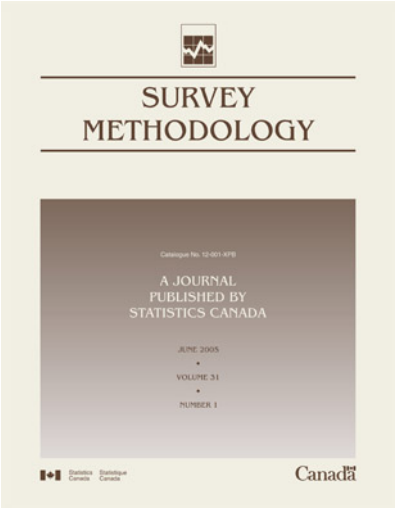
Catalogue no. 12-001-XIE

Survey Methodology

December 2004



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2004

Published by authority of the Minister responsible for Statistics Canada

April 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 30, No. 2, pp. 209-218
Statistics Canada, Catalogue No. 12-001

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

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

Abstract

Samplers often distrust model-based approaches to survey inference because of concerns about misspecification when models are 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. 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 (2003, 2004) 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 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 (BRR) methods. Simulation studies on simulated data and 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 for a common 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.

Key Words: Weighting; REML; Empirical Bayes estimation.

1. Introduction

In a sample survey, let y_i denote the value of an outcome Y for unit i, and let S denote the set of sampled units. The Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952) $\hat{Y}_{\mathrm{HT}} = \sum_{i\in S} y_i/\pi_i$, where π_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 π_i:

$$y_i = \beta\pi_i + \pi_i\varepsilon_i,$$

where ε_i is assumed to be i.i.d. normally distributed with mean zero and variance σ^2.

In Zheng and Little (2003, 2004), we proposed a nonparametric model

$$y_i = f(\pi_i) + \varepsilon_i, \varepsilon_i \sim \text{ind } N\left(0, \pi_i^{2k}\sigma^2\right),$$

using penalized splines to model mean of outcome y_i as a smoothly-varying function f of the inclusion probabilities π_i. We showed in Zheng and Little (2003) 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.

In this article we consider two-stage sampling. In the first stage, a subset of m primary sampling units (PSUs) is drawn from a population with H PSUs with unequal probabilities $\pi_{1,h}$, $h = 1, ..., H$. Let us number the included PSUs from 1 to m. In the second stage, a simple random sample (srs) of n_h out of N_h secondary sampling units (SSUs) is drawn from the sampled PSU labeled h with probability $\pi_{2,h}$. The overall selection probability for unit i in PSU h is $\pi_h = \pi_{1,h}\pi_{2,h}$, and the HT estimator of the mean of an outcome Y is $\bar{y}_w = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi}/(\pi_{1,h}\pi_{2,h})/N$, where y_{hi} is the value of Y for unit i in PSU h and N is the known total number of units (SSUs) 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 π_h are equal for all SSUs.

1. Hui Zheng, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115. E-mail: zheng@hcp.med.harvard.edu; Roderick J.A. Little, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109. E-mail: rlittle@umich.edu.

The inverse probability weighted mean in this case equals the simple sample mean $\bar{y} = \sum_{h=1}^{m}\sum_{i=1}^{n_h} y_{hi} / \sum_{h=1}^{m} n_h$.

We assume throughout this article that the selection probabilities $\pi_{1,h}$ are known for all the PSUs $h = 1, \ldots, H$. In sections 2 and 3, we assume the PSU counts N_h are also known for all the PSUs in the population. In section 4, we discuss the common situation where N_h is only known for sampled PSUs, but the N_h for nonsampled PSUs can be estimated using a regression model based on auxiliary variables known for all PSUs in the population.

Särndal, Swensson and Wretman (1992) discussed model-assisted alternatives to the HT estimator for two-stage samples with auxiliary information available at the PSU or SSU level. In the first case, let x_h denote a vector of PSU-level auxiliary variables for PSU h. The PSU totals $t_h = \sum_{i=1}^{N_h} y_{hi}$ are assumed to be related to x_h according to a linear model:

$$E(t_h \mid x_h) = x_h^T \beta, \mathrm{Var}(t_h) = \sigma_h^2, \ h = 1, \ldots, H$$

(Särndal *et al.* 1992). β is estimated by the probability-weighted regression

$$\hat{B} = \left(\sum_{h=1}^{m} x_h x_h^T / \left(\sigma_h^2 \pi_{1,h}\right)\right)^{-1} \sum_{h=1}^{m} x_h t_h^* / \left(\sigma_h^2 \pi_{1,h}\right),$$

where $t_h^* = \sum_{i=1}^{n_h} y_{hi} / \pi_{2,h}$, leading to the projected totals $\hat{t}_h = x_h^T \hat{B}$, $h = 1, \ldots H$. In practice, estimates $\hat{\sigma}_h^2$, either simply assumed (*e.g.*, σ_h proportional to a measure of size of stratum h) or estimated, replace σ_h^2 in the above formula. The generalized regression (GR) estimator of the grand total is

$$\hat{T}_A = \sum_{i=1}^{H} \hat{t}_h + \sum_{h=1}^{m} \frac{(t_h^* - \hat{t}_h)}{\pi_{1,h}},$$

and the estimate for the mean is $\hat{T}_A / N$. The term $\sum_{h=1}^{m} (t_h^* - \hat{t}_h) / \pi_{1,h}$ is a bias calibration term that makes the estimator design- consistent.

In the second case where auxiliary information is known at the SSU level, let x_{hi} denote the set of auxiliary variables for SSU i in PSU h, $h = 1, \ldots, H; i = 1, \ldots, N_h$. The relationship between the outcome and the auxiliary information is modeled by

$$E(y_{hi} \mid x_{hi}) = x_{hi}^T \beta, \ldots, \mathrm{Var}(y_{hi}) = \sigma_{hi}^2, \ h = 1, \ldots, H, \ i = 1, \ldots, N_h.$$

The probability weighted regression estimate for β is

$$\hat{B} = \left(\sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi} x_{hi}^T / \left(\sigma_{hi}^2 \pi_{hi}\right)\right)^{-1} \sum_{h=1}^{m}\sum_{i=1}^{n_h} x_{hi}\, y_{hi} / \left(\sigma_h^2 \pi_{hi}\right),$$

where π_{hi} is the probability for unit (h, i) to be included in the sample. The GR estimator for the grand total is

$$\hat{T}_B = \sum_{h=1}^{H}\sum_{i=1}^{N_h} \hat{y}_{hi} + \sum_{h=1}^{m}\sum_{i=1}^{n_h} \frac{(y_{hi} - \hat{y}_{hi})}{\pi_{hi}},$$

where $\hat{y}_{hi} = x_{hi}^T \hat{B}$. The estimator for the mean is $\hat{T}_B / N$.

These two methods do not account for the within-PSU correlations of outcome. These correlations can be modeled by treating PSU means as random effects in a hierarchical model. For the case where PSU-level information x_h is available for all PSUs, one such model is:

$$y_{hi} \mid \mu_h \overset{\text{ind}}{\sim} N(\mu_h, \ \sigma^2)$$
$$\mu \sim N_H(\varphi, \ D) \qquad (1)$$

where $\mu = (\mu_1, \ldots, \mu_H)^T$, $\varphi = (\varphi_1, \ldots, \varphi_H)^T$ where μ_h is the mean outcome in PSU h, $\varphi_h = x_h^T \beta$, and D is the covariance matrix of the PSU means. The model-based estimator of $\bar{Y}$ is given by

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

where $\hat{\mu}_h = \hat{E}(y_{hi} \mid \mathbf{y}, x_h)$, and $\mathbf{y}$ is the vector of outcomes in the sample.

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

Exchangeable random effects (XRE): (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \ldots, H$ and $D = \tau^2 I_H$;

Autoregressive (AR1): (Lazzaroni and Little 1998): $\varphi_h \equiv \mu_o, h = 1, \ldots, H$ and $D = r^2 \{\rho^{|i-j|}\}$;

Linear (LIN): (Lazzaroni and Little 1998): $\varphi_h = \alpha + \beta x_h, h = 1, \ldots, H$ and $D = \tau^2 I_H$;

Nonparametric: (Elliott and Little 2000): $\varphi_h = f(x_h), h = 1, \ldots, H$ and $D = 0$.

The nonparametric models in Elliott and Little (2000) assume nonparametric mean function relating the outcome to the design variables. By assuming $D = 0$, the PSU means are modeled to equal the mean function f instead of varying around it. Nonparametric mixed models relax the assumptions on D (*e.g.*, $D = \tau^2 I_H$) and serve as a natural extension of the Elliott and Little (2000) model and linear mixed models with a parametric mean structure.

It is worth pointing out that some estimators in the above family of models correspond to standard design-based estimators. For example, in an equal-probability design where n_h are approximately constant across PSUs, the unweighted mean corresponds to the special model-based estimator that assumes φ_h is constant.

2. Estimation for the *P*-spline Mixed Model

The linear structure of φ 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 effects part in the LIN model. *P*-spline nonparametric mixed models (Lin and Zhang 1999; Brumback, Ruppert and Wand 1999; Coull, Schwartz and Wand 2001) are even more flexible, since they replace polynomials by smooth nonparametric functions. We propose the following *p*-spline nonparametric mixed model for inference about the population mean:

***P*-spline nonparametric mixed model** (PMM):

$$\varphi_h = f(x_h),\ h = 1, ..., H,\ D = \tau^2 I_H,$$

where f is a nonparametric degree p spline function:

$$f(x;\beta) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \sum_{l=1}^{K} \beta_{l+p} (x - \kappa_l)_+^p,$$

where $\kappa_1 < ... < \kappa_K$ are K fixed knots, $\beta_0, ..., \beta_{p+K}$ are coefficients to be estimated and $(x)_+^p = x^p \mathbf{I}(x \geq 0)$.

A naive way of estimating $\beta_0, ..., \beta_{p+K}$ is to treat them as fixed and estimate them together with the variance components σ^2 and τ^2 by fitting a linear mixed model. However this method can yield estimates of f with too much roughness and variability. To avoid overfitting, the roughness of the estimation $\hat{f}$ can penalized by adding a penalty term to the sum of squared deviations, so that the solution $\hat{\beta}_0, ..., \hat{\beta}_p$ is minimizes

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

This is achieved in the context of the model by assigning $\beta_0, ..., \beta_p$ flat priors, $(\beta_{p+1}, ..., \beta_{p+K})$ a normal prior $N_m(0, \sigma_\beta^2)$, and letting $\alpha = \tau^2 / \sigma_\beta^2$. The result is a penalized spline (p-spline) model.

When $p = 1$, $\hat{f}$ is piecewise linear and the coefficients $\beta_0, ..., \beta_{K+1}$ and σ^2, σ_β^2 and τ^2 are estimated by fitting the linear mixed model:

$$y = X_1 \beta + X_2 u + \varepsilon, \qquad (2)$$

where $y = (y_{11}, y_{12}, ..., y_{mn_m})^T$, $\beta = (\beta_0, \beta_1)^T$, $u = (\beta_2, ..., \beta_{K+1}, u_1, ..., u_m)^T$,

$$X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_1 \\ . & . \\ . & x_1 \\ . & x_2 \\ . & . \\ . & . \\ . & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_m \end{bmatrix},$$

$$X_2 = \begin{bmatrix} (x_1 - \kappa_1)_+ & ... & (x_1 - \kappa_K)_+ & 1 & 0 & . & 0 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_1 - \kappa_1)_+ & ... & (x_1 - \kappa_K)_+ & 1 & 0 & . & . \\ (x_2 - \kappa_1)_+ & ... & (x_2 - \kappa_K)_+ & 0 & 1 & . & . \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_2 - \kappa_1)_+ & ... & (x_2 - \kappa_K)_+ & 0 & 1 & . & 0 \\ . & ... & . & 0 & 0 & . & 1 \\ . & ... & . & . & . & . & . \\ . & ... & . & . & . & . & . \\ (x_m - \kappa_1)_+ & ... & (x_m - \kappa_K)_+ & 0 & 0 & . & 1 \end{bmatrix},$$

where x_h in X_1 and $(x_h - \kappa_l)_+$ in X_2 are both repeated n_h times. The random terms u and ε are mutually independent with

$$u = (\beta_2, ..., \beta_{K+1}, u_1, ..., u_m)^T \sim N_{K+m}(0, G),$$
$$G = \begin{bmatrix} \sigma_\beta^2 I_K & 0 \\ 0 & \tau^2 I_m \end{bmatrix}.$$

Variance components σ^2, σ_β^2 and τ^2 can be estimated by fitting model (2) by restricted maximum likelihood (REML).

The predicted means of PSUs included in the sample are given by: $\hat{\mu} = X_1 \hat{\beta} + X_2 \hat{u}$, where $\hat{\beta} = (X_1^T \hat{V}^{-1} X_1)^{-1} X_1^T \hat{V}^{-1} \bar{y}$, $\hat{u} = \hat{G} X_2^T \hat{V}^{-1} (\bar{y} - X_1 \hat{\beta})$, where $\hat{V} = X_2 \hat{G} X_2^T + \hat{\sigma}^2 \Sigma$, $\Sigma = \text{diag}[\{1/n_h\}_{h=1}^m]$ and $\bar{y} = (\bar{y}_1, ..., \bar{y}_m)^T$. The predicted mean for a PSU h that is not selected in the first stage is $\hat{\mu}_h = x_h^T \hat{\beta}^*$, where

$$x_h = [1 \;\; x_h \;\; (x_h - \kappa_1)_+ \ldots (x_h - \kappa_K)_+]^T$$

and

$$\hat{\beta}^* = [\hat{\beta}_0 \;\; \hat{\beta}_1, \;\ldots,\; \hat{\beta}_{K+1}]^T.$$

Combining the predictions, we obtain the model-based estimator of the population mean

$$\hat{E}(\bar{Y} \mid y, x_h) = \frac{1}{N}\Big(\sum_{h=1}^{m}[n_h \bar{y}_h + (N_h - n_h)\,\hat{\mu}_h] + \sum_{h=m+1}^{H} N_h\,\hat{\mu}_h\Big).$$

3. Variance Estimation Methods

3.1 Empirical Bayes Model-based Variance

Model (2) can be interpreted as a Bayes model in which the parameters $u = (\beta_2, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ have multivariate normal prior $N_{K+m}(0, G)$, and β_0, β_1, σ^2, σ_β^2 and τ^2 all have the flat priors. This leads to the Bayes posterior variance for the vector $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ conditional on σ^2, σ_β^2 and τ^2 as

$$\mathrm{Var}((\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T \mid \sigma^2, \sigma_\beta^2, \tau^2, y) = \sigma^2 (X^T X + \Delta)^{-1}$$

where $X = [X_1 \; X_2]$ and

$$\Delta = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2/\sigma_\beta^2\, I_K & 0 \\ 0 & 0 & 0 & \sigma^2/\tau^2\, I_m \end{bmatrix},$$

where I_K and I_m are $(K \times K)$ and $(m \times m)$ identity matrices, respectively.

The empirical Bayes posterior variance for $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ is calculated by replacing σ^2, σ_β^2 and τ^2 with their maximum likelihood (ML) or restricted maximum likelihood (REML) estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$, respectively. The empirical Bayes method underestimates the true posterior variance, but the underestimation is not severe for the sample sizes encountered in many survey settings. A fully Bayes solution is also possible, but is not covered here.

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

$$\hat{T}_2 = \sum_{h=1}^{m}(N_h - n_h)\,\hat{\mu}_h + \sum_{h=m+1}^{H} N_h \hat{\mu}_h = N_P X_P \left[\hat{\beta}_0 \; \hat{\beta}_1 \ldots \hat{\beta}_{K+1} \; \hat{\mu}_1 \ldots \hat{\mu}_m\right]^T, \qquad (3)$$

where

$$N_P = \left[(N_1 - n_1)\ldots(N_m - n_m)\, N_{m+1} \ldots N_H\right],$$

and

$$X_P = \begin{bmatrix} 1 & x_1 & (x_1-\kappa_1)_+ & \cdots & (x_1-\kappa_K)_+ & 1 & 0 & \cdots & 0 \\ . & . & . & \cdots & . & 0 & 1 & 0 & 0 \\ . & . & . & \cdots & . & . & . & 0 & . \\ . & . & . & \cdots & . & 0 & 0 & 1 & 0 \\ 1 & x_m & (x_m-\kappa_1)_+ & \cdots & (x_m-\kappa_K)_+ & 0 & \cdots & 0 & 1 \\ 1 & x_{m+1} & (x_{m+1}-\kappa_1)_+ & \cdots & (x_{m+1}-\kappa_K)_+ & 0 & \cdots & \cdots & 0 \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ . & . & & \cdots & . & . & \cdots & \cdots & . \\ 1 & x_H & (x_H-\kappa_1)_+ & \cdots & (x_H-\kappa_K)_+ & 0 & \cdots & \cdots & 0 \end{bmatrix}.$$

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

$$\mathrm{Var}(\hat{\bar{Y}} \mid \sigma^2, \sigma_\beta^2, \tau^2, X, X_P) = \sigma^2 (N_P X_P (X^T X + \Delta)^{-1} X_P^T N_P^T)/N^2.$$

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 PSUs into G equal-sized subgroups 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 obtained by excluding the elements from the PSUs in the g^{th} subgroup.

The jackknife 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 $\hat{\bar{\bar{Y}}} = \sum_{g=1}^{G} \hat{\bar{Y}}_g / G$. 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 first stage inclusion probabilities, and the G subgroups are constructed by randomly selecting one element from each stratum. To save computation, estimates $\hat{\sigma}^2, \hat{\sigma}_\beta^2$ and $\hat{\tau}^2$ are not recomputed for each replicate. That is, we compute pseudovalues of $(\beta_0, \beta_1, \ldots, \beta_{K+1}, u_1, \ldots, u_m)^T$ based on the variance components estimated from the whole sample.

Miller (1974) and Shao and Wu (1987, 1989) proved asymptotic properties of the jackknife estimator and jackknife variance estimation in the case of multiple linear

regression. Zheng and Little (2004) provided a theoretical justification for the jackknife method for the *p*-spline model-based estimator in the case of one-stage designs. Numerical simulations in section 4 suggest the above described jackknife method also works well for the two-stage design. Improved performance might be achieved using the weighted jackknife proposed by Hinkley (1977).

3.3 The Balanced Repeated Replication Method

The BRR method can be applied in stratified designs with two units sampled in each stratum. For designs with one PSU per stratum, strata are often collapsed (Kalton 1977) for BRR variance estimation. In our application we assume the PSUs are sampled systematically from a randomly ordered list. This can be viewed approximately as a stratified design with *n* strata each consisting of PSUs with cumulative measures of approximate size $\sum_{i=1}^{H} z_i / n$, where z_i are the measures of size for the PSUs . 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 $2\sum_{i=1}^{N} z_i / n$, and two units sampled per stratum. Balanced repeated half samples are constructed by selecting one PSU from each stratum, with the selection scheme based on Hadamard matrices (Plackett and Burman 1946). Let $\overline{Y}_b$ 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 given by $v_{\text{BRR}}(\overline{Y}) = 1/B \sum_{b=1}^{B} (\overline{Y}_b - \overline{Y})^2$. This estimate of the variance is subject to some bias, because it treats the design as if it was stratified with two PSUs per stratum.

4. When Some PSU Counts Are Not Known

In sections 2 and 3 we assumed that the PSU counts N_h are known for sampled and non-sampled PSUs. In this section we discuss the situation where N_h is only known exactly for the sampled PSUs (labeled 1 through *m*). We also assume that values $M_h, h = 1, ..., H$ of an auxiliary variable predictive of N_h are known for the whole population. For example, the M_h may be PSU counts estimated from outside sources such as a census. We conduct a regression of N_h on M_h using the sampled PSUs and replace the counts N_h in (3) for nonsampled PSUs with predictions $\hat{N}_h, h = m+1, \ldots, H$ from this regression. The resulting estimate of the total is

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

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

$$\begin{aligned}\mathrm{Var}(\tilde{T} \mid \pi_h, M_h) = {} & \mathrm{Var}(E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)) \\ & + E(\mathrm{Var}(\tilde{T} \mid \hat{N}_h, \pi_h, M_h)),\end{aligned}$$

where

$$E(\tilde{T} \mid \hat{N}_h, \pi_h, M_h) = \sum_{h=1}^{m} (N_h - n_h)\,\mu_h + \sum_{h=m+1}^{H} \hat{N}_h\,\mu_h$$

and

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

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

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

5. Simulations

5.1 Simulation Design

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

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

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

The population consists of 500 PSUs, and in the first stage 48 PSUs are sampled systematically with probability proportional to size (PPS) from a randomly-ordered list. The PSU sizes are uniformly distributed with values ranging from 4 to about 400. The SSU count in each PSU is generated from a distribution with mean equal to 1.05 times the measure of size and log-normal errors with standard deviation 30.

Two types of second-stage sampling plans are studied: 1) within-PSU 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 SSUs.; 2) within-PSU simple random sampling with the same sampling rate across sampled PSUs, so that the resulting inclusion probabilities for the SSUs in PSU *h* are proportional to $\pi_{1,h}$.

For each sample drawn under both sampling plans, the following methods are applied:

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 $\pi_{1,h}$ as the covariate. We use 20 equal percentiles of $\pi_{1,h}$ of the sampled PSUs as the knots for *p*-spline regression.
D. The PMM method with the PSU means μ_h estimated the same way as in C, but using estimated PSU counts from a simple linear regression of N_h on the measures of size, which are proportional to $\pi_{1,h}$. This part of the simulation is conducted to study the method described in section 4.

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 the PMM estimator, 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 and the coverage rate of the corresponding 95% confidence interval are used to judge the quality of inference. For method D, we study the empirical bias of the model-based variance estimator described in section 4, together with coverage rates of associated confidence intervals.

In the second simulation study, we draw samples of household income data from the 5% public use microdata sample (PUMS) for the State of Michigan in the 1990 US Census, which we treat as a finite population. This simulation is more realistic than the previous simulation in that the outcome values are drawn from a real rather than simulated distribution. The PSUs we simulate are based on the natural geographical clusters called "Public Use Microdata Areas" (PUMAs),which are typically counties and places. There are 67 PUMAs in the Michigan 5% PUMS, with counts of families ranging from around 1,300 to over 10,000. We increase the number of available PSUs by dividing each PUMA into 5, resulting in 335 PSUs. The PSU counts ranges from 134 to 3,058. Figure 1 gives the scatter plot of one sample of the average household income versus sampled PSU sizes together with the regression curve $\hat{f}(x)$.

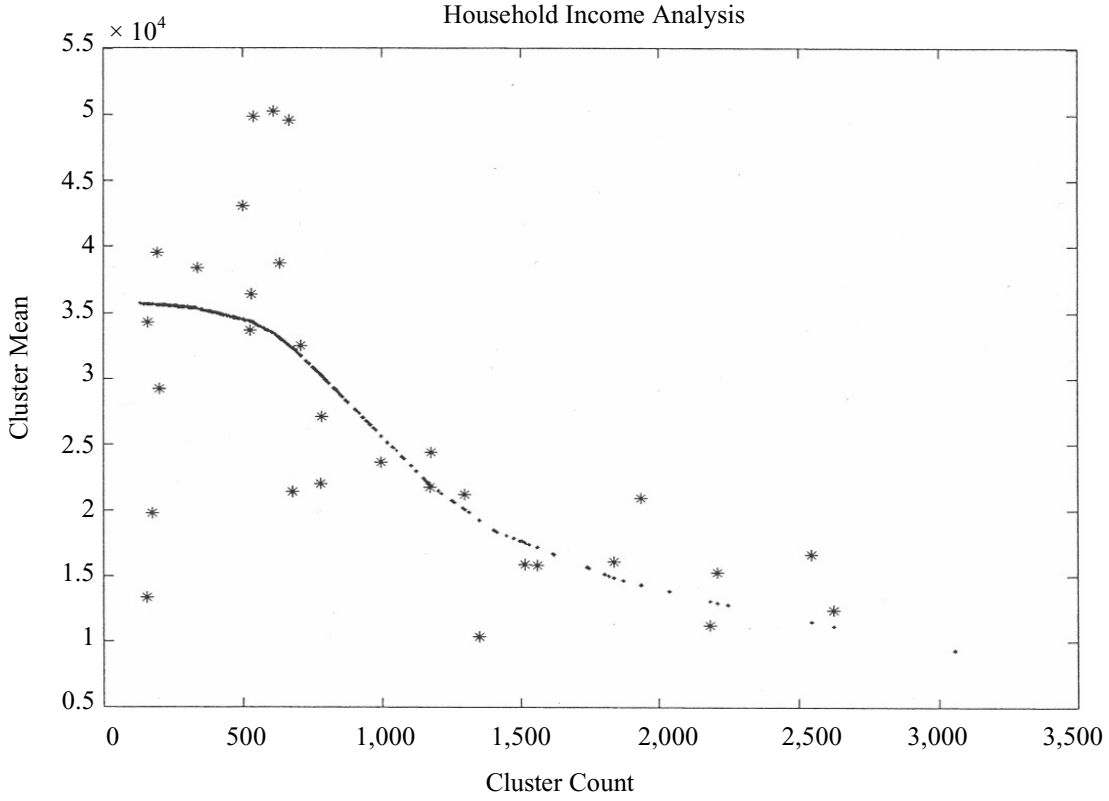


Figure 1. *P*-spline Regression Curve (dotted line) and the Average Household Income (stars) in Sampled PSUs.

Five hundred two-stage samples are drawn, each consisting of 30 PSUs and 20 SSUs (families) from each selected PSU. The first stage sampling is systematic PPS where the measures of size are equal to the PSU counts. The second stage sample is simple random sampling with inclusion probabilities proportional to the inverse of the first stage inclusion probabilities. In the estimation of the mean, we use the true PSU counts as variable x_h, with values proportional to the first-stage inclusion probabilities. We apply the *p*-spline nonparametric mixed model formulated in (2). We use 10 equally spaced sample percentiles of the PSU counts as the knots in the *p*-spline.

5.2 Results

Table 1 gives the empirical bias and root mean squared error (RMSE) from four estimation methods of the finite population mean applied to equal probability sample from populations generated with both normal and log-normal within-PSU errors and two (σ, τ) combinations. The empirical bias and RMSE are estimated by the mean bias and squared error from the 500 repeated samples.

Table 1 suggests the PMM based methods give estimators with small biases. 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 NULL and LINDOWN, where f is linear but not constant, the linear model-assisted and PMM method are comparable and both are more efficient than the HT estimator in terms of root mean squared error. For populations EXP and SINE, whose mean functions are not linear, the PMM method is superior to both the HT and the linear model-assisted estimators. The improvement of efficiency requires the knowledge of complete design information including probabilities $\pi_{1,h}$ and PSU counts N_h for the whole population. When using estimated PSU counts $\hat{N}_h$ in the place of N_h, the resulting estimator is less efficient than in the case with known N_h, but the PMM estimator can still outperform the HT when the mean function is non-constant. Comparisons on populations with normal or log-normal within-PSU errors result in similar findings.

Similar gains for the PMM method are seen in Table 2, for the case of unequal probability sampling. This suggests that the key to improved efficiency is the better prediction given by the nonparametric models. Tables 1 and 2 both suggest 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 PSU means.

Table 3 compares point estimation and coverage of 95% confidence intervals from three variance estimation methods for PMM: the empirical Bayes model-based method, the Jackknife method and the BRR method. The empirical Bayes method is generally satisfactory but tends to underestimate 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 yields estimates with improved efficiency over the traditional HT and linear model-assisted estimators and satisfactory design-based inferences.

Table 1
Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Equal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	$(\times 10^{-3})$	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Normal	NULL	1.1	29.7	0.8	30.0	0.8	29.9	1.3	30.1
Errors	LINDOWN	3.5	30.7	3.6	36.4	3.7	30.7	2.3	30.4
$\tau = 0.2$	EXP	−4.4	29.1	−9.4	53.0	−9.5	36.7	−4.3	29.1
$\sigma = 0.1$	SINE	4.8	32.5	2.1	42.0	−0.3	35.9	5.2	34.3
Normal	NULL	5.7	22.0	6.6	22.5	6.6	22.1	5.5	22.3
Errors	LINDOWN	0.5	20.4	−0.6	27.1	−0.3	20.5	1.6	20.6
$\tau = 0.1$	EXP	0.9	23.1	1.9	50.3	−4.2	31.7	0.4	23.4
$\sigma = 0.2$	SINE	7.0	22.3	6.5	34.9	3.8	26.4	8.0	26.4
Log-normal	NULL	1.7	32.3	0.9	32.3	0.7	32.3	1.5	32.5
Errors	LINDOWN	2.9	31.9	3.8	39.4	2.7	32.1	3.2	32.0
$\tau = 0.2$	EXP	−0.6	28.4	−5.9	51.5	−6.9	36.4	−0.3	28.5
$\sigma = 0.1$	SINE	6.9	33.8	1.5	43.7	−1.9	39.0	−3.1	35.0
Log-normal	NULL	8.5	30.5	9.6	31.3	9.2	31.0	9.1	30.8
Errors	LINDOWN	3.6	32.3	1.9	37.5	3.6	32.1	6.4	33.1
$\tau = 0.1$	EXP	3.9	29.0	6.8	53.8	1.0	34.4	3.7	29.4
$\sigma = 0.2$	SINE	−2.9	30.1	−8.9	44.7	−12.0	38.4	−3.8	35.9

Table 2
Empirical Biases and RMSE of PMM, HT, GR and PMM with Estimated N_h for Samples Under Unequal Probability Designs

		PMM		Horvitz-Thompson		Linear Model-Assisted		PMM with Estimated N_h	
	($\times 10^{-3}$)	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
Normal	NULL	−4.5	29.3	−3.7	33.6	−3.2	30.5	−4.5	29.3
Errors	LINDOWN	−0.9	27.0	3.7	35.5	1.8	27.7	−0.7	26.9
$\tau = 0.2$	EXP	5.8	32.0	1.9	56.8	0.4	39.4	14.1	34.4
$\sigma = 0.1$	SINE	7.1	30.1	6.1	39.5	3.6	32.8	5.3	30.4
Normal	NULL	−7.7	21.3	−7.7	24.9	−6.6	21.1	−7.6	21.2
Errors	LINDOWN	1.1	20.7	3.2	30.6	1.2	20.7	3.5	21.1
$\tau = 0.1$	EXP	−2.3	20.9	−6.5	53.3	−7.2	30.0	−3.0	20.9
$\sigma = 0.2$	SINE	5.6	20.9	6.9	36.2	4.0	28.6	4.3	21.1
Log-normal	NULL	−0.5	28.5	−2.0	30.6	−2.1	29.5	−0.3	28.5
Errors	LINDOWN	5.4	32.6	5.0	39.0	3.7	34.1	6.0	32.7
$\tau = 0.2$	EXP	−1.3	28.6	−7.6	62.6	−7.1	36.8	−9.3	30.3
$\sigma = 0.1$	SINE	3.7	31.2	2.3	43.1	0.1	36.1	1.6	31.0
Log-normal	NULL	3.6	22.8	5.7	28.8	5.7	24.2	3.6	22.7
Errors	LINDOWN	6.0	26.8	9.3	37.5	7.5	27.3	2.5	26.0
$\tau = 0.1$	EXP	0.8	26.3	−2.3	50.8	−3.5	33.1	11.5	29.0
$\sigma = 0.2$	SINE	3.7	26.9	2.9	37.6	−0.1	30.2	2.2	27.8

Table 3
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using the Model-based, Jackknife and BRR Methods

		Empirical variance	Empirical Bayes Model-based		Jackknife($K = 8$)		BRR	
			Estimate		Estimate		Estimate	
	Shape	($\times 10^{-5}$)	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%	($\times 10^{-5}$)	%
Normal	NULL	88	74	92.8	94	96.4	96	94.4
Errors	LINDOWN	94	73	89.6	94	94.6	98	94.2
$\tau = 0.2$	EXP	85	70	91.4	88	94.6	85	93.4
$\sigma = 0.1$	SINE	83	67	91.6	90	95.8	85	94.4
Normal	NULL	48	45	93.8	48	96.0	49	93.8
Errors	LINDOWN	42	45	96.8	51	96.2	51	96.8
$\tau = 0.1$	EXP	53	54	95.0	61	97.2	59	95.2
$\sigma = 0.2$	SINE	44	46	95.8	55	96.6	49	96.0
Log-normal	NULL	104	83	91.8	104	94.8	100	93.6
Errors	LINDOWN	102	98	93.6	106	95.6	107	95.0
$\tau = 0.2$	EXP	81	77	93.4	97	96.4	89	94.8
$\sigma = 0.1$	SINE	92	99	94.8	97	95.2	92	93.4
Log-normal	NULL	93	97	94.2	100	96.2	99	95.2
Errors	LINDOWN	104	101	93.6	106	96.0	102	92.8
$\tau = 0.1$	EXP	84	81	94.6	84	95.2	82	95.0
$\sigma = 0.2$	SINE	110	96	94.4	98	95.6	92	93.0

Tables 4 and 5 give the empirical variance of the PMM estimator when the non-sampled PSU counts N_h are estimated. They also give the mean estimated variance of this estimator and corresponding coverage rates by the 95% C.I. The confidence intervals are calculated by the usual normal theory intervals based on our point and variance estimators. 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 occasion under-coverage of the population mean. It remains to be studied in the future whether the JRR and BRR methods also yield satisfactory inferences for this method.

For the simulation study using 5% PUMS data, the simple mean has bias = −50.9 and RMSE = 2,600 and the *p*-spline nonparametric mixed model based method has bias = −41.9 and RMSE = 2,153.Thus both methods have small bias and the model-based estimator has a 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 PSUs increases (figure 1). The PMM method exploits this relationship in its predictions.

Table 4
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Normal Errors

	$\sigma = 0.1$ and $\tau = 0.2$			$\sigma = 0.2$ and $\tau = 0.1$		
	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
LINDOWN	93	74	90.4	43	46	95.6
EXP	85	72	93.0	55	56	96.2
SINE	110	98	94.8	50	55	97.6

Table 5
Variance Estimation and Empirical Coverage Rates of 95% C.I. Using *P*-spline and Estimated PSU Counts, Population Simulated with Log-normal Errors

	$\sigma = 0.1$ and $\tau = 0.2$			$\sigma = 0.2$ and $\tau = 0.1$		
	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
LINDOWN	103	98	94.4	110	102	94.4
EXP	81	79	94.6	87	83	94.2
SINE	110	150	96.4	91	130	95.8

6. Discussion

Previous parametric model-based inference methods have been criticized mainly for their potentially large design biases when the model 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, having small biases for a variety of population shapes.

Design information such as inclusion probabilities plays a key role in the model-based inference. Inverse-probability weighted methods imply simple assumptions about the relationship between the outcome variables and the design variables. With 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 estimators can be more efficient than the simple mean for 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 good confidence coverage 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-PSU variances). When the PSU 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 PSU counts precisely enough. The model relating these counts to the auxiliary variable was treated parametrically here, but this could also be specified nonparametrically without much difficulty.

We believe *p*-spline nonparametric mixed models can be applied to more complex designs such as stratified and multi-stage designs. We also believe without much more effort our methods can be generalized for binary or ordinal 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. *Biostatistics*, 2(3), 337-349.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, 16, 191-209.

Firth, D., and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society*, B, 60, 3-21.

Ghosh, 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*, A, 142, 33-46.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, 47, 663-685.

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, J., and Wu, C.F.J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Annals of Statistics*, 15, 1563-1579.

Shao, J., and Wu, C.F.J. (1989). A general theory for jackknife variance estimation. *Annals of Statistics*, 17, 1176-1197.

U.S. Census (1990). Dept. of Commerce. Census of Population and Housing, [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. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19, 99-117.

Zheng, H., and Little, R.J.A. (2004). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. To appear in *Journal of Official Statistics*.