Estimation of A Finite Population Mean - An Empirical Restricted Best Prediction Approach

Jiming Jiang and P. Lahiri
University of California, Davis
and
University of Maryland, College Park

Abstract

We propose a model-based restricted best (RB) predictor of a finite population mean that minimizes the mean square prediction error (MSPE) among the class of all predictors that depend on the sample only through the traditional design-unbiased estimator of the finite population mean. We then consider an empirical restricted best (ERB) predictor derived from the RB that does not require the knowledge of an explicit model for the unobserved units of the finite population. The proposed ERB converges in probability (with respect to a model) to the customary design-consistent estimator, irrespective of the assumed model. Unlike recent papers on design consistency, this paper presents a general methodology that can be applied to a wide class of models, including mixed linear models and generalized linear mixed models in common use, and provides a formal proof of design-consistency. The case of unknown model parameters is undertaken and a second-order accurate model-based MSPE of the proposed ERB predictor is obtained to measure the precision of the proposed predictor. The method proposed here will find applications in small-area estimation problems.

KEYWORDS: asymptotics, mean square prediction error, robustness, small-area estimation, survey weights.

1 Introduction

In this paper, we revisit the model-based estimation in finite population sampling where a large finite population is stratified into several strata, each assumed to be a realization from a superpopulation generated by a multilevel mixed model. Applications of multi-level mixed models in finite population sampling can be traced back to the well-celebrated paper by Ericson (1969) who put forward a subjective Bayesian approach to finite population inference. Ghosh and Meeden (1986) considered an empirical Bayes estimation of a finite population mean using a simple random effects normal model. Ghosh and Lahiri (1987) extended their method to a non-normal situation using the assumption of posterior linearity. These three papers and several other follow-up papers, including Nandram and Sedransk (1993), Arora, Lahiri & Mukherjee (1997), Datta & Ghosh (1991), essentially use unit level mixed models and assume equal sampling weights within each stratum. The resulting model-based estimators are not design-consistent, i.e., they do not converge to the corresponding true stratum mean when the stratum sample size is large unless the design is self-weighing.

In this paper, our goal is to obtain an *efficient* model-based estimator of a stratum mean which retains information contained in sampling weights. A sampling weight attached to a sampling unit represents a certain number of population units. It is the basic inclusion probability of the unit adjusted for various factors such as post-stratification, non-response, etc. We first obtain the restricted best (RB) predictor which has the least mean square prediction error (MSPE) among the class of all predictors which depend on the data only through the traditional design-unbiased sample survey estimator. This is one way of incorporating the sampling weights in the model-based procedure. This RB estimator of the stratum mean requires distributional assumptions for the unobserved units of the finite population. To circumvent the problem, we consider an empirical restricted best (ERB) predictor that estimates the RB unbiasedly with respect to the sampling design. This is yet another way of incorporating the sampling weights in the estimation procedure. Although we need an explicit model for the sampled units to compute our proposed predictor, it enjoys an appealing robustness property in that it is design-consistent even if the model fails. We consider the case when the model parameters (also known as hyperparameters) of our assumed model are unknown.

We discuss the estimation of precision of our proposed ERB. In this paper, we propose a model MSPE estimator that accounts for various sources of variabilities. One might wonder about a possible design-based precision measure. But, Prasad-Rao (1999) showed that their model-based precision measure is much more stable than an alternate design-based precision measure proposed by Kott (1989).

Our main application is small-area estimation where we treat each stratum as a small-area. For a review on small-area estimation, see Ghosh & Rao (1994), Lahiri and Meza (2001), Pfeffermann (2002), among others. Importance of design-consistency when using unit level models has been stressed in a number of papers on small-area estimation. See Arora and Lahiri (1997), Prasad & Rao (1999), Kott (1989), You and Rao (2000), Folsom, Shah & Vaish (1999), among others. However, most of the papers deal with mixed linear normal models, except for the paper by Folsom et al. (1999). In contrast with other papers, we provide a general methodology for producing ERB predictors which are also design-consistent. Our method works for a general class of models including mixed linear models and generalized linear mixed models in common use.

Section 2 presents a formulation of the problem and the ERB. In Section 3, we present the design consistency property of our proposed predictor. Some examples are considered in section 4 to verify the conditions of design-consistency. A MSPE estimator of the proposed estimator, whose bias is of the order $o(m^{-1})$, where m is the number of strata, is presented in Section 5. All technical proofs are deferred to the Appendix.

2 An Empirical Restricted Best Predictor

Consider a finite population stratified into m strata and let N_i be the size of the ith stratum $(i = 1, \cdots, m)$. Let Y_{ij} denote the value of a characteristic of interest for the jth observation in the ith stratum $(i = 1, \cdots, m; j = 1, \cdots, N_i)$. In this paper, we consider the estimation of the ith stratum mean given by $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} Y_{ij}$ based on a sample, say Y_{ij} $(i = 1, \cdots, m; j = 1, \ldots, n_i)$, obtained using a complex sampling scheme. Let $\tilde{w}_{ij}$ denote the corresponding sampling weight $(i = 1, \cdots, m; j = 1, \ldots, n_i)$.

We assume that Y_{ij} $(i = 1, \ldots, m; j = 1, \ldots, N_i)$ are realized values from a hypothetical superpopulation characterized by a probability distribution. We use the the notation $\mathrm{E_d}$ for the expectation with respect to the sampling design and $\mathrm{E_m}$ the expectation with respect to the mixed model. Thus, $\mathrm{E} = \mathrm{E_d}\mathrm{E_m}$ and we assume that the expectations are interchangeable and that the sampling design is non-informative, i.e., does not depend on the assumed model.

Define the traditional survey weighted estimator of $\bar{Y}_i$ as $T_i = \sum_{j=1}^{n_i} w_{ij} Y_{ij}$, where $w_{ij} = \tilde{w}_{ij} / \sum_{j=1}^{n_i} \tilde{w}_{ij}$. Throughout the paper we assume that T_i is a design-unbiased estimator of $\bar{Y}_i$, i.e., $\mathrm{E_d}(T_i) = \bar{Y}_i$. We shall first find the best predictor of $\bar{Y}_i$ based on T_i. This optimal predictor will be referred to as the restricted best (RB) predictor of $\bar{Y}_i$.

We need an explicit model for sampled units. However, for the unobserved units of the finite population, we only need to assume the existence of a random effect v_i associated with the ith stratum such that

$$\mathrm{E_m}(Y_{ij}|v_i, T_i) = \mathrm{E_m}(Y_{ij}|v_i).$$

The assumption holds for the mixed linear normal model (e.g., Prasad and Rao 1999) and the mixed logistic model (Jiang and Lahiri 2001). We define the mean square prediction error (MSPE) of an arbitrary predictor $\hat{\bar{Y}}_i$ of $\bar{Y}_i$ as

$$\mathrm{MSPE}(\hat{\bar{Y}}_i) = \mathrm{E}(\hat{\bar{Y}}_i - \bar{Y}_i)^2,$$

where the expectation is taken over both the sampling design and the assumed mixed model. The following theorem yields RB predictor of $\bar{Y}_i$ defined as $\hat{\bar{Y}}_i^{RB} = N_i^{-1} \sum_{j=1}^{N_i} \mathrm{E_m}[\mathrm{E_m}(Y_{ij}|v_i)|T_i]$.

Theorem 1. Under the assumed mixed model and the sampling design, $\hat{\bar{Y}}_i^{RB}$ minimizes the MSPE among the class of all predictors which depend on the data only through T_i's.

Note that $\mathrm{E}[\hat{\bar{Y}}_i^{RB} - \bar{Y}_i] = 0$ so that $\hat{\bar{Y}}_i^{RB}$ is also an unbiased predictor of $\bar{Y}_i$. In this paper, we shall assume that an explicit model for the unobserved unit of the finite population is not available and thus $\mathrm{E_m}[\mathrm{E_m}(Y_{ij}|v_i)|T_i]$ are unknown for all unobserved units of the ith stratum. Hence, generally $\hat{\bar{Y}}_i^{RB}$ cannot be computed. Note that for fixed values of $\mathrm{E_m}[\mathrm{E_m}(Y_{ij}|v_i)|T_i]$ $(i = 1, \cdots, N_i; i = 1, \cdots, m)$, we can treat $\hat{\bar{Y}}_i^{RB}$ as an unknown finite population mean which can be estimated unbiasedly (with respect to the sampling design) by

$$\hat{\bar{Y}}_i^{ERB} = \sum_{j=1}^{n_i} w_{ij} \mathrm{E_m}[\mathrm{E_m}(Y_{ij}|v_i)|T_i],$$

assuming that $\mathrm{E_m}[\mathrm{E_m}(Y_{ij}|v_i)|T_i]$ are fully specified for all sampled units $j = 1, \cdots, n_i; i = 1, \cdots, m$.

The following theorem provides a tool for computing $\hat{\bar{Y}}_i^{ERB}$ and justifies it as an empirical restricted best, or empirical restricted Bayes, (ERB) predictor of $\bar{Y}_i$. Define $\zeta_i = \mathrm{E_m}(T_i|v_i)$.

Theorem 2. Under the assumed model and sampling design, we have

(i) $\hat{\bar{Y}}_i^{ERB} = \tilde{\zeta}_i$, where $\tilde{\zeta}_i = \mathrm{E_m}[\zeta_i|T_i]$;

(ii) $\mathrm{E}[\tilde{\zeta}_i - \hat{\bar{Y}}_i^{RB}] = 0$.

Remark 1: Define $\zeta_i(v) = \zeta_i|_{v_i=v}$. Then, we can obtain $\tilde{\zeta}_i$ as

$$\tilde{\zeta}_i = \frac{\int \zeta_i(v) f(T_i, v) f(v) dv}{\int f(T_i, v) f(v) dv}, \tag{1}$$

where $f(T_i, v)$, $f(v)$ are nonnegative functions such that, under the model, $f(T_i, v) = f(T_i|v_i)|_{v_i=v}$, the conditional density of T_i given v_i, and $f(v)$ is the density of v_i, which may depend on ψ, a vector of dispersion parameters.

The proposed ERB $\tilde{\zeta}_i$ depends on T_i and often on θ, a vector of unknown model parameters which include ψ and possibly other parameters, i.e.,

$$\tilde{\zeta}_i = u_i(T_i, \theta) \tag{2}$$

for some function $u_i(\cdot, \cdot)$.

When θ is unknown, it is replaced by $\hat{\theta}$, a model-consistent estimator. This gives the following ERB predictor of $\bar{Y}_i$ for the most practical situation:

$$\hat{\zeta}_i = u_i(T_i, \hat{\theta}). \tag{3}$$

3 Asymptotic behavior of ERB

The following theorems states that, as long as n_i is large, $\tilde{\zeta}_i$ agrees asymptotically with T_i, regardless of the model and θ. If a design-unbiased estimator of $\bar{Y}_i$ is not found, we may choose T_i to be at least design-consistent so we can make our ERB predictor close to a design-consistent estimator for large n_i. We shall prove the result both with respect to the model and to the design. So, as before, $\mathrm{P_m}$, $\mathrm{E_m}$ and $\mathrm{var_m}$ represent probability, expected value and variance with respect to the assumed model, and $\mathrm{P_d}$, $\mathrm{E_d}$ and $\mathrm{var_d}$ those with respect to the design. Write $n = n_i$ for notational simplicity. Theorem 3 below is regarding consistency with respect to the model, or model-consistency.

Theorem 3. Suppose that 1) $\mathrm{E_m}|Y_{ij}|$ and $\int |\zeta_i(v)|f(v)dv$ are bounded; 2) there is a sequence $\lambda_n \in (0,\infty]$ such that, as $n \to \infty$,

$$\left(\int f(T_i,v)f(v)dv\right)^{-1} \sup_{|\zeta_i(v)-T_i|>\epsilon, |v|<\lambda_n} f(T_i,v)$$

$\longrightarrow 0$ in $\mathrm{P_m}$ for any $\epsilon > 0$; and 3)

$$\left(\int f(T_i,v)f(v)dv\right)^{-1} \int_{|v|\geq\lambda_n} (|\zeta_i(v)|+1)f(T_i,v)f(v)dv$$

$\longrightarrow 0$ in $\mathrm{P_m}$, where $\int_{|v|\geq\infty}\cdots dv$ is understood as 0. Then, $\tilde{\zeta}_i - T_i \to 0$ in $\mathrm{P_m}$.

When θ is unknown, it is customary to replace it by $\hat{\theta}$, a model-consistent estimator. A similar result is expected to hold when θ is replaced by $\hat{\theta}$, i.e., $\tilde{\zeta}_i$ by $\hat{\zeta}_i$. Note that, in many cases, $\hat{\theta}$ remains consistent even if the model fails. For example, it is known that in linear mixed models the restricted maximum likelihood (REML) estimators of the variance components obtained assuming normality remain consistent even if normality fails [Richardson & Welsh (1994), Jiang (1996)]. Note that in this case $f(T_i, v_i)$ is the (restricted) normal conditional density which is not the true (restricted) conditional density, and θ is the vector of variance components. To see another example, consider the estimation of the parameters in a generalized linear mixed model (GLMM). Suppose that the usual exponential-family assumption fails for the conditional density but the conditional mean is correctly specified. Jiang and Zhang (2001) showed that some GEE type estimators remain consistent. Note that in this case $f(T_i, v_i)$ is the exponential-family conditional density (e.g., binomial) which is not the true conditional density, and θ is the vector of parameters in the GLMM which are still well-defined under the conditional mean model. However, Theorem 3 shows that the consistency of $\hat{\theta}$ is important only when n is small. Such a case will be considered in section 5.

A design consistency analogue of Theorem 3 is the following.

Theorem 4. Suppose that 1) there is a constant $c > 0$ such that $|y_{ij}| \leq c$ and $\int |\zeta_i(v)|f(v)dv \leq c$; 2) there is a sequence $\lambda_n \in (0,\infty]$ such that, as $n \to \infty$,

$$\left(\int f(T_i,v)f(v)dv\right)^{-1} \sup_{|\zeta_i(v)-T_i|>\epsilon, |v|<\lambda_n} f(T_i,v)$$

$\longrightarrow 0$ in $\mathrm{P_d}$ for any $\epsilon > 0$; and 3)

$$\left(\int f(T_i,v)f(v)dv\right)^{-1} \int_{|v|\geq\lambda_n} (|\zeta_i(v)|+1)f(T_i,v)f(v)dv$$

$\longrightarrow 0$ in $\mathrm{P_d}$, where $\int_{|v|\geq\infty}\cdots dv$ is understood as 0. Then, $\tilde{\zeta}_i - T_i \to 0$ in $\mathrm{P_d}$.

The proof of the Theorem 3 is given in Appendix. The proof of Theorem 4 is similar and therefore omitted. In the next section we consider some examples.

4 Examples

Example 1 (Normal data). In the case of mixed linear normal model, one has $Y_{ij} = x_{ij}^{\mathrm{T}}\beta + z_{ij}^{\mathrm{T}}v_i + e_{ij}$, where v_i and e_{ij} are independent with $v_i \sim N(0,\sigma_v^2)$, $e_{ij} \sim N(0,\sigma_i^2)$. Then, we have $T_i|v_i \sim N(\zeta_i, \tau_i^2)$, where $\zeta_i = b_i^{\mathrm{T}}\beta + d_i^{\mathrm{T}}v_i$, $b_i = \sum_{j=1}^n w_{ij}x_{ij}$, $d_i = \sum_{j=1}^n w_{ij}z_{ij}$, and $\tau_i^2 = \sigma_i^2 \sum_{j=1}^n w_{ij}^2$. Therefore,

$$f(T_i,v) = \frac{1}{\sqrt{2\pi\tau_i^2}} \exp\left[-\frac{\{T_i - \zeta_i(v)\}^2}{2\tau_i^2}\right]. \quad (4)$$

Let the x_{ij}s and z_{ij}s be bounded. Then, Y_{ij}s are bounded in L^1, and we have $\int |\zeta_i(v)|f(v)dv \leq |b_i||\beta| + |d_i|\int |v|f(v)dv$, which is bounded, provided that the first moment of v_i under the model is bounded. Thus, condition 1) is satisfied.

If $|\zeta_i(v) - T_i| > \epsilon$, one has

$$f(T_i,v) \leq (2\pi\tau_i^2)^{-1/2}\exp(-\epsilon^2/2\tau_i^2).$$

On the other hand, under the model, $T_i \sim N(\mu_i, \delta_i + \tau_i^2)$ with $\mu_i = b_i^{\mathrm{T}}\beta$ and $\delta_i = \sigma_v^2 d_i^{\mathrm{T}} d_i$. Thus,

$$\begin{aligned} & \int f(T_i,v)f(v)dv \\ = \; & f(T_i) \\ = \; & \frac{1}{\sqrt{2\pi(\delta_i+\tau_i^2)}} \exp\left\{-\frac{(T_i-\mu_i)^2}{2(\delta_i+\tau_i^2)}\right\}. \quad (5) \end{aligned}$$

It follows that, as $n \to \infty$,

$$\begin{aligned}&\left(\int f(T_i,v)f(v)dv\right)^{-1} \sup_{|\zeta_i(v)-T_i|>\epsilon} f(T_i,v)\\ \le\ & \exp\left\{-\frac{1}{2}\left(\frac{\epsilon^2}{\tau_i^2}+\log\tau_i^2\right)\right.\\ &\left.+\frac{(T_i-\mu_i)^2}{2(\tau_i^2+\delta_i)}+\frac{\log(\tau_i^2+\delta_i)}{2}\right\}, \qquad (6)\end{aligned}$$

which converges to 0 in $\mathrm{P_m}$, provided that $\mathrm{E_m}|Y_{ij}|$ is bounded and, as $n \to \infty$, $\sum_{j=1}^n w_{ij}^2 \to 0$ and δ_i is bounded away from 0. Thus, condition 2) is satisfied.

Finally, by taking $\lambda_n = \infty$, condition 3) is obvious.

Remark. Consider a special case: $Y_{ij} = \mu + v_i + e_{ij}$. It is easy to show that, in this case, $\zeta_i = \mu + v_i$, and $\tilde{\zeta}_i = (\sigma_v^2/(\sigma_v^2+\tau_i^2))T_i + (\tau_i^2/(\sigma_v^2+\tau_i^2))\mu$. Thus, our result reduces to that of Prasad & Rao (1999, pp 74).

Example 1 (continued). In case of a finite population, Y_{ij} will not be normally distributed. However, (4) — (6) continue to hold. Note that (4) and (5) are derived under the normal model. Furthermore, the right side of (6) converges to 0 in $\mathrm{P_d}$, provided that all the assumptions hold except that the assumption that $\mathrm{E_m}|Y_{ij}|$ is bounded is replaced by that $|y_{ij}|$ is bounded.

In conclusion, all the assumptions of Theorem 4 hold provided that all the assumptions previously made in Example 1 hold except that $\mathrm{E_m}|Y_{ij}|$ is bounded is replaced by that $|y_{ij}|$ is bounded.

Example 2 (Binary responses). Suppose that, given v_i, Y_{ij}, $1 \le j \le n$ are independent *Bernoulli* with $\mathrm{P}(Y_{ij} = 1|v_i) = p_i$, and $\mathrm{logit}(p_i) = x_i^{\mathrm{T}}\beta + v_i$. It follows that $\zeta_i = p_i$, and $nT_i|v_i \sim \mathrm{Bi}(n, p_i)$, hence $f(T_i, v) = \binom{n}{nT_i} p_i^{nT_i}(1-p_i)^{n-nT_i}|_{v_i=v}$. We first verify the conditions of Theorem 3.

Since $|T_i|$ and $\int |\zeta_i(v)| f(v)dv \le 1$, 1) is satisfied.

Let $\varphi(p) = n\{T_i\mathrm{logit}(p) + \log(1-p)\}$. By Taylor expansion, it is easy to show that, if $|p - T_i| > \epsilon$, $\varphi(p) \le \varphi(T_i) - 8\epsilon^2 T_i(1-T_i)n$. On the other hand, if $|p - T_i| < n^{-1/2}$, i.e., $a_i < v < b_i$, where $a_i = \mathrm{logit}(T_i - n^{-1/2}) - x_i^{\mathrm{T}}\beta$, $b_i = \mathrm{logit}(T_i + n^{-1/2}) - x_i^{\mathrm{T}}\beta$, then, again by Taylor expansion, there is $M > 0$ such that $\Pr(\varphi(p_i) \ge \varphi(T_i) - M) \ \ge\ 1 - \epsilon$, provided that $|x_i|$ is bounded. Thus, with probability $\ge 1 - \epsilon$,

$$\begin{aligned}&\left(\int f(T_i,v)f(v)dv\right)^{-1} \sup_{|p-T_i|>\epsilon} f(T_i,v)\\ \le\ & Cn^{1/2}\exp(-8\epsilon^2 T_i(1-T_i)n)\end{aligned}$$

for some constant C, provided that $f(v)$ is continuous and positive. Therefore, 2) is satisfied.

Finally, by taking $\lambda_n = \infty$, 3) is obvious.

Note that, unlike Example 1, there is no closed form expression for $\tilde{\zeta}_i$ in Example 2. Nevertheless, in this case $\tilde{\zeta}_i$ can be computed easily using a Monte-Carlo method [Jiang (1998)].

Example 2 (continued). Note that Y_{ij} do not have to satisfy the model. In particular, in case of a finite population, it can be shown, using almost exactly the same arguments, that the conditions of Theorem 4 are satisfied as long as T_i is bounded away from 0 and 1 in $\mathrm{P_d}$.

Example 3 (Exponential-Gamma). We assume that, given the random effect v_i, responses Y_{ij}, $j = 1, \ldots, n_i$ are independent with $Y_{ij} \sim \mathrm{Gamma}(\alpha_{ij}, v_i^{-1})$, where α_{ij} is a known positive constant. Furthermore, $v_i \sim \mathrm{Exponential}(\lambda)$. Here $\mathrm{Gamma}(\alpha, \beta)$ has density

$$(\Gamma(\alpha)\beta^\alpha)^{-1} y^{\alpha-1} e^{-y/\beta}, \quad y > 0$$

and $\mathrm{Exponential}(\lambda)$ has density $\lambda e^{-\lambda y}$, $y > 0$. It is easy to see that $T_i|v \sim \mathrm{Gamma}(\alpha., (nv)^{-1})$, where $\alpha. = \sum_{j=1}^n \alpha_j$, so $f(T_i, v) = \{(nv)^{\alpha.}/\Gamma(\alpha.)\} T_i^{\alpha.-1} e^{-nvT_i}$. By Stirling's approximation, it can be shown that

$$\begin{aligned}&f(T_i,v)\\ \approx\ & \sqrt{\frac{\alpha.}{2\pi}} T_i^{-1}\left(\frac{env}{\alpha.}\right)^{\alpha.} T_i^{\alpha.-1} e^{-nvT_i}\\ =\ & \sqrt{\frac{\alpha.}{2\pi}} T_i^{-1} \exp\left(n\bar{\alpha}\left\{1+\log\left(\frac{vT_i}{\bar{\alpha}}\right)-\frac{vT_i}{\bar{\alpha}}\right\}\right),\end{aligned}$$

where $\bar{\alpha} = \alpha./n$. Suppose that, as $n \to \infty$, $\bar{\alpha}$ is bounded from above and away from 0. For any $\eta > 0$, chose a, $A > 0$ such that $\Pr(a \le T_i \le A) \ge 1 - \eta$. Then, there is $\delta > 0$ such that when $0 < v < \delta$, the quantity inside the above $\{\cdots\}$ is less than -1 for all $0 < T_i \le A$. On the other hand, if $v \ge \delta$ and $|T_i - \zeta_i(v)| = |T_i - \bar{\alpha}v^{-1}| > \epsilon$, we have $|(vT_i/\bar{\alpha}) - 1| > \epsilon(v/\bar{\alpha}) \ge \epsilon\delta/\bar{\alpha}$, which is bounded away from 0. Since the function $\varphi(x) = 1 + \log(x) - x$ is strictly negative outside any neighborhood of 1, the quantity inside the above $\{\cdots\}$ is, again, less than some negative constant. It follows that, with probability $\ge 1 - \eta$, we have $\sup_{|\zeta_i(v)-T_i|>\epsilon} f(T_i, v) \le c\sqrt{n}e^{-bn}$ for some constants b, $c > 0$. Furthermore, it is easy to show that, with probability $\ge 1 - \eta$,

$$\left(\int f(T_i,v)f(v)dv\right)^{-1}$$

$$= \frac{T_i^2}{\lambda\bar{a}}\left(1+\frac{\lambda}{nT_i}\right)\left(\left(1+\frac{\lambda}{nT_i}\right)^{\frac{nS}{\lambda}}\right)^{\frac{\lambda\bar{a}}{S}}$$
$$\approx \frac{S^2}{\lambda\bar{a}}\exp\left(\frac{\lambda\bar{a}}{S}\right) \leq c$$

for some constant c. Since η can be arbitrarily small as n increases, condition 2) of Theorem 3 is satisfied with $\lambda_n = \infty$. Condition 3) is obvious.

Example 3 (continued). Again, Y_{ij} do not need to satisfy the model. In particular, in case of finite population, it can be shown, using almost eaxctly the same arguments, that the conditions of Theorem 4 are satisfied as long as T_i is bounded from above and away from zero in $\mathrm{P_d}$.

5 Estimation of MSPE

In this section, we assume that n_i is bounded for all i. Also, we assume that the model holds so that (1) correspond to the ERB predictor. An appropriate precision measure of $\hat{\zeta}_i$ is given by $\mathrm{MSPE}(\hat{\zeta}_i) = \mathrm{E}(\hat{\zeta}_i - \bar{Y}_i)^2$, where the expectation is taken with respect to both the sampling design and the assumed mixed model. Ideally, we would like to obtain an estimator of $\mathrm{MSPE}(\hat{\zeta}_i)$ whose bias (with respect to both the sampling design and the model) is of the order $o(m^{-1})$. However, in this paper we follow the approach of Prasad and Rao (1999) and obtain an estimator of $\mathrm{MSPE}(\hat{\zeta}_i)$ whose bias is of order $o(m^{-1})$ with respect to the model only. The bias will be still of the order $o(m^{-1})$ when an additional expectation is taken with respect to the sampling design, under some mild conditions. In the following we mainly focus on the model-based approximation, so the probability statements, expectations, variances and MSPE are with respect to the assumed model. A remark is made at the end of the section on extension of the results with respect to both model and design.

Throughout the section, we assume $\zeta_i = E(\bar{Y}_i|v_i)$ which holds, for example, the mixed linear normal model of Prasad and Rao (1999) and the mixed logistic model of Jiang and Lahiri (2001). Using this assumption, certain regularity conditions on $f(v_i)$ and the central limit theorem, it can be shown that $\bar{Y}_i - \zeta_i = O_{\mathrm{P}}(N_i^{-1/2})$. Thus, $(\hat{\zeta}_i - \bar{Y}_i)^2 = (\hat{\zeta}_i - \zeta_i)^2 + O_{\mathrm{P}}(N_i^{-1/2})$. Because of the above fact, we approximate $\mathrm{E}(\hat{\zeta}_i - \bar{Y}_i)^2$ by $\mathrm{E}(\hat{\zeta}_i - \zeta_i)^2$. It is possible to claim the order of the neglected terms as $O(N_i^{-1/2})$. However, the proof is technical and is not attempted here. Note that here we are assuming that the population size N_i is much larger than m so we can ignore any term of the order $N_i^{-1/2}$ or lower.

Turning to the model MSPE of $\hat{\zeta}_i$, we have

$$\begin{aligned}\mathrm{MSPE}(\hat{\zeta}_i) &= \mathrm{MSPE}(\tilde{\zeta}_i) + \mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 \\ &\quad + 2\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)(\tilde{\zeta}_i - \zeta_i) . \end{aligned} \qquad (7)$$

First, we have

$$\begin{aligned}\mathrm{MSPE}(\tilde{\zeta}_i) &= \mathrm{E}\zeta_i^2 - \mathrm{E}\tilde{\zeta}_i^2 \\ &= \mathrm{E}\left(\sum_{j=1}^{n_i} w_{ij}\mathrm{E}(Y_{ij}|v_i)\right)^2 - \mathrm{E}u_i^2(T_i,\theta) \\ &\equiv d_i(\theta) . \end{aligned} \qquad (8)$$

Second, by the same arguments as in section 3 of Jiang (1999), we have

$$\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 = e_i(\theta)m^{-1} + o(m^{-1}) , \qquad (9)$$

where $e_i(\theta) = \mathrm{E}\left(\partial u_i/\partial\theta^{\mathrm{T}}\right) V(\theta)\left(\partial u_i/\partial\theta\right)$ with $V(\theta) = m\mathrm{E}(\hat{\theta}-\theta)(\hat{\theta}-\theta)^{\mathrm{T}}$.

Third, to obtain an approximation for the third term on the right side of (6), we make further assumptions on $\hat{\theta}$. Suppose that $\hat{\theta}$ is a solution to an estimating equation of the following type

$$M(\theta) = \frac{1}{m}\sum_{i=1}^{m} a_i(Y_i,\theta) = 0 , \qquad (10)$$

where $Y_i = (Y_{ij})_{1\le j\le n_i}$, $a_i(\cdot,\cdot)$ is vector-valued such that $\mathrm{E}a_i(Y_i,\theta) = 0$ if θ is the true vector of parameters, $1 \le i \le m$. For example, it is easy to see that the maximum likelihood estimator of θ satisfies the above. It can be shown that (see Appendix) if $\hat{\theta}$ satisfies the above, then

$$\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)(\tilde{\zeta}_i - \zeta_i) = g_i(\theta)m^{-1} + o(m^{-1}) , \qquad (11)$$

where $g_i(\theta) = \mathrm{E}\omega_i(Y,\theta)(\mathrm{E}(\zeta_i|T_i) - \mathrm{E}(\zeta_i|Y))$ with

$$\begin{aligned}\omega_i(Y,\theta) &= -\left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)A^{-1}a_i(Y_i,\theta) + m\delta_i(Y,\theta) , \\ \delta_i(Y,\theta) &= \left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)A^{-1}\left(\frac{\partial M}{\partial\theta^{\mathrm{T}}} - A\right)A^{-1}M(\theta) \\ &\quad + \frac{1}{2}\left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)A^{-1}\left(M^{\mathrm{T}}(\theta)A^{-\mathrm{T}}\right.\end{aligned}$$

$$\times \mathrm{E}\left(\frac{\partial^2 M_j}{\partial\theta\partial\theta^{\mathrm{T}}}\right) A^{-1} M(\theta)\Big) + \frac{1}{2} M^{\mathrm{T}}(\theta) A^{-\mathrm{T}} \frac{\partial^2 u_i}{\partial\theta\partial\theta^{\mathrm{T}}} A^{-1} M(\theta) ,$$

$A = \mathrm{E}(\partial M/\partial\theta^{\mathrm{T}})$ and $A^{-\mathrm{T}} = (A^{-1})^{\mathrm{T}}$. Here M_j represents the jth component of M, and (b_j) a vector with components b_j.

Thus, under suitable conditions (see Appendix), we obtain the approximation

$$\mathrm{MSPE}(\hat{\zeta}_i) = d_i(\theta) + m^{-1}(e_i(\theta) + 2g_i(\theta)) + o(m^{-1}) . \quad (12)$$

Finally, if we define

$$\widehat{\mathrm{MSPE}}(\hat{\zeta}_i) = d_i(\hat{\theta}) + m^{-1}(\widehat{e_i(\theta)} + 2g_i(\hat{\theta}) - \widehat{B_i(\theta)}) , \quad (13)$$

where

$$B_i(\theta) = m\left(\left(\frac{\partial d_i}{\partial\theta^{\mathrm{T}}}\right)\mathrm{E}(\hat{\theta}-\theta) + \frac{1}{2}\mathrm{E}(\hat{\theta}-\theta)^{\mathrm{T}}\left(\frac{\partial^2 d_i}{\partial\theta\partial\theta^{\mathrm{T}}}\right)(\hat{\theta}-\theta)\right),$$

and $\widehat{e_i(\theta)}$, $\widehat{B_i(\theta)}$ are estimators of $e_i(\theta)$, $B_i(\theta)$ discussed in the next section, then, under suitable conditions (see Appendix), we have

$$\mathrm{E}(\widehat{\mathrm{MSPE}}(\hat{\zeta}_i) - \mathrm{MSPE}(\hat{\zeta}_i)) = o(m^{-1}) . \quad (14)$$

The quantities $d_i(\theta)$ and $g_i(\theta)$ can be estimated by $d_i(\hat{\theta})$ and $g_i(\hat{\theta})$, respectively. As for $e_i(\theta)$ and $B_i(\theta)$, they involve quantities such as $\mathrm{E}(\hat{\theta}-\theta)$ and $\mathrm{E}(\hat{\theta}-\theta)(\hat{\theta}-\theta)^{\mathrm{T}}$. Their estimation will be discussed in the next section.

Remark. The results of this section, that is, (12) and (14), continue to hold when another expectation is taken with respect to the design, provided that the bounds on the derivatives and moments as well as $\|A^{-1}\|$ given in the Appendix can be made not dependent on the sampling design, which typically is possible. The argument is straightforward from the definition of expectation with respect to the design.

6 Estimation of $e_i(\theta)$ and $B_i(\theta)$

We now return to the remaining issue of the previous section on the estimation of two quantities, $e_i(\theta)$ and $B_i(\theta)$. Here, again, the expected values are with respect to the model. First, we have the following alternative expressions:

$$e_i(\theta) = \mathrm{tr}(V(\theta)G_1(\theta)), \quad (15)$$

$$B_i(\theta) = \left(\frac{\partial d_i}{\partial\theta}\right)^{\mathrm{T}} b(\theta) + \frac{1}{2}\mathrm{tr}(V(\theta)G_2(\theta)), \quad (16)$$

where $b(\theta) = m\mathrm{E}(\hat{\theta}-\theta)$, $V(\theta) = m\mathrm{E}(\hat{\theta}-\theta)(\hat{\theta}-\theta)^{\mathrm{T}}$, $G_1(\theta) = \mathrm{E}(\partial u_i/\partial\theta)(\partial u_i/\partial\theta)^{\mathrm{T}}$ and $G_2(\theta) = \partial^2 d_i/\partial\theta\,\partial\theta^{\mathrm{T}}$. $G_j(\theta)$ can be estimated by a plug-in estimator, that is, $G_j(\hat{\theta})$, $j = 1, 2$. As for $b(\theta)$ and $V(\theta)$, we propose to use the following sandwich-type estimators:

$$\widehat{V(\theta)} = m\left(\sum_{i=1}^m \left.\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}}\right)^{-1}\left(\sum_{i=1}^m \hat{a}_i\hat{a}_i^{\mathrm{T}}\right) \times \left(\sum_{i=1}^m \left.\frac{\partial a_i^{\mathrm{T}}}{\partial\theta}\right|_{\theta=\hat{\theta}}\right)^{-1}, \quad (17)$$

$$\widehat{b(\theta)} = \frac{1}{m}\sum_{i=1}^m \hat{A}^{-1}\left.\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}} \hat{A}^{-1}\hat{a}_i - \frac{1}{2}\hat{A}^{-1} \times \left(\frac{1}{m}\sum_{i=1}^m \hat{a}_i^{\mathrm{T}}\hat{A}^{-\mathrm{T}}\hat{H}_j\hat{A}^{-1}\hat{a}_i\right), \quad (18)$$

where $\hat{a}_i = a_i(Y_i, \hat{\theta})$,

$$\hat{A} = m^{-1}\sum_{i=1}^m \left.\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}},$$

$$\hat{H}_j = m^{-1}\sum_{i=1}^m \left.\frac{\partial^2 a_{i,j}}{\partial\theta\partial\theta^{\mathrm{T}}}\right|_{\theta=\hat{\theta}},$$

and, as before, (b_j) represents a vector whose jth component is b_j. The derivations of (17) and (18) are given in Appendix.

The estimators of $e_i(\theta)$ and $B_i(\theta)$, denoted by $\widehat{e_i(\theta)}$ and $\widehat{B_i(\theta)}$, are obtained by (14) and (15) with $V(\theta)$ and $b(\theta)$ replaced by $\widehat{V(\theta)}$ and $\widehat{b(\theta)}$, respectively, and $G_j(\theta)$ replaced by $G_i(\hat{\theta})$, $j = 1, 2$.

Note. Some derivatives are involved in the expressions of $\widehat{V(\theta)}$ and $\widehat{b(\theta)}$. Sometimes, computing the analytic forms of the derivatives, especially second derivatives, can be tedious and errors are often made. Alternatively, one may use the numerical differentiation method as follows. Let $h(x_1, \ldots, x_k)$ be a twice continously differentiable function, then

$$\begin{aligned}\frac{\partial h}{\partial x_i} &\approx \frac{1}{\Delta x_i}\{h(x_1, \ldots, x_i + \Delta x_i, \ldots, x_k) \\ &\quad -h(x_1, \ldots, x_k)\}, \\ \frac{\partial^2 h}{\partial x_i \partial x_j} &\approx \frac{1}{\Delta x_i \Delta x_j}\{h(x_1, \ldots, x_i + \Delta x_i, \\ &\quad \ldots, x_j + \Delta x_j, \ldots, x_k) \\ &\quad -h(x_1, \ldots, x_i + \Delta x_i, \ldots, x_k) \\ &\quad -h(x_1, \ldots, x_j + \Delta x_j, \ldots, x_k) \\ &\quad +h(x_1, \ldots, x_k)\},\end{aligned}$$

as $\Delta x_i, \Delta x_j \to 0$. In practice, one may take $\Delta x_i = \Delta \cdot x_i$, where Δ is a small number (e.g., $\Delta = 10^{-4}$).

We now use a specific example to illustrate the MSPE estimation.

7 An example

Consider the following simple nested error regression model: $Y_{ij} = \beta x_{ij} + v_i + e_{ij}$, $i = 1, \cdots, m$, $j = 1, \cdots, n_i$, where v_i's and e_{ij}'s are independent normal with means 0 and variances σ_v^2 and $\sigma_e^2 k_{ij}$, respectively, and k_{ij}'s are known. For simplicity and without loss of generality, we let $k_{ij} = 1$. Then, it is a special case of Example 1 in section 4 (see the notation therein). Let $\theta = (\beta, \sigma_v^2, \sigma_e^2)'$ denote the parameters involved.

If we choose $w_{ij} = n_i^{-1}$, then $T_i = \bar{Y}_i$. Furthermore, we have $b_i = \bar{x}_i = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$, $d_i = 1$; $\delta_i = \sigma_v^2$, $\mu_i = \beta\bar{x}_i$, and $\zeta_i = \beta\bar{x}_i + v_i$ so that $\zeta_i(v) = \beta\bar{x}_i + v$. It is easy to show that

$$\tilde{\zeta}_i = \frac{\sigma_e^2}{\sigma_e^2 + n_i\sigma_v^2}\beta\bar{x}_i + \frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}T_i,$$

and $\hat{\zeta}_i$ is $\tilde{\zeta}_i$ with θ replaced by $\hat{\theta}$.

The estimator of θ is the solution to (10), where

$$\begin{aligned}a_{i,1} &= b_{i,1}(\psi)(Y_{i\cdot} - x_{i\cdot}\beta) + \sum_{j=1}^{n_i} x_{ij}(Y_{ij} - x_{ij}\beta), \\ a_{i,2} &= b_{i,2}(\psi)\{(Y_{i\cdot} - x_{i\cdot}\beta)^2 - n_i(\sigma_e^2 + n_i\sigma_v^2)\}, \\ a_{i,3} &= b_{i,3}(\psi)\{(Y_{i\cdot} - x_{i\cdot}\beta)^2 - n_i(\sigma_e^2 + n_i\sigma_v^2)\} \\ &\quad -\sigma_e^{-4}\sum_{j=1}^{n_i}\{(Y_{ij} - x_{ij}\beta)^2 - (\sigma_v^2 + \sigma_e^2)\}\end{aligned}$$

with $\psi = (\sigma_v^2, \sigma_e^2)'$,

$$\begin{aligned}b_{i,1}(\psi) &= \frac{\sigma_v^2 x_{i\cdot}}{\sigma_e^2 + n_i\sigma_v^2}, \\ b_{i,2}(\psi) &= -(\sigma_e^2 + n_i\sigma_v^2)^{-2}, \\ b_{i,3}(\psi) &= \frac{\sigma_v^2(2\sigma_e^2 + n_i\sigma_v^2)}{\sigma_e^4(\sigma_e^2 + n_i\sigma_v^2)^2}.\end{aligned}$$

According to sections 5 and 6, he MSPE estimator is given by $\widehat{\text{MSPE}}(\hat{\zeta}_i) = d_i(\hat{\theta}) + m^{-1}\{\widehat{e_i(\theta)} + 2g_i(\hat{\theta}) - \widehat{B_i(\theta)}\}$, where $\widehat{e_i(\theta)}$ and $\widehat{B_i(\theta)}$ are given at the end of section 6. More specifically, we have the following expression: $d_i(\theta) = \sigma_v^2\sigma_e^2/(\sigma_e^2 + n_i\sigma_v^2)$. Furthermore, in this case, we have $\text{E}(\zeta_i|Y) = \text{E}(\zeta_i|Y_i) = \text{E}(\zeta_i|T_i) = \tilde{\zeta}_i$, so $g_i(\theta) = 0$. The estimation of $e_i(\theta)$ and $B_i(\theta)$ is discussed in section 6.

Appendix

Proof of Theorem 1: Note that for an arbitrary predictor of the form $\hat{\bar{Y}}_i = \hat{\bar{Y}}_i(T_i)$

$$\begin{aligned}&\text{E}_M(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^{RB})(\hat{\bar{Y}}_i^{RB} - \bar{Y}_i) \\ =\ &\text{E}_M\left((\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^{RB})\text{E}_M[(\hat{\bar{Y}}_i^{RB} - \bar{Y}_i)|T_i]\right) \\ =\ &0\end{aligned}$$

since $\text{E}_M[\bar{Y}_i|T_i] = \text{E}_M[\text{E}_M\{\bar{Y}_i|T_i, v_i\}|T_i] = \hat{\bar{Y}}_i^{RB}$.

Using the above fact, we have $\text{MSPE}(\hat{\bar{Y}}_i) = \text{E}(\hat{\bar{Y}}_i - \hat{\bar{Y}}_i^{RB})^2 + \text{MSPE}(\hat{\bar{Y}}_i^{RB})$. This proves Theorem 1. ∎

Proof of Theorem 2: The proof of part (i) follows immediately from the definitions of $\hat{\bar{Y}}_i^{ERB}$, ζ_i and T_i. To prove part (ii) note that

$$\begin{aligned}\text{E}[\tilde{\zeta}_i] &= \text{E}_d\text{E}_M\left[\left\{\sum_{j=1}^{n_i} w_{ij}\text{E}_M(\text{E}_M(Y_{ij}|v_i)|T_i)\right\}\right] \\ &= \text{E}_M\left[\text{E}_d\left\{\sum_{j=1}^{n_i} w_{ij}\text{E}_M(\text{E}_M(Y_{ij}|v_i)|T_i)\right\}\right] \\ &= \text{E}_M[\hat{\bar{Y}}_i^{RB}].\end{aligned}$$

This completes the theorem. ∎

Proof of Theorem 3: We have

$$\tilde{\zeta}_i - T_i = \frac{\int(\zeta_i(v) - T_i)f(T_i, v)f(v)dv}{\int f(T_i, v)f(v)dv}. \quad (A.1)$$

For any $\epsilon > 0$, the numerator in (A.1) =

$$\int_{|\zeta_i(v)-T_i|\le\epsilon}\cdots+\int_{|\zeta_i(v)-T_i|>\epsilon,|v|<\lambda_n}\cdots + \int_{|\zeta_i(v)-T_i|>\epsilon,|v|\ge\lambda_n}\cdots = \sum_{k=1}^{3} I_k . \qquad (A.2)$$

Observe that

$$|I_1| \le \epsilon\int f(T_i,v)f(v)dv , \qquad (A.3)$$

$$|I_2| \le \left\{\sup_{|\zeta_i(v)-T_i|>\epsilon,|v|<\lambda_n} f(T_i,v)\right\} \left(|T_i|+\int|\zeta_i(v)|f(v)dv\right) , \qquad (A.4)$$

$$|I_3| \le (|T_i|+1)\int_{|v|\ge\lambda_n}(|\zeta_i(v)|+1)f(T_i,v)f(v)dv . \qquad (A.5)$$

Thus, by (A.1) — (A.5), and the assumptions, we have $|\tilde{\zeta}_i - T_i| \le \epsilon + \xi_1\delta_1 + \xi_2\delta_2$, where ξ_k are bounded in probability, and $\delta_k \to 0$ in $\mathrm{P_m}$, $k = 1, 2$. Therefore, $\Pr(|\tilde{\zeta}_i - T_i| \ge 3\epsilon) \to 0$. ∎

Derivation of (11): First note that, under regularity conditions, $M(\theta) = O_P(m^{-1/2})$, $\partial M/\partial\theta^{\mathrm{T}} = O_P(1)$, $\partial^2 M/\partial\theta\partial\theta^{\mathrm{T}} = O_P(1)$, $(\partial M/\partial\theta^{\mathrm{T}}) - A = O_P(m^{-1/2})$, and $\hat{\theta}-\theta = O_P(m^{-1/2})$. Hereafter θ represents the true vector of parameters. By Taylor expansion, it is easy to show that

$$\hat{\theta}-\theta = -\left\{\frac{\partial M}{\partial\theta^{\mathrm{T}}}+\frac{1}{2}\left((\hat{\theta}-\theta)^{\mathrm{T}}\frac{\partial^2 M_j}{\partial\theta\partial\theta^{\mathrm{T}}}\right)\right\}^{-1}M(\theta) + o_P(m^{-1}) , \qquad (A.6)$$

where (b_j) represents a matrix whose jth row is b_j.

Denote the matrix in $\{\cdots\}$ by B. Then, we have

$$B = A + O_P(m^{-1/2}),$$

hence

$$\begin{aligned} B^{-1} &= A^{-1} - A^{-1}(B-A)B^{-1} \\ &= A^{-1} + O_P(m^{-1/2}). \end{aligned}$$

It follows, by (A.6), that

$$\hat{\theta}-\theta = -A^{-1}M(\theta) + O_P(m^{-1}).$$

Now, coming back once again, we have

$$\begin{aligned} B^{-1} &= A^{-1} - A^{-1}(B-A)B^{-1} \\ &= A^{-1} - A^{-1}(B-A)(A^{-1} \\ &\quad -A^{-1}(B-A)B^{-1}) \\ &= A^{-1} - A^{-1}(B-A)A^{-1} \\ &\quad +(A^{-1}(B-A))^2B^{-1} \\ &= A^{-1} - A^{-1}\left(\frac{\partial M}{\partial\theta^{\mathrm{T}}}-A\right)A^{-1} \\ &\quad -\frac{1}{2}A^{-1}\left(M^{\mathrm{T}}(\theta)A^{-\mathrm{T}}\frac{\partial^2 M_j}{\partial\theta\partial\theta^{\mathrm{T}}}\right)A^{-1} \\ &\quad +O_P(m^{-1}) . \end{aligned}$$

Thus, once again by (A.6), we have

$$\begin{aligned} \hat{\theta}-\theta &= -A^{-1}M + A^{-1}\left(\frac{\partial M}{\partial\theta^{\mathrm{T}}}-A\right)A^{-1}M \\ &\quad +\frac{1}{2}A^{-1}\left(M^{\mathrm{T}}(\theta)A^{-\mathrm{T}}\mathrm{E}\left(\frac{\partial^2 M_j}{\partial\theta\partial\theta^{\mathrm{T}}}\right)\right) \\ &\quad \times A^{-1}M + o_P(m^{-1}) . \end{aligned}$$

By the above asymptotic expansion and Taylor expansion, we have

$$\begin{aligned} &u_i(T_i,\hat{\theta}) - u_i(T_i,\theta) \\ &= \left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)^{\mathrm{T}}(\hat{\theta}-\theta) + \frac{1}{2}(\hat{\theta}-\theta)^{\mathrm{T}}\frac{\partial^2 u_i}{\partial\theta\partial\theta^{\mathrm{T}}}(\hat{\theta}-\theta) \\ &\quad +o_P(m^{-1}) \\ &= -\left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)^{\mathrm{T}}A^{-1}M + \delta_i(Y,\theta) + o_P(m^{-1}) . \end{aligned}$$

Note that $\delta_i(Y,\theta) = O_P(m^{-1})$.

Since $\mathrm{E}a_{i'}(Y_{i'},\theta)(u_i(T_i,\theta)-\zeta_i) = 0$ if $i' \ne i$, we have $\mathrm{E}M(\theta)(u_i(T_i,\theta)-\zeta_i) = m^{-1}\mathrm{E}a_i(Y_i,\theta)(u_i(T_i,\theta)-\zeta_i)$. Therefore, we have

$$\begin{aligned} &\mathrm{E}(\hat{\zeta}_i-\tilde{\zeta}_i)(\tilde{\zeta}_i-\zeta_i) \\ &= \frac{1}{m}\mathrm{E}\left(-\left(\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}\right)^{\mathrm{T}}A^{-1}a_i(Y_i,\theta)+m\delta_i(Y,\theta)\right) \\ &\quad \times(u_i(T_i,\theta)-\zeta_i) + o(m^{-1}) . \end{aligned}$$

Finally, note that for any function $\omega(y,\theta)$, we have

$$\begin{aligned} &\mathrm{E}\omega(Y,\theta)(u_i(T_i,\theta)-\zeta_i) \\ &= \mathrm{E}\omega(Y,\theta)(\mathrm{E}(\zeta_i|T_i)-\mathrm{E}(\zeta_i|Y)). \end{aligned}$$

Derivation of (16) and (17): By (A.6), we have

$$\hat{\theta}-\theta \;\approx\; -A^{-1}M. \qquad (A.7)$$

Thus, we have

$$\begin{aligned} V(\theta) &\approx mA^{-1}\mathrm{E}(MM^{\mathrm{T}})A^{-\mathrm{T}} \\ &= mA^{-1}\left(\frac{1}{m^2}\sum_{i=1}^{m}\mathrm{E}(a_i a_i^{\mathrm{T}})\right)A^{-\mathrm{T}} \\ &\approx m\left(\sum_{i=1}^{m}\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\hat{\theta}}\right)^{-1}\left(\sum_{i=1}^{m}\hat{a}_i\hat{a}_i^{\mathrm{T}}\right) \\ &\quad\times m\left(\sum_{i=1}^{m}\frac{\partial a_i^{\mathrm{T}}}{\partial\theta}\bigg|_{\theta=\hat{\theta}}\right)^{-1}. \end{aligned}$$

Here the approximation is in the sense that the difference between the two sides is $o_{\mathrm{P}}(1)$. Similarly, by (A.6) (see the expansion of B^{-1} below it) and the previous approximation (A.7), we have a second-order expansion

$$\begin{aligned} \hat{\theta}-\theta &\approx -A^{-1}M \\ &\quad +A^{-1}\left(\frac{\partial M}{\partial\theta^{\mathrm{T}}}-A-\frac{1}{2}(M^{\mathrm{T}}A^{-\mathrm{T}}H_j)\right)A^{-1}M. \end{aligned}$$

It follows that

$$\begin{aligned} b(\theta) &\approx \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left(A^{-1}\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}A^{-1}a_i\right) \\ &\quad -\frac{1}{2}A^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}(a_i^{\mathrm{T}}A^{-\mathrm{T}}H_jA^{-1}a_i)\right) \\ &\approx \frac{1}{m}\sum_{i=1}^{m}\hat{A}^{-1}\frac{\partial a_i}{\partial\theta^{\mathrm{T}}}\bigg|_{\theta=\hat{\theta}}\hat{A}^{-1}\hat{a}_i \\ &\quad -\frac{1}{2}\hat{A}^{-1}\left(\frac{1}{m}\sum_{i=1}^{m}\hat{a}_i^{\mathrm{T}}\hat{A}^{-\mathrm{T}}\hat{H}_j\hat{A}^{-1}\hat{a}_i\right). \end{aligned}$$

Again, the approximation is in the sense that the difference between the two sides is $o_{\mathrm{P}}(1)$.

Conditions for (12) and sketch of proof: The following are some technical conditions for (12) to hold:

(i) u_i is three-times continuously differentiable with respect to θ,

$$\left|\frac{\partial u_i}{\partial\theta}\right| \le b_1(T_i)(1+|\theta|^{\lambda})$$

for some $\lambda>0$, and there is a neighborhood of the true θ, in which

$$\left|\frac{\partial^3 u_i}{\partial\theta_r\partial\theta_s\partial\theta_t}\right| \le b_3(T_i) \quad \text{for all } r,s,t,$$

such that $\mathrm{E}(b_1(T_i))^k$ and $\mathrm{E}(b_3(T_i))^k$ are bounded for any given $k>0$;

(ii) there is a set B with $\mathrm{P}(B^c)=o(m^{-d_1})$ for some $d_1>2$ such that on B we have $|\hat{\theta}-\theta|\le m^{-d_2}$, where $d_1>1/3$, and the following asymptotic expansion holds.

$$\begin{aligned} \hat{\theta}-\theta &= -A^{-1}M+A^{-1}\left(\frac{\partial M}{\partial\theta^{\mathrm{T}}}-A\right)A^{-1}M \\ &\quad +\frac{1}{2}A^{-1}(M^{\mathrm{T}}A^{-\mathrm{T}}H_jA^{-1}M) \\ &\quad +o(m^{-1})D, \end{aligned}$$

where $\mathrm{E}(|D|^k)$ is bounded for any given $k>0$; furthermore, there are c_1 and $c_2>0$ such that $|\hat{\theta}|\le c_1(\log m)^{c_2}$;

(iii) the following are bounded for any given $k>0$: $\mathrm{E}(|a_i|^k)$, $\mathrm{E}(|\partial a_i/\partial\theta|^k)$, $\mathrm{E}(\|\partial^2 a_i/\partial\theta\partial\theta^{\mathrm{T}}\|^k)$, $\mathrm{E}(|\zeta_i|^k)$ and $\mathrm{E}(|\tilde{\zeta}_i|^k)$; and

(iv) $A^{-1}=O(m^{-1})$.

Note 1. The expansion in (ii) is derived by Taylor expansion, and then replacing the first two derivatives of M with respect to θ by their expected values. Sufficient conditions for such an expansion can be found in Das *et al.* (2004, Theorem 2.1 and Theorem 4.1). The last condition of (ii) holds, for example, if $\hat{\theta}$ is truncated (e.g., Das *et al.* 2004, pp. 826).

Note 2. In (i) — (iii), the bounds for the moments may depend on k. Such conditions are satisfied, for example, if Y_{ij}'s are normal, or if Y_{ij}'s are bounded such as in the case of binary responses.

Below is a sketch of the the proof.

First, by Taylor expansion, we have

$$\begin{aligned} \hat{\zeta}_i-\tilde{\zeta}_i &= \frac{\partial u_i}{\partial\theta^{\mathrm{T}}}(\hat{\theta}-\theta)+\frac{1}{2}(\hat{\theta}-\theta)^{\mathrm{T}}\frac{\partial^2 u_i}{\partial\theta\partial\theta^{\mathrm{T}}}(\hat{\theta}-\theta) \\ &\quad +o(m^{-1})D_1, \end{aligned}$$

where $\mathrm{E}|D_1|^k$ is bounded for any $k>0$. Using a first order expansion $\hat{\theta}-\theta=-A^{-1}M+o(m^{-1/2})D_2$, where $\mathrm{E}|D_2|^k$ is bounded for any $k>0$, for the second term, and the expansion in (ii) for the first term, we have

$$\hat{\zeta}_i-\tilde{\zeta}_i = -\frac{\partial u_i}{\partial\theta^{\mathrm{T}}}A^{-1}M+\delta_i+o(m^{-1})D_3, \qquad (A.8)$$

where $\mathrm{E}|D_3|^k$ is bounded for any $k > 0$. The above expansion holds on B, which is assumed in (ii). It follows that

$$\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 1_B = m^{-1} e_i(\theta) + o(m^{-1}).$$

Here we have used, again, the first order expansion of $\hat{\theta} - \theta$ and an argument similar to Jiang and Lahiri (2001, pp. 224). On the other hand, we have $\mathrm{E}(\hat{\zeta}_i - \tilde{\zeta}_i)^2 1_{B^c} = o(m^{-1})$. Using (A.8) and similar argument, it can be shown that (11) holds.

Sufficient conditions for (14); sketch of proof: The following are conditions, in addition to (i) — (iv) above, for (14) to hold.

(v) d_i is three-times continuously differentiable, g_i, G_1 are continuously differentiable, and there are constants c_3, $c_4 > 0$ such that $|d_i|$, $|g_i|$, $|G_1|$ and $|G_2|$ are bounded by $c_3(1 + |\theta|^{c_4})$;

(vi) $\widehat{V(\theta)}$ and $\widehat{b(\theta)}$ are bounded by $c_5(\log m)^{c_6}$ for some c_5, $c_6 > 0$.

Note. Condition (vi) is satisfied if, for example, the estimators $\widehat{V(\theta)}$ and $\widehat{b(\theta)}$ are truncated (e.g., Das *et al.* 2004, pp. 831).

Below is a sketch of the proof.

It suffices to show (a) $\mathrm{E}(d_i(\hat{\theta}) - d_i(\theta) - m^{-1}\widehat{B_i(\theta)}) = o(m^{-1})$; (b) $\mathrm{E}(\widehat{e_i(\theta)} - e_i(\theta)) = o(1)$; and (c) $\mathrm{E}(g_i(\hat{\theta}) - g_i(\theta)) = o(1)$.

(a): By Taylor expansion it can be shown that

$$\mathrm{E}(d_i(\hat{\theta}) - d_i(\theta)) = m^{-1} B_i(\theta) + o(m^{-1}).$$

Furthermore, we have

$$\begin{aligned} \widehat{B_i(\theta)} - B_i(\theta) &= \left(\left.\frac{\partial d_i}{\partial \theta^{\mathrm{T}}}\right|_{\hat{\theta}} - \frac{\partial d_i}{\partial \theta^{\mathrm{T}}} \right) \widehat{b(\theta)} \\ &\quad + \frac{\partial d_i}{\partial \theta^{\mathrm{T}}} (\widehat{b(\theta)} - b(\theta)) \\ &\quad + \frac{1}{2} \mathrm{tr}(\widehat{V(\theta)}(G_2(\hat{\theta}) - G_2(\theta))) \\ &\quad + \frac{1}{2} \mathrm{tr}((\widehat{V(\theta)} - V(\theta)) G_2(\theta)). \end{aligned}$$

Thus, it can be shown that $\mathrm{E}(\widehat{B_i(\theta)} - B_i(\theta)) = o(1)$. In addition, note that

$$\begin{aligned} & \mathrm{E}(d_i(\hat{\theta}) - d_i(\theta) - m^{-1}\widehat{B_i(\theta)}) \\ = \; & \mathrm{E}(d_i(\hat{\theta}) - d_i(\theta)) - m^{-1} B_i(\theta) - m^{-1}\mathrm{E}(\widehat{B_i(\theta)} \\ & - B_i(\theta)). \end{aligned}$$

(b): This follows by similar arguments and the following expression

$$\begin{aligned} \widehat{e_i(\theta)} - e_i(\theta) &= \mathrm{tr}(\widehat{V(\theta)}(G_1(\hat{\theta}) - G_1(\theta)) \\ &\quad + \mathrm{tr}((\widehat{V(\theta)} - V(\theta)) G_1(\theta)). \end{aligned}$$

(c): Write $\mathrm{E}(g_i(\hat{\theta}) - g_i(\theta)) = \mathrm{E}(\cdots)1_B + \mathrm{E}(\cdots)1_{B^c}$. Use Taylor expansion for the first term and the probability of B^c to evaluate the second term.

Acknowledgments

The research of J. Jiang was supported in part by National Science Foundation Grant SES-9978101. The research of P. Lahiri was supported in part by National Science Foundation Grant SES-9978145 and a grant from the Center of Excellence in Health Statistics of the University of Michigan, Ann Arbor.

References

[1] Arora, V. and Lahiri, P. (1997), On the superiority of the Bayesian method over the BLUP in small area estimation problems, *Statist. Sinica*, 7, 1053-1063.

[2] Arora, V., Lahiri, P., and Mukherjee, K. (1997), Empirical Bayes estimation of finite population means from complex surveys, *J. Amer. Statist. Assoc.* 92, 1555-1562.

[3] Battese, G.E., Harter,R., and Fuller, W. (1988), An error component model for prediction of county crop areas using survey and satellite data, *J. Amer. Statist. Assoc.*, 83, 28-36.

[4] Das, K., Jiang, J. and Rao, J. N. K. (2004), Mean squared error of empirical predictor, *Ann. Statist.* 32, 818-840.

[5] Datta, G.S. and Ghosh, M. (1991), Bayesian prediction in linear models: applications to small area estimation, *Ann. Statist.*, 19, 1746-1770.

[6] Ericson, W.A. (1969), Subjective Bayesian Models in sampling Finite Populations (with Discussion), *J. Roy. Statist. Soc.*, Ser B, 31, 195-233.

[7] Folsom, R., Shah, B., and Vaish, A. (1999), Substance abuse in states: a methodological report on model based estimates from the 1994-1996 National Household Survey on Drug Abuse, *Proc. Surv. Res. Meth. Sec. of the Amer. Statist. Assoc.*, 371-375.

[8] Ghosh, M., and Lahiri, P., (1987), Robust Empirical Bayes Estimation of Means from Stratified Samples, *J. Amer. Statist. Assoc.*, 82, 1153-1162.

[9] Ghosh, M., and Meeden, G. (1986), Empirical Bayes Estimation in Finite Population Sampling, *J. Amer. Statist. Assoc.*, 81, 1058-1062.

[10] Ghosh, M. and Rao, J. N. K. (1994), Small area estimation: An appraisal (with discussion), *Statist. Sci.* 9, 55-93.

[11] Jiang, J. (1996), REML estimation: Asymptotic behavior and related topics, *Ann. Statist.* 24, 255-286,

[12] Jiang, J. (1998), Consistent estimators in generalized linear mixed models, *J. Amer. Statist. Assoc.* 93, 720-729.

[13] Jiang, J. (1999), Empirical best prediction for small area inference based on generalized linear mixed models, to appear in *J. Statist. Plann. Inf.*.

[14] Jiang, J. and Lahiri, P. (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Math. Statist.*, 53(2): 217-243.

[15] Jiang, J. and Zhang, W. (2001), Robust estimation in generalized linear mixed models, *Biometrika* 88, 753-765.

[16] Kott, P. (1989), Robust small domain estimation using random effects modelling, *Surv. Meth.*, 15, 3-12.

[17] Lahiri, P. and Meza, J. (2001), Small-Area estimation in environment, to appear in *Encyclopedia of Environmetrics*, Wiley.

[18] Nandram, B., and Sedransk, J. (1993), Empirical Bayes Estimation for the Finite Population Mean on the Current Occasion, *J. Amer. Statist. Asso.*, 88, 994-1000.

[19] Pfeffermann, D. (2002), Small area estimation - new developments and directions, *Int. Statist. Rev.*, **70**, 125-143 .

[20] Prasad, N.G.N., and Rao, J.N.K. (1999), On robust small area estimation using a simple random effects model, *Surv. Meth.*, 25, 67-72.

[21] Richardson, A. M. and Welsh, A. H. (1994), Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models, *Austral. J. Statist.* 36, 31-43.

[22] You, Y. and Rao, J.N.K. (2000), Pseudo Hierarchical Bayes small area estimation combining unit-level models and survey weights, *Tech. Report.*, 171-185.