Small Area Estimation under Density Ratio Model

Jiahua Chen[*] Yukun Liu[†]

Abstract

Sample surveys are widely used to obtain information about totals, means and other parameters of finite populations. In many applications, the same information are also desired for subpopulations such as individuals in specific geographic areas and socio-demographic groups. Often, the surveys are conducted at national or similarly high levels. The random nature of the probability sampling can result in no sampling units from many sub-populations of interest. Estimating parameters of these sub-populations with satisfactory precision and evaluating their accuracy pose serious challenges to statisticians. Lacking sufficient amount of direction information, statisticians resort to suitable models to pool the information across small areas. Most existing discussions have focused on estimating small area means under some models corresponding to imaginary scenarios. They are likely less effective if utilized for estimating small area quantiles. In this paper, we postulate that the small area population distributions have some linear structure with error distributions satisfying a density ratio model. That is, the small area error distributions are all tilted distributions from a common basis. Under this model, we employ empirical likelihood to pool information in samples across all small areas. The resulting approach not only allows us to estimate small area means, but also small area quantiles. We give a comprehensive discussion on this method and provide some preliminary simulation results to illustrate its potential.

Key Words: Empirical best linear unbiased predictors, empirical likelihood, nested error model, population quantile, survey sampling.

1. Introduction

It has been an honour to present our research at the joint statistical meeting this year in celebration of the 75th birthday of Professor J.N.K. Rao. Professor Rao is an iconic researcher in Canada and over the world. He has made tremendous impact to the theory and practice of statistics, particularly in survey sampling. The Rao-Hartley-Cochran method (Rao, Hatley and Cochran, 1962) for unequal probability sampling plan remains the most efficient and practical scheme. The Rao-Wu bootstrap (Rao and Wu, 1988) is an indispensable tool in daily operations of Statistics Canada. His pioneer paper on scale-loading likelihood, Hartley and Rao (1968), precedents the invention of the famous empirical likelihood methodology (Owen 1988). His latest book “Small Area Estimation” (Rao, 2003) is an immediate classic. Reading his papers and listening to his advices have always given us additional insight and motivations.

This paper contains some preliminary results based on our recent research which fits well with the contributions of Professor Rao. We aim to develop empirical likelihood based approaches for small area estimation in this paper. Sample surveys are widely used to obtain information about totals, means and other parameters of finite populations. In many applications, the same information are also desired for subpopulations such as individuals in specific geographic areas or in socio-demographic groups. Estimating finite subpopulation parameters is referred to as small area estimation problem (Rao, 2003). While the geographic areas may not be small, direct information from individual areas are often in severe shortage. Often, the surveys are conducted at national or similarly high levels. The

[*] Department of Statistics, University of British Columbia, jhchen@stat.ubc.ca

[†] School of Finance and Statistics, East China Normal University, ykliu@sfs.ecnu.edu.cn

random nature of the probability sampling can result in no sampling units from many sub-populations of interest. Estimating parameters of these sub-populations with satisfactory precision and evaluating their accuracy pose serious challenges to statisticians.

Due to the scarcity of direct information from small areas, reliable estimates are possible only if some indirection information from rest areas are available and effectively utilized. This leads to a common thread "borrowing strength". At least conceptually, statisticians seek common characteristics that can be accurately estimated based on population level samples. Together with other characteristics learned from other sources, an indirect estimate for the sub-population parameter is obtained. This estimate is then combined "optimally" with the direct estimate if available.

Some pioneer work on small area estimation include Fay and Herriot (1979), Prasad and Rao (1990), Lahiri and Rao (1995). The research in this area has received increasing attention from both public and private sectors (Fay and Herriot,1979; Schaible, 1993; Kriegler and Berk, 2010). There are increasing number of publications related to this topic (Pfeffermann, 2002; Jiang and Lahiri, 2006; Ghosh, Maiti and Roy, 2008; Jiang, Nguyen and Rao, 2010) in recent years.

In this paper, we propose a new model that is suitable not only for mean estimation but also for quantile estimation of small areas. In the next section, we review a popularly used model and motivate the new model. In Section 3, we discuss the inference issue based on the new model. In Section 4, we provide some simulation results. Our preliminary results indicate that the model together aided with the empirical likelihood leads us to a promising new approach for small area estimation. We end the paper with a brief summary and discussion section.

2. Reviewing a commonly used model and introducing a new model

We use the nested-error regression model (NER) by Battese, Harter and Fuller (1988) for illustration now and for comparison later. Consider the situation where the population is made of $m+1$ small areas and n_k sampling units are obtained from the kth area ($k = 0, 1, 2, \ldots, m$). Under this model, the univariate response value and its vector covariates on these sampling units are regarded as independent observations and satisfy

$$y_{kj} = \mathbf{x}_{kj}^{\tau}\beta + v_k + \varepsilon_{kj}, \tag{1}$$

with area-specific random effect $v_k \sim N(0, \sigma_b^2)$ and random error $\varepsilon_{kj} \sim N(0, \sigma^2)$. Under this model, the vector-valued regression coefficient β remains unchanged across all areas, namely it is a common characteristic. Therefore, samples from all areas contain its information and can be pooled to estimate β. Hence, when the overall sample size $\sum n_k$ is large, an estimate $\hat{\beta}$ with satisfactory precision of β can be easily obtained. Suppose the area totals $\mathbf{X}_k$ are known from, say, administrative records. Sensible indirect estimates of the area total of y would be $\hat{\mathbf{Y}}_k = \mathbf{X}_k^{\tau}\hat{\beta}$. Direct estimate of $\mathbf{Y}_k$, if available, can be combined to catch some information about v_k.

In addition to the above model, the model proposed by Fay and Herriot (1979) has received even more attention. Various estimation strategies are proposed based on these models. To our best knowledge, most existing methods focus on estimating small area totals or means. It is curious that the small area median or quantiles are not addressed at all. We suspect that it is not because the small area median and quantiles are less important characteristics, but the commonly used models are un-suitable for constructing quantile estimators. To fill up this gap, we propose a density ratio model as the platform for small area estimation of quantiles. In spite of some obvious differences, the new model shares the spirit of Fay and Herriot (1979) and Battese, Harter and Fuller (1988).

We now use the same notation for response variable and covariates as in (1). Assume that we have a random sample from the target finite population with n_k units from the kth small area, and there are $m+1$ small areas in the population. We postulate that

$$y_{kj} = \mathbf{x}_{kj}^{\tau}\beta_k + \varepsilon_{kj}. \tag{2}$$

Some specifications of this model are as follows. First, we allow a more flexible linear relationship with area specific regression coefficient β_k but forgo the area specific random effect in other models. To avoid excessive number of parameters in this model, we seek a way to link β_k to some auxiliary information. There are many potential choices, but we tentatively populate that

$$\beta_k = \beta + a\bar{\mathbf{X}}_k \tag{3}$$

for some vector β and scalar a, where $\bar{\mathbf{X}}_k$'s are the known area specific means of covariates. Apparently, when $a = 0$, the regression part of our model resembles that of (1).

Second, we regard ε_{kj}, for each k, as a random sample from some distribution $G_k(\cdot)$. We do not impose a parametric form, but postulate a density ratio model (DRM, Anderson, 1979) such that for $k = 1, 2, \ldots, m$,

$$\log\{dG_k(y)/dG_0(y)\} = \boldsymbol{\theta}_k^{\tau}\mathbf{q}(y), \tag{4}$$

for a known vector valued function $\mathbf{q}(y)$ and an area specific tilting parameter $\boldsymbol{\theta}_k$. The baseline distribution $G_0(y)$ is left unspecified. The above model (4) includes Normal, Gamma and many other distribution families as special cases. Currently, we allow one set of parameter for each small area which can be excessive. Instilling an appropriate structure into $\boldsymbol{\theta}_k$ is under investigation but no results are available now.

Equations (2), (3) and (4) together form a platform for our proposed inference on small area quantiles. The key difference of our new model lies in (4). Consider the extreme case when the linear coefficients $\beta_k = 0$ for all k in both (1) and (2). Under model specified by (1) with $\beta = 0$, a sensible quantile estimation for area k will be heavily dependent on the normality assumption imposed on ε_{kj} and the random effect v_k. Under (2)-(4) with $\beta_k = 0$, quantile estimates are linked to G_0 which is nonparametric. A decent nonparametric estimate of G_0, when available, results in decent quantile estimates for all small areas and this approach is likely robust against some degree of model mis-specifications.

3. Inference under DRM

Let us first consider an artificial situation where the values of all regression coefficients are known. In this case, we are provided $m+1$ independent random samples $\{\varepsilon_{kj}\}$, $j = 1, 2, \ldots, n_k$ from a DRM. These observations are the basis for inference on G_k. The inference method comes from literature and it will be part of our method for small area estimation.

Recently, Qin and Zhang (1997), Qin (1998), Zhang (1997) and others find that the DRM can be viewed as the commonly-used logistic regression model. Empirical likelihood (EL, Owen, 1988) is found handy for statistical inference under DRM. They find the resulting maximum empirical likelihood estimators are consistent and asymptotically normal under mild conditions. Fokianos et al. (2001) develop a new one-way analysis-of-variance method based on DRM. Zhang (2000) investigates the asymptotical normality of the EL quantile estimator when $m+1 = 2$. The case for general m is not different in principle but the investigation becomes substantially more technical. Chen and Liu (2012) succeed at proving that for general $m+1$ the EL quantile estimator admits the Bahadur representation and find that the EL quantile estimators are more efficient than the empirical quantiles that only use direct information.

3.1 Empirical likelihood estimate of G_k

Following the idea of Owen (2001), we confine the form of the candidate G_0 to $G_0(y) = \sum_{k,j} p_{kj} I(\epsilon_{kj} \leq y)$ where $I(\cdot)$ is the indicator function. Under this setting, we have $p_{kj} = dG_0(\varepsilon_{kj})$ and

$$dG_k(\varepsilon_{kj}) = \exp\{\boldsymbol{\theta}_k^\tau \mathbf{q}(\varepsilon_{kj})\} dG_0(\varepsilon_{kj}).$$

Then the EL is defined as

$$\begin{aligned} L_n(G_0, G_1, \ldots, G_m) &= \prod_{k=0}^{m} \prod_{j=1}^{n_k} dG_k(\varepsilon_{kj}) \\ &= [\prod_{k=0}^{m} \prod_{j=1}^{n_k} p_{kj}] \cdot \exp[\sum_{k=1}^{m} \sum_{j=1}^{n_k} \{\boldsymbol{\theta}_k^\tau \mathbf{q}(\varepsilon_{kj})\}]. \end{aligned}$$

Note that $G_1, \ldots, G_m$ are fully determined by $\boldsymbol{\theta}^\tau = (\boldsymbol{\theta}_1^\tau, \ldots, \boldsymbol{\theta}_m^\tau)$ and G_0. We may hence write the empirical log-likelihood as

$$\ell_n(\boldsymbol{\theta}, G_0) = \sum_{k=0}^{m} \sum_{j=1}^{n_k} \log(p_{kj}) + \sum_{k=1}^{m} \sum_{j=1}^{n_k} \{\boldsymbol{\theta}_k^\tau \mathbf{q}(\varepsilon_{kj})\}$$

where the parameter $\boldsymbol{\theta}$ and p_{kj}'s satisfy $p_{kj} \geq 0$ and that for all $r = 0, 1, \ldots, m$,

$$\sum_{k=0}^{m} \sum_{j=1}^{n_k} p_{kj} \exp\{\boldsymbol{\theta}_r^\tau \mathbf{q}(\varepsilon_{kj})\} = 1. \tag{5}$$

We remark here that we have used convention $\boldsymbol{\theta}_0 = 0$ for simpler presentation. It is now a routine to reveal that maximizing $\ell_n(\boldsymbol{\theta}, G_0)$ with respect to G_0 under the constraints (5) results in fitted probabilities

$$\hat{p}_{kj} = n^{-1}\{1 + \sum_{t=1}^{m} \nu_t [\exp\{\boldsymbol{\theta}_t^\tau \mathbf{q}(\varepsilon_{kj})\} - 1]\}^{-1} \tag{6}$$

and the profile EL

$$\tilde{\ell}_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m} \sum_{j=1}^{n_k} \log\{1 + \sum_{t=1}^{m} \nu_t [\exp\{\boldsymbol{\theta}_t^\tau \mathbf{q}(\varepsilon_{kj})\} - 1]\} + \sum_{k=1}^{m} \sum_{j=1}^{n_k} \{\boldsymbol{\theta}_k^\tau \mathbf{q}(\varepsilon_{kj})\} \tag{7}$$

with $(\nu_1, \nu_2, ..., \nu_m)$ being the solution to

$$\sum_{k=0}^{m} \sum_{j=1}^{n_k} \frac{\exp\{\boldsymbol{\theta}_r^\tau \mathbf{q}(\varepsilon_{kj})\} - 1}{1 + \sum_{t=1}^{m} \nu_t [\exp\{\boldsymbol{\theta}_t^\tau \mathbf{q}(\varepsilon_{kj})\} - 1]} = 0$$

for $r = 1, \ldots, m$. The stationary points of $\tilde{\ell}_n(\boldsymbol{\theta})$ coincides with those of a dual form of the empirical log-likelihood function (Kezioua and Leoni-Aubina, 2008)

$$\ell_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m} \sum_{j=1}^{n_k} \log[\rho_0 + \sum_{t=1}^{m} \rho_t \exp\{\boldsymbol{\theta}_t^\tau \mathbf{q}(\varepsilon_{kj})\}] + \sum_{k=1}^{m} \sum_{j=1}^{n_k} \boldsymbol{\theta}_k^\tau \mathbf{q}(\varepsilon_{kj}), \tag{8}$$

with $\rho_r = n_r/n$, $r = 0, 1, \ldots, m$ and $n = \sum_{k=0}^{m} n_k$.

For the purpose of point estimation, it is simpler to work with $\ell_n(\boldsymbol{\theta})$ which is convex and free from constraints. Once the values of ε_{kj} are provided, it is relatively simple to find its maximum point, which serves as the maximum EL estimates of $\boldsymbol{\theta}$. They are then used to compute the fitted values defined by (6) with ν_k replaced by ρ_k. We then subsequently obtain estimator $\hat{G}_k$ and other parameters of interest via invariance principle.

There is only one issue not addressed in this part of the inference. We do not have observed values of ε_{kj}. However, this difficulty can be easily resolved. A natural solution is to replace them by residuals obtained from fitting (2). This is the topic of the next subsection.

3.2 Fitting linear model (2)

Given $(y_{kj}, \mathbf{x}_{kj})$ for $k = 0, 1, \ldots, m$ and $j = 1, \ldots, n_k$, we may estimate (β, a) in (3) through least sum of squares. That is, let

$$(\hat{\beta}, \hat{a}) = \arg\min_{\beta, a} \sum_{k,j} \{y_{kj} - \mathbf{x}_{kj}^{\tau}(\beta + a\bar{\mathbf{X}}_k)\}^2. \tag{9}$$

The residuals ($k = 0, 1, \ldots, m;\ k = 1, 2, \ldots, n_k$) are hence given by

$$\hat{\varepsilon}_{kj} = y_{kj} - \mathbf{x}_{kj}^{\tau}(\hat{\beta} + \hat{a}\bar{\mathbf{X}}_k).$$

Substitute $\hat{\varepsilon}_{kj}$ into the dual likelihood (8), we get the maximum EL estimator $\hat{\boldsymbol{\theta}}$. Subsequently, we have

$$\hat{p}_{kj} = n^{-1}\{1 + \sum_{t=1}^{m} \rho_t[\exp\{\hat{\boldsymbol{\theta}}_t^{\tau}\mathbf{q}(\hat{\varepsilon}_{kj})\} - 1]\}^{-1}$$

and

$$\hat{G}_r(\varepsilon) = \sum_{k=0}^{m}\sum_{j=1}^{n_k} \hat{p}_{kj}\exp\{\hat{\boldsymbol{\theta}}_r^{\tau}\mathbf{q}(\hat{\varepsilon}_{kj})\} I(\hat{\varepsilon}_{kj} < \varepsilon). \tag{10}$$

The availability of $\hat{G}_k$ provides a new tool for small area estimation. If the small area mean $\bar{y}_k$ is the parameter of interest, we readily estimate it by

$$\hat{\bar{y}}_k = \bar{\mathbf{X}}_k^{\tau}(\hat{\beta} + \hat{a}\bar{\mathbf{X}}_k) + \int \varepsilon d\hat{G}_k(\varepsilon). \tag{11}$$

If the size of $\bar{\mathbf{X}}_k$ is not available, we may also use $\bar{\mathbf{X}}_k$. The small area distribution of y can be comprehensively estimated as

$$\hat{F}_k(y) = n_k^{-1}\sum \hat{G}_k(y - \mathbf{x}_{kj}^{\tau}\{\hat{\beta} + \hat{a}\bar{\mathbf{X}}_k\}). \tag{12}$$

We may hence estimate the small area quantiles by those of $\hat{F}_k(y)$.

We have yet to provide some results on the large sample properties of the above proposed method. In a related paper, Chen and Liu (2012) study the large sample properties of $\hat{G}_k$ in the situation where $m+1$ independent random samples are available. Theoretical properties of $\hat{F}_k$ cannot be directly derived from these of $\hat{G}_k$. We will leave it as a future topic. In the next section, we provide some simulation results.

4. Simulation study

In this section, we provide a small scale numerical simulation results to investigate the performance of the proposed estimators (11) and (12) for small area means and quantiles. For the sake of comparison, our simulation has included representative estimators of small area means and quantiles designed in the literature under two commonly used models. The chosen estimators for means are the so-called empirical best linear unbiased prediction (EBLUP). There are no existing small area quantile estimators. Our simulation explored the performance of two artificial made-at-request estimators based on these two models. The first model is the nested-error regression model (1) introduced earlier in which the variances of the error terms are assumed equal over all small areas.

4.1 Some existing small area estimators

Suppose we have n_k observations of $(y_{ki}, \mathbf{x}_{ki})$ from small area k according to model (1). Let $\tilde{\beta}$ be the maximum likelihood estimator (MLE) of β and $\bar{y}_k$ be the sample mean of y of sampling units obtained in area k. Assume the small area population mean $\bar{\mathbf{X}}_k$ of covariate **x** is known. We then have two predictions of the small area means from two angles. One is based on the linear model given by $\bar{\mathbf{X}}_k^\tau \tilde{\beta}$ which borrows strength from the samples over all areas through $\tilde{\beta}$. The other is the direction predictor, area sample mean $\bar{y}_k$. Under the model assumption and if the values of σ_b^2 and σ^2 are given, then the best linear combination of these two predictors is given by

$$\tilde{\theta}_k^* = \bar{\mathbf{X}}_k^\tau \tilde{\beta} + \frac{n_k \sigma_b^2}{\sigma^2 + n_k \sigma_b^2} (\bar{y}_k - \bar{\mathbf{X}}_k^\tau \tilde{\beta}), \tag{13}$$

which is referred to as the best linear unbiased estimator (BLUP). This above scenario, knowing the values of σ_b^2 and σ^2, is implausible. Instead, it motivates another predictor. Let $\tilde{\sigma}^2$, $\tilde{\sigma}_b^2$ and $\tilde{\beta}$ be maximum likelihood estimators of σ^2, σ_b^2 and β under NER assumption. It then leads to the following predictor which is often referred to as EBLUP

$$\tilde{\theta}_k = \bar{\mathbf{X}}_k^\tau \tilde{\beta} + \frac{n_k \tilde{\sigma}_b^2}{\tilde{\sigma}^2 + n_k \tilde{\sigma}_b^2} (\bar{y}_k - \bar{\mathbf{X}}_k^\tau \tilde{\beta}). \tag{14}$$

We remark that the above EBLUP is not exactly linearly unbiased.

Most recently, Jiang and Nguyen (2012) investigate the small area estimation problem under the heteroscadastic NER (HNER) model. The HNER model assumes the same linear structure as (1), but postulates area specific variance of ε_{kj}, σ_k^2 and area specific variance of v_k, $\gamma \sigma_k^2$. Apparently, with unequal error variances over the small areas, the EBLUP (14) potentially loses its presumed optimality. Let $\breve{\beta}$, $\breve{\gamma}$ and $\breve{\sigma}_k^2$ be the MLEs of β, γ and σ_k^2 under the HNER. Jiang and Nguyen (2012) find the following estimator

$$\breve{\theta}_k = \bar{\mathbf{X}}_k^\tau \breve{\beta} + \frac{n_k \breve{\gamma}}{1 + n_k \gamma} (\bar{y}_k - \bar{\mathbf{X}}_k^\tau \breve{\beta}) \tag{15}$$

is more accurate under their new model assumption.

We have not seen any discussions on the estimation of small area quantiles under these two models. In the spirit of (12), we can easily give two made-at-request distribution estimators based on the NER and HNER models. They are given respectively by

$$\tilde{F}_k(y) = \frac{1}{n_k} \sum_{j=1}^{n_k} \Phi\Big(\frac{y - \mathbf{x}_{kj}^\tau \tilde{\beta}}{\sqrt{\tilde{\sigma}^2 + \tilde{\sigma}_b^2}} \Big), \tag{16}$$

$$\breve{F}_k(y) = \frac{1}{n_k} \sum_{j=1}^{n_k} \Phi\Big(\frac{y - \mathbf{x}_{kj}^\tau \breve{\beta}}{\sqrt{(1 + \breve{\gamma}) \breve{\sigma}_k^2}} \Big) \tag{17}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. The corresponding quantiles of these distributions serve as natural respective small area quantile estimates.

In this simulation, we examine the small area mean and quantile estimations under three models with the corresponding estimators/predictors. For quantile estimation, we take the 5%, 25%, 50%, 75% and 95% small area quantiles as parameters of interest.

4.2 Simulation settings

We simulated data from five models with the number of small areas $m + 1 = 10$. They do not necessarily conform to any of three models we introduced earlier. The choice is based on the notion that "all models are wrong". In each case, we follow the common practice to make every small area sample size $n_k = 10$ and 20, rather than having them randomly decided by a higher level sampling plan. For each simulated data set, we compute estimates/predictions (14), (15) and (11). The function $\mathbf{q}(y)$ in (11) is set to $(1, y)^\tau$, which has the simplest form among all nondegenerate choices. The process is repeated independently $N = 2000$ times. Let $\hat{\theta}_k^{(j)}$ denote a generic predictor of small area mean $\bar{y}_k$ in the jth repetition. We report the average mean squared error (AMSE) defined as follows:

$$\text{AMSE} = \{N(m+1)\}^{-1} \sum_{k=0}^{m} \sum_{j=1}^{N} (\hat{\theta}_k^{(j)} - \bar{y}_k)^2.$$

We also compute the AMSE in the same way for small area quantile estimators.

Next, we specify five models used in this simulation. In these models, the covariates $\mathbf{x}$ and response value y are linked as follows,

$$\begin{aligned}
y_{kj} &= \mathbf{x}_{kj}^\tau \beta + v_k + \varepsilon_{kj} \qquad &\text{(A)}\\
y_{kj} &= \mathbf{x}_{kj}^\tau \beta + \mathbf{x}_{kj}^\tau \bar{\mathbf{X}}_k/100 + v_k + \varepsilon_{kj} \qquad &\text{(B)}\\
y_{kj} &= \mathbf{x}_{kj}^\tau \beta + \sin(\mathbf{x}_{kj}^\tau \mathbf{x}_{kj}) + v_k + \varepsilon_{kj} \qquad &\text{(C)}\\
y_{kj} &= \sin(\mathbf{x}_{kj}^\tau \beta) + \sin(\mathbf{x}_{kj}^\tau \mathbf{x}_{kj}) + v_k + \varepsilon_{kj} \qquad &\text{(D)}\\
y_{kj} &= \ln(1 + 4|\mathbf{x}_{kj}^\tau \beta|) + \sin(\mathbf{x}_{kj}^\tau \mathbf{x}_{kj}) + v_k + \varepsilon_{kj} \qquad &\text{(E)}
\end{aligned}$$

Model A is specified the same as the heteroscadastic model given by Jiang and Nguyen (2012). They put regression coefficient $\beta = (1, -1)^\tau$ and covariates $\mathbf{x}_{kj}$ equal either $(1, 0)^\tau$ or $(1, 0)^\tau$. In addition, they considered three variance specifications under this model, (I) $\sigma = \sigma_k = 0.2$ for all small areas; (II) half of small areas have $\sigma = \sigma_k = 0.2$ and the rest half of them have $\sigma = \sigma_k = 0.8$; (III) half number of small areas have their σ_k generated from uniform distribution $U[0.2, 0.3]$, and the rest σ_k from $U[0.8, 0.9]$.

Design matrix structure in model (A) does not allow us to fit the full density ratio model with structure specified by (2) and (3). In the simulation, we choose to use a simplified density ratio model in which $a = 0$.

For Models (B)-(E), we use three component covariate $\mathbf{x}_{kj}$ with its first component being one. We generated the other two components of $\mathbf{x}_{kj}$ from a gamma distribution $\Gamma(\lambda_k, 1)$, with shape parameter $\lambda_k = 2 + \sin(e^k)$ and scale parameter 1. We set $\beta = (0, 1, -1)^\tau$ in Model (B) and $\beta = (0, 1, 1)^\tau$ in the rest three models. We generated random effects v_k and errors ε_{kj} according to the following three settings. (I) Both v_k and ε_{kj} have standard normal distribution; (II) The random effect v_k is normally distributed with mean 0 and standard variance σ_k generated from uniform distribution $U(0.4, 0.8)$. The error term ε_{kj} is normal with mean 0 and standard variance ζ_k generated from $U(0.8, 1.4)$; (III) The random effect v_k has centralized $\Gamma(\xi_k, 1)$ distribution with ξ_k generated from $U(1, 2)$. The error term ε_{kj} has centralized Weibull distribution $W(\zeta_k, 1)$ with shape parameter ζ_k generated from $U(0.8, 1.4)$.

Here are considerations behind these models. Model (A) is chosen to match NER and HNER closely. Under the error distribution specification (I), NER model assumptions are fully satisfied. Otherwise, HNER model assumptions are practically satisfied. Thus, the EBLUP based on NER or HNER for estimating small area means are expected to work well. We hope that the DRM based EL estimator is not far behind.

Model (B) matches our DRM specification best although the error distributions under (II) and (III) do not perfectly fit the density ratio model specification. We hope to see whether the EL estimator will outperform here. Model (C) has some non-linear component instilled. The last two models loss all linear relationships between the response variable and covariates. We are curious on how these estimators fetch when their model assumptions are completely violated.

4.3 Simulation results

The simulation results on average mean square errors of three estimators for small area mean are presented in Table 1. Column EL corresponds to empirical likelihood estimator based on density ratio model assumptions. Columns NER and HNER correspond to EBLUP under NER and HNER model assumptions.

Let us first examine the simulation result under Model (A). As expected, the EBLUP under both NER and HNER outperform the DRM based EL estimator. See the AMSE values in columns EL, NER and HNER. Under variance structure scenarios II and III, there are no visible difference in AMSEs in NER and HNER columns. The DRM based EL estimator remains behind but not by a big margin.

Under models (B), there are still no visible difference in AMSEs in NER and HNER columns. The DRM based EL estimator has lower AMSEs as expected under scenarios I and II. The performance comparison is reversed under scenario III. We take the excuse that the EL estimator is not specifically designed to perform for small area mean. We will see that it works nicely for small area medians.

Table 1: AMSE of three small area mean estimators.

		Scenario I			Scenario II			Scenario III		
n_k	Model	EL	NER	NERH	EL	NER	NERH	EL	NER	NERH
10	(A)	0.006	0.003	0.003	0.061	0.038	0.033	0.144	0.112	0.110
	(B)	0.159	0.211	0.212	0.126	0.149	0.151	0.247	0.211	0.215
	(C)	0.173	0.324	0.337	0.145	0.212	0.217	0.240	0.481	0.491
	(D)	0.207	0.241	0.246	0.201	0.389	0.395	0.265	1.009	1.035
	(E)	0.182	0.842	0.869	0.145	1.294	1.329	0.259	4.298	4.450
20	(A)	0.006	0.002	0.002	0.049	0.019	0.018	0.110	0.056	0.056
	(B)	0.129	0.132	0.133	0.083	0.088	0.088	0.207	0.151	0.152
	(C)	0.116	0.331	0.338	0.086	0.156	0.157	0.184	0.411	0.414
	(D)	0.128	0.145	0.145	0.130	0.366	0.362	0.192	0.848	0.846
	(E)	0.125	0.998	1.012	0.086	2.134	2.159	0.202	4.765	4.834

The linear structure is either weakened or totally destroyed in the remaining models. All three estimators are used under wrong model assumptions. The simulation results suggest that the DRM based EL small area mean estimator is a clear winner.

When the sample size for each small area increases from 10 to 20, the DRM based EL estimator has in general a 20%-30% gain in AMSE; while it is surprising that the AMSEs of the EBLUPs under NER and HNER increase under Model (E), although they decrease in the rest models.

In conclusion, even though our new method, the DRM based EL approach does not target small area mean. It has a lot of potential when there is only a weak or no linear relationship between the response variable and the covariates.

We now turn to small area quantile estimations. We generated data from the same models in the same way as for the small area mean estimations. For each data set generated for small area mean estimation, we computed the quantile estimates defined by (12), (16) and (17). We subsequently computed the AMSE of five quantiles estimates. The results are given in Tables 2 and 3 in columns EL, NER and HNER.

With four exceptions (Scenario (II) of Models (A) and (B) at both $n_k = 10$ and 20), the DRM based EL small area quantile estimator has uniformly and substantially lower AMSEs than the other two estimators. As the sample size n_k increases, the EL small area quantile estimator benefits most with reduced AMSEs broadly. This is not the case for two EBLUPs when data are generated from Models (D) and (E). Because the estimators (16) and (17) based on NER and HNER models are not specifically designed for quantile estimation, having a superior performance compared to these two estimators is not a solid evidence for the excellence of EL. However, we notice the variances are often reduced by a factor of 2 and 3 or even 4 for medians. They at least provide a strong support to the new approach.

5. Conclusions and discussions

In this paper, we point out that the small area estimation of population quantiles are not discussed in the literature. We further suggest that the currently used models for small area estimation are not suitable as platforms for addressing this issue. Motivated by our recent work on density ratio models, we propose to use them for the purpose of small area quantile estimation. We develop an empirical likelihood based estimation method and study its properties through simulation. The outcomes are encouraging although the scale of our study is limited. There are a lot of issues to be addressed. They include refining the model structures, more elaborative combination between linear component in the model assumption, and the density ratio structure in error distributions. We hope to present a more complete report in the near future.

REFERENCES

Anderson, J. A. (1979), Multivariate logistic compounds," *Biometrika*, 66, 17-26.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 80, 28-36.

Chen, J. and Liu, Y. (2012), "Quantile and quantile-function estimations under density ratio model," Manuscript.

Fay, R. E. and Herriot, R. A. (1979), "Estimates of income for small places: An application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269-277.

Fokianos, K., Kedem, B., Qin, J. and Short, D. A. (2001), "A semiparametric approach to the one-way layout," *Technometrics*, 43, 56-65.

Ghosh, M., Maiti, T. and Roy, A. (2008), "Influence functions and robust Bayes and empirical Bayes small area estimation," *Biometrika*, 95, 573-585.

Hartley, H. O. and Rao, J. N. K. (1968), "A new estimation theory for sample surveys," *Biometrika*, 55, 547-57.

Jiang, J. and Lahiri P. S. (2006), "Estimation of finite population domain means: a model-assisted empirical best prediction approach," *Journal of the American Statistical Association*, 101 301-311.

Jiang, J. and Nguyen, T. (2012), "Small area estimation via heteroscedastic nested-error regression," *Canada Journal of Statistics*, 40, 588–603.

Jiang, J., Nguyen, T. and Rao, J. S. (2010), "Fence method for nonparametric small area estimation," *Survey Methodology*, 36, 3-11.

Kezioua, A., and Leoni-Aubina, S. (2008), "On empirical likelihood for semiparametric two-sample density ratio models," *Journal of Statistical Planning and Inference*, 138, 915-928.

Kriegler, B. and Berk, R. (2010), "Small area estimation of the homeless in Los Angeles: An application of cost-sensitive stochastic gradient boosting," *The Annals of Applied Statistics*, 4, 1234-1255.

Table 2: AMSE of three estimators of the αth small area quantiles with $n_k = 10$.

		Scenario I			Scenario II			Scenario III		
Model	α	EL	NER	NERH	EL	NER	NERH	EL	NER	NERH
(A)	5%	0.015	0.031	0.039	0.527	0.442	0.209	0.473	0.794	0.918
	25%	0.008	0.026	0.029	0.148	0.164	0.120	0.197	0.596	0.639
	50%	0.007	0.025	0.025	0.062	0.095	0.095	0.155	0.534	0.534
	75%	0.007	0.026	0.025	0.052	0.128	0.119	0.160	0.556	0.565
	95%	0.008	0.033	0.034	0.144	0.270	0.213	0.220	0.740	0.832
(B)	5%	1.700	1.998	2.959	1.574	1.657	2.612	1.960	2.363	3.437
	25%	0.631	1.071	1.297	0.582	0.699	0.904	0.862	1.460	1.775
	50%	0.503	0.946	1.027	0.475	0.605	0.703	0.623	1.267	1.500
	75%	0.638	1.124	1.299	0.604	0.727	0.951	0.737	1.477	1.453
	95%	1.896	2.260	3.095	1.847	1.863	2.907	2.068	2.362	3.169
(C)	5%	0.618	0.881	1.820	0.504	0.655	1.507	1.042	1.702	2.554
	25%	0.451	0.715	0.874	0.441	0.591	0.665	0.552	1.223	1.476
	50%	0.512	0.738	0.994	0.501	0.603	0.795	0.598	1.327	1.624
	75%	0.834	0.990	1.260	0.783	0.837	1.242	0.889	1.562	1.635
	95%	2.796	2.473	3.667	2.680	2.409	3.942	2.841	2.881	4.050
(D)	5%	0.330	0.497	3.747	0.268	0.334	3.318	0.741	1.416	4.203
	25%	0.219	0.447	1.154	0.218	0.328	0.807	0.355	1.010	1.956
	50%	0.224	0.467	0.489	0.245	0.358	0.352	0.266	1.031	1.083
	75%	0.236	0.481	0.960	0.255	0.382	1.213	0.287	1.144	1.078
	95%	0.307	0.570	3.776	0.315	0.431	4.989	0.579	1.298	3.562
(E)	5%	0.318	0.489	3.250	0.222	0.311	2.899	0.890	1.616	3.455
	25%	0.213	0.422	0.867	0.167	0.280	0.561	0.429	1.037	1.711
	50%	0.200	0.404	0.417	0.175	0.266	0.277	0.289	0.977	0.986
	75%	0.209	0.420	0.737	0.193	0.275	0.897	0.291	1.112	0.855
	95%	0.260	0.463	2.205	0.266	0.296	2.869	0.581	1.191	2.231

Table 3: AMSE of three estimators of the αth small area quantiles with $n_k = 10$.

		Scenario I			Scenario II			Scenario III		
Model	α	EL	NER	NERH	EL	NER	NERH	EL	NER	NERH
(A)	5%	0.015	0.030	0.035	0.501	0.434	0.188	0.444	0.774	0.830
	25%	0.008	0.026	0.027	0.138	0.159	0.109	0.163	0.582	0.597
	50%	0.006	0.024	0.024	0.050	0.091	0.091	0.112	0.522	0.522
	75%	0.005	0.025	0.025	0.034	0.124	0.112	0.109	0.545	0.555
	95%	0.007	0.032	0.033	0.113	0.304	0.187	0.152	0.728	0.799
(B)	5%	1.058	1.452	2.311	0.934	1.039	1.987	1.279	1.859	2.896
	25%	0.368	0.840	1.064	0.310	0.462	0.675	0.608	1.215	1.558
	50%	0.287	0.761	0.796	0.249	0.419	0.461	0.407	1.078	1.178
	75%	0.352	0.880	0.952	0.305	0.478	0.615	0.437	1.244	1.246
	95%	1.074	1.611	2.168	0.994	1.179	2.134	1.249	1.812	2.286
(C)	5%	0.463	0.775	1.812	0.330	0.516	1.470	0.957	1.702	2.552
	25%	0.260	0.573	0.761	0.249	0.445	0.509	0.391	1.074	1.350
	50%	0.273	0.557	0.662	0.263	0.418	0.510	0.377	1.136	1.229
	75%	0.444	0.684	0.861	0.396	0.522	0.745	0.517	1.269	1.325
	95%	1.474	1.535	2.560	1.385	1.364	3.084	1.592	2.018	2.940
(D)	5%	0.264	0.483	4.236	0.190	0.293	3.673	0.763	1.461	4.511
	25%	0.144	0.447	1.231	0.137	0.313	0.827	0.322	1.023	1.992
	50%	0.127	0.474	0.490	0.150	0.348	0.345	0.197	1.041	1.086
	75%	0.144	0.482	1.012	0.148	0.365	1.278	0.193	1.145	1.120
	95%	0.166	0.558	4.112	0.177	0.382	5.372	0.287	1.291	3.760
(E)	5%	0.276	0.476	3.263	0.169	0.286	2.934	0.891	1.662	3.476
	25%	0.162	0.411	0.865	0.109	0.262	0.556	0.415	1.028	1.689
	50%	0.133	0.392	0.395	0.102	0.249	0.260	0.239	0.964	0.953
	75%	0.130	0.405	0.683	0.107	0.253	0.824	0.217	1.099	0.839
	95%	0.141	0.449	2.340	0.140	0.264	3.093	0.285	1.168	2.370

Lahiri, P.S. and RAO, J. N. K. (1995), "Robust estimation of mean squared error of small area estimators," *Journal of the American Statistical Association*, 90, 758-766.

Owen, A. B. (1988), "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

Owen, A. B. (2001), *Empirical Likelihood.* Chapman and Hall/CRC, New York.

Pfeffermann, D. (2002), "Small Area Estimation-New Developments and Directions," *International Statistical Review*, 70, 125-143.

Prasad, N. G. N. and Rao, J. N. K. (1990), "The estimation of mean squared errors of small area estimators," *Journal of the American Statistical Association*, 85, 163-171.

Qin, J. (1998), "Inferences for case-control and semiparametric two-sample density ratio models," *Biometrika*, 85, 619-630.

Qin, J. and Zhang, B. (1997), "A goodness-of-fit test for logistic regression models based on case-control data," *Biometrika*, 84, 609-618.

Rao, J. N. K. (2003), *Small Area Estimation.* Wiley, New York.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962), "On a simple procedure of unequal probability sampling without replacement," *Journal of the Royal Statistical Society*, Ser. B, 24, 482-491.

Rao, J. N. K. and Wu, C.-F. J. (1988), "Resampling inference with complex survey data," *Journal of the American Statistical Association*, 83, 231-241.

Schaible, W. L. (1993), "Use of small area estimators in U.S. federal programs. In Small Area Statistics and Survey Designs," ed. by G. Kalton, J. Kordos, and R. Platek, 1, 95-114. Central Statistical Office, Warsaw.

Zhang, B. (1997), "Assessing goodness-of-fit of generalized logit models based on case-control data. *Journal of Multivariate Analysis*, 82, 17–38.

Zhang, B. (2000), "Quantile estimation under a two-sample semi-parametric model," *Bernoulli*, 6, 491–511.