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

Survey Methodology

Small area quantile estimation via spline regression and empirical likelihood

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canada

by Zhanshou Chen, Jiahua Chen and Qiong Zhang

Release date: May 7, 2019

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Vol. 45, No. 1, pp. 81-99
Statistics Canada, Catalogue No. 12-001-X

Small area quantile estimation via spline regression and empirical likelihood

Zhanshou Chen, Jiahua Chen and Qiong Zhang[1]

Abstract

This paper studies small area quantile estimation under a unit level non-parametric nested-error regression model. We assume the small area specific error distributions satisfy a semi-parametric density ratio model. We fit the non-parametric model via the penalized spline regression method of Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008). Empirical likelihood is then applied to estimate the parameters in the density ratio model based on the residuals. This leads to natural area-specific estimates of error distributions. A kernel method is then applied to obtain smoothed error distribution estimates. These estimates are then used for quantile estimation in two situations: one is where we only have knowledge of covariate power means at the population level, the other is where we have covariate values of all sample units in the population. Simulation experiments indicate that the proposed methods for small area quantiles estimation work well for quantiles around the median in the first situation, and for a broad range of the quantiles in the second situation. A bootstrap mean square error estimator of the proposed estimators is also investigated. An empirical example based on Canadian income data is included.

Key Words: Small area quantile; Penalized spline; Empirical likelihood; Density ratio model; Nested-error regression model.

1 Introduction

Sample surveys are widely used to obtain information about totals, means, medians and other quantities of finite populations. Likewise, similar information on sub-populations such as individuals in specific areas and socio-demographic groups are also of interest. Often, a survey is designed to collect information of interest at the population level but leads to insufficient direct information on sub-populations. Because of this, estimating sub-population parameters with satisfactory precision and evaluating their accuracy pose serious challenges to statisticians. Statisticians must resort to suitable models to pool the information across small areas in order to properly estimate parameters for small areas when only small samples or no samples in these areas are available from the sample survey.

Research on small area estimation has received increased attention from both public and private sectors. As historical remarks, we refer to Fay and Herriot (1979), Battese, Harter and Fuller (1988), Prasad and Rao (1990), and Lahiri and Rao (1995) among many others. For a general review of the developments in small area estimation, we refer to Pfeffermann (2002) and Pfeffermann (2013) and the books of Rao (2003) and Rao and Molina (2015). See also Jiang and Lahiri (2006a), Jiang and Lahiri (2006b) and Jiang (2010) for recent publications.

Compared to quantiles, there are relatively more research activities on estimating small area means. Studies on small area quantile estimation are gaining ground. The M-quantile approach of Chambers and Tzavidis (2006) has achieved substantial success. This approach uses the M-quantile approach to

[1] Zhanshou Chen, School of Mathematics and Statistics, Qinghai Normal University, Xining 810008, P.R. China. E-mail: chenzhanshou@126.com; Jiahua Chen and Qiong Zhang, Department of Statistics, University of British Columbia, Vancouver, BC, Canada.

characterize the conditional distributions of the response variable y given covariates $\mathbf{x}$. This information is then used to predict unobserved response values based on which the small area population distributions are estimated. Small area quantile estimation is a natural and welcome side-benefit. See Tzavidis and Chambers (2005), Pratesi, Ranalli and Salvati (2008), Tzavidis, Salvati and Pratesi (2008), and Salvati, Tzavidis and Pratesi (2012) for these developments.

Another approach for small area quantile estimation is proposed by Molina (2010). Let s and r be the sets of sampled and non-sampled units in a survey and y_s and y_r be vectors of corresponding response values. Under a parametric assumption on the joint distribution of y_s and y_r (or the transformed responses) they proposed to work out the conditional distribution of y_r given y_s (and other information). After having the joint distribution and therefore the conditional distribution properly estimated, they suggested sampling from the estimated conditional distribution to create an artificial but complete population with unobserved y_r filled up. The population distribution is estimated based on the completed population. This approach works well for estimating small area means and quantiles. Other methods we are aware of include Tzavidis, Marchetti and Chambers (2010), Chaudhuri and Ghosh (2011) and Chen and Liu (2018). Tzavidis et al. (2010) proposed a general framework for robust small area estimation, based on representing a small area estimator as a function of a predictor of this small area cumulative distribution function. Chaudhuri and Ghosh (2011) proposed an empirical likelihood based Bayesian method. Chen and Liu (2018) proposed an approach for populations admitting a nested-error linear regression model combined with error distributions satisfying a semi-parametric density ratio model (DRM). Simulations indicate that the DRM-based method stands out when the error distributions are skewed.

In this paper, we are interested in the situation where the regression function is not linear, although the nested-error regression model remains appropriate similar to Opsomer et al. (2008). Clearly, methods derived under linear models may lead to substantial bias if the linearity assumption is violated. To reduce the potential risk of serious bias, Opsomer et al. (2008) proposed an Empirical Best Linear Unbiased Prediction (EBLUP) for the small area means under a non-parametric regression model via penalized splines (P-splines); Jiang, Ngueyen and Rao (2010) developed an adaptive fence approach employing a non-parametric model selection technique; Sperlich and José Lombardía (2010) used the local polynomial inference method in the context of small area estimation; Rao, Sinha and Dumitrescu (2014) proposed a robust EBLUP under a P-splines approximated mixed model; Torabi and Shokoohi (2015) proposed a unified analysis of both discrete and continuous responses under P-spline regression models.

We follow their lead and extend their results to allow non-normal error distributions in the nested-error non-parametric regression model. More specifically, we assume the nested-error non-parametric regression model but relax the small area error distribution assumption from normal to a flexible semi-parametric DRM. We use the P-splines regression approach of Opsomer et al. (2008) to fit the nonlinear regression. Empirical likelihood is then applied to estimate the parameters in the DRM based on the residuals. This leads to natural area specific error distribution estimation. A kernel method is then applied to obtain smoothed estimates of error distributions and small area quantiles. We construct quantile estimates in two situations: one is where we have knowledge of only covariate power means at the population level, the other is where we have covariate values of all sample units in the population. Our approach should inherit the merits of working under a non-parametric regression model, and gain from avoiding a parametric error distribution assumption. The resulting small area quantile estimates are hence more robust. Simulations indicate that when the regression function is approximately linear, the performance of the proposed approach is competitive. The proposed approach outperforms when the regression relationship is quadratic or exponential.

The rest of the paper is organized as follows. Section 2 introduces the model and assumptions. Section 3 presents the proposed approach. Section 4 proposes a bootstrap procedure for estimating mean squared errors. In Section 5, we use Monte Carlo methods to evaluate the performance of the proposed method and compare it with some existing methods. An application example is reported in Section 6. Section 7 contains some concluding remarks.

2 Model and assumptions

Consider a finite population containing $N = \sum_{i=0}^{m} N_i$ sample units partitioned into $m+1$ small areas $\{(x_{ij}, y_{ij}): j = 1, 2, \ldots, N_i\}, i = 0, 1, \ldots, m.$ Consider a nested-error non-parametric regression model with one covariate:

$$y_{ij} = m_0(x_{ij}) + v_i + \varepsilon_{ij}, \tag{2.1}$$

where x_{ij} is an auxiliary variable, v_i denotes an area-specific random effect and ε_{ij} are random errors. The regression function $m_0(\cdot)$ is unspecified, but can be approximated sufficiently well by a spline function

$$m_0(x; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p. \tag{2.2}$$

Here p is the degree of the spline, $x_+^p = x^p$ when $x > 0$ and 0 otherwise, $\kappa_k, k = 1, \ldots, K$ are a set of fixed constants called knots, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ is a coefficient vector of the parametric portion of the model, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$ is the vector of spline coefficients, K is the number of spline knots. If knot locations cover the range of x and K is sufficiently large, the class of P-spline (2.2) can approximate any smooth function $m_0(\cdot)$ with a high degree of accuracy, even with a small p (Boor, 2001). Ruppert, Wand and Carroll (2003) recommended using the number of spline knots K as the minimum of 40 and the number of unique x's divided by 4.

We assume that a random sample from the population is obtained under an uninformative sampling plan such that (2.1) remains valid for the sampled units. Our immediate task is to fit this model based on the sampled data and we follow the approach of Opsomer et al. (2008). For ease of presentation, we first introduce some matrix notation. Let n_i be the number of units sampled from small area i. The response values from the i^{th} area will be denoted as $\mathbf{y}_i = \left(y_{i1}, y_{i2}, \ldots, y_{in_i}\right)'$. We then pile them up to form the response vector of length n: $\mathbf{Y}_n' = \left(\mathbf{y}_0', \mathbf{y}_1', \ldots, \mathbf{y}_m'\right)$. We similarly define $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\epsilon}_n$ for error term. We use $\mathbf{v} = (v_0, \ldots, v_m)'$ for area specific random effects and create a matrix $\mathbf{D}$ such that

$$\mathbf{D}\mathbf{v} = (v_0 \mathtt{1}_{n_0}', v_1 \mathtt{1}_{n_1}', \ldots, v_m \mathtt{1}_{n_m}')$$

with $\mathtt{1}_k$ being a length k vector of 1's. We further construct matrices $\mathbf{X}_n$ and $\mathbf{Z}_n$ so that their rows are made up of

$$\mathbf{x}_{ij}' = \left(1, x_{ij}, \ldots, x_{ij}^p\right), \quad \mathbf{z}_{ij}' = \left(\left(x_{ij} - \kappa_1\right)_+^p, \ldots, \left(x_{ij} - \kappa_K\right)_+^p\right)$$

in a proper order. With these matrices and vectors, the data in the sample under model (2.1) are connected by

$$\mathbf{Y}_n = \mathbf{X}_n \boldsymbol{\beta} + \mathbf{Z}_n \boldsymbol{\gamma} + \mathbf{D}\mathbf{v} + \boldsymbol{\epsilon}_n. \tag{2.3}$$

Opsomer et al. (2008) fitted this model under the assumption that the components of $\boldsymbol{\gamma}$, of $\mathbf{v}$ and $\boldsymbol{\epsilon}$ are all independent and identically normally distributed with variances σ_γ^2, σ_v^2 and σ_ϵ^2 respectively. The solutions to the fit are given by

$$\begin{aligned}
\hat{\mathbf{V}} &= \mathbf{Z}_n \boldsymbol{\Sigma}_\gamma \mathbf{Z}_n' + \mathbf{D}\hat{\boldsymbol{\Sigma}}_v \mathbf{D}' + \hat{\boldsymbol{\Sigma}}_\varepsilon, \\
\hat{\mathbf{v}} &= \hat{\boldsymbol{\Sigma}}_v \mathbf{D}' \hat{\mathbf{V}}^{-1} \left(\mathbf{Y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}\right), \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}_n' \hat{\mathbf{V}}^{-1} \mathbf{X}_n\right)^{-1} \left(\mathbf{X}_n' \hat{\mathbf{V}}^{-1} \mathbf{Y}_n\right), \\
\hat{\boldsymbol{\gamma}} &= \hat{\boldsymbol{\Sigma}}_\gamma \mathbf{Z}_n' \hat{\mathbf{V}}^{-1} \left(\mathbf{Y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}\right)
\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_\gamma$, $\hat{\boldsymbol{\Sigma}}_v$, $\hat{\boldsymbol{\Sigma}}_\varepsilon$ are restricted maximum likelihood estimates of the covariance matrices of $\boldsymbol{\gamma}$, $\mathbf{v}$ and $\boldsymbol{\epsilon}$, and $\hat{\mathbf{V}}$ is the estimate of $\mathbf{V} \equiv \operatorname{var}(\mathbf{Y}_n)$.

Opsomer et al. (2008) then gave the empirical best linear unbiased predictor of the small area mean:

$$\hat{\bar{Y}}_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_i + \ldots + \hat{\beta}_p \bar{X}_i^p + \bar{\mathbf{z}}_i \hat{\boldsymbol{\gamma}} + \hat{v}_i, \tag{2.4}$$

where $\bar{X}_i, \ldots, \bar{X}_i^p$ are the means of the powers of population units x_{ij} in area i, i.e., $\bar{X}_i^s = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}^s$ for $s = 1, \ldots, p$, and $\bar{\mathbf{z}}_i \hat{\boldsymbol{\gamma}}$ stands for the true means of the spline basis functions over the small area i. Clearly, the above discussion easily extends to non-parametric additive models with two or more covariates (Lin and Zhang (1999), Ruppert et al. (2003) and Wood (2006)).

In this paper, we follow Opsomer et al. (2008) to get all the fitted values. For small area quantile estimation, we remove the normality assumption on ϵ_{ij}. Instead, we assume that their distributions $G_i(u)$ satisfy a DRM so that for $i = 1, \ldots, m$,

$$\log\{dG_i(u) / dG_0(u)\} = \theta_i' \mathbf{q}(u), \tag{2.5}$$

with a pre-specified basis function $\mathbf{q}(u)$ and an area-specific tilting parameter θ_i. One may include $i = 0$ in the above equation by setting $\theta_0 = 0$. We require the first element of $\mathbf{q}(u)$ to be one, so that the first element of θ_i is a normalization parameter. The DRM includes normal, Gamma, and many other distribution families as special cases. Discussions about DRM can be found in Anderson (1979), Qin and Zhang (1997), Kezioua and Leoni-Aubina (2008) and Chen and Liu (2013).

Equations (2.1), (2.2) and (2.5) together form the platform of this paper for small area quantile estimation. Our work differs from Opsomer et al. (2008) in that we focus on small area quantile estimation without a normality assumption on $G_i(\cdot)$. At the same time, this paper differs from Chen and Liu (2018) by postulating a non-parametric regression relationship between y_{ij} and x_{ij} instead of a linear one.

3 Proposed approach

For any $\alpha \in (0, 1)$, the α^{th} quantile of a distribution F is defined to be

$$\xi_\alpha = \inf\{u: F(u) \geq \alpha\}.$$

If $\hat{F}(u)$ is an estimate of $F(u)$, its α- quantile is naturally estimated by

$$\hat{\xi}_\alpha = \inf\left\{u: \hat{F}(u) \geq \alpha\right\}. \tag{3.1}$$

Under the distributional assumption on ϵ_{ij}, we have

$$\begin{aligned} P\left(y_{ij} \leq u\right) &= \mathbb{E}\left\{P\left(\varepsilon_{ij} \leq u - m_0\left(x_{ij}\right) - v_i \mid x_{ij}, v_i\right)\right\} \\ &= \mathbb{E}\left\{G_i\left(u - m_0\left(x_{ij}\right) - v_i\right)\right\}. \end{aligned}$$

Hence, the population distribution of the i^{th} small area is given by

$$F_i(u) = N_i^{-1}\sum_{j=1}^{N_i} G_i\left(u - m_0\left(x_{ij}\right) - v_i\right).$$

Once G_i and $m_0(\cdot)$ are suitably estimated, so will be the small area quantiles.

We follow the empirical likelihood idea of Chen and Liu (2018) for estimating $G_i(\cdot)$. Suppose the values of ε_{ij} in the sample are known. Consider a candidate G_0 of the form

$$G_0(u) = \sum_{i,j} p_{ij} I\left(\varepsilon_{ij} \leq u\right),$$

where $I(\cdot)$ is an indicator function and $\sum_{i,j} = \sum_{i=0}^{m}\sum_{j=1}^{n_i}$. We hence have $p_{ij} = dG_0\left(\varepsilon_{ij}\right)$ and under DRM $dG_i(\varepsilon_{st}) = p_{st}\exp\left\{\boldsymbol{\theta}_i'\mathbf{q}(\varepsilon_{st})\right\}$ for $i = 0, 1, \ldots, m$ which implies

$$G_i(u) = \sum_{s,t} p_{st}\exp\left\{\boldsymbol{\theta}_i'\mathbf{q}(\varepsilon_{st})\right\} I(\varepsilon_{st} \leq u). \tag{3.2}$$

By Owen (2001), we obtain the empirical likelihood function

$$L_n\left(G_0, G_1, \ldots, G_m\right) = \prod_{i,j} dG_i\left(\varepsilon_{ij}\right) = \left\{\prod_{i,j} p_{ij}\right\} \exp\left[\sum_{i,j}\left\{\ \boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{ij}\right)\right\}\right],$$

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

$$\sum_{i,j} p_{ij} \exp\left\{\boldsymbol{\theta}_s'\mathbf{q}\left(\varepsilon_{ij}\right)\right\} = 1. \tag{3.3}$$

Note that we have used the convention $\boldsymbol{\theta}_0 = 0$ for simpler presentation. Because $G_1, \ldots, G_m$ are fully determined by $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_m'\right)$ and $G_0,$ we write the empirical log-likelihood as

$$\ell_n\left(\boldsymbol{\theta}, G_0\right) = \sum_{i,j} \log\left(p_{ij}\right) + \sum_{ij} \boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{ij}\right).$$

Maximizing $\ell\left(\boldsymbol{\theta}, G_0\right)$ with respect to G_0 under the constraints (3.3) results in fitted probabilities

$$\hat{p}_{ij} = n^{-1}\left\{1 + \sum_{s=1}^{m} \lambda_s\left[\exp\left\{\boldsymbol{\theta}_s'\mathbf{q}(\varepsilon_{ij})\right\} - 1\right]\right\}^{-1} \tag{3.4}$$

and the profile log EL

$$\ell_n\left(\boldsymbol{\theta}\right) = -\sum_{i,j} \log\left\{1 + \sum_{s=1}^{m} \lambda_s\left[\exp\left\{\boldsymbol{\theta}_s'\mathbf{q}\left(\varepsilon_{ij}\right)\right\} - 1\right]\right\} + \sum_{i,j} \boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{ij}\right)$$

with $\left(\lambda_1, \ldots, \lambda_m\right)$ being the solution to

$$\sum_{s,t} \frac{\exp\left\{\boldsymbol{\theta}_i'\mathbf{q}\left(\varepsilon_{st}\right)\right\} - 1}{1 + \sum_{l=1}^{m} \lambda_l\left[\exp\left\{\boldsymbol{\theta}_l'\mathbf{q}\left(\varepsilon_{st}\right)\right\} - 1\right]} = 0.$$

Since the values of ε_{ij} are not available, we replace them by the residuals obtained from fitting model (2.1) under assumption (2.2):

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{m}_0\left(x_{ij}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) - \hat{v}_i$$

where

$$\hat{m}_0(x; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_p x^p + \sum_{k=1}^{K} \hat{\gamma}_k\left(x - \kappa_k\right)_+^p. \tag{3.5}$$

Let $\hat{\ell}_n\left(\boldsymbol{\theta}\right)$ be the log EL function $\tilde{\ell}_n\left(\boldsymbol{\theta}\right)$ after ε_{ij} are replaced by $\hat{\varepsilon}_{ij}$. We define the maximum EL estimator of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}} = \operatorname{argmax}\hat{\ell}_n\left(\boldsymbol{\theta}\right)$ and estimate $G_i\left(u\right)$ by

$$\tilde{G}_i\left(u\right) = \sum_{s,t} \hat{p}_{st} \exp\left\{\hat{\boldsymbol{\theta}}_i'\mathbf{q}\left(\hat{\varepsilon}_{st}\right)\right\} I\left(\hat{\varepsilon}_{st} \leq u\right) \tag{3.6}$$

with

$$\hat{p}_{st} = n^{-1}\left\{1 + \sum_{l=1}^{m}(n_l/n)\left[\exp\left\{\boldsymbol{\theta}_l'\mathbf{q}(\hat{\varepsilon}_{st})\right\} - 1\right]\right\}^{-1}$$

and $\hat{\boldsymbol{\theta}}_0 = 0$. The R package **drmdel** can be used to compute $\hat{\boldsymbol{\theta}}$ and $\hat{p}_{ij}$ which has 11 choices of basis function $\mathbf{q}(u)$.

Because $\tilde{G}_i(u)$ is discrete, the following kernel smoothed distribution $\hat{G}_i(u)$ leads to better quantile estimation:

$$\hat{G}_i(u) = \sum_{j=1}^{n_i}\hat{w}_{ij}\Phi\left(\frac{\hat{\varepsilon}_{ij} - u}{b}\right), \tag{3.7}$$

where the weights are chosen to be $\hat{w}_{ij} = \tilde{G}_i(\hat{\varepsilon}_{ij}) - \tilde{G}_i(\hat{\varepsilon}_{ij}-)$, b is a bandwidth parameter, and $\Phi(\cdot)$ is the distribution function of standard normal. As suggested by Chen and Liu (2013), we choose $b = 1.06n^{-1/5}\min\{\hat{\sigma}, \hat{Q}/1.34\}$ where $\hat{\sigma}$ is the standard deviation of the distribution $\hat{G}_i$ and $\hat{Q}$ is its interquartile range.

In some applications, only population power means of covariates are known and can be used for statistical inference. In other applications, covariates of all members of the population are known. This leads two possible quantile estimates. In the first case, we estimate F_i by

$$\hat{F}_i^{(a)}(u) = n_i^{-1}\sum_{j=1}^{n_i}\hat{G}_i\left(u - \hat{\bar{Y}}_i - \left\{\hat{m}_0\left(x_{ij}; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) - \hat{m}_0\left(\bar{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)\right\}\right), \tag{3.8}$$

where we use $\hat{m}_0\left(\bar{x}_i; \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ specified in (3.5).

When the census information about x is available, we estimate F_i by

$$\hat{F}_i^{(b)}(u) = N_i^{-1}\left\{\sum_{j\in s_i} I\left(y_{ij} \le u\right) + \sum_{j\in r_i}\hat{G}_i\left(u - \hat{m}_0\left(x_{ij}\right) - \hat{v}_i\right)\right\}, \tag{3.9}$$

where s_i and r_i are sets of observed and unobserved units in small area i. The rest of the specifications are the same as in (3.8).

The proposed estimates resemble those of Chen and Liu (2018) but we use a non-parametric regression. Because collecting population power means of covariates is easier than collecting covariates values of all units in the population $\hat{F}_i^{(a)}(u)$ is more broadly applicable than $\hat{F}_i^{(b)}(u)$. It is also computationally more efficient. Because $\hat{F}_i^{(b)}(u)$ uses covariate values of all units in the population, it should statistically outperform when both are applicable.

4 Bootstrap estimation of the mean squared errors

The proposed small area quantile estimators are assembled with many intermediate steps. It is difficult to analytically evaluate the variances or mean squared error (MSE) of such estimators. We follow others

(Sinha and Rao (2009), Tzavidis et al. (2010) and Chen and Liu (2018)) to develop a bootstrap procedure as follows:

Step 1 Obtain estimates $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{\sigma}_v^2$ and $\hat{m}_0\left(x, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ based on Model (2.1), and calculate $\hat{G}_i(u)$ as in (3.7).

Step 2 Generate a bootstrap finite population $H^* = \left\{y_{ij}^*, x_{ij}\right\}, i = 0, \ldots, m, j = 1, \ldots, N_i$ with

$$y_{ij}^* = \hat{m}_0\left(x_{ij}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right) + v_i^* + \varepsilon_{ij}^*,$$

where the bootstrap residuals ε_{ij}^* are sampled from CDF $\hat{G}_i(u)$, and v_i^* are generated from $N(0, \hat{\sigma}_v^2)$.

Step 3 From the bootstrap population H^*, we select $n_i^* = n_i$ sample units from small area i by simple random sampling without replacement, and repeat it L times to get $h_l^*, l = 1, \ldots, L.$ For each sample h_l^*, compute the estimates $\hat{F}_i^{(a)*l}(u)$ and $\hat{F}_i^{(b)*l}(u)$ as in (3.8) and (3.9) respectively.

Step 4 Compute the empirical MSE estimator of $\hat{\tau}$ as

$$\operatorname{mse}(\tau^*) = L^{-1}\sum_{l=1}^{L}\left(\hat{\tau}^{*l} - \tau^*\right)^2,$$

where $\hat{\tau}^{*l} = \tau\left(\hat{F}^{*l}(u)\right)$ denotes any functional of $\hat{F}^{(a)*l}(u)$ or $\hat{F}^{(b)*l}$ and $\tau^* = \tau(F^*(u))$ with $F^*(u)$ being the known CDF of the bootstrap populations.

Step 5 Repeat Steps 2 to 4, B times, and define the bootstrap MSE estimate as

$$B^{-1}\sum_{b=1}^{B}\operatorname{mse}(\tau^*)_b,$$

where $\operatorname{mse}(\tau^*)_b$ is the $\operatorname{mse}(\tau^*)$ calculated in the b^{th} repetition.

The performance of the bootstrap MSE estimator will be examined and reported in the simulation section.

5 Monte Carlo simulations

In this section, we use simulation to evaluate the performances of the proposed penalized spline regression model based empirical likelihood estimators (PEL) and their MSE estimates. When only the covariate population means are known the proposed estimators are compared with only the nested-error linear regression model based empirical likelihood estimator (LEL) of Chen and Liu (2018), and the direct estimator (DE). When covariate values are known for all sample units, the comparison is extended to also include six estimators of Tzavidis et al. (2010), denoted as EBLUP/naïve, EBLUP/CD, EBLUP/RKM, M-quantile/naïve, M-quantile/CD and M-quantile/RKM. Here, EBLUP/CD and M-quantile/CD denote the EBLUP and M-quantile estimator are obtained based on the CDF proposed by Chambers and Dunstan

(1986), and corresponding estimators based on the CDF proposed by Rao, Kovar and Mantel (1990) write as RKM.

Similar to Chen and Liu (2018), we must choose $\mathbf{q}(u)$ in the DRM. Two candidates are $\mathbf{q}_1(u) = (1, u)'$ and $\mathbf{q}_2(u) = \left(1, \operatorname{sign}(u)\sqrt{|u|}\right)'$. Some preliminary simulation results indicate that $\mathbf{q}_1(u) = (1, u)'$ works well for the P-splines fitted non-parametric regression model, but $\mathbf{q}_2(u)$ does not. Instead, the choice of $\mathbf{q}_2^*(u) = (1, u, u^2)'$ leads to competitive performance. So, we use $\mathbf{q}_1(u)$ and $\mathbf{q}_2^*(u)$ in our simulation.

Following Rao et al. (2014) and Torabi and Shokoohi (2015), we generated data from the following three models:

$$\begin{aligned} \text{A}:\ & y_{ij} = 1 + x_{ij} + v_i + \varepsilon_{ij}, \\ \text{B}:\ & y_{ij} = 1 + x_{ij} + x_{ij}^2 + v_i + \varepsilon_{ij}, \\ \text{C}:\ & y_{ij} = 1 - x_{ij} + 0.5\exp\left(x_{ij}\right) + v_i + \varepsilon_{ij}. \end{aligned}$$

They lead to linear, quadratic and exponential regression functions respectively. We set the number of small areas to be 30 and area population sizes $N_i = 500(i+1)$, $i = 0, 1, \ldots, 29$. We generated covariate x_{ij} from $N(0, 1)$. Once x_{ij} are generated, we treated them as fixed in the simulation. The area-specific random effect v_i were generated from $N(0, 1)$, and the errors ε_{ij} were generated from the following four distributions.

$$\begin{aligned} \text{(i)}:\ & N(0, 1), \\ \text{(ii)}:\ & t(3), \\ \text{(iii)}:\ & \text{normal mixture } 0.5N(-1, 1) + 0.5N(1, 1), \\ \text{(iv)}:\ & N(0, \sigma_i^2), \text{ with } \sigma_i \sim U(0.5, 2), i = 0, \ldots, 29. \end{aligned}$$

Distribution (ii) has a heavy tail, distributions (ii) and (iii) are symmetric, and distribution (iv) is heteroscedastic.

We used $R = 1{,}000$ repetitions in the simulation and drew random samples of size $n = 500$ without replacement from the population in each repetition. To avoid the possibility that some small areas have too few sample units, we drew $n - 60$ units at the population level and allocated an additional 2 units in each small area. We used R package **mgcv** for the REML method with default options for values of p and K when fitting the P-spline function (2.4). We calculated estimates of the 5%, 25%, 50%, 75%, and 95% small area quantiles denoted as DE, LEL1, LEL2, PEL1, PEL2, for direct estimator, estimators of Chen and Liu (2018) and the proposed estimators using $\mathbf{q}_1(\cdot)$ and $\mathbf{q}_2(\cdot)$. We report their average mean squared error (AMSE) and absolute biases (ABIAS) defined below:

$$\begin{aligned} \text{AMSE} &= \{R(m+1)\}^{-1}\sum_{i=0}^{m}\sum_{r=1}^{R}\left(\hat{\xi}_i^{(r)} - \xi_i^{(r)}\right)^2, \\ \text{ABIAS} &= (m+1)^{-1}\sum_{i=0}^{m}\left| R^{-1}\sum_{r=1}^{R}\hat{\xi}_i^{(r)} - R^{-1}\sum_{r=1}^{R}\xi_i^{(r)}\right|, \end{aligned}$$

where $\hat{\xi}_i^{(r)}$ is either one of the quantile estimates of for the i^{th} small area in the r^{th} repetition. The results under Models A, B, and C are given in Tables 5.1-5.3 respectively. Both PEL and LEL are based on $\hat{F}_i^{(a)}$ and its mirror version in Chen and Liu (2018).

Under Model A, the linear model is valid. Hence, we expect LEL to be superior. According to Table 5.1, two methods are similar for the 25%, 50% and 75% quantiles. LELs outperform PELs for the 5% quantile while the comparison reverses for the 95% quantile. Both PEL and LEL outperform DE for the 25%, 50% and 75% quantiles with big margins. An overall impression is that the proposed methods still work satisfactorily.

Under Model B, the linear model breaks down mildly. Results in Table 5.2 show that the PEL estimators have lower AMSE for lower quantiles. The LELs still have low AMSE in spite of have higher ABIAS. The advantage of the proposed PEL under the non-parametric nested-error regression models focus for quantiles in middle levels. With fewer observations near extreme quantiles, the non-parametric model is hard to fit.

The linearity is seriously violated under Model C. LEL is expected to have poor performance and this is evident as shown in Table 5.3. At the same time, PELs work well for the 25%, 50% and 75% quantiles. The choice of $\mathbf{q}_2^*(u)$ also helps in general. For extreme quantiles, PELs remain unworth the trouble compared with DE.

Table 5.1
AMSE and ABIAS of small area quantile estimators under Model A

		AMSE					ABIAS				
	α	DE	LEL1	LEL2	PEL1	PEL2	DE	LEL1	LEL2	PEL1	PEL2
Error distribution (i)	5%	0.470	0.120	0.142	0.121	0.162	0.346	0.022	0.028	0.024	0.032
	25%	0.219	0.074	0.080	0.074	0.082	0.081	0.006	0.006	0.006	0.006
	50%	0.187	0.067	0.067	0.067	0.068	0.011	0.005	0.005	0.006	0.006
	75%	0.218	0.074	0.079	0.074	0.082	0.081	0.007	0.005	0.008	0.006
	95%	0.470	0.121	0.142	0.123	0.165	0.340	0.024	0.031	0.023	0.033
Error distribution (ii)	5%	1.287	0.249	0.786	0.276	1.726	0.352	0.011	0.023	0.011	0.089
	25%	0.297	0.196	0.217	0.178	0.186	0.084	0.022	0.036	0.021	0.031
	50%	0.238	0.187	0.182	0.167	0.154	0.011	0.010	0.010	0.010	0.009
	75%	0.304	0.197	0.233	0.179	0.189	0.081	0.023	0.038	0.023	0.032
	95%	1.344	0.249	1.919	0.319	2.297	0.349	0.013	0.034	0.015	0.100
Error distribution (iii)	5%	0.636	0.165	0.199	0.163	0.234	0.408	0.008	0.013	0.008	0.019
	25%	0.340	0.132	0.147	0.133	0.152	0.109	0.010	0.007	0.011	0.008
	50%	0.306	0.128	0.128	0.130	0.132	0.014	0.007	0.007	0.007	0.007
	75%	0.340	0.133	0.151	0.134	0.156	0.108	0.011	0.009	0.012	0.008
	95%	0.651	0.168	0.205	0.166	0.243	0.410	0.010	0.016	0.010	0.022
Error distribution (iv)	5%	1.225	2.589	0.787	2.679	0.651	0.504	0.220	0.028	0.222	0.071
	25%	0.574	0.681	0.380	0.652	0.349	0.114	0.174	0.047	0.157	0.017
	50%	0.488	0.273	0.277	0.241	0.291	0.017	0.010	0.010	0.009	0.010
	75%	0.571	0.700	0.383	0.670	0.349	0.121	0.183	0.057	0.166	0.012
	95%	1.251	2.611	0.795	2.709	0.655	0.519	0.207	0.037	0.210	0.082

Table 5.2
AMSE and ABIAS of small area quantile estimators under Model B

		AMSE					ABIAS				
	α	DE	LEL1	LEL2	PEL1	PEL2	DE	LEL1	LEL2	PEL1	PEL2
Error distribution (i)	5%	0.524	2.998	2.991	0.404	0.439	0.382	1.520	1.502	0.017	0.019
	25%	0.474	0.182	0.183	0.259	0.262	0.177	0.118	0.123	0.018	0.017
	50%	0.865	0.907	0.951	0.215	0.219	0.092	0.785	0.791	0.031	0.031
	75%	1.963	0.985	1.170	0.817	0.825	0.132	0.602	0.616	0.021	0.021
	95%	7.850	3.083	3.783	9.163	9.193	1.200	1.159	1.185	0.251	0.251
Error distribution (ii)	5%	1.227	2.768	3.065	0.492	1.691	0.352	1.430	1.423	0.067	0.143
	25%	0.562	0.280	0.268	0.331	0.327	0.189	0.087	0.087	0.027	0.024
	50%	0.976	0.924	0.957	0.287	0.281	0.098	0.728	0.733	0.046	0.046
	75%	2.119	1.023	1.231	0.817	0.854	0.129	0.557	0.572	0.034	0.034
	95%	8.392	2.989	4.864	8.405	9.180	1.250	1.140	1.147	0.112	0.119
Error distribution (iii)	5%	0.842	2.171	2.207	0.425	0.491	0.500	1.252	1.238	0.013	0.014
	25%	0.657	0.209	0.209	0.292	0.296	0.176	0.076	0.077	0.010	0.011
	50%	0.935	0.791	0.805	0.244	0.249	0.082	0.679	0.682	0.026	0.027
	75%	1.983	0.981	1.086	0.739	0.752	0.131	0.588	0.597	0.024	0.024
	95%	8.020	2.782	3.251	8.344	8.385	1.219	1.059	1.078	0.144	0.145
Error distribution (iv)	5%	1.458	3.913	3.066	2.414	0.814	0.557	1.195	1.172	0.226	0.053
	25%	0.919	0.460	0.397	0.474	0.472	0.206	0.154	0.137	0.058	0.017
	50%	1.183	0.913	0.920	0.398	0.416	0.071	0.629	0.640	0.048	0.023
	75%	2.195	1.223	1.209	1.022	0.902	0.163	0.471	0.511	0.033	0.031
	95%	8.043	2.954	3.420	7.476	7.639	1.268	0.975	1.042	0.104	0.115

Table 5.3
AMSE and ABIAS of small area quantile estimators under Model C

		AMSE					ABIAS				
	α	DE	LEL1	LEL2	PEL1	PEL2	DE	LEL1	LEL2	PEL1	PEL2
Error distribution (i)	5%	0.279	1.340	1.258	0.092	0.151	0.267	0.997	0.978	0.051	0.031
	25%	0.146	0.316	0.263	0.087	0.098	0.068	0.282	0.280	0.035	0.046
	50%	0.152	0.326	0.403	0.094	0.096	0.011	0.215	0.227	0.019	0.015
	75%	0.335	0.868	1.368	0.225	0.244	0.029	0.665	0.700	0.043	0.044
	95%	7.011	0.890	6.818	27.97	27.81	0.291	0.206	0.301	1.398	1.384
Error distribution (ii)	5%	1.180	1.181	1.355	0.278	1.776	0.286	0.849	0.836	0.090	0.174
	25%	0.205	0.461	0.395	0.201	0.208	0.063	0.317	0.327	0.085	0.098
	50%	0.201	0.450	0.502	0.201	0.191	0.024	0.226	0.235	0.013	0.012
	75%	0.528	0.943	1.422	0.390	0.422	0.017	0.641	0.681	0.096	0.104
	95%	7.478	0.890	6.306	23.33	25.01	0.479	0.089	0.107	1.055	1.084
Error distribution (iii)	5%	0.438	1.063	1.004	0.157	0.240	0.349	0.826	0.803	0.065	0.034
	25%	0.299	0.328	0.289	0.158	0.181	0.120	0.158	0.161	0.009	0.020
	50%	0.305	0.364	0.409	0.174	0.179	0.013	0.151	0.157	0.035	0.029
	75%	0.428	0.709	1.035	0.275	0.308	0.077	0.499	0.524	0.015	0.017
	95%	6.718	0.974	4.704	24.79	25.04	0.232	0.321	0.378	1.336	1.325
Error distribution (iv)	5%	1.078	4.146	2.303	3.378	0.685	0.444	0.918	0.803	0.409	0.035
	25%	0.530	0.829	0.531	0.668	0.380	0.107	0.105	0.156	0.147	0.071
	50%	0.490	0.526	0.565	0.297	0.344	0.021	0.177	0.188	0.054	0.017
	75%	0.718	1.454	1.412	1.149	0.542	0.076	0.438	0.542	0.061	0.048
	95%	6.430	2.492	4.002	22.54	21.92	0.462	0.364	0.242	1.258	1.042

Next, we study estimators applicable when covariate values are known for all sample units. The simulation includes EB0, EB1, EB2, MQ0, MQ1 and MQ2 stand for EBLUP/naïve, EBLUP/CD, EBLUP/RKM, M-quantile/naïve, M-quantile/CD and M-quantile/RKM respectively. We set relatively small population sizes $N_i = 500$ to save some computation. Table 5.4 contains the AMSE of these estimators under Models A, B and C with $N(0, 1)$ error distribution. To save space, we do not present the corresponding bias results. The simulation results show that the proposed method has lower AMSE and ABIAS (not presented) in general. It works well even for quantiles at rather extreme levels.

To save space, we pool the AMSE results for all 5 levels of quantiles in Table 5.5. The entry corresponding to A_i is the average AMSE for estimating quantiles at levels 5%, 25%, 50%, 75%, and 95% when data are generated from Model A with error distribution (i). We notice that with more detailed information on covariates, the LEL and PEL estimators are substantially more accurate compared to results in Tables 5.1-5.3. From Model A to Model C, the regression line becomes less linear. Correspondingly, the proposed quantile estimators have greater advantages against other estimators.

Now we evaluate the bootstrap MSE estimator proposed in Section 4. Because this method involves heavy computation, we confined the simulation to the estimator based on $\hat{F}_i^{(b)}(u)$ with basis function $\mathbf{q}_1(u) = (1, u)'$ and put $B = 100$, $L = 100$. We report the average ratios of the estimated MSEs and the simulated MSEs across all the small areas. The closer the ratio to one, more accurate the bootstrap MSE estimate. From Table 5.6 we can see that the average ratios close to one in majority situations except for error distribution (iv) on extreme levels of quantiles. We conclude that the bootstrap MSE estimator is generally satisfactory.

Table 5.4
AMSE of 10 quantile estimators when all covariance values are known with *N*(0, 1) error distribution

	α	**EB0**	**EB1**	**EB2**	**MQ0**	**MQ1**	**MQ2**	**LEL1**	**LEL2**	**PEL1**	**PEL2**
Model A	5%	0.477	0.123	0.501	0.536	0.127	0.499	0.128	0.146	0.078	0.110
	25%	0.139	0.073	0.154	0.198	0.074	0.154	0.073	0.078	0.065	0.073
	50%	0.061	0.066	0.124	0.119	0.066	0.124	0.066	0.066	0.064	0.064
	75%	0.145	0.074	0.149	0.204	0.074	0.149	0.074	0.080	0.066	0.073
	95%	0.491	0.125	0.394	0.552	0.129	0.395	0.126	0.146	0.079	0.113
Model B	5%	1.270	2.500	0.928	1.682	2.575	0.946	2.965	2.949	0.079	0.110
	25%	0.351	0.152	0.239	0.262	0.149	0.239	0.193	0.193	0.069	0.069
	50%	0.834	0.723	0.285	0.631	0.722	0.284	0.899	0.944	0.071	0.073
	75%	0.314	0.634	0.532	0.257	0.644	0.530	0.986	1.160	0.082	0.084
	95%	3.710	2.095	3.690	4.209	2.059	3.685	3.235	3.900	0.154	0.156
Model C	5%	0.346	0.830	0.415	0.708	0.307	0.351	1.087	1.028	0.075	0.130
	25%	0.345	0.173	0.169	0.388	0.110	0.154	0.263	0.224	0.066	0.075
	50%	0.340	0.170	0.142	0.207	0.150	0.136	0.291	0.349	0.065	0.067
	75%	0.288	0.577	0.211	0.191	0.376	0.227	0.731	1.088	0.068	0.087
	95%	2.578	11.47	8.087	5.194	14.64	11.96	0.868	4.215	0.148	0.156

Table 5.5
Average AMSE over 5 quantiles when all covariate values are known

Model	EB0	EB1	EB2	MQ0	MQ1	MQ2	LEL1	LEL2	PEL1	PEL2
A_{i}	0.263	0.092	0.264	0.322	0.094	0.264	0.093	0.103	0.070	0.087
A_{ii}	0.810	1.379	1.822	0.810	1.381	1.796	0.217	0.370	0.203	0.744
A_{iii}	0.754	0.183	0.408	0.819	0.183	0.407	0.149	0.168	0.135	0.168
A_{iv}	0.687	0.186	0.399	0.746	0.188	0.399	0.281	0.196	0.256	0.164
B_{i}	1.296	1.221	1.135	1.408	1.230	1.138	1.832	1.829	0.091	0.098
B_{ii}	1.442	1.714	2.348	1.496	1.718	2.343	1.596	1.812	0.230	0.504
B_{iii}	1.270	1.081	1.357	1.348	1.088	1.351	1.399	1.521	0.163	0.179
B_{iv}	1.346	1.177	1.315	1.436	1.183	1.317	1.565	1.701	0.205	0.166
C_{i}	0.799	2.645	1.805	1.339	3.117	2.566	0.648	1.381	0.084	0.103
C_{ii}	1.441	3.439	3.368	2.232	3.967	3.898	0.725	1.168	0.241	0.377
C_{iii}	1.141	2.516	1.898	1.834	2.937	2.572	0.595	1.133	0.153	0.186
C_{iv}	1.149	2.499	1.909	1.821	2.933	2.639	0.767	1.176	0.280	0.179

Table 5.6
Average ratios of bootstrap MSEs and simulated MSEs

α	A_{i}	A_{ii}	A_{iii}	A_{iv}	B_{i}	B_{ii}	B_{iii}	B_{iv}	C_{i}	C_{ii}	C_{iii}	C_{iv}
5%	1.01	1.03	1.05	0.36	1.05	0.98	1.01	0.39	0.99	1.19	1.10	0.27
25%	1.00	0.99	1.05	0.74	1.03	0.99	0.95	1.03	1.03	0.97	0.99	0.73
50%	1.06	1.04	0.97	1.10	1.01	1.03	0.96	0.99	1.09	0.96	0.97	1.03
75%	1.01	0.99	1.06	0.76	1.10	1.01	0.98	0.90	1.06	0.96	1.03	0.52
95%	1.04	1.20	1.10	0.33	0.89	1.02	1.13	1.02	0.95	1.37	1.13	0.69

6 Empirical application

We now illustrate the proposed estimators based on the data set *Survey of Labour and Income Dynamics* (SLID) provided by Statistics Canada (2014) downloaded from University of British Columbia library data centre. The data contain 147 variables and 47,705 sample units. We are grateful to Statistics Canada for making the data set available, but we do not address the original goal of the survey here. Instead, we use it as a superpopulation to study the effectiveness of the proposed small area quantile estimator.

In this study, we singled out 9 of the 147 variables. They are `ttin`, `gender`, `spouse`, `edu`, `age`, `yrx`, `tweek`, `jobdur` and `tpaid`, standing respectively for: total income, gender, whether living with the spouse, the highest level of education, age, years of experience, number of weeks employed, education level, months of duration of current job and total hours paid at this job. After removing units containing missing values in these 9 variables as well as those with `ttin` $\leq 0,$ we obtained a data set containing 28,302 sample units. The covariates power means at the population level are still calculated based on all available observations. We created 28 sub-populations (namely small areas) labeled as $4(k-1)+i,\ k = 1, 2, \ldots, 7,$

$i = 1, 2, 3, 4$ based on gender-spouse-edu combinations. Here k denotes education level and $i = 1, 2, 3, 4$ denote male living with the spouse, female living with spouse, male not living with spouse and female not living with spouse respectively. The education levels are given as follows.

k	Highest education level
1	No more than 10 years elementary and secondary school
2	11-13 years of elementary and secondary school (but did not graduate)
3	Graduated high school
4	Some university or non-university postsecondary with no certificate
5	Non-university postsecondary or university certificate below Bachelor's
6	Bachelor's degree
7	University certificate above Bachelor's

We regarded $\log(\texttt{ttin})$ as the response variable and fitted linear and additive non-parametric regressions with respect to other 5 variables. Based on the whole data, the adjusted R-square of the non-parametric fit is 0.482 which is much larger than 0.370 obtained by fitting the linear regression. This suggests that a non-parametric mixed model is a good choice. Figure 6.1 shows the fitted curves of log(`ttin`) with respect to these two covariates. Also, the R-square is as high as 0.483 even if the model includes only covariates `age` and `tpaid` and a random effect. These exploratory analyses prompt us to use only these two covariates in our simulation. We carried the simulation with sample sizes $n = 200$; 500 and 1,000. To make sampling proportions in small areas close to their sizes, we let $n_i = a_i + 2$, $i = 1, \ldots, 28$ with a_i generated from the multinomial distribution with $p_i = N_i / N$.

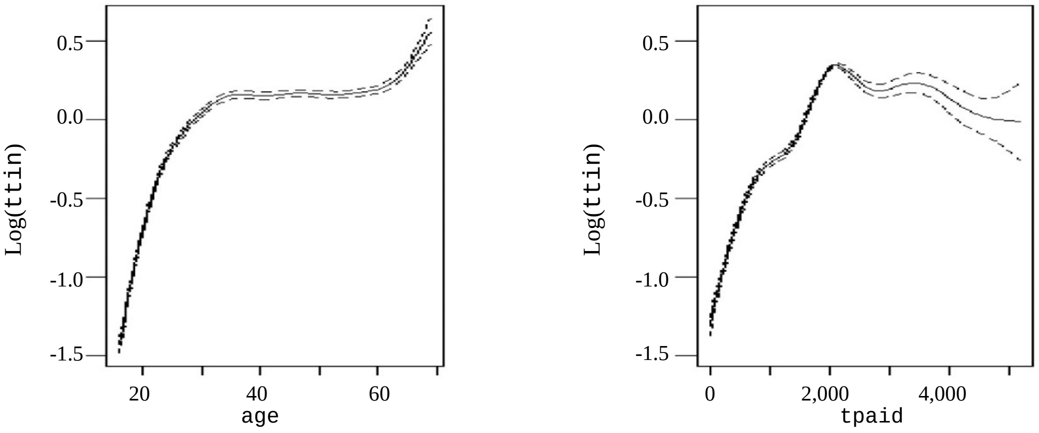

Figure 6.1 Fitted curves of log(`ttin`) with respect to `age` and `tpaid`.

The simulated AMSE of 10 estimators based on 1,000 repetitions are reported in Table 6.1. We first notice that both our PEL estimators outperform the other estimators, in general, indicating the advantage of our non-parametric DRM based small area estimation technique. The PEL1 compared to PEL2 has the lower AMSE for 5%, 25%, and 50% quantiles, but slightly higher AMSE for 75% and 95% quantiles indicating the heteroscedasticity of data is not serious. Regardless the PEL estimators, we notice the LEL estimators outperform other estimators for 5% quantile, and have similar performance for other quantiles. Increasing the sample size reduces the AMSE of all estimators. Clearly, it is hard to estimate the 5% quantile with a good precision because the data are skewed toward the left so there are few observations for estimating the lower quantiles. Interestingly, LEL1 is not affected as much by the skewness. We feel that the kernel smoothing step (3.7) is helpful here. Without this smoothing step, LEL1 would perform much worse. Unreported simulations show that the ABIAS of all estimators decreases in general as the sample size increases and this is most apparent for DE.

To check the performance of the proposed first estimator which using only covariate average information. In Figures 6.2, we depict the 2.5%, 50%, and 97.5% quantiles of 1,000 small area median estimates by the DE, LEL1, LEL2, PEL1, PEL2 with sample size $n = 200$ with the true medians marked by dots. The y-axis is the total income and x-axis is the education level. It is seen that the PEL2 boxes are the shortest for most small areas.

Table 6.2 reports the bootstrap MSE estimates as well as the average ratios of bootstrap and simulated MSEs of the small area median estimators based on $\hat{F}_i^{(a)}(u)$ and $\hat{F}_i^{(b)}(u)$ with sample size $n = 200.$ The number of simulation repetition is 500 with basis function $\mathbf{q}_1(u) = (1, u)'$ and $B = 100,$ $L = 100.$ We can see the estimator $\hat{F}_i^{(a)}(u)$ has higher MSE than $\hat{F}_i^{(b)}(u),$ and most average ratios close to one.

Table 6.1
AMSE of small area quantile estimators based on real data

	α	EB0	EB1	EB2	MQ0	MQ1	MQ2	LEL1	LEL2	PEL1	PEL2
n = 200	5%	0.784	0.769	0.901	0.714	0.763	0.885	0.245	0.421	0.242	0.336
	25%	0.107	0.256	0.488	0.102	0.261	0.467	0.115	0.131	0.097	0.152
	50%	0.080	0.119	0.236	0.064	0.116	0.223	0.076	0.095	0.056	0.102
	75%	0.122	0.100	0.142	0.085	0.102	0.138	0.085	0.076	0.069	0.068
	95%	0.233	0.190	0.280	0.141	0.138	0.266	0.217	0.179	0.117	0.096
n = 500	5%	0.793	0.603	0.826	0.710	0.579	0.805	0.173	0.345	0.210	0.301
	25%	0.072	0.110	0.207	0.076	0.119	0.197	0.069	0.127	0.063	0.091
	50%	0.049	0.050	0.074	0.036	0.050	0.072	0.053	0.076	0.040	0.043
	75%	0.108	0.044	0.060	0.055	0.046	0.058	0.054	0.047	0.046	0.043
	95%	0.257	0.128	0.152	0.109	0.058	0.148	0.138	0.125	0.086	0.077
n = 1,000	5%	0.792	0.397	0.542	0.706	0.377	0.528	0.078	0.130	0.095	0.144
	25%	0.054	0.056	0.098	0.066	0.067	0.095	0.041	0.043	0.038	0.056
	50%	0.034	0.026	0.032	0.027	0.026	0.031	0.019	0.028	0.018	0.024
	75%	0.102	0.024	0.030	0.043	0.026	0.030	0.037	0.033	0.019	0.023
	95%	0.270	0.088	0.090	0.095	0.114	0.090	0.074	0.067	0.053	0.057

Table 6.2
Bootstrap MSE estimates and average ratios of the estimated and simulated MSEs

	$\hat{F}_i^{(a)}(u)$					$\hat{F}_i^{(b)}(u)$				
	5%	**25%**	**50%**	**75%**	**95%**	**5%**	**25%**	**50%**	**75%**	**95%**
MSE	0.542	0.196	0.117	0.098	0.165	0.204	0.093	0.068	0.062	0.102
Ratio	0.843	0.959	1.014	0.988	0.871	0.969	0.994	1.003	0.996	0.975

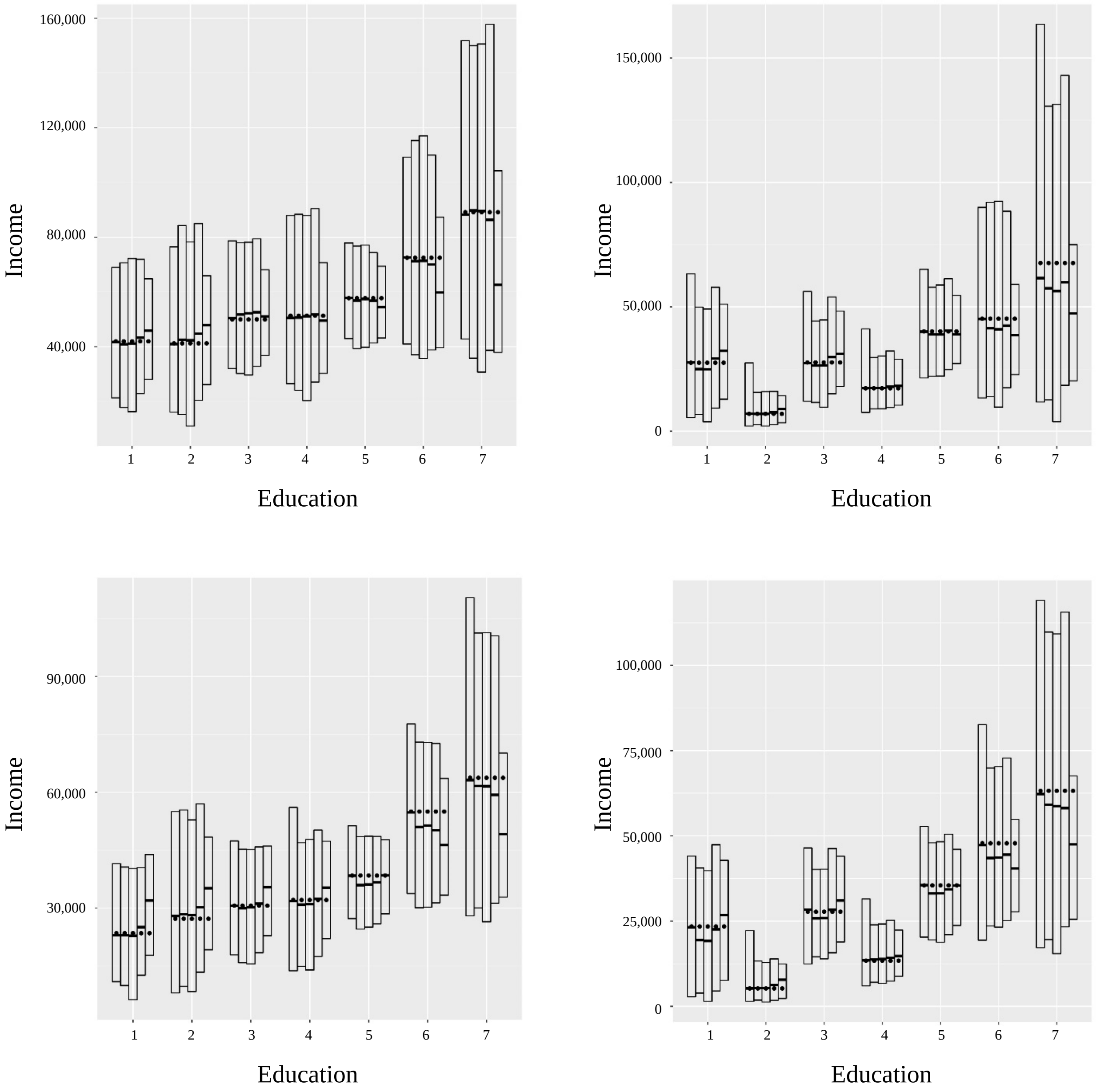

Figure 6.2 The bottom, middle and top lines of each bar denote 2.5%, 50% and 97.5% quantiles of 1,000 small area estimates of the total income. The dot in each bar denotes true small area median. Five bars in each cluster are formed by DE, LEL1, LEL2, PEL1, PEL2 estimates. Top two plots: male living (left) and not living (right) with spouse; Bottom two plots: female living (left) and not living (right) with spouse. Seven clusters in each plot correspond to 7 education levels.

7 Conclusion

We studied the small area quantile estimation under the nested-error non-parametric regression model and a semi-parametric DRM assumption on error distributions. We proposed two quantile estimators based on P-splines and empirical likelihood approach. Simulation results show that the proposed estimators are robust and have respectable efficiency under both linear and non-parametric regression functions for mid-range quantiles. The proposed approach can be extended to non-parametric regression models with multiple covariates in principle, though it will lead to many more parameters to be estimated. This problem will be investigated in a future work.

Acknowledgements

We thank Professors Simon Wood, Matt Wand, Mahmoud Torabi and Song Cai for their helpful suggestions on R packages used in this paper. This work was supported by the National Natural Science Foundation of China (No.11661067), Natural Science Foundation of Qinghai Province (No.2015-ZJ-717,2019-ZJ-920), “Western Light” talent program of Chinese Academy of Science (2017) and the funding through the Canadian Statistical Sciences Institute.

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.

Boor, C.D. (2001). *A Practical Guide to Splines.* New York: Springer.

Chambers, R., and Dunstan, R. (1986). Estimating distribution functions from survey data. *Biometrika*, 73, 597-604.

Chambers, R., and Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, 93, 255-268.

Chaudhuri, S., and Ghosh, M. (2011). Empirical likelihood for small area estimation. *Biometrika*, 98, 473-480.

Chen, J., and Liu, Y. (2013). Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41, 1669-1692.

Chen, J., and Liu, Y. (2018). Small area quantile estimation. *International Statistics Review*. In print. arXiv:1705.10063.

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.

Jiang, J. (2010). *Large Sample Techniques for Statistics.* New York: Springer.

Jiang, J., and Lahiri, P. (2006a). 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 Lahiri, P. (2006b). Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Jiang, J., Ngueyen, T. and Rao, J.S. (2010). Fence method for nonparametric small area estimation. *Survey Methodology*, 36, 1, 3-11. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2010001/article/11244-eng.pdf.

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.

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.

Lin, X., and Zhang, D. (1999). Inference in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society, Series B*, 61, 381-400.

Molina, I., and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, 38, 369-385.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: B*, 70, 265-286.

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

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.

Pfeffermann, D. (2002). Small area estimation-New developments in small area estimation. *International Statistical Review*, 70, 125-143.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, 28, 40-68.

Pratesi, M., Ranalli, M.G. and Salvati, N. (2008). Semiparametric M-quantile regression for estimation for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics*, 19, 687-701.

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.* New York: John Wiley & Sons, Inc.

Rao, J.N.K., Kovar, J.G. and Mantel, H.J. (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. *Biometrika*, 77, 365-375.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation, 2nd Edition.* New York: John Wiley & Sons, Inc.

Rao, J.N.K., Sinha, S.K. and Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *The Canadian Journal of Statistics*, 42, 126-141.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Salvati, N., Tzavidis, N. and Pratesi, M. (2012). Small area estimation via M-quantile geographically weighted regression. *Test*, 21, 1-28.

Sinha, S.K., and Rao, J.N.K. (2009). Robust small area estimation. *The Canadian Journal of Statistics*, 37(3), 381-399.

Sperlich, S., and José Lombardía, M. (2010). Local polynomial inference for small area statistics: Estimation, validation and prediction. *Journal of Non-parametric Statistics*, 22, 633-648.

Statistics Canada (2014). Survey of labour and income dynamics, 2011. Access: http://tinyurl.com/y2ys2zzs.

Torabi, M., and Shokoohi, F. (2015). Non-parametric generalized linear mixed models in small area estimation. *The Canadian Journal of Statistics*, 43, 82-96.

Tzavidis, N., and Chambers, R. (2005). Bias adjusted estimation for small areas with M-quantile models. *Statistics in Transition*, 7, 707-713.

Tzavidis, N., Salvati, N. and Pratesi, M. (2008). M-quantile models with application to poverty mapping. *Statistical Methods and Applications*, 17, 393-411.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust prediction of small area means and quantiles. *Australian and New Zealand Journal of Statistics*, 52, 167-186.

Wood, S.N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, Florida: Chapman & Hall/CRC.