M., and Arcos A. (2011). Nonparametric Regression with Mixed Data Types in Survey Sampling. Preprint submitted to Revista Matematica Complutense.

Särndal, C.-E. (2007). The Calibration Approach in Survey Theory and Practice. Survey Methodology, 33, 99–119.

Särndal, C.-E. and Lundström, S. (2005). Estimation in Surveys with Nonresponse. Wiley: Chichester.

Särndal, C.-E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Journal of Official Statistics, 24, 167–191.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling. Berlin, New York: Springer.

Silva, D.N.D. and Opsomer, J.D. (2006). A Kernel Smoothing Method of Adjusting for Unit Non-response in Sample Surveys. The Canadian Journal of Statistics/La Revue Canadienne de Statistique, 34, 563–579.

Wu, C. and Sitter, R.R. (2001). A Model-Calibration to Using Complete Auxiliary Information from Survey data. Journal of the American Statistical Association, 96, 185–193.

Received November 2009
Revised January 2012