Journal of Official Statistics, Vol. 19, No. 2, 2003, pp. 99–117

Penalized Spline Model-Based Estimation of the Finite Populations Total from Probability-Proportional-to-Size Samples

Hui Zheng[1] *and Roderick J.A. Little*[2]

The Horvitz-Thompson (HT) estimator is a simple design-unbiased estimator of the finite population total for sample designs with unequal probabilities of inclusion. Viewed from a modeling perspective, the HT estimator performs well when the ratios of the outcome values y_i and the selection probabilities π_i are approximately exchangeable. When this assumption is far from met, the Horvitz-Thompson estimator can be very inefficient. We consider alternatives to the HT estimator that posit a smoothly-varying relationship between y_i (or a function of y_i) and the inclusion probability π_i (or a function of π_i), and that model this relationship using penalized splines. The methods are intended for situations with probability-proportional-to-size (PPS) sampling and continuous survey outcomes. Simulation studies are conducted to compare the spline-based predictive estimators and parametric alternatives with the HT estimator and extensions such as the generalized regression (GR) estimator. These studies show that the p-spline model-based estimators are generally more efficient than the HT and GR estimators in terms of the root mean squared error. In situations that most favor the HT or GR estimators, the p-spline model-based estimators have comparable efficiency.

The p-spline model-based estimators and the Horvitz-Thompson estimator are compared on a Block Statistics data set from a U.S. census.

Key words: Horvitz-Thompson estimator; spline regression; linear mixed model; bias calibration; design consistency.

1. Introduction

We consider probability sampling designs where a random sample S with elements $y_1, \ldots, y_n$ is drawn from the finite population according to the inclusion probabilities π_i, $i = 1, \ldots, N$. Of main interest is statistical inference for the total T of an outcome Y for the finite population P with N elements. In probability-proportional-to-size sampling, the values of π_i are proportional to the values x_i of size variable X and are usually known for the whole population before S is drawn. In these cases π_i's are considered fixed.

The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) is a standard design-unbiased linear estimator of T, and weights cases by the inverse of their inclusion

[1] Harvard Medical School, Department of Health Care Policy, 180 Longwood Ave., Boston 02446, U.S.A. Email: zheng@hcp.med.harvard.edu
[2] University of Michigan, Department of Biostatistics, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A. Email: rlittle@umich.edu
Acknowledgments: This research was supported by NSF grant DMS9803720. We appreciate useful comments from Trivellore Raghunathan and Christopher Skinner and a referee. These comments led to valuable theoretical and numerical advances in our study.

probabilities. However, the predictive estimator for a good statistical model relating the outcomes y_i and π_i is potentially more efficient than the HT estimator. The HT estimator itself can be regarded as a model-based estimator for the following linear model relating y_i to π_i:

$$y_i = \beta\pi_i + \pi_i \varepsilon_i \tag{1}$$

or equivalently,

$$z_i = y_i/\pi_i = \beta + \varepsilon_i \tag{2}$$

where ε_i in Equations (1) and (2) are assumed to be i.i.d. normally distributed with mean zero and variance σ^2. Regressions for (1) and (2) both lead to $\hat{\beta} = n^{-1}\sum_{i\in S} y_i/\pi_i$ where n is the sample size. The corresponding projective estimator for T,

$$\hat{T}_{proj} = \sum_{i\in P}\hat{y}_i = \sum_{i\in P}\pi_i \times n^{-1}\left(\sum_{i\in S} y_i/\pi_i\right) = \sum_{i\in S} y_i/\pi_i \tag{3}$$

equals the HT estimator, since $\sum_{i\in P}\pi_i = n$. The predictive estimator based on model (1), namely

$$\hat{T}_{pred} = \sum_{i\in S} y_i + \sum_{i\notin S}\hat{y}_i = \sum_{i\in S} y_i + \frac{1}{n}\sum_{i\in S}\frac{y_i}{\pi_i}\times\sum_{i\notin S}\pi_i = \sum_{i\in S} y_i + \left(1 - n^{-1}\sum_{i\in S}\pi_i\right)\hat{T}_{proj} \tag{4}$$

differs from the HT estimator by a quantity that tends to zero with the sampling fraction n/N.

Models other than (1) can also yield estimates with design consistency. Model-assisted methods (Deville and Särndal 1992; Särndal et al. 1992) improve efficiency of estimation via modeling and achieve design-consistency by bias calibration. For example, assisted by a statistical model, the sum of a projective estimator $\hat{T}_{PROJ} = \sum_{i\in P}\hat{y}_i$ and its corresponding bias calibration $\sum_{i\in S} 1/\pi_i(y_i - \hat{y}_i)$ is design-consistent. A predictive estimator $\hat{T}_{PRED} = \sum_{i\in S} y_i + \sum_{i\in P-S}\hat{y}_i$ has the bias calibration $\sum_{i\in S}(1/\pi_i - 1)(y_i - \hat{y}_i)$. In Firth and Bennett (1998), parametric and nonparametric regression models such as simple linear regression models of y_i on a function of π_i (e.g., $1/\pi_i$) are shown to have the ''internally bias calibrated'' (IBC) property. With this property, bias calibration terms vanish in the GR estimators, hence the predictive or projective estimators are design-consistent. In Little (1983) and Elliot and Little (2000), the population is divided into subclasses according to π_i and the subclass means are treated as random effects. The model-based estimators are shown to outperform design-based estimation in terms of root mean squared error (RMSE) if the model is reasonable.

On the other hand, model-based methods may yield biased estimates when the underlying regression model is misspecified, motivating consideration of more flexible mean structures for $E(Y_i|\,\pi_i)$. In Dorfman (1992), finite population totals are estimated by the nonparametric model-based method using an auxiliary variable. In Chambers, Dorfman, and Wehrly (1993), nonparametric model-based estimation using kernel smoothing is proposed for estimating finite population quantities. In this article, we consider using the unequal probabilities of inclusion π_i as the independent variable in the nonparametric model. For example, the relationship between an outcome and some population size

variable used in PPS sampling (such as the areas of counties) may often be considered nonlinear, and not fit well using standard polynomial models. We seek flexible mean structures $E(Y_i|\pi_i)$ that are more robust to misspecification than parametric models and still provide more efficient estimation of T than the HT estimator.

Nonparametric regression using splines has undergone extensive development in recent years. Smoothing splines (Eubank 1988; Wahba 1990) use a knot at each distinct value (except the boundary values) of the X-variable and control overfitting by applying a roughness penalty. Penalized splines (p-splines), formally introduced by Eilers and Marx (1996), are in general computationally inexpensive and allow flexible knot selection, yet yield sound performance. P-splines are also easy to implement: there is a close relationship between p-spline regression models and mixed-effects models, which implies that they can be fitted using widely-available statistical software such as SAS Proc Mixed (SAS 1992) and S-Plus (Pinheiro and Bates 2000) function lme ( ).

In this article, we model the conditional mean $E(Y_i|\pi_i)$ for a continuous outcome Y_i given the selection probabilities by penalized splines. Model-based prediction of T is then based on the predictions from the spline regression. Simulation studies based on probability-proportional-to-size (PPS) sampling are conducted on a variety of smoothly-varying mean structures. Our simulations suggest that for estimation of the finite population total, p-spline model-based predictive estimators are in general more efficient than the HT estimator and the generalized regression (GR) estimator, a common design-based modification of the HT estimator used to improve its precision. In situations that favor the HT estimator, the nonparametric model-based estimators are only slightly less efficient. P-spline models relating the outcome to π_i's lead to many model-based estimators that are IBC and hence design-consistent.

2. Nonparametric Model-based Estimation

2.1. P-spline model

We consider the following general model:

$$y_i = f(\pi_i, \beta) + \varepsilon_i,\ \ \varepsilon_i \sim \text{iid}\ N(0,\ \pi_i^{2k}\sigma^2) \tag{5}$$

where $N(\cdot\,,\cdot)$ denotes the normal distribution, and f is a function of π_i that is continuous to the $(p-1)$th derivative with unknown parameters β. The exponent k (usually taking values 0, 1/2, or 1) models error heteroscedasticity; for simplicity we assume here k is known, although simultaneous estimation of this parameter is also possible. The special case of (5) where $f(\pi_i) = \beta\pi_i$ and $k = 1$ coincides with Model (1), leading to the HT estimator (3) or (4).

The function f is estimated by splines, which are piecewise polynomial functions that are smooth to a certain degree. For example, a cubic spline function is a piecewise cubic polynomial that is continuous in its second derivative. The splines can be expressed as a linear combination of a set of basis functions defined with respect to a set of knots. In Eilers and Marx (1996), a set of B-splines is used as the basis function. In Ruppert and Carroll (2000), truncated polynomials are used. Although theoretically equivalent to truncated polynomials, B-splines are more ''balanced'' and are numerically more stable

in some extreme cases. In this article, we are more interested in the cumulative (sum over a population) performance of the estimator than in the pointwise performance of regression curves. We choose to use truncated polynomials because of their simplicity although other methods may have better numerical properties.

Specifically, the function f is estimated by the p-spline written as a linear combination of truncated polynomials:

$$\begin{aligned} &\hat{f}(\pi_i, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{l=1}^{m} \beta_{l+p} (\pi_i - \kappa_l)_+^p, \; i = 1, \ldots, N \\ &\beta_{l+p} \underset{iid}{\sim} N(0, \tau^2), \; l = 1, \ldots, m \end{aligned} \tag{6}$$

where the constants $\kappa_1 < \ldots < \kappa_m$ are selected fixed knots and $(u)_+^p = u^p \mathbf{I}(u \geqslant 0)$. The truncated polynomial is continuous to the $(p-1)$th derivative and has a change of $p!$ in the pth derivative at the corresponding knot. The coefficients $\beta_{p+1}, \ldots, \beta_{p+m}$ are simply proportional to the amount of discontinuity by the regression function in the pth derivative. The effect of treating $\{\beta_l, l = p+1, \ldots, p+m\}$ as normal random effects is to add a penalty term $\sum_{l=p+1}^{p+m} \hat{\beta}_l^2 / \tau^2$ to the sum of squares that is minimized in a least squares fit, thus smoothing their estimates towards zero.

In Eilers and Marx (1996), the penalty on roughness can be applied to an order different from that of the basis functions. Here we do not include those variants because simulations show they do not have a big impact in the estimation of population totals.

The variance τ^2 is an additional parameter estimated from the data. Models (5) and (6) can be written in the matrix form

$$Y = X\beta^{(1)} + Z\beta^{(2)} + \varepsilon \tag{7}$$

where $Y = (y_1, \ldots, y_n)^T$, $\beta^{(1)} = (\beta_1, \ldots, \beta_p)^T$, $\pi^l = (\pi_1^l, \ldots, \pi_n^l)^T$, $l = 1 \ldots, p$

$\beta^{(2)} = (\beta_{p+1}, \ldots, \beta_{p+m})^T \sim N_m(0,\ \tau^2 I_m)$,

$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T \sim N_n(0, \sigma^2 Diag(\pi_1^{2k}, \ldots, \pi_n^{2k}))$

$$X = \begin{pmatrix} 1 & \pi_1 & \ldots & \pi_1^p \\ 1 & \pi_2 & \ldots & \pi_2^p \\ . & . & \ldots & . \\ & & & \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 1 & \pi_n & \ldots & \pi_n^p \end{pmatrix}, \quad Z = \begin{pmatrix} (\pi_1 - \kappa_1)_+^p & \ldots & (\pi_1 - \kappa_m)_+^p \\ . & \ldots & . \\ . & \ldots & . \\ . & \ldots & . \\ (\pi_n - \kappa_1)_+^p & \ldots & (\pi_n - \kappa_m)_+^p \end{pmatrix}$$

and $N_m(\mu, \Sigma)$ denotes the multivariate normal distribution with mean vector μ and covariance matrix Σ.

Assuming constant error variance (i.e., $k = 0$), the ML estimate of the regression parameters conditional on $\alpha = \sigma^2/\tau^2$ is $(\hat{\beta}_0, \ldots, \hat{\beta}_{m+p})^T = (\Pi^T \Pi + D)^{-1} \Pi^T Y$, where $\Pi = [X \quad Z]$ and the ith row of Π is $\Pi_i = (1, \pi_i, \ldots, \pi_i^p, (\pi_i - \kappa_1)_+^p, \ldots, (\pi_i - \kappa_m)_+^p)$

and the matrix D is diagonal with the first $p+1$ elements equal to 0 and the remaining m elements equal to $\alpha = \sigma^2/\tau^2$. A widely used approach to estimate α is through the minimization of the generalized cross validation (GCV) statistic. A simple alternative is to calculate $\hat{\sigma}^2$ and $\hat{\tau}^2$ using restricted maximum likelihood (REML) algorithms with standard software such as SAS Proc Mixed (SAS 1992) and the S-plus function lme ( ) (Pinheiro and Bates 2000). For the scenarios simulated in this study, we found the numerical difference between the GCV criterion and the REML method to be small and only report the results from the REML method.

When assuming error variances are $\pi_i^{2k}\sigma^2$, $k \neq 0$, Models (5) and (6) are fitted by replacing matrices Y, Z and π^l, $l = 1, \ldots, p$ by $Y^* = W^{1/2}Y$, $Z^* = W^{1/2}Z$ and $\pi^{l*} = W^{1/2}\pi^l$, $l = 1, \ldots, p$ respectively, where $W = diag(\pi_1^{-2k}, \pi_2^{-2k}, \ldots, \pi_n^{-2k})$. Conditional on $\alpha = \sigma^2/\tau^2$, the estimates $(\hat{\beta}_0, \ldots, \hat{\beta}_{m+p})^T = (\Pi^{*T}\Pi^* + D)^{-1}\Pi^{*T}Y^* = (\Pi^T W \Pi + D)^{-1}\Pi^T WY$, where $\Pi^* = W^{1/2}\Pi$.

Assuming that the π_i's are known for the whole population, the model-based predictive estimator is given by $\hat{T}_{PRED} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} E(Y_i|\pi_i)$, where

$$E(Y_i|\pi_i) = f(\pi_i, \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 \pi_i + \ldots + \hat{\beta}_p \pi_i^p + \sum_{j=1}^{m} \hat{\beta}_{j+p}(\pi_i - \kappa_j)_+^p$$

An important modeling issue here is the number and positioning of the knots. In our simulation study, we deliberately choose simple approaches that avoid subjective choices based on prior knowledge or looks at the data. We choose two fixed numbers of knots (5 or 15), and place knots at the m evenly spaced sample percentiles, that is, $100j(m+1)^{-1}$, $j = 1, \ldots, m$, of the sample distribution of π_i. This choice of knots works well for our simulated populations. Alternative choices such as equally spaced knots may be better for some datasets. The number of knots can be chosen subjectively according to the complexity of the observed relationship between the outcome and the selection probabilities. Also, in cases with severe spatial inhomogeneity one may consider more sophisticated methods of knot placement, such as stepwise knot selection (Friedman and Silverman 1989; Friedman 1991; Stone et al. 1997) or Bayesian knot selection using reversible jump Markov Chain Monte Carlo methods (Green 1995; Denison, Mallick, and Smith 1998). Ruppert and Carroll (2000) suggest applying different roughness penalties at different knots to adapt to the spatial heterogeneity. Such methods are harder to assess by simulation and are not considered here.

2.2. *Design Consistency of the P-spline Model-based Estimators*

The p-spline model can be chosen to yield estimators that have desirable design-based properties. The reason for this is that fitted p-splines based on Equations (5) and (6) have the following property (denoted as property A):

$$\sum_{i \in S} y_i/\pi_i^l = \sum_{i \in S} \hat{y}_i/\pi_i^l \text{ for } l = (2k-p), (2k-p+1), \ldots, 2k$$

Proof for $k = 0$: It suffices to show that $X^T(Y - \hat{Y}) = 0$, where $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)^T$.

From Section 2.1, we have $(\hat{\beta}_0, \ldots, \hat{\beta}_{m+p})^T = (\Pi^T \Pi + D)^{-1} \Pi^T Y$. It follows that

$$\begin{aligned}
\Pi^T(Y - \hat{Y}) &= \Pi^T(Y - \Pi\hat{\boldsymbol{\beta}}) \\
&= \Pi^T(Y - \Pi(\Pi^T\Pi + D)^{-1}\Pi^T Y) \\
&= (\Pi^T - \Pi^T\Pi(\Pi^T\Pi + D)^{-1}\Pi^T)Y \\
&= (\Pi^T - (\Pi^T\Pi + D)(\Pi^T\Pi + D)^{-1}\Pi^T + D(\Pi^T\Pi + D)^{-1}\Pi^T)Y \\
&= D(\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= D\hat{\boldsymbol{\beta}} \\
&= (0, \ldots, 0, \alpha\hat{\beta}_{p+1}, \ldots, \alpha\hat{\beta}_{p+m})^T
\end{aligned}$$

where the first $p+1$ entries of the vector are zeros. Since $\Pi = [X \quad Z]$, we have $X^T(Y - \hat{Y}) = 0$.

The proof for the case where k is nonzero follows when replacing Y, X and Z with $Y^* = W^{1/2}Y$, $X^* = W^{1/2}X$ and $Z^* = W^{1/2}Z$, respectively.

Property A can be utilized to prove that some p-spline estimators are IBC, and hence design-consistent (Firth and Bennett 1998). For example, assuming $k = 1/2$ and $p = 1$, we have $\sum_{i\in S} y_i/\pi_i = \sum_{i\in S} \hat{y}_i/\pi_i$, and $\sum_{i\in S} y_i = \sum_{i\in S} \hat{y}_i$. Hence the predictive estimator $\hat{T}_{PRED} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{y}_i$ is IBC, because the calibration term $\sum_{i\in S}(1/\pi_i - 1)(y_i - \hat{y}_i)$ equals zero for all possible S.

Similarly, the p-spline for a model with $1/\pi_i$ as the X – variable and a constant variance is also IBC, and hence design-consistent. Since IBC p-spline models are far from unique, we can choose one that fits the data best. On the other hand, the simulations in the next section indicate that bias for p-spline models that do not have the IBC property is minor, so design consistency may not be of paramount importance.

3. Simulation Study

3.1. Design of the simulation study

Simulation studies are conducted to study the performance of the predictive estimator based on spline models compared with the HT estimator for a variety of populations. We compare seven estimators:

1. HT, the projective HT estimator, Equation (3).
2. HT PRED, the predictive HT estimator, Equation (4).
3. P0_5, the p-spline model-based estimator with $p = 1$ and $k = 0$ and using 5 knots.
4. P0_15, the p-spline model-based estimator with $p = 1$ and $k = 0$ and using 15 knots.
5. P1_5, the p-spline model-based estimator with $p = 1$ and $k = 1$ and using 5 knots.
6. P1_15, the p-spline model-based estimator with $p = 1$ and $k = 1$ and using 15 knots.
7. GR, the generalized regression estimator $\hat{T}_{GR} = \sum_{i\in P} \hat{y}_i + \sum_{i\in S}(y_i - \hat{y}_i)/\pi_i$, where $\hat{y}_i = E(Y_i|\boldsymbol{\pi}_i)$ based on a simple linear regression of y_i on $\boldsymbol{\pi}_i$.

8. PROJ2, $\hat{T}_{PROJ2} = \sum_{i \in S} \hat{y}_i/\pi_i$, where $\hat{y}_i = E(Y_i|\,\boldsymbol{\pi}_i)$ based on the p-spline model, assuming $k = 0$ and using 15 knots. This extension to the projective estimator only requires that π_i's are known for the sampled units and can be used when π_i are not known for the whole population.

For estimation of finite population totals, simulations suggest linear splines perform as well as (if not better than) higher order ones such as quadratic or cubic splines. Here we only report the results from linear p-splines ($p = 1$). We would like to point out that higher order splines may have better pointwise performance in regression, which is not of our primary interest here.

First, we simulate finite populations with different mean structures $E(Y_i|\,\pi_i)$ and constant error variance structure. Random probability-proportional-to-size samples are then drawn from the finite populations with inclusion probabilities $\pi_i \propto x_i$. The values x_i of the size variable X vary so that the maximum inclusion probability is approximately 30 times as large as the minimum. The PPS samples are drawn systematically from a randomly ordered list.

We simulated three different population sizes, 300, 1,000, and 2,000, with sample sizes 30, 100, and 200 respectively. For population size 300, X takes the consecutive integer values 11, 12, ..., 310. For population size 1,000, X takes the values 35, 36, ..., 1,034. For sample size 2,000, X takes the values 71, 72, ..., 2,070. The corresponding inclusion probabilities have a ratio of about 30 between the maximum and the minimum.

For each population and sample size combination, 500 samples are obtained and the eight above-mentioned estimators are compared.

For each population and sample size, six different mean structures are simulated for $f(\pi_i) = E(Y_i|\,\pi_i)$: constant function (NULL) $f(\pi_i) \equiv 0.30$, linearly increasing function with zero intercept (LINUP) $f(\pi_i) = 3\pi_i$, linearly decreasing function with positive intercept (LINDOWN) $f(\pi_i) = 0.58 - 3\pi_i$, exponentially increasing function (EXP) $f(\pi_i) = \exp(-4.64 + 26\pi_i)$, sine function (SINE) $f(\pi_i) = \sin(35.69\pi_i)$, and the ''S'' shaped function (ESS) $y_i = 0.6\, logit^{-1}(50 * \pi_i - 5 + \varepsilon_i), \varepsilon_i \overset{iid}{\sim} N(0, 1)$. Of these, LINUP most favors the HT estimator. For the first five populations, independent random errors (with constant variances and zero mean) are added to the mean structure. For population ESS, the independent errors are added inside the inverse logit to simulate the situations where there are lower and upper limits for the values of Y. The constants in the above functions are chosen so that $T \approx 90$ for N = 300, $T \approx 300$ for N = 1,000 and $T \approx 600$ for N = 2,000.

To compare normal and skewed error distributions, we simulate both normal and lognormal errors for all but the ESS populations, which already have non-normal errors. Lognormal errors are generated according to the expression $\sigma\sqrt{64/151} * (\exp(\delta_i) - 13/8)$, $\delta_i \underset{iid}{\sim} N(0, 1)$, so that the errors have mean zero and variance σ^2. The error variances for populations NULL, LINUP, LINDOWN, EXP and SINE are all 0.04. Figures 1 and 2 give the plot of the populations with $N = 300$.

We fit Model (6) with $p = 1$, $k = 1$, and both 5 and 15 equally spaced percentiles as knots, using an REML algorithm. Results on bias are shown in Tables 1 through 3 and on root mean squared error (RMSE) are shown in Tables 4 through 6.

We then change the sampling rate for a fixed population size $N = 1{,}000$ with normal

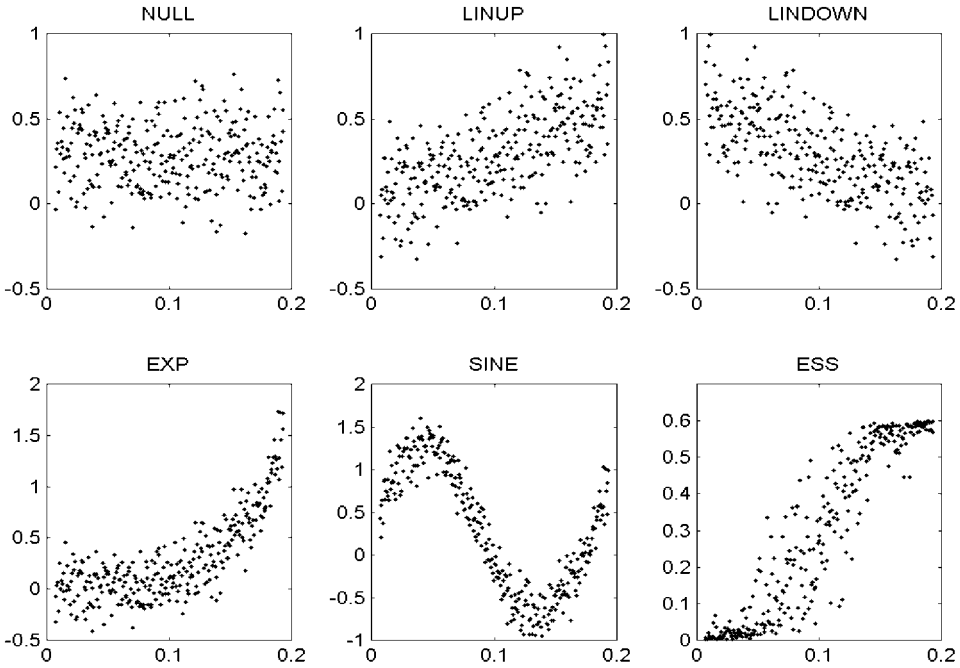

Fig. 1. Six simulated populations (N = 300) X-axis: pi(i); Y-axis: y(i) with normal errors

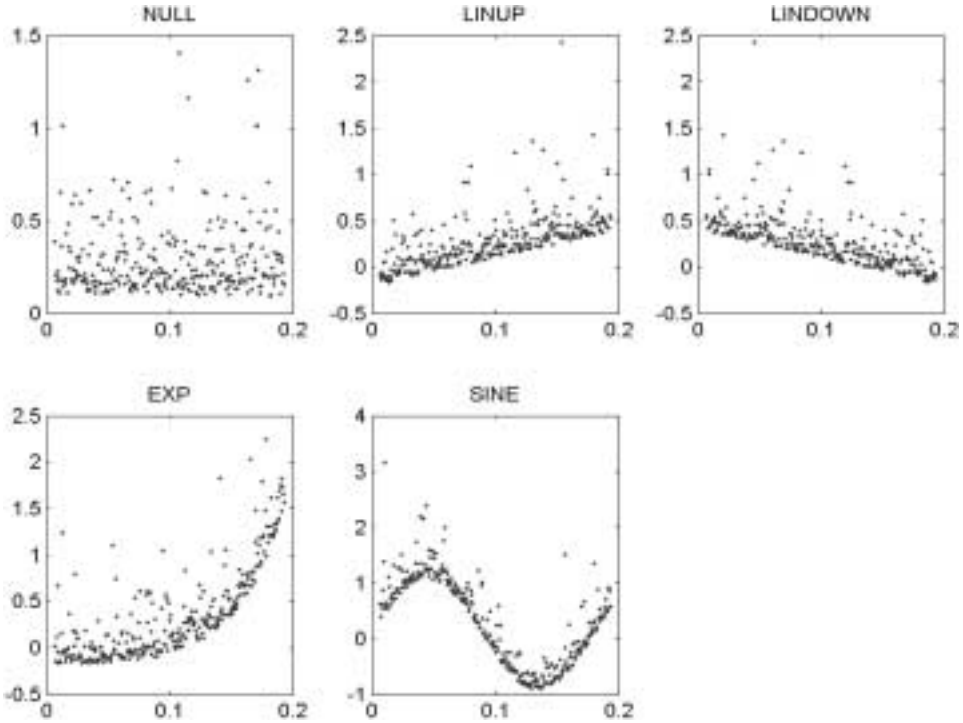

Fig. 2. Five simulated populations (N = 300) X-axis: pi(i); Y-axis: y(i) with lognormal errors

Table 1. Empirical bias ×10 of eight estimators for $N = 300$, $n = 30$ (Minimum absolute bias for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	−10	−7	**1**	3	−6	−30	−4	−14
	Lognormal	−13	−12	−8	−7	**−3**	−26	−9	−7
LINUP	Normal	−7	−4	8	11	3	3	**−3**	−4
	Lognormal	**−4**	−6	−13	−13	6	9	−7	11
LINDOWN	Normal	−11	−9	−9	−9	**−1**	−49	−9	−4
	Lognormal	31	30	10	11	**−2**	−52	17	12
SINE	Normal	272	261	176	216	−46	−140	−194	**−19**
	Lognormal	286	266	236	215	−26	−127	−102	**8**
EXP	Normal	−25	−22	−19	−7	7	30	−92	**−3**
	Lognormal	−29	−31	−20	−26	**−1**	29	−41	−19
ESS		−10	−7	−5	−5	**−1**	7	7	−6

Table 2. Empirical bias ×10 of eight estimators for $N = 1{,}000$, $n = 100$ (Minimum absolute bias for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	−4	−5	−7	−10	9	−75	**−3**	3
	Lognormal	53	46	37	35	40	−53	**19**	76
LINUP	Normal	13	11	**3**	3	−8	−8	−4	11
	Lognormal	5	**1**	5	−3	−7	−4	−33	−6
LINDOWN	Normal	21	22	−11	−12	**5**	−174	13	24
	Lognormal	42	46	37	41	42	−133	**9**	67
SINE	Normal	464	101	140	95	**−48**	−379	−189	52
	Lognormal	496	254	183	73	50	−294	**−41**	155
EXP	Normal	9	**−1**	16	19	−19	64	−73	−25
	Lognormal	**7**	18	18	36	25	121	−81	15
ESS		−35	−29	−5	−5	**−4**	22	8	−29

Table 3. Empirical bias ×10 of eight estimators for $N = 2{,}000$, $n = 200$ (Minimum absolute bias for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	−18	−19	−19	−17	−24	−186	**−8**	−36
	Lognormal	−22	−21	−22	−26	**−14**	−191	−15	−18
LINUP	Normal	45	38	−25	−26	3	8	**1**	41
	Lognormal	**−13**	−15	44	78	−39	−61	−16	−41
LINDOWN	Normal	20	28	6	9	**−1**	−357	7	21
	Lognormal	67	61	**5**	9	21	−310	13	76
SINE	Normal	811	252	134	80	**−23**	−652	−55	148
	Lognormal	731	261	98	35	**15**	−665	−56	205
EXP	Normal	−5	−0	−5	**−3**	−29	140	−16	−30
	Lognormal	−36	**−5**	41	54	−17	178	−131	−30
ESS		−59	−36	−1	2	**−2**	56	−4	−33

Table 4. Empirical RMSE ×10 of eight estimators for $N = 300$, $n = 30$ (Minimum RMSE for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	**126**	130	127	131	220	207	130	212
	Lognormal	**108**	109	117	118	188	174	134	171
LINUP	Normal	**124**	127	124	127	143	133	134	134
	Lognormal	118	118	122	125	116	**110**	120	114
LINDOWN	Normal	115	120	**115**	118	316	295	130	312
	Lognormal	190	190	**175**	176	302	279	191	300
SINE	Normal	451	470	**393**	430	656	622	572	673
	Lognormal	491	479	482	**472**	738	697	541	758
EXP	Normal	151	154	151	**149**	185	183	326	190
	Lognormal	**142**	148	174	196	200	190	300	184
ESS		70	68	71	68	70	**67**	78	75

Table 5. Empirical RMSE ×10 of eight estimators for $N = 1{,}000$, $n = 100$ (Minimum RMSE for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	200	**200**	211	212	328	309	208	316
	Lognormal	318	324	337	333	441	407	**308**	482
LINUP	Normal	224	227	282	289	242	222	252	**204**
	Lognormal	**247**	**247**	293	298	296	275	297	249
LINDOWN	Normal	294	297	315	315	659	629	**292**	652
	Lognormal	**232**	243	321	335	650	612	280	658
SINE	Normal	671	**347**	591	421	1,340	1,298	899	1,354
	Lognormal	758	525	545	**402**	1,297	1,234	836	1,318
EXP	Normal	262	**257**	272	281	326	322	566	322
	Lognormal	**371**	403	416	470	407	400	584	384
ESS		98	94	99	**93**	130	125	115	133

Table 6. Empirical RMSE ×10 of eight estimators for $N = 2{,}000$, $n = 200$ (Minimum RMSE for each population is in bold print)

Population	Error distribution	P0_5	P0_15	P1_5	P1_15	HT	HT PRED	GR	PROJ2
NULL	Normal	365	**361**	397	400	528	520	377	502
	Lognormal	**386**	401	478	488	657	629	450	613
LINUP	Normal	286	287	321	318	314	291	317	**283**
	Lognormal	523	545	806	934	611	556	644	**510**
LINDOWN	Normal	256	**248**	289	288	749	769	287	743
	Lognormal	430	418	**400**	414	764	761	426	780
SINE	Normal	1,011	498	640	**462**	1,490	1,526	1,008	1,503
	Lognormal	1,039	684	772	**630**	1,613	1,633	1,120	1,644
EXP	Normal	**340**	363	361	365	435	430	637	422
	Lognormal	501	585	692	812	504	498	910	**482**
ESS		140	135	133	**131**	150	155	153	149

Table 7. Empirical RMSE ×10 of five estimators for $N = 1{,}000$, $n = 50$ (Minimum RMSE for each population is in bold print)

Population	P0_15	P1_15	HT	HT PRED	GR
NULL	**314**	362	485	465	328
LINUP	**392**	452	454	438	432
LINDOWN	**359**	385	837	809	386
SINE	**994**	1,002	1,849	1,796	1,395
EXP	**422**	444	529	522	815
ESS	176	**160**	181	178	189

errors. In addition to the 10% sampling rate, we simulate 5% and 20% sampling rates for the six mean structures. The comparisons among the RMSE of P0_15, P1_15, HT, HT PRED and GR are shown in Tables 7 and 8.

Finally we simulate data with heteroscedastic independent errors with variances $\pi_i^2\sigma^2$ where $\sigma = 1$ for mean structures NULL, LINUP, LINDOWN, SINE, and EXP with population size 1,000 and sample size 100. PPS samples are obtained the same way as described before. Comparisons are made among five estimators:

1. Constant variance p-spline model-based estimator P0_15;
2. Nonconstant variance p-spline model-based estimator P1_15;
3. The Horvitz-Thompson estimator HT;
4. The predictive HT estimator HT PRED;
5. The generalized regression estimator $\hat{T}_{GR} = \sum_{i\in P}\hat{y}_i + \sum_{i\in S}(y_i - \hat{y}_i)/\pi_i$.

Table 9 gives the simulation results for this case.

3.2. Results

HT PRED is generally quite similar to HT in these simulations, with slightly larger empirical bias and slightly smaller root mean squared error. From Tables 1–3, we see that the p-spline based estimators P0_5, P0_15, P1_5 and P1_15 are slightly more biased than the GR estimator in general. However, Tables 4–6 suggest that in terms of root mean squared error, the p-spline model-based estimators P0_5, P0_15, P1_5, P1_15 outperform the HT estimator when the mean structure is not linearly increasing or with nonzero intercept and perform about as well as the HT and GR estimators when the mean is linearly increasing without intercept (e.g., population LINUP). GR performs well when the mean structure is linear with or without intercept (e.g., populations NULL, LINUP, and LINDOWN). GR has larger RMSE than the spline model-based estimators when the

Table 8. Empirical RMSE ×10 of five estimators for $N = 1{,}000$, $n = 200$ (Minimum RMSE for each population is in bold print)

	P0_15	P1_15	HT	HT PRED	GR
Population	**117**	129	244	262	134
LINUP	173	178	174	**152**	180
LINDOWN	214	223	518	545	**208**
SINE	210	**184**	865	1,003	449
EXP	**165**	200	221	254	394
ESS	61	**59**	81	88	72

mean structure is not linear (populations SINE, EXP, and ESS). The above comparison holds for the populations simulated with normal or lognormal errors.

Despite assuming the wrong variance structure, P1_5 and P1_15 also outperform HT and HT PRED in populations NULL, LINDOWN, SINE, EXP, and ESS.

The PROJ2 estimator, which does not make use of the selection probabilities of non-sampled units, is comparable in performance to the HT estimator; hence gains in efficiency of the spline methods require knowledge of the selection probabilities for non-sampled units.

In Table 9, HT is seen to perform the best in the population LINUP, which is generated using the Horvitz-Thompson Model (1). The next best method is P1_15, which is nearly as good as HT in RMSE. In all other cases the spline estimates are much better than HT. The p-spline model with the correct variance structure (i.e., P1_15) outperforms the p-spline model with incorrectly specified variance structure (i.e., P0_15) in populations LINUP, LINDOWN, and SINE.

4. A Real Dataset

We demonstrate the p-spline model-based estimation on a dataset in Kish (1965). The data is a list of the 270 blocks in Ward 1 of Fall River, MA, from its volume of Block Statistics of the 1950 U.S. Census. Let variable Z_i represent the number of dwellings in block i and variable Y_i represent the number of dwellings occupied by renters in that block. We want to estimate the total number of dwellings occupied by renters in the 270 blocks.

We use the size variable X_i = ranking of Z_i among the blocks so that a larger value of Z_i corresponds to a larger value of X_i.

The following scatter plot shows the distribution of Y_i vs. X_i:

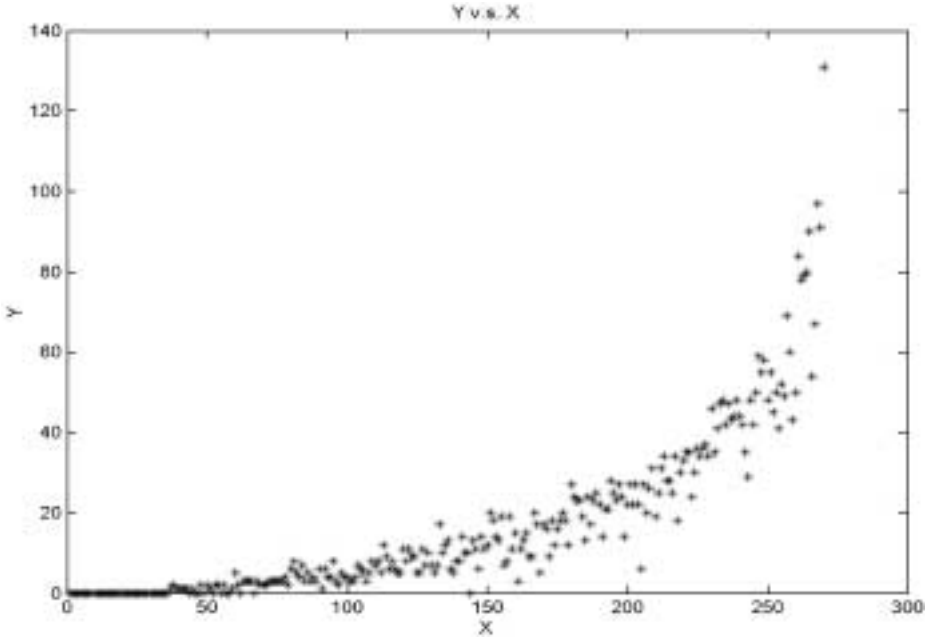

Table 9. Empirical RMSE ×10 of five estimators for data with heteroscedastic errors $N = 1{,}000$, $n = 100$ (Minimum RMSE for each population is in bold print)

Population	P0_15	P1_15	HT	HT PRED	GR
NULL	**81**	88	284	275	91
LINUP	74	67	**63**	64	72
LINDOWN	120	**98**	565	546	103
SINE	408	**270**	1,311	1,267	838
EXP	**106**	108	208	217	504

Clearly the relationship between outcome values and inclusion probabilities is not linear.

Five hundred repeated systematic samples of size 27 are drawn without replacement from random lists.

We apply the HT estimator and the P0_10 and P1_10 estimators on these samples (P0_10 assumes constant variance and P1_10 assumes variance proportional to π_i^2, both using 10 knots).

The true value of T is 4,559. The empirical biases of the three estimators are HT: 14.38, P0_10: -60.48 and P1_10: −34.48. The empirical root mean squared errors are: HT: 534.64, P0_10: 274.84 and P1_10: 285.31.

The advantage of using p-spline model-based estimators is obvious. The spline model-based estimators have small p-biases while having smaller root mean squared errors than the HT or parametric model-assisted estimator GR.

5. Discussion

Survey samplers are hesitant to use model-based prediction because of potential lack of robustness to modeling assumptions. However, prediction methods based on p-splines make relatively weak ''nonparametric'' assumptions, and are becoming more accessible, being readily fitted using widely available software packages. Our simulations show that such methods can yield large gains in mean squared error over the design-unbiased HT estimator when the data mean structure violates the Model (1) implied by the HT estimator, with little loss in efficiency when conditions favor the HT estimator. In general the p-spline nonparametric model fits the mean structure $E(Y_i|\pi_i)$ more flexibly than parametric models such as Model (1), and parametric models assumed when using generalized regression estimators. As shown in the simulation study, for moderate population sizes and moderate sampling rates, the gain in precision more than offsets the slight increase in design bias in the simulated populations NULL, LINDOWN, EXP, SINE, and ESS.

A plot of sampled y_i vs π_i can aid in choosing estimators of T. If the plot shows a no-intercept linear relationship, the HT estimator works well. If the relationship is non-linear, we believe that spline-based estimators are more appropriate. The p-spline estimates in our simulations worked well using simple untailored methods of knot placement, but their performance might be further enhanced by a more careful choice of the number and placement of knots.

An interesting finding from our simulations is that the gains of the model-based prediction estimators were not realized by the model-based projection estimator PROJ2, which performed very much like the HT estimator. This finding suggests that gains from

modeling require prediction for the nonsampled cases, and hence knowledge of the selection probabilities for these cases. The selection probabilities of nonsampled cases are known for systematic PPS design under study, but they are often not included as part of the data file for estimation. Hence transmission of this information to the data user is important if the gains in efficiency of modeling are to be realized.

An important issue not discussed here is inference. That is, variance estimation and coverage of confidence intervals for the population total. Such issues are discussed in Zheng and Little (2002). Simulations show that the p-spline estimators with their variances estimated using jackknife techniques yield better design-based inference than alternatives based on the HT estimator. In a future study we plan to make extensions of p-spline methods to be applied for nonnormal or binary outcomes, and extensions to multistage designs involving nonconstant sampling probabilities at more than one stage of selection.

6. References

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations Using Nonparametric Calibration. Journal of the American Statistical Association, 88, 268–277.

Denison, D.G.T., Mallick, B.K., and Smith, F.M. (1998). Automatic Bayesian Curve Fitting. Journal of the Royal Statistical Society, Series B, 60, 333–350.

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

Dorfman, A.H. (1992). Non-parametric Regression for Estimating Totals in Finite Populations. Proceedings of the American Statistical Association, Section on Survey Research Methods, 622–625.

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

Elliott, M.R. and Little, R.J.A. (2000). Model Based Alternatives to Bayesian Trimming Survey Weights. Journal of Official Statistics, 16, 191–209.

Eubank, R.L. (1988). Spline Smoothing and Nonparametric Regression. New York and Basel: Marcel Dekker.

Firth, D. and Bennett, K.E. (1998). Robust Models in Probability Sampling. Journal of the Royal Statistical Society, Series B, 60, 3–21.

Friedman, J.H. and Silverman, B.W. (1989). Flexible Parsimonious Smoothing and Additive Modeling. Technometrics, 31, 3–21.

Friedman, J.H. (1991). Multivariate Adaptive Regression Splines (with discussion). Annals of Statistics, 19, 1–141.

Green, P.J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika, 82, 711–732.

Hastie, T.J. and Tibshirani, R. (1990). Generalized Additive Models. London: Chapman & Hall.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Little, R.J.A. (1983). Estimating a Finite Population Mean from Unequal Probability Samples. Journal of the American Statistical Association, 78, 596–604.

Pinheiro, J.C. and Bates, D.M. (2000). Mixed-Effects Models in S and S-PLUS, Springer: New York.

Ruppert, D. and Carroll R.J. (2000). Spatially Adaptive Penalties for Spline Fitting. Australia and New Zealand Journal of Statistics, 42, 205–223.

Särndal, C.E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. Springer Verlag.

SAS (1992). The Mixed Procedure, in SAS/STAT Software: Changes and Enhancements, Release 6.07. Technical Report P-229, SAS Institute, Inc., Cary, NC.

Stone, C.J., Hansen, M., Kooperberg, C., and Truong, Y.K. (1997). Polynomial Splines and Their Tensor Products in Extended Linear Modeling (with discussion). Annals of Statistics, 25, 1371–1470.

Wahba, G. (1990). A Comparison of GCV and GML for Choosing the Smoothing Parameters in the Generalized Spline Smoothing Problem. Annals of Statistics, 4, 1378–1402.

Zheng, H. and Little, R.J.A. (2002). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. Forthcoming.

Received July 2001
Revised October 2002