Journal of Official Statistics, Vol. 21, No. 1, 2005, pp. 1–20

Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model

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

Inference about the finite population total from probability-proportional-to-size (PPS) samples is considered. In a previous article (Zheng and Little 2003), penalized spline (*p*-spline) nonparametric model-based estimators were shown to generally outperform the Horvitz-Thompson (HT) and generalized regression (GR) estimators in terms of the root mean squared error. In this article we develop model-based, jackknife and balanced repeated replicate variance estimation methods for the *p*-spline based estimators. Asymptotic properties of the jackknife method are discussed. Simulations show that *p*-spline point estimators and their jackknife standard errors lead to inferences that are superior to HT or GR based inferences. This suggests that nonparametric model-based prediction approaches can be successfully applied in the finite population setting by avoiding strong parametric assumptions.

Key words: Jackknife; balanced repeated replication; Horvitz-Thompson estimator; sampling weights; variance estimation.

1. Introduction

Survey sampling is perhaps unique in being the only area of current statistical activity where inferences are primarily based on the randomization distribution rather than on statistical models for the survey outcomes. This so-called design-based approach to survey inference is described in standard survey texts such as Hansen, Hurwitz, and Madow (1953), Kish (1965) and Cochran (1977). For a population with N units, let $Y = (Y_1, \ldots, Y_N)^T$, the vector of survey variables for unit i, and let $I = (I_1, \ldots, I_N)^T$ denote the vector of *inclusion indicator variables,* where $I_i = 1$ if unit i is included in the sample and $I_i = 0$ if it is not included. The main characteristic of design-based inference is that it is based on the distribution of I, with the survey variables Y treated as fixed quantities.

The model-based approach to survey sampling inference posits a model for the survey outcomes Y, which is then used to predict the nonsampled values of the population, and hence finite population quantities Q. There are two variants of the modeling approach: superpopulation modeling and Bayesian modeling. In superpopulation modeling

[1] Partners AIDS Research Center and Harvard Medical School, 90 Staniford Street 9th Floor, Boston, MA 02114, U.S.A. Email: hzheng1@partners.org
[2] Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A. Email: rlittle@umich.edu.
Acknowledgements: This research was supported by grant DMS 0106914 from the National Science Foundation. We thank the associate editor and three referees whose valuable comments led to many improvements to this article.

(Brewer 1963; Royall 1970; Valliant, Dorfman, and Royall 2000), the population values of Y are assumed to be a random sample from a "superpopulation," and assigned a probability distribution $p(Y|\theta)$ indexed by fixed parameters θ. The Bayesian approach (Ericson 1969; Rubin 1987; Ghosh and Meeden 1997) adds a prior for the parameters and bases inference for finite population quantities on the posterior predictive distribution of Y. In general, inferences under either variant are based on the joint distribution of Y and I. However, in probability sampling, where the distribution of I given Y does not depend on the values of Y after conditioning on survey design variables, inferences can be based on the distribution of Y alone provided the design variables are included in the model (Rubin 1987).

An advantage of the model-based approach is that it provides a unified approach to survey inference, aligned with mainline statistics approaches in other application areas such as biostatistics and econometrics. Also, the Bayesian variant may yield better inferences for small sample problems where exact frequentist solutions are not available, by propagating error in estimating parameters. Model-based inferences will generally outperform design-based inferences if the model is correctly specified. However, all models are simplifications and hence subject to misspecification error. The major drawback with model-based inference is that if the model is seriously misspecified it can lead to inferences that are worse (and potentially much worse) than design-based inferences (Hansen, Madow, and Tepping 1983; Holt, Smith, and Winter 1980; Pfeffermann and Holmes 1985). A key to robust models for sample surveys is to account for aspects of the survey design, such as stratification, clustering and weighting. In this article we focus on survey weights, a particularly interesting survey design feature since it is handled somewhat differently by the model- and design-based paradigms.

Specifically, we consider the case of sampling with probability proportional to size (PPS), where a size measure X is known for all units in the population, and unit i is selected with probability π_i proportional to its size x_i. PPS samples can be selected in a number of ways that lead to different joint selection probabilities for pairs of units (Hanif and Brewer 1980). We consider here the practical and common fixed sample size design. From a random starting point, units are selected systematically from a randomly ordered list, at regular intervals on a scale of cumulated sizes (Kish 1965, Ch. 7); units that would be selected with probability one are moved into a certainty stratum. We consider statistical inference for the finite population total T of a continuous outcome Y; our methods can be modified to handle ordinal or nominal outcomes.

The standard design-based approach to PPS samples is to weight sampled units by the inverse of their probability of selection, yielding the Horvitz-Thompson (HT) estimator

$$\hat{T}_{HT} = \sum_{i=1}^{n} y_i/\pi_i \tag{1}$$

(Horvitz and Thompson 1952), where the summation is over n sampled units. This is also the projective estimator (Firth and Bennett 1998) for an "HT model," where y_i given π_i is assumed to have mean $\beta\pi_i$ and variance $\sigma^2\pi_i^2$. It is well known that the HT estimator is design unbiased, but can be inefficient when the "HT model" is not a good approximation to reality. A parody of this situation is the famous "circus elephant" example in Basu (1971).

Modelers who ignore the design weights do so at their peril: results are highly vulnerable to model misspecification. However, a number of authors (Rubin 1983;

Little 1983a) have argued that from a modeling perspective, the weights should be used as predictors in a model rather than used to weight the sampled cases. In the case of PPS sampling, this suggests basing inferences on the predictions of a regression model relating Y to X. Recently, several authors have argued for models in survey settings that make relatively weak assumptions of the form of the relationship, since sample sizes are often large and strong models are viewed with skepticism. In particular, Dorfman (1992) and Chambers, Dorfman, and Wehrly (1993) estimate a finite population total by a nonparametric model relating Y to an auxiliary variable, using kernel smoothing. Rizzo (1992) discusses inference of finite population quantities conditioned on the selection probabilities. Breidt and Opsomer (2000) use the local polynomial kernel as the smoothing tool and develop a design-consistent model-assisted estimator of the total.

A modification of the prediction approach is to base the estimate of T on predictions from a model, but then adjust the estimator to achieve design consistency. In particular, generalized regression estimators (GR) achieve this by adding a calibration term consisting of a design-weighted sum of residuals to the predictions $\hat{y}_i$ from the model:

$$\hat{T}_{GR} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{n} (y_i - \hat{y}_i)/\pi_i \tag{2}$$

This estimator is design consistent for the total, and more efficient than the HT estimator if the auxiliary variables are good predictors of Y. For discussions of this "model assisted" approach, see Särndal, Swensson, and Wretman (1989; 1992).

Some have argued that the calibration correction in (2) is unnecessary if the model is chosen so that the prediction or projection estimator is design consistent, a condition that is relatively easy to achieve (Little 1983b; Firth and Bennett 1998). Zheng and Little (2003) compared prediction estimates of the population total based on p-splines with the HT and the GR estimates. These simulations, which are briefly summarized in Section 4, indicate that nonparametric models lead to prediction estimators of T with negligible bias and improved efficiency over HT or GR estimators, for a wide range of simulated populations.

Even if the spline-based prediction estimators were more efficient than design-based competitors, the latter might still be preferred if they yielded better inferences, that is had better confidence coverage, or tests closer to their nominal significance levels. Hence, the goal of the current article is to consider variance estimation and inference properties of the estimators compared in Zheng and Little (2003). A variety of approaches to variance estimation, based on the information matrix, balanced repeated replication and the jackknife, are considered for both the spline-based estimator and competitors. A simulation study indicates that the spline-based estimator is not only more efficient, but yields design-based inferences that are as good as, or better than, inferences based on the HT and GR estimators. We view this as further evidence that a model-based prediction approach can be successfully applied in the finite population setting, providing strong parametric assumptions are avoided and attention is paid to modeling the features of the survey design.

The rest of the article is organized as follows. In Section 2 we describe penalized spline model-based point estimation and three associated variance estimators. In Section 3 we present a simulation study that compares inferences under the various approaches for a variety of simulated populations and situations. Conclusions and suggestions for future work are presented in Section 4.

2. Inference About a Finite Population Total Based on Penalized Spline Model

2.1. Penalized spline model-based estimation

A model-based alternative to HT given by Zheng and Little (2003) predicts nonsampled values of $y_i, i \in P - S$ using the following penalized spline (Ruppert, Wand, and Carroll 2003) nonparametric regression model:

$$y_i = f(\pi_i, \beta) + \varepsilon_i, \quad \varepsilon_i \sim \text{ind } N(0, \pi_i^{2k}\sigma^2) \tag{3}$$

where the function f is a spline:

$$f(\pi_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{l=1}^{m} \beta_{l+p}(\pi_i - \kappa_l)_+^p, i = 1, \ldots, N \tag{4}$$

Here $k \geq 0$ is a constant reflecting the knowledge of error variance and the constants $\kappa_1 < \ldots < \kappa_m$ are selected fixed knots, and $(u)_+^p = u^p$ if $u > 0$ and 0, otherwise. In the spirit of Ruppert and Carroll (2000), Ruppert (2002), and Ruppert, Wand, and Carroll (2003), we favor a modeling strategy that places a large number of knots (for example, 15 or 30) at prespecified locations, and then achieves smoothness by treating $\beta_{p+1}, \ldots, \beta_{p+m}$ as random effects centered at 0. The least squares estimation of coefficients $\beta_{p+1}, \ldots, \beta_{p+m}$ tends to result in over-fitting. To penalize the roughness of the regression function f, a penalization term $\alpha\Sigma_{l=1}^m \beta_{l+p}^2$ is added to the log-likelihood. This is equivalent to giving $(\beta_{p+1}, \ldots, \beta_{p+m})^T$ a normal prior $N_m(0, \tau^2 I_m)$, where $\tau^2 = \sigma^2/\alpha$. The degree of smoothing is then based empirically on the estimate of the variance ratio $\alpha = \sigma^2/\tau^2$. Assuming constant error variance (that is, $k = 0$), the maximum likelihood (ML) estimate of the regression parameters conditional on $\alpha = \sigma^2/\tau^2$ is

$$(\hat{\beta}_0, \ldots, \hat{\beta}_{m+p})^T = (\Pi^{*T}\Pi^* + D(\alpha))^{-1}\Pi^{*T}Y^* = (\Pi^T W \Pi + D(\alpha))^{-1}\Pi^T WY \tag{5}$$

where $Y = (y_1, \ldots, y_n)^T$, the ith row of Π is $\Pi_i = (1, \pi_{i,} \ldots, \pi_i^p, (\pi_i - \kappa_1)_+^p, \ldots, (\pi_i - \kappa_m)_+^p)$, the matrix $D(\alpha)$ is diagonal with first $p+1$ elements equal to 0 and remaining m elements equal to $\alpha = \sigma^2/\tau^2$, $W = diag(\pi_1^{-2k}, \pi_2^{-2k}, \ldots, \pi_n^{-2k})$, $\Pi^* = W^{1/2}\Pi$ and $Y^* = W^{1/2}Y$. For the constant variance model $k = 0$, $W = I$ and $\Pi^* = \Pi$.

For unknown σ^2 and τ^2, restricted maximum likelihood (REML) estimates of β are obtained by replacing $D(\alpha)$ in (5) by $D(\hat{\alpha})$, where $\hat{\alpha} = \hat{\sigma}^2/\hat{\tau}^2$ and $\hat{\sigma}^2$ and $\hat{\tau}^2$ are REML estimates of σ^2 and τ^2. We consider the predictive estimator of the total based on this model

$$\hat{T}_{PRED} = \sum_{i=1}^{n} y_i + \sum_{i=n+1}^{N} \hat{E}(Y_i|\pi_i) \tag{6}$$

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

$$\hat{T}_{PROJ} = \sum_{i=1}^{N} \hat{E}(Y_i|\pi_i) \tag{7}$$

is also considered by some survey samplers, but makes less sense from a model-based perspective.

2.2. *Model-based variance estimation*

The empirical Bayes posterior variance of β in (3), when conditioned on $\hat{\sigma}^2$ and $\hat{\alpha} = \hat{\sigma}^2/\hat{\tau}^2$, is $\hat{\sigma}^2\{\Pi^{*T}\Pi^* + D(\hat{\alpha})\}^{-1}$. It follows that the estimated variance for the projective estimator is

$$Var(\hat{T}_{PROJ}) = \hat{\sigma}^2 1_N^T \Pi_P^T \{\Pi^{*T}\Pi^* + D(\hat{\alpha})\}^{-1} \Pi_P 1_N \tag{8}$$

where 1_N is an $(N \times 1)$ vector with elements equal to 1 and Π_P is the analogous quantity to Π for the whole population P. The empirical Bayes posterior variance for the predictive estimator is

$$Var(\hat{T}_{PRED}) = \hat{\sigma}^2 1_{N-n}^T \Pi_{P-S}^T \{\Pi^{*T}\Pi^* + D(\hat{\alpha})\}^{-1} \Pi_{P-S} 1_{N-n} \tag{9}$$

where 1_{N-n} is an (N-n) by 1 vector of elements equal 1 and Π_{P-S} is the analogous quantity to Π for the nonsampled population P-S. The estimates (6) and (7) and associated variance estimates (8) and (9) can be computed with standard software such as SAS Proc Mixed and S-plus function lme. Since the empirical Bayes posterior variance does not incorporate the variability in estimating the variance components, it tends to underestimate the variance of the point estimator. However, the underestimation is not big enough to seriously bias the variance estimation. This assessment was mentioned in Ruppert and Carroll (2000) and is empirically validated in our simulation study.

2.3. *Replication-based variance estimation methods*

The variance estimators (8) and (9) rely on model assumptions, and might fail when the model (specifically, the assumed variance structure) is incorrect. In this section we propose replication-based methods that are less reliant on the model and hence are more consistent with design-based perspectives.

2.3.1. The jackknife method

Originally introduced by Quenouille (1949), the jackknife method is a broadly useful method for both finite and infinite population inference (Shao and Wu 1989).

The jackknife method involves the following procedure. The sample S is divided into G subgroups with equal number of units and the gth pseudovalue is computed as $\hat{T}_g = G\hat{T} - (G-1)\hat{T}_{(g)}$, where $\hat{T}$ is the original p-spline model-based estimator and $\hat{T}_{(g)}$ is the same estimator calculated from the reduced sample not including the elements in the gth subgroup.

The jackknife variance estimate of $\hat{T}$ is

$$v(\hat{T}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{T}_g - \hat{\bar{T}})^2 \tag{10}$$

where $\hat{\bar{T}} = \Sigma_{g=1}^{G} \hat{T}_g/G$. In order to balance the distribution of the selection probabilities across the subgroups, sampled units are stratified into n/G strata each of size G with similar values of π_i, and the G subgroups are then constructed by removing one element at a time without replacement from each stratum.

We suggest the following strategy for choosing the smoothing parameter $\hat{\alpha} = \hat{\sigma}^2/\hat{\tau}^2$. When the sample size is not large and repeatedly fitting the p-spline model does not require an unreasonably large amount of computation, $\hat{\alpha}$ should be estimated for each replicate sample when computing the replicate estimates. When the sample size is very large, the portion of observations removed in each replicate is not large, and repeated computing of $\hat{\alpha}$ becomes burdensome, we suggest that $\hat{\alpha}$ not be recomputed for each replicate sample. That is, we compute pseudovalues as $\beta^T_{(g)} = (\Pi^T_{(g)}\Pi_{(g)} + D(\hat{\alpha}))^{-1}\Pi^T_{(g)}Y$, where $\Pi_{(g)}$ is constructed in the same way as Π but omitting the gth subgroup, but the estimate $\hat{\alpha}$ is computed for the full sample. Asymptotic theories similar to that in the Appendix imply that doing so will not hurt the consistency of the variance estimator.

Miller (1974) proved the asymptotic properties of the jackknife estimator in the case of multiple regression. In the sample survey setting, Shao and Wu (1987; 1989) discussed the properties of jackknife variance estimation in linear regression models. In our case, the p-spline regression is a form of ridge regression conditioned on $\hat{\alpha}$. If the p-spline is a low dimensional smoother, that is, the dimension of the "design matrix" Π^* is small compared with the sample size n, then the jackknife method has asymptotic properties similar to linear regression. In the Appendix, we give a brief proof of the asymptotic consistency of the jackknife variance estimator in the delete-one case and under regularity conditions similar to those in Miller (1974). Simulations in Section 3 study the performance of the jackknife method for moderate-sized samples.

2.3.2. The balanced repeated replicate method

The BRR method was developed for stratified designs with two units sampled in each stratum. It is the most computationally efficient technique when the half-samples are fully balanced. In practical applications of BRR, clusters (PSUs or small strata) are often grouped into pairs, and units within large strata are randomly split.

The systematic PPS design can be viewed approximately as a stratified design with n strata each consisting of units with cumulative measures of approximate size $\Sigma_{i=1}^N x_i/n$. One unit is sampled from each of the n strata. Assuming n is even, the design can also be approximated by a stratified design with $n/2$ strata with cumulative measures of size $2\Sigma_{i=1}^N x_i/n$, and two units are sampled per stratum. Balanced repeated half-samples are then constructed by selecting one unit from each stratum, with the selection scheme based on Hadamard matrices (Plackett and Burman 1946). Let $\hat{T}_b$ be the p-spline estimator computed from the bth half-sample, using the same knots as used in the computation using the full sample – the number and placement of knots needs to allow the spline model to be fitted on each half-sample. The BRR estimator is then given by

$$v_{BRR}(\hat{T}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{T}_b - \hat{T})^2 \tag{11}$$

For smoothing parameter $\hat{\alpha}$, we suggest a replicate estimate be computed for each replicate sample.

To construct half-samples, pairs of units are selected according to the sequence in which they are selected from the random list. This BRR method with two units sampled per stratum does not fully reflect the improved efficiency from the systematic PPS sampling

method, and hence can be expected to overestimate the true variance of the p-spline estimator. This conjecture is consistent with simulation results reported in the next section.

3. Simulation Study

3.1. The simulated populations

Artificial populations are simulated according to six different mean functions relating to outcome y_i and the inclusion probabilities π_i. We simulate the inclusion probabilities as proportional to consecutive integer values: 11, 12, . . . , 310 for $N = 300$; 35, 36, . . . , 1,034 for $N = 1,000$; and 71, 72, . . . , 2,070 for $N = 2,000$.

Five of the simulated populations are generated by adding independent errors with variance 0.2 to the following mean functions:

(NULL) $f(\pi_i) \equiv 0.30$

(LINUP) $f(\pi_i) = 3\pi_i$, linearly increasing function with a zero intercept

(LINDOWN) $f(\pi_i) = 0.58 - 3\pi_i$, linearly decreasing function with positive intercept

(EXP) $f(\pi_i) = \exp(-4.64 + 26\pi_i)$, an exponentially increasing function

(SINE) $f(\pi_i) = \sin(35.69\pi_i)$.

A sixth population is generated to yield an "S" shaped function:

(ESS) $y_i = 0.6\,\text{logit}^{-1}(50^*\,\pi_i - 5 + \varepsilon_i), \varepsilon_i \overset{iid}{\sim} N(0, 1)$.

Since the errors in ESS lie inside the logit function, this population has heteroscedastic errors. Plots of samples from these populations are provided in Figure 1. Population sizes 300, 1,000, and 2,000 with respective sample sizes 32, 96, and 192 are simulated for each mean function. For each simulated population, 1,000 repeated samples are drawn using the systematic PPS sampling design. Numerical comparisons of various methods are all based on the empirical results from the repeated samples.

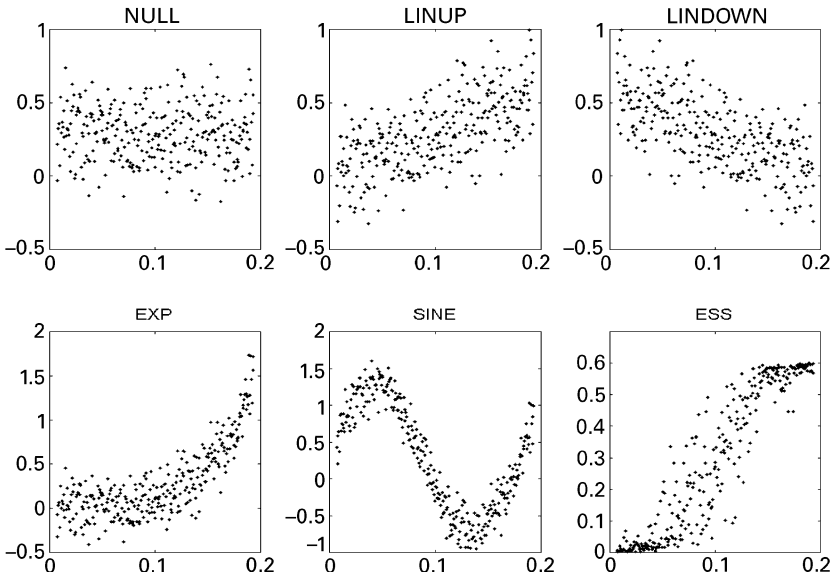

Fig. 1. Six simulated populations ($N = 300$) X-axis: pi(i); Y-axis: y(i) with normal errors

3.2. Bias and mean squared error of alternative point estimators

A detailed discussion of bias and mean squared error properties of the p-spline, HT, and GR estimators is presented elsewhere (Zheng and Little 2003). We illustrate those findings in Table 1, which presents empirical bias and root mean squared error (RMSE) of point estimates from the following methods:

(a) P0_15, a *p*-spline prediction estimator (6) with $k = 0$ and the knots selected to be 15 equal sample percentiles of π_i and with degree of spline $p = 1$.
(b) HT, the Horvitz-Thompson estimator (1).
(c) GR, a generalized regression estimator (2) assisted by a simple linear with intercept regression model that regresses y_i on π_i, assuming a constant error variance.

For each of the six mean structures described in Section 3.1, the estimates were computed for 500 systematic PPS samples of size 96. Table 1 suggests that P0_15 has smaller empirical RMSE than HT or GR for the populations with nonlinear mean structures (SINE EXP and ESS). P0_15 has similar RMSE as GR when the mean function is linear (NULL, LINUP and LINDOWN). P0_15 has similar RMSE as HT for the population with a linearly increasing mean function without an intercept (LINUP), which favors HT. The empirical bias of P0_15 is small and in most cases P0_15 has empirical bias comparable to HT and GR. Similar findings are presented in the more extensive simulations in Zheng and Little (2003).

3.3. Variance estimation for alternative methods

In this section we compare the inferences for the *p*-spline prediction and projection estimators, the variances being estimated by (8)–(11), with inferences based on the HT and GR estimators. For HT, we show results for two variance estimation methods:

A) the random groups variance estimator

$$v_{RG} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \left(\hat{T}_i - \hat{T}_{HT}\right)^2 \tag{12}$$

where the sample is divided into K random subsamples, each of size $m = n/K$, and $\hat{T}_i = \Sigma_{l=1}^{m} y_l/(mp_l)$ with $p_l = \pi_l/n$ is the HT estimator from the ith subsample;

Table 1. Empirical bias and root mean squared error of three point estimators: P0_15, HT and GR
$N = 1,000$, $n = 96$

	P0_15		HT		GR	
	Empirical bias	RMSE	Empirical bias	RMSE	Empirical bias	RMSE
NULL	0.27	21.79	−1.93	35.11	0.99	23.69
LINUP	3.24	25.89	1.49	27.32	−2.79	34.29
LINDOWN	0.87	26.71	2.04	63.29	−1.63	35.33
SINE	22.01	45.48	4.85	112.71	−3.63	94.61
EXP	0.15	27.39	1.09	34.74	−0.57	54.34
ESS	−4.41	10.22	0.82	11.20	0.92	30.24

B) the with-replacement PPS variance estimator

$$v_{WR} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \hat{T}_{HT} \right)^2 \tag{13}$$

which ignores the effect of sampling without replacement on the variance. This is also a model-based variance estimator for the projective estimator, assuming the "HT model."

We also considered three other variance estimators suggested in Wolter (1985), a Yates-Grundy estimator with joint inclusion probabilities approximated as in Hartley and Rao (1962), a paired units estimator and a consecutive differences estimator. These did less well in our simulations, and their performance is reported elsewhere (Zheng 2003).

For GR, we apply the formula given by Särndal et al. (1989) for a regression on a covariate X:

$$v_{GR} = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\Delta_{kl}}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l} \tag{14}$$

where

$$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l, \pi_{kk} = \pi_k, g_k = 1 + \left(\sum_{k=1}^{N} x_k^T - \sum_{i=1}^{n} \frac{x_k^T}{\pi_k} \right) \left(\sum_{k=1}^{n} \frac{x_k x_k^T}{\pi_k} \right)^{-1} x_k$$

$$e_k = y_k - \hat{y}_k, \quad k = 1. \; . \; .n$$

where the covariate is $x_k = [1 \; \pi_k]^T$.

We use the Hartley-Rao approximation

$$\pi_{ij} = \frac{n-1}{n} \pi_i \pi_j + \frac{n-1}{n^2} \left(\pi_i^2 \pi_j + \pi_j^2 \pi_i \right) - \frac{n-1}{n^3} \pi_i \pi_j \sum_{k=1}^{N} \pi_k^2$$

for estimating the pairwise joint inclusion probabilities. This approximation for the joint inclusion probability is valid when $\max(\{\pi_i, i = 1. \; . \; .n\}) = O(N^{-1})$, which is satisfied by our simulation sampling design.

First, 1,000 repeated PPS samples are drawn from each artificial population using the systematic sampling method. For each repeated sample, the proposed inference method (p-spline point estimation and empirical Bayes, JRR, and BRR variance estimators) as well as inference methods associated with HT and GR are computed. The coverages of these inference methods are then compared on the basis of their empirical performances.

Next, we consider the robustness of the model-based and replication-based methods in the presence of misspecification of the variance structure, by assessing their performance for populations with heteroscedastic errors. We apply the total estimator P0_15, which assumes constant error variance, on two groups of populations. The first group of populations is generated with constant-variance error and the second group generated with the same mean structure as the first group but with error variances proportional to π_i^2.

Thus, P0_15 assumes the correct error variance for the first group while it misspecifies the error variance for the second group.

Last, we study how the number of knots influences the coverage in population SINE, whose mean function requires more knots than the other populations. We study the relationship between the coverage of 95% C.I. and the number of knots employed.

4. Results

Table 2 gives a comparison of six variance estimators in terms of the mean estimates of the variance. The six variance estimators are: v_{RG} and v_{WR} for HT; design-based variance estimator for GR; and empirical Bayes, JRR, and BRR for P0_15. The empirical variances of HT, GR, and P0_15 are also listed in Table 2. The averages of the two variance estimators for HT track the empirical mean squared errors reasonably well, particularly for the larger sample sizes. This table also suggests that the design-based estimator for the variance of GR can seriously underestimate the variance for small to moderate-size samples.

For populations other than SINE and for the two larger sample sizes ($n = 96$ and 192), the average estimated variances from the jackknife and empirical Bayes methods track the empirical mean squared errors well, and the BRR method tends to yield conservative estimated variances. For the small sample size ($n = 32$) and populations other than SINE, the empirical Bayes variance tends to underestimate the variances of the *p*-spline point estimators for populations other than ESS, and to overestimate the variance for the population ESS, perhaps because the variance structure for that population is hetereoscedastic and hence misspecified by the model; the jackknife and BRR methods tend to have upward biases for these cases. For the SINE population the average of the empirical Bayes variance estimates seriously underestimates the empirical mean squared error. As discussed later, this finding appears to reflect the fact that there are not enough knots in the *p*-spline regression to estimate the SINE function well for these populations. The jackknife and BRR methods overestimate the variance for the SINE population, the BBR method severely so. Besides the replication methods' conservative tendency, the estimated smoothing parameters also contribute to the overestimation by the BRR.

In Table 3, three inference approaches are compared: HT with the random groups variance estimator (9), GR with the design-based variance estimator (13), and the *p*-spline with the jackknife variance estimator. From this table, it is clear that the *p*-spline method gives confidence intervals that are shorter than those given by the HT method when the mean function is not linear-with-no-intercept. It also gives C.I.s that are shorter than those from the GR method when the mean function is not linear. When the data are in favor of HT or GR, p-spline based inferences yield comparable coverage. With the exception of population SINE, the p-spline method generates C.I.s with satisfactory coverage rates for the simulated populations. There is some undercoverage by the C.I.s from the HT method for the populations NULL and LINDOWN, which seriously violate the "HT model" assumption. In terms of coverage rate, the C.I.s given by the GR method are quite unsatisfactory for small (32) to moderate (96) sample sizes and only become better for a large sample size (192).

Table 2. Empirical variance of HT, GR and spline estimates, and means of associated variance estimators. Values are presented as ratios to the empirical variance of the HT estimator for the associated population

	Population	Horvitz-Thompson			GR		*P*-spline(P0_15) Predictive estimator			BRR
		Empirical variance	Mean v_{RG} ($K = 10$)	Mean v_{WR}	Empirical variance	Mean v_{GR}	Empirical variance	Empirical Bayes	Jackknife ($G = 10$)	
	NULL	1	1.11	1.11	0.45	0.32	0.38	0.40	0.49	0.49
	LINUP	1	1.14	1.05	1.76	1.26	0.80	0.91	1.04	1.06
$N = 300$	LINDOWN	1	0.90	0.89	0.28	0.19	0.17	0.13	0.20	0.19
$n = 32$	SINE	1	1.13	1.13	0.86	0.55	0.27	0.11	0.41	0.44
	EXP	1	1.12	1.09	3.09	2.22	0.79	0.92	1.07	1.32
	ESS	1	1.52	1.06	7.06	5.65	1.26	1.03	1.65	2.29
	NULL	1	1.09	1.09	0.46	0.43	0.39	0.41	0.45	0.47
	LINUP	1	1.17	1.12	1.57	1.37	0.89	0.77	0.91	0.94
$N = 1,000$	LINDOWN	1	1.03	1.02	0.31	0.26	0.18	0.15	0.16	0.16
$n = 96$	SINE	1	1.10	1.11	0.70	0.60	0.12	0.06	0.15	0.32
	EXP	1	1.10	1.06	2.45	2.09	0.62	0.60	0.64	0.79
	ESS	1	1.27	1.06	7.31	6.29	0.68	0.91	0.80	1.13
	NULL	1	0.96	0.93	0.47	0.43	0.38	0.39	0.39	0.41
	LINUP	1	1.11	1.13	1.68	1.61	0.82	0.81	0.89	0.92
$N = 2,000$	LINDOWN	1	0.99	1.01	0.31	0.29	0.16	0.15	0.16	0.16
$n = 192$	SINE	1	1.13	1.13	0.84	0.71	0.09	0.05	0.11	0.21
	EXP	1	1.01	1.01	2.60	2.42	0.54	0.56	0.58	0.65
	ESS	1	1.10	1.05	7.53	6.86	0.57	0.79	0.58	0.74

Table 3. Comparison of three approaches to inference: HT with Random Group estimator and C.I. constructed with 9 degrees of freedom; GR with Yates-Grundy estimator; P-spline with Jackknife estimator. Average width (A.W.) of 95% CI and coverage rate (%) of 95% C.I.

	Population	HT with Random Group method ($K = 10$, $df = 9$)		GR with Yates-Grundy method		P0_15 with Jackknife method	
		A.W.	%	A.W.	%	A.W.	%
$N = 300$, $n = 32$	NULL	68	89	40	88	48	95
	LINUP	48	98	53	87	47	96
	LINDOWN	98	82	52	87	51	95
	SINE	223	92	161	84	114	80
	EXP	63	95	89	86	57	94
	ESS	26	97	51	89	24	95
$N = 1,000$, $n = 96$	NULL	131	93	88	92	89	97
	LINUP	109	96	123	94	98	95
	LINDOWN	230	92	124	92	94	94
	SINE	446	94	340	93	145	91
	EXP	135	96	193	90	105	95
	ESS	48	98	109	92	37	93
$N = 2,000$, $n = 192$	NULL	196	92	137	95	129	96
	LINUP	142	97	178	95	130	97
	LINDOWN	317	92	180	94	129	94
	SINE	611	94	497	92	182	93
	EXP	184	95	289	93	138	95
	ESS	63	97	161	92	45	95

For the SINE population, the coverage rates of the C.I.s corresponding to the three variance estimators for the p-spline with 15 knots are unsatisfactory. Figure 2 displays these coverage rates as a function of the number of knots, and indicates that for this population at least 30 knots are needed for valid inference. This figure also shows that the jackknife method has quite robust coverage, while the BRR method tends to be conservative and yields 95% confidence intervals that over-cover the population quantity.

Table 4 provides more information on the effect of misspecification of the variance structure. We compare model-based and jackknife variance estimators of P0_15, which corresponds to a *p*-spline with 15 knots and assuming constant error variance, on populations with homoscedastic and heteroscedastic errors with variance proportional to π_i^2. This table suggests that the model-based variance estimator is sensitive to misspecification of the variance structure while the jackknife method is more robust to this type of misspecification.

5. Discussion

The HT estimator is design-unbiased, and it can be used with an appropriate variance estimator to yield valid large-sample inferences. However, its efficiency and its

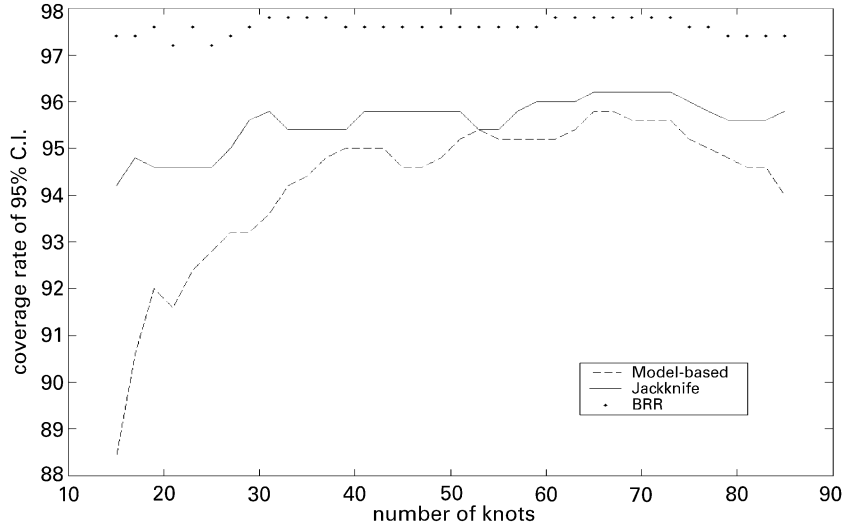

Fig. 2. Coverage rate (percentage) of 95% C.I. vs number of knots for population SINE $N = 2,000$, $n = 192$, Coverage rate computed from 500 repeated samples (target $= 93-97\%$)

performance in moderate-sized samples depend on the validity of the underlying "HT model." The GR estimator can yield increases in efficiency but is sub-optimal if based on a poorly chosen model and can yield anti-conservative inferences in moderate-sized samples. We note that coverage of GR-based inferences may be improved by using usually more conservative variance estimators such as the jackknife estimator. However, using an inflexible mean function can be expected to result in wide confidence intervals for populations with nonlinear mean structures, such as ESS and EXP.

Our proposed nonparametric model based on *p*-splines assumes a more flexible mean structure than that implied by the HT or GR model. Since these models give a close approximation to the mean function, bias calibration as in Equation (2) is either automatic (Zheng and Little 2003) or unnecessary. In particular, Breidt, Claeskens, and Opsomer (2004) propose the GR estimator with the linear model replaced by a penalized spline on covariates. We expect this approach (with the selection probabilities as

Table 4. Inferences based on P0_15 and model-based and jackknife variance estimators, when the variance is correctly and incorrectly specified. Coverage rate (%) of 95% C.I. and relative bias, the ratio of empirical bias to the empirical variance. ($N = 1,000$, $n = 100$)

Population	Variance structure incorrectly specified				Variance structure correctly specified			
	Model based		Jackknife		Model based		Jackknife	
	Relative bias	%	Relative bias	%	Relative bias	%	Relative bias	%
NULL	1.64	99	0.20	96	−0.18	93	−0.06	93
LINUP	1.82	99	0.35	97	0.04	95	0.23	96
LINDOWN	1.50	99	0.17	97	−0.11	94	0.00	97
SINE	−0.52	82	0.32	88	0.02	95	0.13	96
EXP	2.54	100	0.15	97	0.00	95	0.05	94

covariate) will perform similarly to our method, but do not think the calibration component of the estimator is needed. The p-spline method with the jackknife variance estimate yields shorter confidence intervals than the design-based methods while achieving coverage rates that are superior to those of traditional methods. An exception is its performance in the SINE populations, where more than the chosen number of 15 knots is needed for inference. Our jackknife method might be improved by using adjustments of the type considered by Hinkley (1977); this remains a topic for future research.

The model-based empirical Bayes variance estimator is valid if the model is correctly specified. However, our simulations suggest that it is vulnerable to misspecification of the variance structure. One possible solution is to estimate parameters for the variance structure, such as the parameter k in Equation (3), from the data by parametric (e.g., likelihood-based) or nonparametric methods. Here we adopted the less efficient but simpler approach of fixing k and using a robust variance estimator based on the jackknife. We imagine more general methods for estimating the variance structure will be developed in the future.

The jackknife method of variance estimation worked well in our simulations, whereas the BRR method tended to yield conservative standard errors. The bootstrap might also be expected to work well if the bootstrapping was done in a way that balanced the distribution of the selection probabilities in the bootstrap samples.

Survey samplers favor simple estimation methods that can be applied to large samples in a production setting. Thus, we deliberately chose a relatively straightforward parametric approach to spline regression with fixed knots, which can be readily implemented with existing software. Our simulations suggested that this approach worked well in most cases, but yielded unsatisfactory confidence coverage in the SINE population when an insufficient number of knots were used. Numerous authors (Friedman and Silverman 1989; Friedman 1991; Stone et al. 1997; Denison, Mallick, and Smith 1998; Ruppert and Carroll 2000; Ruppert 2002) have proposed sophisticated knot selection methods that might be profitably applied to complex mean functions. An advantage of our parametric penalized spline approach is that it extends in an obvious way to nonnormal outcomes, by replacing the linear mixed model (3) – (4) by a generalized linear mixed model. Specifically, the normal specification in (3) could be replaced by a generalized linear model, with a canonical link relating the mean outcome to the spline function $f(\pi_i, \beta)$.

Many other useful extensions of the proposed approach can be envisaged. For example, here we have confined attention to regression on a single covariate, namely the selection probability. In practice, there may be a set of covariates $X_1, \ldots, X_p$ available to predict Y, and the selection probabilities may be determined by all or a subset of these variables, say $\pi = \pi(X_1, \ldots, X_q)$ for $q \leq p$. The relationship between Y and $X_1, \ldots, X_p$ might then be modeled by a p-dimensional spline, but for moderate values of p (say, three or more), this approach is vulnerable to the so-called "curse of dimensionality." We suggest in such settings modeling the relationship between Y and $\pi(X_1, \ldots, X_q)$ by a spline as described here, and including the other covariates parametrically. For example, linear additive terms in the other covariates might be added to the model defined by Equations (3) and (4), dropping one covariate to avoid multi-collinearity. Even if the linear additive terms in $X_1, \ldots, X_p$ in

this model are not correctly specified, the estimated finite population mean from this prediction model can be design-consistent provided the relationship with the selection probability is captured by the spline on the inclusion probability; the additional covariates may still reduce prediction errors and hence increase precision. His approach is discussed in the context of missing-data adjustments in Little and An (2004). We have focused here on modeling the relationship with the selection probability, without considering covariates, since this is the crucial relationship to specify correctly in the model. Another extension of our proposed approach, to multistage sampling, is discussed in Zheng and Little (2004).

In conclusion, we believe that p-spline models on the selection probabilities provide an attractive approach to survey inference based on probability-proportional-to-size samples.

Appendix

Asymptotic Consistency of the Jackknife Variance Estimator

We prove the jackknife variance estimation method works for the penalized-spline regression whenever it works for the simple linear regression of splines.

First, we define some notation:

A. $\beta = (\beta_0, \ldots, \beta_{m+p})^T$, the coefficients in Model (4).
B. $\hat{\beta}^0 = \left(\Pi^T\Pi\right)^{-1}\Pi^T Y$, the least squares (LS) estimator of β from the whole sample.
C. $\hat{\beta} = \left(\Pi^T\Pi + D(\hat{\alpha})\right)^{-1}\Pi^T Y$, the estimator of β given by (5) from the whole sample, $D(\hat{\alpha})$ is defined as in (5). From here on we replace the notation $D(\hat{\alpha})$ by D for simplicity.
D. $\hat{\beta}^0_{-i} = \left(\Pi^T_{-i}\Pi_{-i}\right)^{-1}\Pi^T_{-i}Y_{-i}$, the LS estimator of β and from the reduced sample with the ith element omitted, Π_{-i} is constructed the same way as Π but omitting the ith observation.
E. $\hat{\beta}_{-i} = \left(\Pi^T_{-i}\Pi_{-i} + D\right)^{-1}\Pi^T_{-i}Y_{-i}$ the estimator of β given by (5) and from the reduced sample.
F. Π_i, the ith row of matrix Π.

We prove the validity of the jackknife method under the following conditions:

(1) $y_i = f(\pi_i, \beta) + \varepsilon_i$, $\varepsilon_i \sim$ ind $N(0, \pi_i^{2k}\sigma^2)$, $f(\pi_i, \beta) = \beta_0 + \Sigma_{j=1}^{p}\beta_j\pi_i^j + \Sigma_{l=1}^{m}\beta_{l+p}(\pi_i - \kappa_l)_+^p$, the knots $\kappa_1 < \ldots < \kappa_m$ are fixed, m does not depend on n, and $\beta_0, \ldots, \beta_{m+p}$ are fixed and nonzero, which is equivalent to $\alpha = 0$. Under this condition, the underlying mean function is fixed rather than varying and the priors of $\beta_{p+1}, \ldots, \beta_{m+p}$ in Model (3) only serve as a roughness control technique for moderate-size samples and do not reflect prior belief on the coefficients. Another implication of this condition is that $\alpha \to 0$ as $n \to \infty$, so $\hat{\alpha}$ is bounded.
(2) $E(\varepsilon_i^4) < \infty$ and $k = 0$; when k is not zero, the proof holds after the transformation $\Pi^* = W^{1/2}\Pi$, $Y^* = W^{1/2}Y$.
(3) Π_i, the ith row in the matrix Π, is bounded for all i and n;
(4) $\frac{1}{n}\Pi^T\Pi \to \Sigma$, as $n \to \infty$ for all i for a positive definite matrix Σ.

With Assumptions (3) and (4), it follows that $1/(n-1)\Pi^T_{-i}\Pi_{-i} \to \Sigma$ as $n \to \infty$ uniformly with respect to i.

Under the above assumptions, $nVar(\hat{\beta}^0) \rightarrow \sigma^2\Sigma^{-1}$ and $nVar(\hat{\beta}) \rightarrow \sigma^2\Sigma^{-1}$.

LEMMA 1 $\hat{\beta}^0 - \hat{\beta} = D(\Pi^T\Pi)^{-1}\hat{\beta}$

Proof

$$\begin{aligned}
\hat{\beta}^0 - \hat{\beta} &= (\Pi^T\Pi)^{-1}\Pi^T Y - (\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= (\Pi^T\Pi)^{-1}\Pi^T Y - (\Pi^T\Pi)^{-1}(\Pi^T\Pi)(\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= (\Pi^T\Pi)^{-1}\Pi^T Y - (\Pi^T\Pi)^{-1}(\Pi^T\Pi + D)(\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&\quad + (\Pi^T\Pi)^{-1}D(\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= (\Pi^T\Pi)^{-1}D(\Pi^T\Pi + D)^{-1}\Pi^T Y = (\Pi^T\Pi)^{-1}D\hat{\beta}
\end{aligned}$$

QED.

LEMMA 2 $\hat{\beta}^0_{-i} - \hat{\beta}^0 = O(n^{-1})$ uniformly for all i.

Proof

$$\begin{aligned}
\hat{\beta}^0_{-i} - \hat{\beta}^0 &= (\Pi^T_{-i}\Pi_{-i})^{-1}\Pi^T_{-i}Y_{-i} - (\Pi^T\Pi)^{-1}\Pi^T Y \\
&= (\Pi^T_{-i}\Pi_{-i})^{-1}\Pi^T_{-i}(Y_{-i} - \Pi_{-i}\hat{\beta}^0) \\
&= (\Pi^T_{-i}\Pi_{-i})^{-1}(\Pi^T(Y - \Pi\hat{\beta}^0) - \Pi^T_i(Y_i - \Pi_i\hat{\beta}^0)) \\
&= -(\Pi^T_{-i}\Pi_{-i})^{-1}\Pi^T_i(Y_i - \Pi_i\hat{\beta}^0)
\end{aligned}$$

$\Pi^T_i(Y_i - \Pi_i\hat{\beta}^0)$ is uniformly *O(1)* and $(\Pi^T_{-i}\Pi_{-i})^{-1}$ is uniformly $O(n^{-1})$. Hence $\hat{\beta}^0_{-i} - \hat{\beta}^0 = O(n^{-1})$ uniformly. QED.

LEMMA 3 $\hat{\beta}_{-i} - \hat{\beta} = (\hat{\beta}^0_{-i} - \hat{\beta}^0) + O(n^{-2})$ uniformly for all i.

Proof

$$\begin{aligned}
\hat{\beta}_{-i} - \hat{\beta} &= (\Pi^T_{-i}\Pi_{-i} + D)^{-1}\Pi^T_{-i}Y_{-i} - (\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= (\Pi^T_{-i}\Pi_{-i} + D)^{-1}\Pi^T_{-i}Y_{-i} - (\Pi^T_{-i}\Pi_{-i} + D)^{-1}(\Pi^T_{-i}\Pi_{-i} + D)(\Pi^T\Pi + D)^{-1}\Pi^T Y \\
&= (\Pi^T_{-i}\Pi_{-i} + D)^{-1}\Pi^T_{-i}(Y_{-i} - \Pi_{-i}\hat{\beta}) - (\Pi^T_{-i}\Pi_{-i} + D)^{-1}D\hat{\beta}
\end{aligned}$$

from Lemma 1, $\hat{\beta} = \hat{\beta}^0 - D(\Pi^T\Pi)^{-1}\hat{\beta}$,

$$= (\Pi_{-i}^T \Pi_{-i} + D)^{-1} \Pi_{-i}^T (Y_{-i} - \Pi_{-i}(\hat{\beta}^0 - D(\Pi^T \Pi)^{-1} \hat{\beta})) - (\Pi_{-i}^T \Pi_{-i} + D)^{-1} D\hat{\beta}$$

$$= (\Pi_{-i}^T \Pi_{-i} + D)^{-1} \Pi_{-i}^T (Y_{-i} - \Pi_{-i}\hat{\beta}^0) + (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_{-i}^T \Pi_{-i}) D(\Pi^T \Pi)^{-1} \hat{\beta}$$

$$- (\Pi_{-i}^T \Pi_{-i} + D)^{-1} D\hat{\beta}$$

$$= (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_{-i}^T \Pi_{-i}) (\Pi_{-i}^T \Pi_{-i})^{-1} \Pi_{-i}^T (Y_{-i} - \Pi_{-i}\hat{\beta}^0)$$

$$+ (\Pi_{-i}^T \Pi_{-i} + D)^{-1} ((\Pi_{-i}^T \Pi_{-i})(\Pi^T \Pi)^{-1} - I) D\hat{\beta}$$

$$= (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_{-i}^T \Pi_{-i})(\hat{\beta}_{-i}^0 - \hat{\beta}^0) - (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_i^T \Pi_i)(\Pi^T \Pi)^{-1} D\hat{\beta}$$

$$= (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_{-i}^T \Pi_{-i} + D)(\hat{\beta}_{-i}^0 - \hat{\beta}^0) - (\Pi_{-i}^T \Pi_{-i} + D)^{-1} D(\hat{\beta}_{-i}^0 - \hat{\beta}^0)$$

$$- (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_i^T \Pi_i)(\Pi\Pi)^{-1} D\hat{\beta}$$

$$= (\hat{\beta}_{-i}^0 - \hat{\beta}^0) - (\Pi_{-i}^T \Pi_{-i} + D)^{-1} D(\hat{\beta}_{-i}^0 - \hat{\beta}^0) - (\Pi_{-i}^T \Pi_{-i} + D)^{-1} (\Pi_i^T \Pi_i)(\Pi^T \Pi)^{-1} D\hat{\beta}$$

By Assumption (3) and that $\hat{\alpha}$ is bounded, $(\Pi_{-i}^T \Pi_{-i} + D)^{-1}$ is $O(n^{-1})$ uniformly; by Lemma 2, $\hat{\beta}_{-i}^0 - \hat{\beta}^0 = O(n^{-1})$ uniformly. So the second term in the last line of the equation is $O(n^{-2})$ uniformly.

By Assumption (3), $(\Pi\Pi)^{-1}$ is $O(n^{-1})$; by Assumption (3) and that $\hat{\alpha}$ is bounded, $(\Pi_{-i}^T \Pi_{-i} + D)^{-1}$ is $O(n^{-1})$ uniformly; by Assumption (3), $\Pi_i^T \Pi_i$ is bounded; $\hat{\beta} \rightarrow \beta$ in probability. So the third term in the last line of the equation is also $O(n^{-2})$ uniformly.

QED.

THEOREM. If Assumptions (1) - (4) are all satisfied, then the delete-one jackknife variance estimator for $\hat{\beta}$, $v_J = (n-1)/(n)\Sigma_{i=1}^n (\hat{\beta}_{-i} - \hat{\beta})^T (\hat{\beta}_{-i} - \hat{\beta})$, is asymptotically consistent, i.e., $nv_J \rightarrow \sigma^2 \Sigma^{-1}$ in probability.

Proof.

$$nv_J = (n-1) \sum_{i=1}^{n} (\hat{\beta}_{-i} - \hat{\beta})^T (\hat{\beta}_{-i} - \hat{\beta})$$

from Lemma 3,

$$= (n-1) \sum_{i=1}^{n} \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0 + O(n^{-2})\right)^T \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0 + O(n^{-2})\right)$$

since $\hat{\beta}_{-i}^0 - \hat{\beta}^0$ is $O(n^{-1})$ for all i,

$$= (n-1) \sum_{i=1}^{n} \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0\right)^T \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0\right) + O(n^{-2}) + O(n^{-1})$$

$$= (n-1) \sum_{i=1}^{n} \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0\right)^T \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0\right) + O(n^{-1})$$

Under the Assumptions (1), (2), and (3), the jackknife estimator for the LSE $v_J^0 = (n-1)/(n)\Sigma_{i=1}^n \left(\hat{\beta}_{-i}^0 - \hat{\beta}^0\right)^T \left(\hat{\beta}_{-J}^0 - \hat{\beta}^0\right)$ satisfies $nv_J^0 \to \sigma^2 \Sigma^{-1}$ in probability, which leads to $nv_J \to \sigma^2 \Sigma^{-1}$ in probability. QED.

The validity of the jackknife variance estimation for $\hat{T}_{PROJ}$ and $\hat{T}_{PRED}$ follows from the validity of jackknife variance estimation for $\hat{\beta}$.

Whether or not the jackknife method is valid for p-spline regression when the number of knots increases with sample size remains a question to be answered.

6. References

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling, Part I. Foundations of Statistical Inference V.P. Godambe and D.A. Sprott (eds). Toronto: Holt, Rinehart, and Winston, 203–242.

Breidt, F.J. and Opsomer, J.D. (2000). Local Polynomial Regression Estimators in Survey Sampling. Annals of Statistics, 28, 1026–1053.

Breidt, F., Claeskens, G. and Opsomer, J. (2004). Model-Assisted Estimation for Complex Surveys Using Penalized Splines. Preprint Series #03–15, Department of Statistics, Iowa State University.

Brewer, K.R.W. (1963). Ratio Estimation and Finite Populations: Some Results Deducible from the Assumptions of an Underlying Stochastic Process. Australian Journal of Statistics, 5, 93–105.

Cochran, W.G. (1977). Sampling Techniques. 3rd Edition. New York: John Wiley.

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.

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.

Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. Journal of the Royal Statistical Society, Series B, 31, 195–234.

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.

Ghosh, M. and Meeden, G. (1997). Bayesian Methods for Finite Population Sampling. London: CRC Press.

Hanif, M. and Brewer, K.R.W. (1980). Sampling with Unequal Probabilities without Replacement: A Review. International Statistical Review, 48, 317–335.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys (with Discussion). Journal of the American Statistical Association, 78, 776–793.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sampling Survey Methods and Theory, Vols. I and II. New York: John Wiley.

Hartley, H.O. and Rao, J.N.K. (1962). Sampling with Unequal Probabilities and without Replacement. Annals of Mathematical Statistics, 33, 350–374.

Hinkley, D.V. (1977). Jackknifing in Unbalanced Situations. Technometrics, 19, 285–292.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980). Regression Analysis of Data from Complex Surveys. Journal of the Royal Statistical Society, Series A, 143, 474–487.

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

Kish, L. (1965). Survey Sampling. New York: John Wiley.

Little, R.J.A. (1983a). Comment on an Evaluation of Model Dependent and Probability Sampling Inferences in Sample Surveys by M.H. Hansen, W.G. Madow and B.J. Tepping. Journal of the American Statistical Association, 78, 797–799.

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

Little, R.J.A. and An, H. (2004). Robust Likelihood-based Analysis of Multivariate Data with Missing Values. Statistical Sinica, 14, 949–968.

Miller, R.G. (1974). An Unbalanced Jackknife. Annals of Statistics, 2, 880–891.

Pfeffermann, D. and Holmes, D.J. (1985). Robustness Considerations in the Choice of Method of Inference for Regression Analysis of Survey Data. Journal of the Royal Statistical Society, Series A, 148, 268–278.

Plackett, R.L. and Burman, J.P. (1946). The Design of Optimum Multifactorial Experiments. Biometrika, 33, 305–325.

Quenouille, M.H. (1949). Approximate Tests of Correlation in Time Series. Journal of the Royal Statistical Society, Series B, 11, 68–84.

Rizzo, L. (1992). Conditionally Consistent Estimators Using Only Probabilities of Selection in Complex Sample Surveys. Journal of the American Statistical Association, 87, 1166–1173.

Royall, R.M. (1970). On Finite Population Sampling under Certain Linear Regression Models. Biometrika, 57, 377–387.

Rubin, D.B. (1983). Comment on An Evaluation of Model Dependent and Probability Sampling Inferences in Sample Surveys by M.H. Hansen, W.G. Madow, and B.J. Tepping. Journal of the American Statistical Association, 78, 803–805.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

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

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

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). Semiparametric Regression. Cambridge, UK: Cambridge University Press.

Shao, J. and Wu, C.F.J. (1987). Heteroscedasticity-robustness of Jackknife Variance Estimators in Linear Models. Annals of Statistics, 15, 1563–1579.

Shao, J. and Wu, C.F.J. (1989). A General Theory for Jackknife Variance Estimation. Annals of Statistics, 17, 1176–1197.

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.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1989). The Weighted Residual Technique for Estimating the Variance of the General Regression Estimator of the Finite Population Total. Biometrika, 76, 527–537.

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

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference. New York: Wiley.

Wolter, K.M. (1985). Introduction to Variance Estimation: Springer-Verlag.

Zheng H. (2003). Penalized Spline Nonparametric Regression Methods for Survey Samples with Potentially Unequal Probabilities of Inclusion. Ph.D. Dissertation, Department of Biostatistics, University of Michigan, Ann Arbor, MI.

Zheng, H. and Little, R.J.A. (2003). Penalized Spline Model-Based Estimation of Finite Population Total from Probability-Proportional-to-Size Samples. Journal of Official Statistics, 19, 99–117.

Zheng, H. and Little, R.J.A. (2004). Penalized Spline Nonparametric Mixed Model for Inference of Finite Population Means from Two-Stage Samples. Survey Methodology, 30, 2, 209–218.

Received July 2003
Revised June 2004