A note on weight smoothing in survey sampling

Jae Kwang Kim* HaiYing Wang †

January 29, 2023

Abstract

Weight smoothing is a useful technique in improving the efficiency of design-based estimators at the risk of bias due to model misspecification. As an extension of the work of Kim and Skinner (2013), we propose using the weight smoothing to construct the conditional likelihood for efficient analytic inference under informative sampling. The Beta prime distribution can be used to build a parameter model for weights in the sample. A score test is developed to test for model misspecification in the weight model. A pretest estimator using the score test can be developed naturally. The pretest estimator is nearly unbiased and can be more efficient than the design-based estimator when the weight model is correctly specified or the original weights are highly variable. A limited simulation study is presented to investigate the performance of the proposed methods.

Key words: conditional maximum likelihood method; analytic inference, score test, pretest estimation.

*Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.
†Department of Statistics, University of Connecticut, Storrs, Connecticut, 06269, U.S.A.

1 Introduction

Suppose that the finite population of (x_i, y_i) is an independent and identically distributed (IID) realization of the superpopulation model with density $f(y \mid x; \theta)g(x)$, where θ is the parameter of interest and the marginal density $g(\cdot)$ is completely unspecified. From the finite population, we obtain a probability sample A with a known first-order inclusion probability π_i. We observe (x_i, y_i) in the sample. We are interested in estimating the model parameter θ from the complex sample, which is the main problem in the area of analytic inference in survey sampling. See Korn and Graubard (1999) and Fuller (2009, Ch. 6) for comprehensive overviews of analytic inference in survey sampling.

For efficient estimation, we can construct the conditional likelihood function from the sample as follows:

$$L_c(\theta) = \prod_{i \in A} \frac{f(y_i \mid x_i; \theta)\tilde{\pi}(x_i, y_i)}{\int f(y \mid x_i; \theta)\tilde{\pi}(x_i, y)d\mu(y)} \tag{1}$$

where

$$\tilde{\pi}(x, y) = E\left(\pi \mid x, y\right) \tag{2}$$

is the conditional inclusion probability and $\mu(\cdot)$ is the dominating measure. See Section 8.2 of Kim and Shao (2021) for some details of the conditional maximum likelihood method.

To compute the conditional inclusion probability in (2), we can use the formula of Pfeffermann and Sverchkov (1999):

$$E(\pi \mid x, y) = \frac{1}{E_s(w \mid x, y)}, \tag{3}$$

where $w = \pi^{-1}$ and $E_s(\cdot)$ is the expectation with respect to the sample distribution, the conditional distribution given the sample.

The conditional inclusion probability obtained from (3) can be used to calculate the smoothed weight $\tilde{w}_i = \{\tilde{\pi}(x_i, y_i)\}^{-1}$. The weight smoothing can reduce the variability of the sampling weight $w_i = \pi_i^{-1}$ in estimating parameters and thus can lead to more efficient estimation, as discussed by Beaumont (2008) and Kim and Skinner (2013). To compute the conditional expectation $E_s(w \mid x, y)$, we need to build a regression model for w, which can be called a weight model.

In this article, we explore some particular parametric classes of weight models. In Section 2, a weight model using the Beta prime distribution is introduced. In Section 3, a score test for correct model specification in the weight model is proposed. In Section 4, results from a limited simulation study are presented. Some concluding remarks are made in Section 5.

2 Weight model

Because the sampling weights satisfy $w_i \geq 1$ $(i = 1, \ldots, n)$, it is assumed that w_i^{-1} are modeled as a Beta distribution $\text{Beta}(m(x_i, y_i)\phi,\ \{1 - m(x_i, y_i)\}\phi)$. Thus, the density function satisfies

$$f(w^{-1} \mid x, y) \propto (w^{-1})^{m\phi-1}(1 - w^{-1})^{(1-m)\phi-1},$$

and the conditional expectation and variance are

$$E(w^{-1} \mid x, y) = m(x, y), \quad \text{and} \quad V(w^{-1} \mid x, y) = \frac{m(x, y)\{1 - m(x, y)\}}{1 + \phi},$$

respectively, where ϕ is the precision parameter. An example of a mean function is the logistic model:

$$m(x, y; \beta) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 y)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 y)}. \tag{4}$$

This is essentially a beta regression model. Further details on beta regression can be found in Ferrari and Cribari-Neto (2004).

Unfortunately, the beta regression approach cannot be applied directly because the regression model does not necessarily hold in the sample due to informative sampling. To avoid this problem, we can derive the distribution of the sampled data. Recall that if $X \sim \text{Beta}(\alpha, \beta)$ then $1 - X$ follows $\text{Beta}(\beta, \alpha)$ and $(1 - X)/X$ follows a Beta prime distribution $\text{Beta}'(\beta, \alpha)$. Therefore, $o = w - 1$ follows $\text{Beta}'(\{1 - m(x_i, y_i)\}\phi,\ m(x, y)\phi)$, and the density function is expressed as

$$f(o \mid x, y) \propto o^{(1-m)\phi-1}(1 + o)^{-\phi}.$$

Based on Bayes' theorem and $w^{-1} = (1+o)^{-1}$, the sampled distribution of o satisfies

$$f_s(o \mid x, y) \propto f(o \mid x, y) P(\delta = 1 \mid x, y, w) = o^{(1-m)\phi - 1}(1+o)^{-\phi-1}, \tag{5}$$

which implies $o \mid (x, y, \delta = 1) \sim \text{Beta}'(\{1 - m(x,y)\}\phi,\ m(x,y)\phi + 1)$. Thus, we obtain the following.

$$E_s(w \mid x, y) = 1 + E_s(o \mid x, y) = \frac{1}{m(x, y; \beta)} \tag{6}$$

and

$$\begin{aligned} Var_s(w \mid x, y) &= \frac{1 - m(x,y)}{m(x,y)} \frac{1}{m(x,y) \cdot \phi - 1} \\ &\cong \frac{1 - m(x,y)}{m(x,y)} \frac{1}{m(x,y) \cdot \phi} \end{aligned}$$

for sufficiently large ϕ. Thus, we obtain the following method of moments estimator of ϕ:

$$\hat{\phi} = \frac{1}{n} \sum_{i \in A} \frac{\left\{ w_i \cdot m(x_i, y_i; \beta) - 1 \right\}^2}{1 - m(x_i, y_i; \beta)} \tag{7}$$

which depends on unknown parameter β.

We can use the following iterative estimation procedure estimate model parameters.

1. Compute

   $$\hat{\phi}^{(0)} = \frac{1}{n} \sum_{i \in A} \frac{\left(w_i / \bar{w} - 1 \right)^2}{1 - 1/\bar{w}}$$

   as an initial estimator of ϕ, where $\bar{w} = n^{-1} \sum_{i \in S} w_i$.

2. Using $\hat{\phi}^{(t)}$, compute $\hat{\beta}^{(t)}$ by finding the maximizer of

   $$\ell_c(\beta \mid \hat{\phi}^{(t)}) = \sum_{i \in S} \log f_s(o_i \mid x_i, y_i; \beta, \hat{\phi}^{(t)})$$

   with respect to β, where

   $$f_s(o \mid x, y; \beta, \phi) = \frac{\Gamma(\phi + 1)}{\Gamma(\phi - m\phi)\Gamma(m\phi + 1)} o^{\phi - m\phi - 1}(1+o)^{-\phi-1},$$

   and $m = m(x, y; \beta)$.

3. Compute $\hat{\phi}^{(t+1)}$ by applying (7) with $\beta = \hat{\beta}^{(t)}$. Iteratively update $\hat{\phi}$ and $\hat{\beta}$ until convergence.

3 Score test for weight model specification

The weight smoothing method in Section 2 is justified under the assumption that the weight model is correctly specified. In practice, we may wish to test for the validity of the weight model before we use the model-based estimator. In this section, we consider a version of score test for model specification.

Let $\hat{\theta}_c$ be the maximizer of the conditional likelihood function in (1). Let $\hat{\theta}_d$ be the design-based estimator of θ that is obtained by maximizing the pseudo log-likelihood function

$$\ell_p(\theta) = \sum_{i \in A} \frac{1}{\pi_i} \log f(y_i \mid \mathbf{x}_i; \theta). \tag{8}$$

The pseudo MLE has been discussed in Chambers and Skinner (2003). Thus, we can develop a test for the following null hypothesis:

$$E\left(\hat{\theta}_d\right) = E\left(\hat{\theta}_c\right). \tag{9}$$

However, developing a Wald-type test statistics for the null hypothesis in (9) can be cumbersome as the variance-covariance matrix of $\hat{\theta}_d - \hat{\theta}_c$ needs to be estimated.

Instead of testing (9), we can consider testing the following null hypothesis

$$H_0: \quad E\left\{\hat{S}_c(\theta_0)\right\} = 0, \tag{10}$$

where θ_0 is the true parameter and $\hat{S}_c(\theta) = n^{-1}\partial \log L_c(\theta)/\partial\theta$ is the score function obtained from the conditional log-likelihood in (1). That is,

$$\hat{S}_c(\theta) = \frac{1}{n} \sum_{i \in A} \left[S(\theta; x_i, y_i) - E_s\left\{S(\theta; x_i, Y) \mid x_i\right\}\right],$$

where $S(\theta; x, y) = \partial \log f(y \mid x; \theta)/\partial\theta$ and

$$E_s\left\{S(\theta; x, Y) \mid x\right\} = \frac{\int S(\theta; x, y)\tilde{\pi}(x, y) f(y \mid x; \theta) dy}{\int \tilde{\pi}(x, y) f(y \mid x; \theta) dy}.$$

Under some regularity conditions (Binder, 1983), we can establish that

$$\sqrt{n}\left[\hat{S}_c(\theta) - E\left\{\hat{S}_c(\theta)\right\}\right] \xrightarrow{\mathcal{L}} N\left[0, \mathcal{I}_c(\theta)\right], \tag{11}$$

as $n \to \infty$, where $\xrightarrow{\mathcal{L}}$ denotes the convergence in distribution and

$$\begin{aligned} \mathcal{I}_c(\theta) &= -E\left\{\frac{\partial}{\partial\theta'}S_c(\theta)\right\} \\ &= n^{-1}\sum_{i=1}^{n}\left[E\left\{S_i S_i'\tilde{\pi}_i \mid \mathbf{x}_i;\theta\right\} - \frac{\left\{E\left(S_i\tilde{\pi}_i \mid \mathbf{x}_i;\theta\right)\right\}^{\otimes 2}}{E\left(\tilde{\pi}_i \mid \mathbf{x}_i;\theta\right)}\right]. \end{aligned} \tag{12}$$

The proposed test statistic is

$$T(\hat{\theta}_d) = n\hat{S}_c(\hat{\theta}_d)'\left\{\mathcal{I}_c(\hat{\theta}_d)\right\}^{-1}\hat{S}_c(\hat{\theta}_d)$$

where $\hat{\theta}_d$ is the pseudo ML estimator of θ_0. Note that

$$T(\hat{\theta}_d) = T(\theta_0) + o_p(1),$$

as $\hat{\theta}_d = \theta_0 + o_p(1)$, regardless of whether the weight model holds or not. Under the null hypothesis in (10), by (11), we can establish that T converges to $\chi^2(q)$ distribution where $q = \dim(\theta)$. If the null hypothesis is rejected, then it implies that $\tilde{\pi}(x, y)$ in constructing the conditional likelihood in (1) is incorrectly specified. Otherwise, we can safely use the conditional ML estimator.

Strictly speaking, the information matrix in (12) ignores the uncertainty of $\hat{\beta}$ in $\tilde{\pi}_i = \tilde{\pi}(x_i, y_i; \hat{\beta})$. To incorporate the uncertainty in $\hat{\beta}$, we can consider another information matrix for β. Ignoring the uncertainty in $\hat{\beta}$ will overestimate the variance and lead to a conservative test. See the simulation study in the next section.

4 Simulation study

To test our theory, we performed a limited simulation study. In the simulation, we generate a finite population of size $N = 10,000$ and use Poisson sampling to select a sample of expected size $n = 1,000$. We repeat this procedure independently $B = 1,000$ times.

In each Monte Carlo sample, we generate (x_i, y_i, π_i) for $i = 1, \ldots, N$ where $x_i \sim U(0, 2)$, $y_i = \theta_0 + \theta_1 x_i + e_i$, $(\theta_0, \theta_1) = (0.5, 0.5)$, $e_i \sim N(0, 0.5^2)$, and $\pi_i \mid x_i, y_i \sim$

Beta$(m(x_i, y_i)\phi,\ \{1 - m(x_i, y_i)\}\phi)$, where

$$m(x, y; \beta) = \frac{\exp(\beta_0 + \beta_1 x + \beta_2 y)}{1 + \exp(\beta_0 + \beta_1 x + \beta_2 y)} \tag{13}$$

with $\beta_1 = 1$, $\beta_2 = 1$, and β_0 being different values for different cases to ensure that $n = 1,000$. We used two different values of ϕ, $\phi = 100$ versus $\phi = 1,000$, in the simulation study. The weight distribution is less skewed for $\phi = 1,000$.

We have four different sampling designs as follows:

Case 1 $\phi = 100$; weight model is specified correctly.

Case 2 $\phi = 100$; lowest 30% π_i's are multiplied by 0.25, i.e., top 30% w_i's in the full data are multiplied by 4. Thus, the weight model (13) is incorrectly specified.

Case 3 $\phi = 1000$; weight model is correctly specified.

Case 4 $\phi = 1000$; the lowest 30% π_i's are multiplied by 4, i.e., the top 30% w_i's in the full data are multiplied by 0.25. Thus, the weight model (13) is incorrectly specified.

We are interested in estimating θ_0 and θ_1. The following three estimators are considered.

1. PMLE: The pseudo maximum likelihood estimator $\hat{\theta}_d$ maximizing (8).

2. CMLE: The conditional maximum likelihood estimator $\hat{\theta}_c$ maximizing (1) with $\tilde{\pi}(x, y) = \{\tilde{w}(x, y)\}^{-1}$ and $\tilde{w}(x, y)$ is the smoothed weight under the specified weight model. To avoid numerical problems, we estimate σ^2 in a design-based way.

3. PreTest: The pretest estimator using the score test in Section 3. That is, the pretest estimator $\hat{\theta}_{pre}$ with $\alpha = 0.05$ is defined as

$$\hat{\theta}_{pre} = \begin{cases} \hat{\theta}_d & \text{if } T(\hat{\theta}) > q_{0.95}(\chi^2_2) \\ \hat{\theta}_c & \text{otherwise,} \end{cases}$$

   where $q_{0.95}(\chi^2_2)$ is the 0.95 quantile of the $\chi^2(2)$ distribution.

Table 1 presents the biases, standard errors, and root mean square errors (RMSE) of the three estimators using Monte Carlo samples. The simulation results can be summarized as follows.

1. The PMLE is nearly unbiased for all cases, but it is less efficient than the other methods in Cases 1 and 3, where the weight model is correctly specified.

2. The CMLE is the most efficient but is subject to significant biases when the weight model is incorrectly specified. The efficiency gain is higher for a smaller ϕ, as the distribution of w_i's is more skewed and the advantage of weight smoothing is more significant.

3. The pretest estimator is nearly unbiased for all cases and can be more efficient than the PMLE when the weight model is correctly specified (Case 1 and Case 3) or the original weights are highly variable (Case 2).

Table 1: Monte Carlo biases, standard errors (SE) and root mean square errors (RMSE) of the three estimators based on 1,000 Monte Carlo samples

		θ_0			θ_1		
Case	**Method**	**SE**	**Bias**	**RMSE**	**SE**	**Bias**	**RMSE**
	PMLE	0.0768	-0.001	0.0768	0.0799	0.001	0.0800
1	CMLE	0.0608	-0.001	0.0608	0.0425	0.001	0.0425
	PreTest	0.0701	0.006	0.0704	0.0672	-0.004	0.0673
	PMLE	0.1198	-0.000	0.1198	0.1182	0.008	0.1185
2	CMLE	0.0750	0.020	0.0777	0.0375	0.066	0.0764
	PreTest	0.1198	0.001	0.1198	0.1179	0.008	0.1182
	PMLE	0.0651	0.000	0.0651	0.0645	0.000	0.0645
3	CMLE	0.0525	0.002	0.0526	0.0413	-0.002	0.0413
	PreTest	0.0561	0.003	0.0563	0.0499	-0.003	0.0500
	PMLE	0.0455	0.001	0.0456	0.0432	0.000	0.0432
4	CMLE	0.0472	0.053	0.0713	0.0432	-0.127	0.1345
	PreTest	0.0456	0.001	0.0456	0.0433	0.000	0.0433

The rejection rates for the score test are 0.119, 0.952, 0.051, and 0.997 for the four cases, respectively, where the level of significance is $\alpha = 0.05$. The high rejection rate of 0.119 in Case 1 is due to the effect of ignoring uncertainty in weight smoothing. The effect of ignoring the uncertainty in weight smoothing is negligible in Case 3, since the effect of weight smoothing is less significant when ϕ is large. The higher rejection rate indicates that the score test is conservative in adopting CMLE using $\tilde{w}_i$ over PMLE.

5 Concluding Remark

This article is dedicated to the memory of Professor Chris Skinner. The first author collaborated on various projects with Chris Skinner and their first research outcome was published in Kim and Skinner (2013). When J.K. Kim visited Chris Skinner at Southampton in the summer of 2011, they first worked on analytic inference under informative sampling, studying the work of Pfeffermann and Sverchkov (1999), but they did not make a connection with weight smoothing at that time. Instead, they mainly focused on the weight smoothing method. About ten years later, we now present a method that connects weight smoothing to the likelihood framework.

Weight smoothing is a potentially useful idea, but the correct model specification is required. The pretest estimator using the score test in Section 3 can be used in practice, as it compromises the efficiency of weight smoothing and the robustness of design-based estimation. How to estimate the variance of the pretest estimator is not explored in this paper and will be investigated in the future.

Acknowledgements

The authors thank the Editor and the Assistant Editor, Cynthia Bocci, for their constructive comments. The research of the first author was supported by a grant from the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. The second author's research was supported by NSF grant CCF 2105571 and UConn CLAS Research Funding in Academic Themes.

References

Beaumont, J. F. (2008). A new approach to weighting and inference in sample surveys. *95*, 539–553.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Reviews 51*, 279–292.

Chambers, R. L. and C. J. Skinner (2003). *Analysis of survey data*. John Wiley & Sons.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*, 407–419.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.

Kim, J. K. and J. Shao (2021). *Statistical Methods for Handling Incomplete Data* (2nd ed.). CRC press.

Kim, J. K. and C. J. Skinner (2013). Weighting in survey analysis under informative sampling. *Biometrika 100*, 358–398.

Korn, E. L. and B. I. Graubard (1999). *Analysis of Health Surveys*. John Wiley & Sons.

Pfeffermann, D. and M. Sverchkov (1999). Parametric and semiparametric estimation of regression models fitted to survey data. *Sankhyā, Series B 61*, 166–186.