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

Year 2003 *Paper* 8

Penalized Spline Nonparametric Mixed Models for Inference About a Finite Population Mean from Two-Stage Samples

Hui Zheng* Rod Little†

*University of Michigan, huizheng@umich.edu
†University of Michigan, rlittle@umich.edu

Penalized Spline Nonparametric Mixed Models for Inference About a Finite Population Mean from Two-Stage Samples

Hui Zheng and Rod Little

Abstract

Samplers often distrust model-based approaches to survey inference due to concerns about model misspecification when applied to large samples from complex populations. We suggest that the model-based paradigm can work very successfully in survey settings, provided models are chosen that take into account the sample design and avoid strong parametric assumptions. The Horvitz-Thompson (HT) estimator is a simple design-unbiased estimator of the finite population total in probability sampling designs. From a modeling perspective, the HT estimator performs well when the ratios of the outcome values and the inclusion probabilities are exchangeable. When this assumption is not met, the HT estimator can be very inefficient. In Zheng and Little (2002a, 2002b) we used penalized splines (p-splines) to model smoothly -varying relationships between the outcome and the inclusion probabilities in one-stage probability proportional to size (PPS) samples. We showed that p-spline model-based estimators are in general more efficient than the HT estimator, and can be used to provide narrower confidence intervals with close to nominal confidence coverage. In this article, we extend this approach to two-stage sampling designs. We use a p-spline based mixed model that fits a nonparametric relationship between the primary sampling unit (PSU) means and a measure of PSU size, and incorporates random effects to model clustering. For variance estimation we consider the empirical Bayes model-based variance, the jackknife and balanced repeated replication. Simulation studies on simulated data and on samples drawn from public use microdata in the 1990 census demonstrate gains for the model-based p-spline estimator over the HT estimator and linear model-assisted estimators. Simulations also show the variance estimation methods yield confidence intervals with satisfactory confidence coverage. Interestingly, these gains can be seen in an equal probability design, where the first stage

selection is PPS and the second stage selection probabilities are proportional to the inverse of the first stage inclusion probabilities, and the HT estimator leads to the unweighted mean. In situations that most favor the HT estimator, the model-based estimators have comparable efficiency.

Draft March 14, 2003

PENALIZED SPLINE NONPARAMETRIC MIXED MODELS FOR INFERENCE ABOUT A FINITE POPULATION MEAN FROM TWO-STAGE SAMPLES

Hui Zheng and Roderick Little

Department of Biostatistics

University of Michigan

Abstract

Samplers often distrust model-based approaches to survey inference due to concerns about model misspecification when applied to large samples from complex populations. We suggest that the model-based paradigm can work very successfully in survey settings, provided models are chosen that take into account the sample design and avoid strong parametric assumptions.

The Horvitz-Thompson (HT) estimator is a simple design-unbiased estimator of the finite population total in probability sampling designs. From a modeling perspective, the HT estimator performs well when the ratios of the outcome values and the inclusion probabilities are exchangeable. When this assumption is not met, the HT estimator can be very inefficient. In Zheng and Little (2002a, 2002b) we used penalized splines (p-splines) to model smoothly –varying relationships between the outcome and the inclusion probabilities in one-stage probability proportional to size (PPS) samples. We showed that p-spline model-based estimators are in general more efficient than the HT estimator, and can be used to provide narrower confidence intervals with close to nominal confidence coverage. In this article, we extend this approach to two-stage sampling designs. We use a p-spline based mixed model that fits a nonparametric relationship between the primary sampling unit (PSU) means and a measure of PSU size, and incorporates random effects to model clustering. For variance estimation we consider the empirical Bayes model-based variance, the jackknife and balanced repeated replication. Simulation studies on simulated data and on samples drawn from public use microdata in the 1990 census demonstrate gains for the model-based p-spline estimator over the HT estimator and linear model-assisted estimators. Simulations also show the variance estimation methods

yield confidence intervals with satisfactory confidence coverage. Interestingly, these gains can be seen in an equal probability design, where the first stage selection is PPS and the second stage selection probabilities are proportional to the inverse of the first stage inclusion probabilities, and the HT estimator leads to the unweighted mean. In situations that most favor the HT estimator, the model-based estimators have comparable efficiency.

Keywords: weighting, REML, empirical Bayes estimation

1. Introduction

In a sample survey, let y_i denote the value of a survey outcome Y for unit i, and let S denote the set of sampled units. The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) $\hat{Y}_{HT} = \sum_{i \in S} y_i / p_i$, where p_i is the probability of selection of unit i, is a design-unbiased estimator of the finite population total (and of the mean when divided by the known population count N). It can also be regarded as a model-based projective estimator (Firth and Bennett 1998) for the following linear model relating y_i to p_i:

$$y_i = b p_i + p_i e_i,$$

where e_i are assumed to be i.i.d. normally distributed with mean zero and variance s^2.

In Zheng and Little (2002a, b), we proposed a nonparametric model

$$y_i = f(p_i) + e_i,\ e_i \sim \text{ind}\ N(0, p_i^{2k} s^2)$$

using penalized splines to model mean of outcome y_i as a smoothly varying function of the inclusion probabilities p_i. We showed in Zheng and Little (2002a) that the nonparametric model-based estimators are more efficient than HT for general one-stage probability-proportional-to-size (PPS) samples and not much less efficient than HT when the data are generated using a model that favors HT.

We now consider the case of two-stage sampling. In the first stage, a subset of m primary sampling units (PSU's) is drawn from a population with H PSU's with unequal probabilities $p_{1,h}$, $h = 1,...,H$. Let us number the included PSU's from 1 to m. In the second stage, a simple random sample (srs) of n_h out of N_h secondary sampling units (SSU's) is drawn from sampled PSU labeled h with probability $p_{2,h}$ for the hth PSU. The overall selection probability for unit i in cluster h is $p_h = p_{1,h}p_{2,h}$, and the Horvitz-Thompson estimator of the mean of an outcome Y is $\bar{y}_w = \frac{1}{N}\sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi}/(p_{1,h}p_{2,h})$, where y_{hi} is the value of Y for unit i in cluster h and N is the known total number of units (SSU's) in the whole population. In a commonly adopted design, the first stage selection probability is proportional to an estimate of the PSU size, and the second stage inclusion probabilities are proportional to the inverse of the first stage inclusion probabilities so that the overall inclusion probabilities p_h are equal for all SSU's. The inverse probability weighted mean in this case becomes the simple sample mean $\bar{y} = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi} \Big/ \sum_{h=1}^{m} n_h$.

In sections 2 and 3, we assume the cluster counts N_h and the values $p_{1,h}, p_{2,h}, h = 1...H$ of the identifying variable are known for all clusters, whether sampled or not. In Section 4, we discuss the situation where $p_{1,h}, p_{2,h}, h = 1...H$ are known while

N_h are only known for sampled PSU's, but can be estimated based on some auxiliary variable.

Särndal et al. (1992) discussed model-assisted alternatives to the HT estimator for two-stage samples. If the auxiliary information is on PSU (cluster) level, a linear model is applied to regress PSU totals t_h on the auxiliary variable, say, z_h (z_h can be a vector). If the auxiliary information is SSU (element) level, a linear model is applied to regress outcome y_{hi} on the auxiliary variables, say, z_{hi}. For example, if the auxiliary information is on the cluster level, that is $z_{hi} = z_h$ for all *i*, then the cluster totals t_h are assumed to be related to z_h according to a linear model:

$$E(t_h \mid z_h) = z_h^T b \ , \ Var(t_h) = s_h^2, \ h = 1...H$$

Särndal et al. then estimate b by the probability weighted regression

$$\hat{B} = \left(\sum_{h=1}^{m} z_h z_h^T / (s_h^2 p_{1,h}) \right)^{-1} \sum_{h=1}^{m} z_h t_h^* / (s_h^2 p_{1,h}), \text{ where } t_h^* = \sum_{i=1}^{n_h} y_{hi} / p_{2,h},$$

leading to the projected totals $\hat{t}_h = z_h^T \hat{B}, h = 1...H$. The generalized regression (GR) estimator of the grand total is $\hat{T}_A = \sum_{i=1}^{H} \hat{t}_h + \sum_{h=1}^{m} \frac{(t_h^* - \hat{t}_h)}{p_{1,h}}$ and the estimate for the mean is $\hat{T}_A / N$. The term $\sum_{h=1}^{m} \frac{(t_h^* - \hat{t}_h)}{p_{1,h}}$ is the bias calibration term that makes the estimator design consistent.

In the case where auxiliary information $\{x_{hi}\}, h = 1,..., H; i = 1,..., n_h$ on the element (SSU) level is known for the whole population, the relationship between the outcome and the auxiliary information is modeled by

$E(y_{hi} \mid x_{hi}) = x_{hi}^T b$, $Var(y_{hi}) = s_{hi}^2, h = 1...H, i = 1...N_h$.

The probability weighted regression estimate for b is

$\hat{B} = \left(\sum_{h=1}^{m} \sum_{i=1}^{n_h} x_{hi} x_{hi}^T / (s_{hi}^2 p_{hi}) \right)^{-1} \sum_{h=1}^{m} \sum_{i=1}^{n_h} x_{hi} y_{hi} / (s_h^2 p_{hi})$, where p_{hi} is the probability for unit *(h, i)* to be included in the sample.

The GR estimator for the grand total is $\hat{T}_B = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \hat{y}_{hi} + \sum_{h=1}^{m} \sum_{i=1}^{n_h} \frac{(y_{hi} - \hat{y}_{hi})}{p_{hi}}$, where $\hat{y}_{hi} = x_{hi}^T \hat{B}$. The estimator for the mean is then $\hat{T}_B / N$.

The linear models discussed by Särndal et al. (1992) do not account for the within-cluster correlations. The following family of models allow for within-cluster correlations by treating cluster means as random effects:

$$\begin{aligned} y_{hi} \mid m_h &\overset{ind}{\sim} N(m_h, s^2) \\ m &\sim N_H(f, D) \end{aligned} \qquad (1)$$

where $m = (m_1,...,m_H)$, $f = (f_1,...,f_H)$, and D is the covariance matrix. A total number of m PSU's are sampled from a total of H PSU's.

The model-based estimator of $\bar{Y}$ is given by

$E(\bar{Y} \mid \mathbf{y}, p_{1,h}) = \frac{1}{N} \left(\sum_{h=1}^{m} [n_h \bar{y}_h + (N_h - n_h) \hat{m}_h] + \sum_{h=m+1}^{H} N_h \hat{m}_h \right)$, where

$\hat{m}_h = E(\bar{Y}_h \mid \mathbf{y}, p_{1,h}) = E(m_h \mid \mathbf{y}, x_{1,h})$.

In an equal probability design, where n_h are approximately constant across PSU's, the unweighted mean $\bar{y}$ corresponds to the special model specification where ε are constant.

Different assumptions about f and D in model (1) lead to the following models:

Exchangeable random effects (XRE): (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998)

$$f_h \equiv m,\ h = 1 \ldots H \ \text{ and } \ D = t^2 I_H$$

Autoregressive (AR1): (Lazzaroni and Little 1998)

$$f_h \equiv m,\ h = 1 \ldots H \ \text{ and } \ D = r^2 \{ r^{|i-j|} \}$$

Linear (LIN): (Lazzaroni and Little 1998)

$$f_h = a + b x_h,\ h = 1 \ldots H \ \text{ and } \ D = t^2 I_H$$

Nonparametric: (Elliott and Little 2000)

$$f_h = f(x_h),\ h = 1 \ldots H \ \text{ and } \ D = 0$$

The nonparametric models in Elliott and Little (2000) assume nonparametric mean function relating the outcome and the design variables. By assuming $D = 0$, no variability is allowed around the mean function. That is, the cluster means are modeled to pass through the overall mean function f instead of vary around it. Nonparametric mixed models (Lin and Zhang 1999; Brumback, Ruppert and Wand 1999; Coull, Schwartz and Wand 2001) relax the assumptions on D (e.g., $D = t^2 I_H$) and serve as a natural extension to both the linear mixed models and the nonparametric model in Elliott and Little (2000).

2. Estimation with the P-spline Mixed Model Method

The linear structure of f in LIN model is subject to misspecification when the actual mean structure is non-linear. The non-linearity problem can be partially solved by adding polynomial terms (e.g., quadratic or cubic terms) to the fixed effect part in the

LIN model. Using nonparametric functions to model the mean structure f, p-spline nonparametric mixed models are even more flexible than polynomial mixed models.

We propose the following p-spline nonparametric mixed model for inference of the population mean.

P-spline nonparametric mixed model (PMM):

$f_h = f(x_h),\ h = 1 \ldots H$ and $D = t^2 I_H$, where f is a nonparametric function

Methods for estimating f are not unique. We use splines of degree $p > 0$ to estimate f:

$\hat{f}(x; ß) = b_0 + \sum_{j=1}^{p} b_j x^j + \sum_{l=1}^{K} b_{l+p} (x - k_l)_+^p$, where $k_1 < \ldots < k_K$ are K fixed knots,

$b_0, \ldots, b_{p+K}$ are coefficients to be estimated and $(x)_+^p = x^p \mathbf{I}(x \geq 0)$.

A simple way of estimating $ß_0, \ldots, ß_{p+K}$ is to treat them as fixed effects and estimate them together with the variance components s^2 and t^2 by fitting a mixed model similar to that used in the LIN model. However this method can yield estimates of f with too much roughness and variability. To avoid overfitting, the roughness of the estimation $\hat{f}$ is penalized by applying a factor a to the least squares so that the solution $\hat{b}_0, \ldots, \hat{b}_p$ is the minimizer of

$$\sum_{h=1}^{m} (\hat{f}(x_h) - \hat{m}_h)^2 + a \sum_{l=1}^{K} b_{l+p}^2 .$$

This is achieved in the context of the model by assigning $b_0, \ldots, b_p$ flat priors, $(b_{p+1}, \ldots, b_{p+K})$ a normal prior $N_m(0, s_b^2)$, and letting $a = t^2 / s_b^2$. The result is a penalized spline (p-spline) model.

In the case of $p = 1$, $\hat{f}$ is piecewise linear and the coefficients $b_0, \ldots, b_{K+1}$ and the

variance components s^2, s_b^2 and t^2 are estimated by fitting the linear mixed model:

$$y = X_1 ß + X_2 u + e,$$

where $y = (y_{11}, y_{12}, ..., y_{mn_m})^T$, $ß = (ß_0, ß_1)^T$, $u = (ß_2, ..., ß_{K+1}, u_1, ..., u_m)^T$,

$$\mathbf{X_1} = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ . & . \\ . & x_1 \\ . & x_2 \\ . & . \\ . & . \\ . & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix}, \text{ and } \mathbf{X_2} = \begin{bmatrix} (x_1 - k_1)_+ & ... & (x_1 - k_K)_+ & 1 & 0 & . & 0 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_1 - k_1)_+ & ... & (x_1 - k_K)_+ & 1 & 0 & . & . \\ (x_2 - k_1)_+ & ... & (x_2 - k_K)_+ & 0 & 1 & . & . \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_2 - k_1)_+ & ... & (x_2 - k_K)_+ & 0 & 1 & . & 0 \\ . & ... & . & 0 & 0 & . & 1 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_m - k_1)_+ & ... & (x_m - k_K)_+ & 0 & 0 & . & 1 \end{bmatrix}, \quad (2)$$

where x_h in X_1 and $(x_h - k_l)_+$ in X_2 are both repeated n_h times. u and e are mutually independent and

$$u = (ß_2, ..., ß_{K+1}, u_1, ..., u_m)^T \sim N_{K+m}(0, G), \; G = \begin{bmatrix} s_ß^2 I_K & 0 \\ 0 & t^2 I_m \end{bmatrix}.$$

Variance components s^2, s_b^2 and t^2 are estimated by fitting model (2) with the restricted maximum likelihood (REML) algorithm.

The predicted means of clusters included in the sample are given by:

$\hat{\mu} = X_1 \hat{ß} + X_2 \hat{u}$, where $\hat{ß} = (X_1^T \hat{V}^{-1} X_1)^{-1} X_1^T \hat{V}^{-1} \bar{y}$, $\hat{u} = \hat{G} X_2^T \hat{V}^{-1} (\bar{y} - X_1 \hat{ß})$, where $V = X_2 G X_2^T + s^2 S$, $S = diag[\{1/n_h\}_{h=1}^m]$ and $\bar{y} = (\bar{y}_1, ..., \bar{y}_m)^T$.

The predicted mean for a cluster *h* that is not selected in the first stage is

$\hat{\mu}_h = x_h^T \hat{ß}^*$, where $x_h = [1 \; x_h \; (x_h - k_1)_+ ... (x_h - k_K)_+]^T$ and $\hat{ß}^* = [\hat{b}_0 \; \hat{b}_1 ... \hat{b}_{K+1}]^T$.

Using the predicted cluster means $\hat{m}_h = E(y_{hi} \mid x_h)$, we have the model-based estimator $E(\bar{Y} \mid x,y) = \frac{1}{N}\left(\sum_{h=1}^{m}[n_h \bar{y}_h + (N_h - n_h)\hat{m}_h] + \sum_{h=m+1}^{H} N_h \hat{m}_h\right)$.

3. Variance Estimation Methods

3.1 Empirical Bayes Model-based Variance

Model (2) can be interpreted as a Bayes model in which the parameters $u = (ß_2,\ldots,ß_{K+1},u_1,\ldots,u_m)^T$ have multivariate normal prior $N_{K+m}(0,G)$,

$$G = \begin{bmatrix} s_ß^2 I_K & 0 \\ 0 & t^2 I_m \end{bmatrix},$$

and b_0, b_1, s^2, s_b^2 and t^2 all have the flat priors. This leads to the Bayes posterior variance for the vector $(b_0, b_1,\ldots, b_{K+1}, u_1,\ldots,u_m)^T$ conditional on s^2, s_b^2 and t^2 as $Var(b_0, b_1,\ldots, b_{K+1}, u_1,\ldots,u_m \mid s^2, s_b^2, t^2, y)^T = s^2(X^T X + \Delta)^{-1}$, where $X = [X_1\ X_2]$ and

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & s^2/s_b^2 I_K & 0 \\ 0 & 0 & 0 & s^2/t^2 I_m \end{bmatrix},$$

where I_K and I_m are K by K and m by m identity matrices, respectively.

The empirical Bayes posterior variance for $(b_0, b_1,\ldots, b_{K+1}, u_1,\ldots,u_m)^T$ is then calculated by replacing s^2, s_b^2 and t^2 by their maximum likelihood (ML) or restricted maximum likelihood (REML) estimators $\hat{s}^2, \hat{s}_b^2$ and $\hat{t}^2$, respectively, in the above formula. The empirical Bayes method underestimates the true posterior variance.

However the underestimation is not severe. A fully Bayes solution is also possible, but is not covered here.

The predicted population mean is $\hat{T}_{pred}/N$, where $\hat{T}_{pred} = T_1 + \hat{T}_2$, where $T_1 = \sum_{h=1}^{H} n_h \bar{y}_h$, the total of the sample, and $\hat{T}_2$ is the estimated total for those units not included in the sample, i.e.,

$$\hat{T}_2 = \sum_{h=1}^{m} (N_h - n_h)\hat{m}_h + \sum_{h=m+1}^{H} N_h \hat{m}_h = N_P X_P [\hat{b}_0 \ \hat{b}_1 \ ... \ \hat{b}_{K+1} \ \hat{u}_1 \ ... \ \hat{u}_m]^T ,$$

where $N_P = [(N_1 - n_1) \ ... \ (N_m - n_m) \ N_{m+1} ... N_H],$ and

$$X_P = \begin{bmatrix} 1 & x_1 & (x_1 - k_1)_+ & ... & (x_1 - k_K)_+ & 1 & 0 & ... & 0 \\ . & . & . & ... & . & 0 & 1 & 0 & 0 \\ . & . & . & ... & . & . & . & 0 & . \\ . & . & . & ... & . & 0 & 0 & 1 & 0 \\ 1 & x_m & (x_m - k_1)_+ & ... & (x_m - k_K)_+ & 0 & ... & 0 & 1 \\ 1 & x_{m+1} & (x_{m+1} - k_1)_+ & ... & (x_{m+1} - k_K)_+ & 0 & ... & ... & 0 \\ . & . & & ... & . & . & ... & ... & . \\ . & . & & ... & . & . & ... & .. & . \\ . & . & & ... & . & . & ... & ... & . \\ 1 & x_H & (x_H - k_1)_+ & ... & (x_H - k_K)_+ & 0 & ... & ... & 0 \end{bmatrix}. \qquad (3)$$

The empirical Bayes posterior variance for $\hat{\bar{Y}} = \hat{T}_{pred}/N$ is

$$Var(\hat{\bar{Y}} \mid s^2, s_b^2, t^2, X, X_P) = s^2 (N_P X_P (X^T X + \Delta)^{-1} X_P^T N_P^T)/N^2 .$$

3.2 The Jackknife Method

A jackknife variance estimator is developed for the PMM estimator. The jackknife replicates are constructed by dividing the set of PSU's into *G* subgroups with the same number of PSU's and computing the *g*th pseudovalue as $\hat{\bar{Y}}_g = G\hat{\bar{Y}} - (G-1)\hat{\bar{Y}}_{(g)}$,

where $\hat{\bar{Y}}$ is the original PMM estimator and $\hat{\bar{Y}}_{(g)}$ is the same estimator calculated from the reduced sample not including the elements from the PSU's in the *k*th subgroup.

The variance estimate of $\hat{\bar{Y}}$ is

$$v(\hat{\bar{Y}}) = \frac{1}{G(G-1)} \sum_{g=1}^{G} (\hat{\bar{Y}}_g - \hat{\bar{\bar{Y}}})^2 ,$$

where $\hat{\bar{\bar{Y}}} = \sum_{g=1}^{G} \hat{\bar{Y}}_g / K$. 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 first stage inclusion probabilities, and the *G* subgroups are then constructed by randomly selecting one element from each stratum. To save computation, estimates $\hat{s}^2, \hat{s}_b^2$ and $\hat{t}^2$ are not recomputed for each replicate. That is, we can compute pseudovalues of $(b_0, b_1, ..., b_{K+1}, u_1, ..., u_m)^T$ based on the variance components estimated from the whole sample.

Miller (1974) proved the asymptotic properties of the jackknife estimator in the case of multiple regressions. Hinkley (1977) gave weighted jackknife with improved performance. Shao and Wu (1987, 1989) discussed the general properties of jackknife variance estimation in linear regression models. In Zheng and Little (2002), we gave a theoretical justification for the jackknife method for the p-spline model-based estimator in the simple case of one-stage designs. Numerical simulations in section 4 suggest the above described jackknife method also works well for the two stage design.

Improvements in the spirit of Hinkley (1977) are possible and will be considered in future work.

3.3 The Balanced Repeated Replicate Method

The BRR method can be applied when the design is stratified with two units sampled in each stratum. In practice, collapsing of strata and random combinations of units within strata (Kalton, 1977) are often employed for BRR variance estimation. In our application we assume in the first stage that the primary sampling units are sampled systematically from a randomly ordered list. This can be viewed approximately as a stratified design with *n* strata each consisting of PSU's with cumulative measures of approximate size $\sum_{i=1}^{N} z_i \Big/ n$, where z_i are the measures of size for the PSU's . One PSU is sampled from each of the *n* strata. Assuming *n* is even, the design can be approximated by a stratified design with *n*/2 strata with measures of size $2\sum_{i=1}^{N} z_i \Big/ n$, and two units are sampled per stratum. Balanced repeated half samples are then constructed by selecting one PSU from each stratum, with the selection scheme based on Hadamard matrices (Plackett and Burman, 1946). Let $\hat{\bar{Y}}_b$ be the p-spline estimator computed from the *b*th 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 $v_{BRR}(\hat{\bar{Y}}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\bar{Y}}_b - \hat{\bar{Y}})^2$.

By treating the design as if it was stratified with two PSU's included per stratum, the BRR method gives biased estimates for the true variance of the p-spline estimator.

4. When the Cluster Counts are Unknown

In sections 2 and 3 we assumed that the cluster counts $N_h, h = 1...H$ for all sampled or non-sampled clusters are known. In this section we discuss the common

situation where $p_{1,h}, p_{2,h}, h = 1...H$ are known and N_h are only known exactly for the sampled clusters (labeled 1 through m). We also assume that values of some auxiliary variable $M_h, h = 1...H$ are known for the whole population and have a close relationship with N_h. This information may be cluster counts estimated from outside sources such as a census.

In this situation, we use an additional regression model to estimate N_h for those non-sampled clusters based on M_h. We then replace the $N_h, h = m+1,...,H$ in (3) by estimates $\hat{N}_h, h = m+1,...,H$. The resulting estimate of the total is

$$\tilde{T} = T_1 + \sum_{h=1}^{m} (N_h - n_h)\hat{m}_h + \sum_{h=m+1}^{H} \hat{N}_h \hat{m}_h \ .$$

The variance estimate of $\tilde{T}$ needs to incorporate the additional variability in $\hat{N}_h$. In particular, a model-based variance for $\tilde{T}$ is

$$Var(\tilde{T} \mid p_h, M_h) = Var(E(\tilde{T} \mid \hat{N}_h, p_h, M_h)) + E(Var(\tilde{T} \mid \hat{N}_h, p_h, M_h)),$$

where the expectations are taken under the distributions described in the superpopulation models. $E(\tilde{T} \mid \hat{N}_h, p_h, M_h) = \sum_{h=1}^{m} (N_h - n_h) m_h + \sum_{h=m+1}^{H} \hat{N}_h m_h$ and

$$Var(\tilde{T} \mid \hat{N}_h, p_h, M_h) \approx s^2 (\tilde{N}_P X_P (X^T X + \Delta)^{-1} X_P^T \tilde{N}_P^T),$$

where $\tilde{N}_P = [(N_1 - n_1)...(N_m - n_m)\ \hat{N}_{m+1} ... \hat{N}_H]$, and X, X_P and Δ are defined in (3).

If both models for estimating m_h and N_h are correctly specified, the above variance can be estimated according to the corresponding models.

5. Simulations

5.1 Simulation Design

Two simulation studies are conducted to compare the inverse probability weighting method, the model-assisted method discussed in Särndal et al. (1992) and the PMM method in the case of two-stage samples.

In our first simulation study, artificial populations are generated with different mean functions $f(p_{1,h})$ of the first stage inclusion probabilities. Four different mean functions are simulated: 1) NULL, a constant function; 2) LINUP, a linearly increasing function; 3) LINDOWN, a linearly decreasing function; and 4) EXP, an exponentially increasing function.

Two combinations of values for variance components are: 1) $s = 0.1$ and $t = 0.2$; 2) $s = 0.2$ and $t = 0.1$. Only normal errors around the mean functions are simulated while both normal and lognormal within-cluster errors are simulated.

The total number of PSU's is 500. The first stage samples are systematic probability-proportional-to-size (PPS) with 48 PSU's included in the sample. The size variables in the PPS sampling take integer values ranging from 4 to about 400. The SSU count in each PSU is generated with the mean equals 1.05 times the measure of size and with log-normal errors with standard deviation 30.

Two types of second stage sampling plans are studied: 1) within-cluster simple random sampling (srs) with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities, resulting in an equal inclusion probability for all SSU's.; 2) within-cluster simple random sampling with the same sampling rate across sampled

PSU's, so that the resulting inclusion probabilities for the SSU's in PSU h are proportional to $p_{1,h}$.

For both sampling plans, the following methods are computed:

A. The HT estimator.

B. The model-assisted estimation method. We use a linear model regressing the outcome y_{hi} on the first stage inclusion probabilities, which are treated as element-level information. The GR estimator is computed by the formula given in Section 1.

C. The PMM method, with the first stage inclusion probabilities $p_{1,h}$ as the covariate. We use 20 equal percentiles of $p_{1,h}$ of the sampled PSU's as the knots for p-spline regression.

D. The PMM method with the cluster means m_h estimated the same way as in C., but use estimated PSU counts from a simple linear regression model regressing N_h on the measures of size, which are proportional to $p_{1,h}$. This simulation is conducted to study the method described in section 5.

Estimates of $\bar{Y}$ from methods A-D are calculated for each of the 500 samples drawn repeatedly from the artificial populations (each artificial population is generated only once). For methods A-C, we also compare the variance estimation methods of the PMM estimator in the first simulation study. We compute the empirical Bayes, the jackknife (K=8) and BRR variance estimators for each repeated sample. The mean estimate for the variance of PMM as well as the coverage rate of the corresponding 95% confidence interval are used to judge the quality of inference. For method D, we study a model-based variance estimator, also judged by empirical bias and coverage rates. The computational method is given in section 5.

In the second simulation study, we use household income data from 5% public use microdata sample (PUMS) in 1990 US Census. We concentrate on the household income data in the state of Michigan and treat the 5% PUMS as the finite population. This population does not necessarily replicate the true distribution of household income in the state of Michigan, but serves as a population whose finite population quantities are known. This simulation is more realistic than the previous simulation in that the outcome values are drawn from a real rather than simulated distribution.

The clusters we simulate are based on the natural geographical clusters, or "Public Use Microdata Areas" (PUMAs). PUMAs are typically counties and places. There are 67 PUMAs in the Michigan 5% PUMS, with counts of families ranging from around 1300 to over 10000. We increase the number of available PSU's by dividing each PUMA into 5, resulting in 335 PSU's. The PSU counts ranges from 134 to 3058. Figure 1 gives the scatter plot of one sample of the average household income versus sampled cluster sizes together with the regression curve $\hat{f}(x)$.

The two-stage sampling is with equal probability. The first stage sampling is PPS drawn with systematic sampling where the measure of size is equal to the PSU counts. The second stage sample is simple random sample with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities. In the estimation of the mean, we use the true cluster counts. We draw 500 repeated two-stage samples. In the first stage, 30 PSU's are drawn from the total 335 PSU's. In the second stage, 20 SSU's (families) are drawn from each selected PSU's. We apply the p-spline nonparametric mixed model formulated as in (2). For the knots of the p-spline function, we use 10

equally spaced sample percentiles of the PSU counts, i.e., the 100/11th, 200/11th, … ,1000/11th percentiles of the sampled PSU counts.

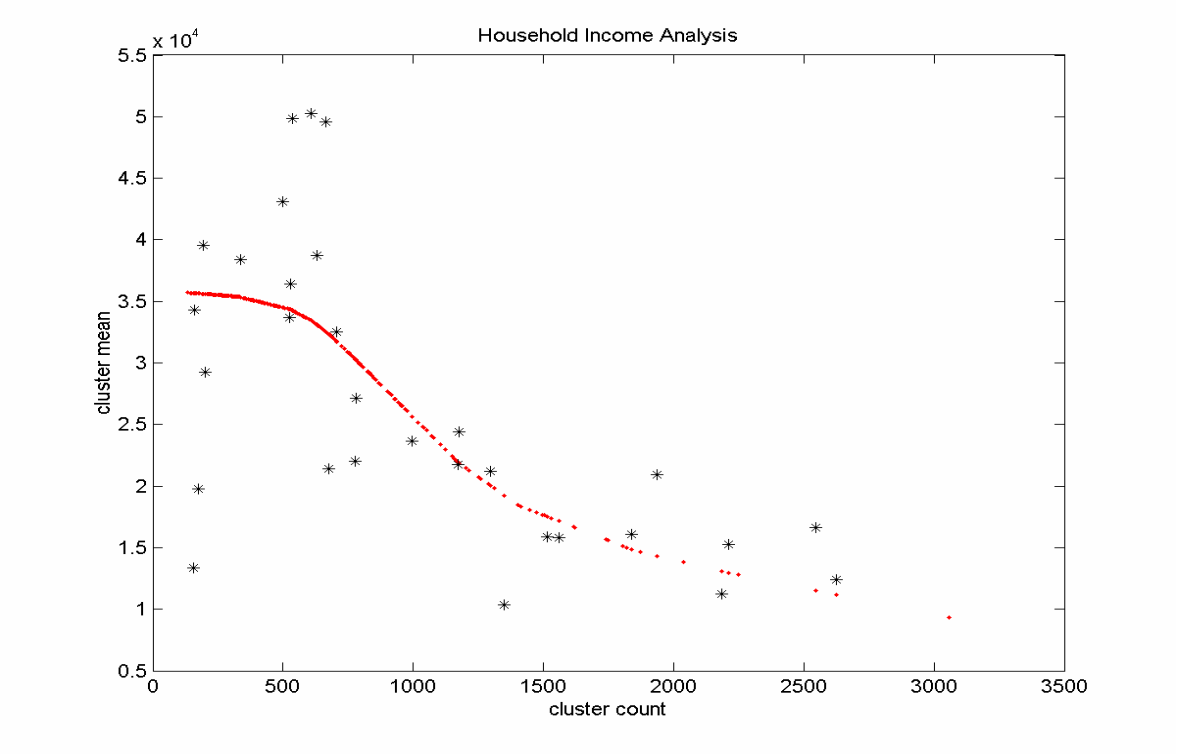

Figure 1. P-spline Regression Curve (dotted line) and the Average Household Income (stars) in Sampled PSU's

5.2 Results

Tables 1-2 give the empirical bias and root mean squared error (RMSE) for populations generated with both normal within-cluster errors and two (s, t) combinations.

Tables 1-2 suggest the PMM based methods give estimators with small biases. From these tables, we also see in the case of equal probability sampling**,** the PMM estimator is roughly as efficient as HT estimator when the mean function f is constant. In the more general cases such as LINUP, LINDOWN, where f is linear but not

constant, the linear model-assisted and PMM method are comparable and both are more efficient in terms of root mean squared error than the HT estimator. For population EXP, whose mean function is not linear, the PMM method is clearly superior to both the HT and the linear model-assisted estimators. The improvement of efficiency requires the knowledge of complete design information including probabilities $p_{1,h}$ and PSU counts N_h for the whole population. When using estimated cluster counts $\hat{N}_h$ in the place of N_h, the resulting estimator is slightly less efficient than in the case with known N_h, but the PMM estimator still outperforms the HT when the mean function is non-constant. Tables 3 and 4 show similar findings for data with log-normal within-cluster errors.

Tables 5-8 show, for unequal probability sampling, a similar pattern of comparison as in the case of equal probability sampling. This suggests that the key to improved efficiency is the better prediction given by the nonparametric models.

Tables 1-8 all show that the p-spline model-based estimators have very small empirical design-biases. We believe this is because the flexible mean functions yield good predictions of the cluster (PSU) means.

Tables 9-12 compare three proposed variance estimation methods: the empirical Bayes model-based method, the Jackknife method and the BRR method. These tables indicate the proposed PMM inference methods have coverage for the true mean close to the nominal value of 95%. The empirical Bayes method tends to underestimates the true variance of PMM estimator, resulting in under-coverage in some cases. The jackknife and the BRR methods tend to yield more robust estimates for the variance. In general, PMM allows us to draw satisfactory inference for the population mean while providing

estimates with improved efficiency over the traditional HT and linear model-assisted estimators.

Tables 13 and 14 give the empirical variance of the PMM estimator when the non-sampled cluster counts N_h are estimated. It also gives the mean estimates of the variance of this estimator and coverage rates. These two tables show the inference method discussed in section 5 tends to underestimate the true variance of PMM estimator using $\hat{N}_h$, giving in occasional under-coverage of the population mean. It remains to be studied in the future whether the JRR and BRR method also give satisfactory inference for this method.

For the simulation study using 5% PUMS data, the p-spline nonparametric mixed model based method has bias= $-41.9 and RMSE=$2153, the simple mean has bias=$-50.9 and RMSE=$2600. Both methods have small biases. The model-based estimator has an RMSE 17% less than the RMSE of the simple mean. This improved efficiency is due to the fact that the average household income decreases for as the number of families in the clusters increases (see figure 1). The PMM method exploits this relationship in its predictions.

6. Discussion

Previous parametric model-based estimators of finite population quantities have been criticized mainly for their potential for large design bias when the mean structure of the models is misspecified. In our nonparametric models, the linearity assumption is replaced by a much weaker assumption of a smoothly-varying relationship. As a result, the model-based estimators are more robust and have small biases.

Design information such as inclusion probabilities and information such as cluster counts play key roles in the estimation of finite population quantities. Inverse probability weighting often corresponds to simple model assumptions about the relationship between the outcome variables and the design variables. In the method we propose, the gain in efficiency is realized by applying nonparametric models that relax these assumptions.

Our study has an interesting finding that the model-based estimation can be more efficient than the simple mean estimation in an equal probability design. In other studies, we also find gains in efficiency from p-spline nonparametric mixed model in estimating post-stratum means in post-stratified samples.

The empirical Bayes method, the jackknife and BRR methods all give sound coverage of the true design-based variance of the proposed estimator. This means we are able to draw valid inference with confidence intervals that are narrower than those given by the traditional methods. However, we expect the empirical Bayes method to be sensitive to model assumptions on the variance components (e.g., constant within-cluster variances). When the cluster counts are not known for the sample but not for the whole population, model-based estimates of the unknown counts can still provide sound estimates of the population mean, if the model tracks the true cluster counts precisely enough. The model between these counts and the auxiliary variable was treated parametrically here, but this could also be specified nonparametrically without much difficulty.

In the future, we plan to apply p-spline nonparametric mixed models to more complex cases such as stratified and multi-stage designs. We also plan to consider generalized p-spline nonparametric mixed models for non-normally distributed outcomes.

Acknowledgements

This research was supported by grant DMS 0106914 from the National Science Foundation.

References

Brumback, B.A., Ruppert, D. and Wand, M.P. (1999). Comment to "Variable selection and function estimation in additive nonparametric regression using data-based prior". Journal of the American Statistical Association, 94, 794-797

Coull, B.A., Schwartz, J. and Wand, M.P. (2001) Respiratory Health and Air Pollution: Additive Mixed Model Analyses.

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

Gosh, M. and Meeden, G. (1986). Empirical Bayes Estimation of Means from Stratified Samples. Journal of the American Statistical Association, 81, 1058-1062

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

Holt, D. and Smith, T.M.F. (1979). Poststratification. Journal of the Royal Statistical Society, A142, 33-46

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

Lazzaroni, L.C. and Little, R.J.A. (1998). Random Effects Models for Smoothing Poststratification Weights. Journal of Official Statistics, 14, 61-78

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models using

smoothing splines. Journal of the Royal Statistical Society, B 61, 381-400

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.

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

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.

U.S. Dept. of Commerce, Bureau of the Census. Census of Population and Housing, 1990 [United States]: public use microdata sample: 5- percent sample Computer file]. 3rd release. Washington, DC: U.S. Dept. of Commerce, Bureau of the Census [producer], 1995. Ann Arbor, MI: Inter-University Consortium for Political and Social Research [distributor], 1996.

Zheng, H and Little, R.J.A. (2002a) Penalized Spline Model-based Estimation of the Finite Population Total From Probability-Proportional-to-Size Samples, to appear on Journal of Official Statistics.

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

Table 1. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with normal within-cluster noise ($s = 0.1$ and $t = 0.2$) and samples under an equal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	1.1	29.7	0.8	30.0	0.8	29.9	1.3	30.1
LINUP	-3.9	29.0	-5.2	34.5	-5.1	28.9	-3.8	29.3
LINDOWN	3.5	30.7	3.6	36.4	3.7	30.7	2.3	30.4
EXP	-4.4	29.1	-9.4	53.0	-9.5	36.7	-4.3	29.1

Table 2. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with normal within-cluster noise ($s = 0.2$ and $t = 0.1$) and samples under an equal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	5.7	22.0	6.6	22.5	6.6	22.1	5.5	22.3
LINUP	-0.9	22.2	-0.2	27.7	-1.8	22.2	-0.5	22.3
LINDOWN	0.5	20.4	-0.6	27.1	-0.3	20.5	1.6	20.6
EXP	0.9	23.1	1.9	50.3	-4.2	31.7	0.4	23.4

Table 3. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with log-normal within-cluster noise ($s = 0.1$ and $t = 0.2$) and samples under an equal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	1.7	32.3	0.9	32.3	0.7	32.3	1.5	32.5
LINUP	-0.6	30.0	-2.6	33.2	-1.4	30.4	-0.6	30.0
LINDOWN	2.9	31.9	3.8	39.4	2.7	32.1	3.2	32.0
EXP	-0.6	28.4	-5.9	51.5	-6.9	36.4	-0.3	28.5

Table 4. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with log-normal within-cluster noise ($s = 0.2$ and $t = 0.1$) and samples under an equal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	8.5	30.5	9.6	31.3	9.2	31.0	9.1	30.8
LINUP	12.8	29.0	14.7	35.3	13.8	29.5	12.7	29.5
LINDOWN	3.6	32.3	1.9	37.5	3.6	32.1	6.4	33.1
EXP	3.9	29.0	6.8	53.8	1.0	34.4	3.7	29.4

Table 5. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with normal within-cluster noise ($s = 0.1$ and $t = 0.2$) and samples under an unequal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	-4.5	29.3	-3.7	33.6	-3.2	30.5	-4.5	29.3
LINUP	3.2	28.6	0.4	38.9	1.3	31.2	4.5	28.7
LINDOWN	-0.9	27.0	3.7	35.5	1.8	27.7	-0.7	26.9
EXP	5.8	32.0	1.9	56.8	0.4	39.4	14.1	34.4

Table 6. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with normal within-cluster noise ($s = 0.2$ and $t = 0.1$) and samples under an unequal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	-7.7	21.3	-7.7	24.9	-6.6	21.1	-7.6	21.2
LINUP	-6.7	21.0	-6.2	35.8	-6.6	21.3	-8.6	21.7
LINDOWN	1.1	20.7	3.2	30.6	1.2	20.7	3.5	21.1
EXP	-2.3	20.9	-6.5	53.3	-7.2	30.0	-3.0	20.9

Table 7. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with log-normal within-cluster noise ($s = 0.1$ and $t = 0.2$) and samples under an unequal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	-0.5	28.5	-2.0	30.6	-2.1	29.5	-0.3	28.5
LINUP	0.4	28.8	-0.3	43.4	1.5	30.1	-3.1	29.0
LINDOWN	5.4	32.6	5.0	39.0	3.7	34.1	6.0	32.7
EXP	-1.3	28.6	-7.6	62.6	-7.1	36.8	-9.3	30.3

Table 8. Empirical Biases and RMSE of PMM, HT, GR and PMM with estimated N_h for data with log-normal within-cluster noise ($s = 0.2$ and $t = 0.1$) and samples under an unequal probability design.

$(\times 10^{-3})$	PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
NULL	15.4	34.1	14.0	35.1	14.9	33.5	15.2	33.8
LINUP	-2.6	23.7	-5.6	33.2	3.7	23.6	-3.3	23.9
LINDOWN	6.0	26.8	9.3	37.5	7.5	27.3	2.5	26.0
EXP	0.8	26.3	-2.3	50.8	-3.5	33.1	11.5	29.0

Table 9. Variance estimation and empirical coverage rates of 95% C.I. using three inference methods for data with normal within-cluster noise ($s = 0.1$ and $t = 0.2$) target coverage (93-97%).

	Empirical variance ($\times 10^{-5}$)	Empirical Bayes Model-based		Jackknife(K=8)		BRR	
		Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%
NULL	88	74	92.8	94	96.4	96	94.4
LINUP	84	64	91.2	80	94.6	82	93.4
LINDOWN	94	73	89.6	94	94.6	98	94.2
EXP	85	70	91.4	88	94.6	85	93.4

Table 10. Variance estimation and empirical coverage rates of 95% C.I. using three inference methods for data with normal within-cluster noise ($s = 0.2$ and $t = 0.1$), target coverage (93-97%).

	Empirical variance ($\times 10^{-5}$)	Empirical Bayes Model-based		Jackknife(K=8)		BRR	
		Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%
NULL	48	45	93.8	48	96.0	49	93.8
LINUP	49	43	92.4	48	95.2	47	93.0
LINDOWN	42	45	96.8	51	96.2	51	96.8
EXP	53	54	95.0	61	97.2	59	95.2

Table 11. Variance estimation and empirical coverage rates of 95% C.I. using three inference methods for data with log-normal within-cluster noise ($s = 0.1$ and $t = 0.2$), target coverage (93-97%).

	Empirical variance ($\times 10^{-5}$)	Empirical Bayes Model-based		Jackknife(K=8)		BRR	
		Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%
NULL	104	83	91.8	104	94.8	100	93.6
LINUP	90	87	93.2	97	95.4	98	94.8
LINDOWN	102	98	93.6	106	95.6	107	95.0
EXP	81	77	93.4	97	96.4	89	94.8

Table 12. Variance estimation and empirical coverage rates of 95% C.I. using three inference methods for data with log-normal within-cluster noise ($s = 0.2$ and $t = 0.1$), target coverage (93-97%).

	Empirical variance ($\times 10^{-5}$)	Empirical Bayes Model-based		Jackknife(K=8)		BRR	
		Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%	Estimate ($\times 10^{-5}$)	%
NULL	93	97	94.2	100	96.2	99	95.2
LINUP	84	71	95.4	78	96.6	76	95.2
LINDOWN	104	101	93.6	106	96.0	102	92.8
EXP	84	81	94.6	84	95.2	82	95.0

Table 13. Variance estimation and empirical coverage rates of 95% C.I. using P-spline and estimated cluster counts, Population simulated with normal errors, target coverage (93-97%).

	$s = 0.1$ and $t = 0.2$			$s = 0.2$ and $t = 0.1$		
	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate
NULL	90	76	91.8	50	46	93.2
LINUP	86	65	90.8	50	45	92.6
LINDOWN	93	74	90.4	43	46	95.6
EXP	85	72	93.0	55	56	96.2

Table 14. Variance estimation and empirical coverage rates of 95% C.I. using P-spline and estimated cluster counts, Population simulated with log-normal errors, target coverage (93-97%).

	$s = 0.1$ and $t = 0.2$			$s = 0.2$ and $t = 0.1$		
	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate	Empirical Variance ($\times 10^{-5}$)	Estimated Variance ($\times 10^{-5}$)	Coverage Rate
NULL	105	84	91.8	95	99	94.8
LINUP	90	89	93.8	87	73	95.0
LINDOWN	103	98	94.4	110	102	94.4
EXP	81	79	94.6	87	83	94.2