University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2003 *Paper* 9

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

Hui Zheng* Rod Little†

*Harvard Medical School, huizheng@umich.edu

†University of Michigan, rlittle@umich.edu

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

Hui Zheng and Rod Little

Abstract

Inference about the finite population total from probability-proportional-to-size (PPS) samples is considered. In previous work (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.

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

Hui Zheng

Post-doctoral Fellow, Department of Health Care Policy, Harvard Medical School

180 Longwood Avenue, Boston, MA 02115, USA. Email: huizheng@umich.edu

Roderick J. A. Little

Professor, Department of Biostatistics, University of Michigan

1420 Washington Heights, Ann Arbor, MI 48109, USA. Email: rlittle@umich.edu

Summary. Inference about the finite population total from probability-proportional-to-size (PPS) samples is considered. In previous work (Zheng and Little, 2002), 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 $Y = (Y_1, ..., Y_N)$ where Y_i is the set of survey variables for unit *i*, and let $I = (I_1, ..., I_N)$ denote the set 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 non-sampled 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 (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 $p(Y \mid \theta)$ 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 paper 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, chapter 7); units that would be selected with probability one are removed 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

$$\hat{T}_{HT} = \sum_{i=1}^{n} y_i / \pi_i \ , \qquad (1)$$

(Horvitz and Thompson, 1952), where the summation is over *n* sampled units. This is also the projective estimator (Firth and Bennett, 1998) for a "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 from 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 Dorfman and Wehrly (1993) estimate a finite population totals by a nonparametric model relating *Y* to an auxiliary variable, using kernel smoothing. 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:

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

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). In particular, in the context of PPS sampling, Zheng and Little (2002) compare prediction estimates of the population total based on p-splines with the HT estimator and the GR estimation based on a simple linear regression model. 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 have better confidence coverage, or tests closer to their nominal significance levels. Hence, the goal of the current paper is to consider variance estimation and inference properties of the estimators compared in Zheng and Little (2002). 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 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 paper 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 (2002) predicts non-sampled values of $y_i, i \in P - S$ using the following 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 penalized spline:

$$\begin{aligned} & f(\pi_i, \beta) = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{l=1}^{m} \beta_{l+p} (\pi_i - \kappa_l)_+^p, \quad i=1,\ldots, N. \\ & \beta_{l+p} \underset{iid}{\sim} N(0, \tau^2), l = 1, ..., m. \end{aligned} \tag{4}$$

Here the constants $\kappa_1 < ... < \kappa_m$ 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 others, we favor a modeling strategy that places a large number of knots (for example, 15 or 30) at pre-specified locations, and then achieves smoothness by treating $\beta_{p+1}, ..., \beta_{p+m}$ as random effects centered at 0. The degree of smoothing is 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

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

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

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

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

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

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

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

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

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

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 non-sampled 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`.

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 *g*th pseudovalue is computed as $\hat{T}_g = K\hat{T} - (K-1)\hat{T}_{(g)}$, 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 *k*th 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 $\hat{\bar{T}} = \sum_{g=1}^{G} \hat{T}_g / G$. 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 randomly selecting one element from each stratum. To save computation, estimates $\hat{\alpha} = \hat{\sigma}^2 / \hat{\tau}^2$ are not recomputed for each replicate. That is, we compute pseudovalues as $\beta_{(g)}{}^T = (\Pi_{(g)}{}^T \Pi_{(g)} + D(\hat{\alpha}))^{-1} \Pi_{(g)}{}^T Y$, where $\Pi_{(g)}$ is constructed in the same way as Π but omitting the *g*-th subgroup, but the estimate $\hat{\alpha}$ is computed for the full sample.

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 Appendix A, 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 application 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 $\sum_{i=1}^{N} x_i \Big/ n$. 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 $2\sum_{i=1}^{N} x_i \Big/ n$, 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

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

Since the BRR method with two units sampled per stratum does not fully reflect efficiency gains from PPS stratification, it can be expected to overestimate the true variance of the p-spline estimator, a conjecture that is consistent with simulation results reported in the next section.

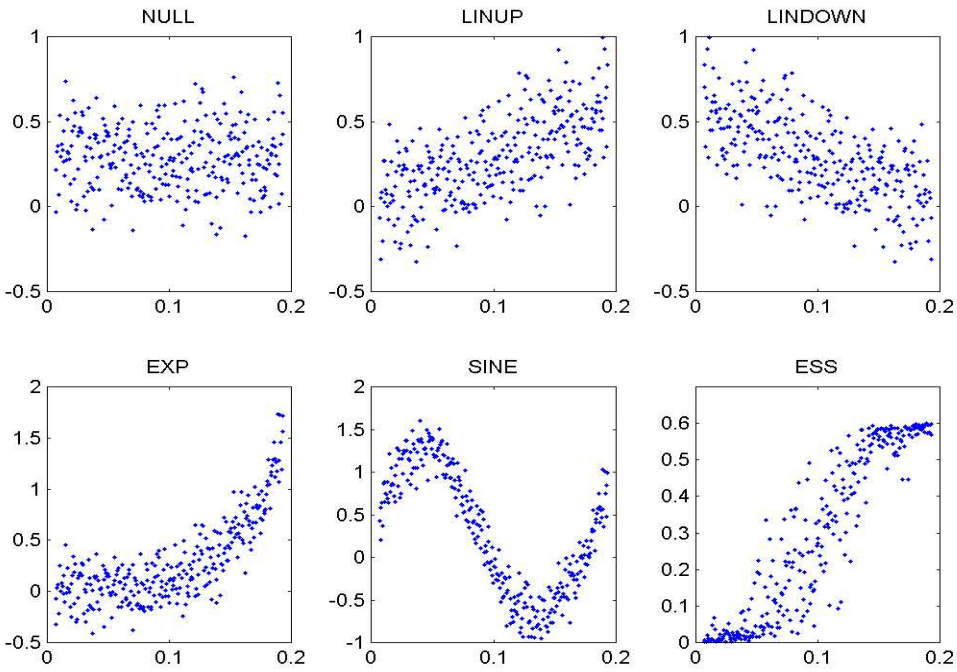

Figure 1. Six simulated populations (N=300) X-axis: pi(i); Y-axis: y(i) with normal errors

3. Simulation Study

3.1 The Simulated Populations

Artificial populations are simulated according to six different mean functions relating outcome y_i and the inclusion probabilities π_i : Five of these 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) $y_i = 0.6\log it^{-1}(50 * \pi_i - 5 + \varepsilon_i), \varepsilon_i \overset{iid}{\sim} N(0,1)$.

Since the errors in ESS lie inside the logit function, this population had heteroscedastic errors. Plots of samples from these populations are provided in Figure 1. Population sizes 300, 1000 and 2000 with respective sample sizes 32, 96 and 192 are simulated for each mean function. For each simulated population, 500 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.

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, 2002). 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 15 knots equally spaced with respect to the percentiles of the distribution of X.

b) HT, the Horvitz-Thompson estimator (1).

c) GR, a generalized regression estimator (2) assisted by a simple linear 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

without an intercept mean function (LINUP), which is in favor of HT. The empirical bias of P0_15 is small and in most cases P0_15 has comparable empirical bias as HT and GR. Similar findings are presented in the more extensive simulations in Zheng and Little (2002).

3.3 Variance Estimation for Alternative Methods

In this section we compare the inferences for the P-spline prediction and projection estimators, with variances estimated by (8)-(11), with inferences based on the HT and GR estimators. For HT, we show results for two variance estimation methods, the random groups variance estimator

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

where the sample is divided into *K* random subsamples, each of size $m = n/K$, and $\hat{T}_k = \sum_{i=1}^{m} y_i / (mp_i)$, $p_i = \pi_i / n$ is the HT estimator from the *k*th subsample; and 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 impact 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 hence are omitted here to save space. Results for all five estimators are given in Zheng (2002).

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

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

where

$$\Delta_{kl} = \pi_{kl} - \pi_k\pi_l, \pi_{kk} = \pi_k, g_k = 1 + (\sum_{k=1}^{N} x_k - \sum_{i=1}^{n}\frac{x_k}{\pi_k})(\sum_{k=1}^{n} x_k^2)^{-1} x_k,$$

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

Here the covariate is $x_k = \pi_k$, and we use the Hartley-Rao approximation

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

for π_{kl}. The approximation formula for joint inclusion probability is valid when $\max(\{\pi_i, i = 1...n\}) = O(N^{-1})$, which is satisfied by our simulated sampling design.

First, 500 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 coverage of these inference methods are then compared based on 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 P1_15, which assumes the error variances are proportional to π_i^2, 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, P1_15

assumes the correct error variance for the second group while it misspecifies the error variance for the first 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 estimate 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.

In Table 3, three inference approaches: 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 are compared. 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 under-coverage 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).

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 coverages 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 yield 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 P1_15, which

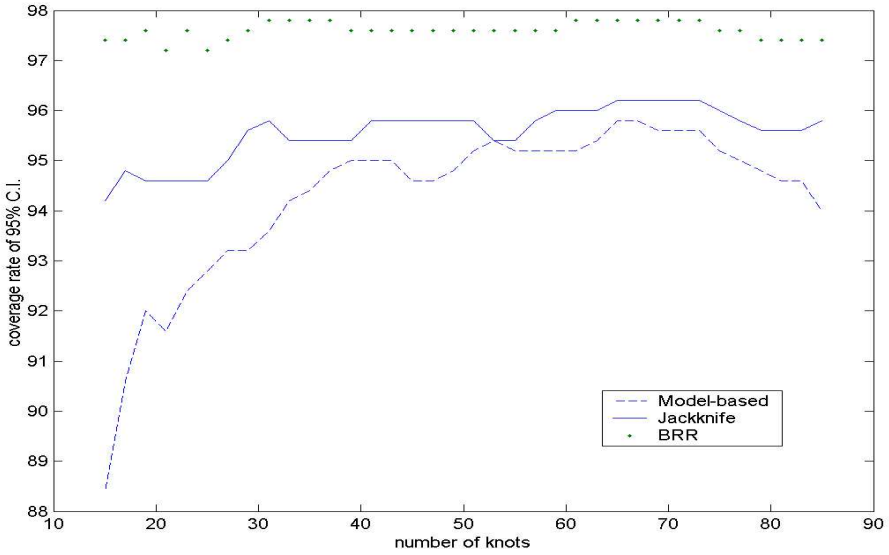

Figure 2.
Coverage rate (percentage) of 95% C.I. vs. number of knots for population SINE N=2000, n=192, Coverage rate computed from 500 repeated samples (target =93-97%)

corresponds to a p-spline with 15 knots and assuming error variance proportional to π_i^2, 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 robust to this form of misspecification.

5. Discussion

Although the HT estimator is design-unbiased, and can be used with an appropriate variance estimator to yield valid large-sample inferences, its efficiency and its performance in moderate-sized samples depend on the validity of the underlying "HT model". The GR estimator can yield increases in efficiency, but is also sub-optimal if based on a poorly chosen model, and can yield anti-conservative inferences in moderate sized samples. Our proposed nonparametric model based on p-splines assumes a more flexible mean structure than that implied by the HT or GR models. Since these models give a close approximation to the mean function, calibration as

in the GR estimator is not necessary. The p-spline method with the jackknife variance estimate yields shorter confidence intervals than the design-based methods, while achieving noncoverage rates that are superior to traditional methods. An exception is its performance in the SINE populations, where more than the chosen number of 15 knots is needed for inference. A referee noted that 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 Eq. (3), from the data. Here we adopted the less efficient but simpler approach of fixing k and use a robust variance estimator based on the jackknife.

Survey samples 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 & Silverman, 1989; Friedman, 1991; Stone et al., 1997; Denison, Mallick and Smith, 1998; Ruppert and Carroll, 2000, Rupert 2002) have proposed sophisticated knot selection methods that might be profitably applied to complex mean functions.

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 boostrapping was done in a way that balanced the distribution of the selection probabilities in the bootstrap samples.

In conclusion, we believe that p-spline models provide an attractive approach to survey inference based on probability-proportional-to-size samples. We are currently considering extensions of the proposed approach to multistage sampling, and to non-normal outcomes.

6. References

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling, Part I. *In Foundations of Statistical Inference* (eds. V.P. Godambe and D.A. Sprott), 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.

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.

Cassel, C-M, Sarndal, C-E. and Wretman, J.H. (1977), *Foundations of Inference in Survey Sampling*, Wiley; New York.

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

Cox, B. G., Binder, D. A. Chinnappa, B. N. (eds.) (1995). Business Survey Methods, John Wiley & Sons, Inc.

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*, B60, 333-350.

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

Dumouchel, W.H. and Duncan, G.J. (1983), "Using sample survey weights in multiple regression analysis of stratified samples," *Journal of the American Statistical Association,* 78, 535-543.

Ericson, W.A. (1969), "Subjective Bayesian models in sampling finite populations," *Journal of the Royal Statistical Society,* B, 31, 195-234.

Firth, D. and Bennett, K.E. (1998). Robust Models in Probability Sampling. *Journal of the Royal Statistical Society*, 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," *Journal of the American Statistical Association,* 78, 776-793 (with discussion).

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,* A, 143, 474-87.

Kalton, G. (1977). Practical Methods for Estimating Survey Sampling Errors. *Bulletin of the International Statistical Institute,* 47, 495-514.

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. (1991), "Inference with Survey Weights," *Journal of Official Statistics*, 7, 405-424.

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*, 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*, B 11, 68-84.

Rao, J. N. K. and Wu, C. F. J. (1988). Resampling Inference with Complex Sample Data. *Journal of the American Statistical Association*, 83, 231-239.

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*. Wiley: New York.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, to appear.

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. 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*, Springer Verlag: New York.

Shao and C. F. J. Wu (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.

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

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

Zheng, H. and Little, R. J. A. (2002). Penalized Spline Model-Based Estimation of Finite Population Total from Probability-Proportional-to-Size Samples. Submitted to *Journal of Official Statistics*.

Appendix A. Asymptotic Consistency of the Jackknife Variance Estimation

Some notation:

A. $\beta = (\beta_0, ..., \beta_{m+p})^T$, the coefficients under 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. $\hat{\beta}_{-i}^{0} = \left(\Pi_{-i}^{T}\Pi_{-i}\right)^{-1}\Pi_{-i}^{T}Y_{-i}$, the LS estimator of β and from the reduced sample with the *i*th element omitted, Π_{-i} is constructed the same way as Π but omitting the *i*th observation.

E. $\hat{\beta}_{-i} = \left(\Pi_{-i}^{T}\Pi_{-i} + D\right)^{-1}\Pi_{-i}^{T}Y_{-i}$ 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 assumptions:

1) Model (3) is correct, the knots $\kappa_1 < ... < \kappa_m$ are fixed and *m* does not depend on *n*.

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) $\hat{\alpha}$ is bounded. This is in fact is necessarily true for a fixed nontrivial mean functions (trivial functions are polynomial functions with degrees no greater than *p*). For trivial mean functions, traditional multiple regression theory holds and is not discussed here.

4) Π_i, the *i*th row in the matrix Π, is bounded for all *i* and *n*;

5) $\frac{1}{n}\Pi^T\Pi \to \Sigma$, as $n \to \infty$ for all *i* for a positive definite matrix Σ.

With assumptions 4) and 5), it follows that $\frac{1}{n-1}\Pi_{-i}^{T}\Pi_{-i} \to \Sigma$ as $n \to \infty$ uniformly with respect to *i.*

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

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

Proof

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

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

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

By assumptions 3) and 4), $(\Pi_{-i}^T \Pi_{-i} + D)^{-1}$ 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 4), $(\Pi_{-i}^T \Pi_{-i})^{-1}$ is $O(n^{-1})$ uniformly; by assumptions 3) and 4), $(\Pi_{-i}^T \Pi_{-i} + D)^{-1}$ is $O(n^{-1})$ uniformly; by assumption 4), $\Pi_i^T \Pi_i$ is bounded; $\hat{\beta} \to \beta$ in probability. So the third term in the last line of the equation is also $O(n^{-2})$ uniformly.

QED.

Theorem 1. If assumptions 1)- 4) are all satisfied, then the delete-one jackknife variance estimator for $\hat{\beta}$, $v_J = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\beta}_{-i} - \hat{\beta}\right)^T \left(\hat{\beta}_{-i} - \hat{\beta}\right)$, is asymptotically consistent, i.e., $nv_J \to \sigma^2 \Sigma^{-1}$ in probability.

Proof.

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

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 4), the jackknife estimator for the LSE

$v_J^0 = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{-i}^{0} - \hat{\beta}^{0}\right)^{T}\left(\hat{\beta}_{-J}^{0} - \hat{\beta}^{0}\right)$ 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 $\hat{\beta}$.

Appendix B.

Inferences for Horvitz Thompson estimator: Var(HT) = empirical variance; meanvar = average estimated variance and N.C. = noncoverage of 95% CI over 1000 samples (target = 30-70), for each of five estimators of variance (V1 – V5) and three population sizes with K=10 for the jackknife method.

	Population	Var(HT)	v1 (Yates-Grundy)		v2 (Random Group)		v3 (With Replacement)		v4 (Paired Differences)		v5 (Successive Differences)	
			meanvar	N.C.	meanvar	N.C.	meanvar	N.C.	meanvar	N.C.	meanvar	N.C.
	NULL	350	370	148	389	106	387	134	379	140	384	136
A	LINUP	156	154	44	179	16	164	28	166	46	166	32
	LINDOWN	1005	858	204	904	182	895	200	892	200	889	196
N=300	SINE	3215	3433	96	3624	80	3639	84	3598	96	3601	98
n=32	EXP	254	247	68	284	48	276	56	275	62	272	62
	ESS	31	31	76	47	34	33	58	34	74	34	70
	NULL	1229	1275	98	1340	68	1334	86	1327	100	1331	96
B	LINUP	744	791	60	870	42	834	50	835	50	833	50
	LINDOWN	4001	3940	110	4109	82	4095	102	4107	100	4110	94
N=1000	SINE	12679	13388	68	13908	60	14089	66	14017	70	14082	70
n=96	EXP	1205	1176	64	1324	42	1278	54	1274	50	1266	50
	ESS	125	122	50	159	14	132	44	133	46	132	48
	NULL	2934	2601	108	2817	76	2725	94	2688	98	2697	98
C	LINUP	1308	1396	50	1450	32	1474	44	1469	40	1466	40
	LINDOWN	7447	7205	104	7386	78	7501	98	7485	98	7484	98
N=2000	SINE	23172	24912	52	26153	56	26251	50	26337	50	26196	50
n=192	EXP	2337	2158	62	2361	46	2357	50	2337	50	2359	48
	ESS	247	239	62	272	34	259	50	258	58	258	52

Table1. Comparison of three point estimators: P0_15, HT and GR
N=1000,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

Table 2. Empirical mean estimates of 6 variance estimators: v_{RG}, v_{WR}, $\hat{V}(GR)$, Empirical Bayes, Jackknife (K=10) and BRR.

	Population	Empirical Var (HT)	Mean(v_{RG})	Mean (v_{WR})	Empirical Var(GR)	Mean $\hat{V}(GR)$	Empirical Var(P0_15)	Empirical Bayes	Jackknife (K=10)	BRR
	NULL	350	389	387	157	113	133	140	171	172
A	LINUP	156	179	164	275	196	125	142	163	165
	LINDOWN	1005	904	895	281	192	169	131	197	192
N=300	SINE	3215	3624	3639	2752	1767	873	349	1326	1426
n=32	EXP	254	284	276	784	566	201	233	273	334
	ESS	31	47	33	219	175	39	32	51	71
	NULL	1229	1340	1334	560	527	475	501	555	576
B	LINUP	744	870	834	1168	1020	660	573	678	702
	LINDOWN	4001	4109	4095	1246	1037	713	608	622	658
N=1000	SINE	12679	13908	14089	8937	7656	1584	769	1890	4006
n=96	EXP	1205	1324	1278	2952	2515	750	726	796	947
	ESS	125	159	132	914	786	85	114	100	141
	NULL	2934	2817	2725	1389	1250	1120	1133	1153	1195
C	LINUP	1308	1450	1474	2197	2107	1070	1056	1170	1204
	LINDOWN	7447	7386	7501	2337	2160	1217	1086	1174	1226
N=2000	SINE	23172	26153	26251	19482	16346	2027	1556	2551	4861
n=192	EXP	2337	2361	2357	6073	5656	1254	1297	1345	1518
	ESS	247	272	259	1860	1696	142	194	144	182

Table 3. Comparison of three approaches to inference: HT with V2, GR with Yates-Grundy, P-spline with Jackknife: A.W. = Average 95% CI width and N.C. = non-coverage rate of 95% C.I. over 1000 samples (target = 30-70)

	Population	HT		GR		P-spline	
		A.W.	N.C.	A.W.	N.C.	A.W.	N.C.
N=300, n=32	NULL	68	106	40	122	48	46
	LINUP	48	16	53	128	47	40
	LINDOWN	98	182	52	134	51	48
	SINE	223	80	161	156	114	204
	EXP	63	48	89	142	57	64
	ESS	26	34	51	112	24	48
N=1000, n=96	NULL	131	68	88	80	89	28
	LINUP	109	42	123	64	98	48
	LINDOWN	230	82	124	82	94	62
	SINE	446	60	340	74	145	86
	EXP	135	42	193	96	105	54
	ESS	48	14	109	84	37	66
N=2000,n=192	NULL	196	76	137	54	129	42
	LINUP	142	32	178	52	130	30
	LINDOWN	317	78	180	64	129	58
	SINE	611	56	497	82	182	74
	EXP	184	46	289	66	138	48
	ESS	63	34	161	76	45	46

Table 4. Inferences for P1_15 with model-based and jackknife standard errors, applied to data with homoscedastic and heteroscedastic errors. Var(P1_15) = empirical variance; meanvar = average estimated variance and N.C. = noncoverage of 95% CI over 1000 samples (target = 30-70). N=1000,n=100.

Population	Variance Structure Incorrectly Specified					Variance Structure Correctly Specified				
	Var (P1_15)	Model Based s.e.		Jackknife s.e.		Var (P1_15)	Model Based s.e.		Jackknife s.e.	
		meanvar	N.C.	meanvar	N.C.		meanvar	N.C.	meanvar	N.C.
NULL	668	383	150	830	56	94	82	56	110	50
LINUP	1012	500	192	1318	56	75	71	56	96	38
LINDOWN	732	461	140	1058	38	96	81	84	104	44
SINE	1070	785	118	2742	46	238	278	54	581	48
EXP	897	529	136	1326	50	98	106	50	126	26