Quantile regression analysis of survey data under informative sampling

Sixia Chen* and Yan Daniel Zhao †

Abstract

For complex survey data, the parameters in a quantile regression can be estimated by minimizing an objective function with units weighted by the original design weights. However, when the complex survey sampling design is informative, i.e., when the design weights are correlated with the study variable even after conditioning on other covariates, the efficiency of this design-weighted estimator may be improved. In this paper, we propose several weight smoothing estimators for quantile regression analysis of complex survey data collected with an informative sampling design. Our new estimators incorporate nonparametric methods for modeling the weight functions and pseudo-population bootstrap methods for variance estimation. We conducted a simulation study to compare our proposed methods with the original design-based method in terms of bias, standard error, mean squared error, and coverage. Our proposed estimators have smaller bias and mean squared error than does the design-based estimator. We further illustrate and compare estimators for the 1988 US National Maternal and Infant Health Survey.

Key words: Complex survey; Informative sampling; Nonparametric; Quantile regression; Weight smoothing.

1 Introduction

Researchers often use data collected from complex survey designs to draw scientific conclusions. For instance, Nelson et al. (2003) compared national estimates of smoking, height, and diabetes, by using the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS). Harrington et al. (2014)

*Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK 73104, U.S.A.

†Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK 73104, U.S.A.

used National Health Nutrition and Examination (NHANES) 2009/2010 to estimate the amount of time that the U.S. population spent sitting by age, sex, ethnicity, education, and body mass index. It is well known that statistical analysis ignoring design features, including stratification, clustering, and unequal weighting, may lead to biased results unless the sampling design is ignorable, see Pfeffermann and Sverchkov (1999), (2003), (2009), among others. Generalized linear and mixed models with complex survey designs were developed in Chambers and Skinner (2003) and Heeringa, West, and Berglund (2010).

The sampling design is informative when sample inclusion is related to the outcome variable conditional on covariates (Fuller, 2009). For such designs, survey weights are often used in regression analysis of survey data to ensure consistent estimation of parameters. For example, the sampling design of NHANES (2013-2014) was informative, since the first stage strata were built by using county-level health characteristics that are correlated with the study variables of interest, given the covariates. The traditional design-based approach, using the original design weights, leads to unbiased estimates, but the efficiency can be improved. One approach is a likelihood-based method that maximizes the conditional sample likelihood by using the joint model of study variable and sampling indicator, see Chambers (2003), Pfeffermann and Sverchkov (2009), Pfeffermann (2011) and Scott and Wild (2011), among others. A second approach replaces the original design weights by predictions from a model for the conditional distribution of the design weights given the data, as in Magee (1998), Pfeffermann and Sverchkov (1999), Beaumont (2008), Fuller (2009), and Kim and Skinner (2013). In particular, Kim and Skinner (2013) proposed optimal weight modifications compared with other methods under generalized linear models.

These statistical methods model the conditional mean values of the study variables by regression. Such models may be suboptimal when the distribution of study variables is skewed or has outliers. In such cases, quantile regression (QR) (Koenker and Basset, 1978, Koenker, 2005) is an effective tool for conditional modeling, providing robustness against outliers and a more comprehensive analysis of the relationship between variables than is offered by the conditional mean model.

There is rich literature on the use of QR for data collected by simple random sampling; for example, see He and Shao (1996), Knight (1998), Mu and He (2007), and references therein. Deaton (1997) and Cameron and Trivedi (2005) applied QR to survey data ignoring the complex sampling scheme, and their estimates may be biased if the original sampling design is informative. Only a few research papers discuss QR estimates that account for a complex survey sampling scheme, including Li et al. (2010) and Geraci (2016). These papers do not discuss QR for data collected using informative sampling, the topic of the present paper. The quantile regression coefficients are defined at the super-population level, and consistency depends on the quantile regression model assumption. Specifically, we extended several weight smoothing estimators, including unsmoothed and smoothed optimal estimators in Kim and Skinner (2013) and estimators proposed by Beaumont (2008) and Pfeffermann-Sverchkov (1999), to our data. Some of our proposed estimators (DW, PS, UOPT) are design consistent, even when the model (1) does not hold. Other estimators (SDW, SPS, SOPT) are consistent if the corresponding weight models are correct.

The remainder of the paper is organized as follows. After preliminaries in Section 2, our weight smoothing estimators are proposed and developed in Section 3. In Section 4, we describe algorithms for computing our proposed estimators. Variance estimation is presented in Section 5. A simulation study is described in Section 6, and a Real-Data-Based Simulation Study using the 1988 US National Maternal and Infant Health Survey is presented in Section 7. In Section 8, we conclude the paper with a discussion.

2 Quantile Regression and the Design-Based Estimator

Suppose the finite population $\mathcal{F}_N = \{(x_i, y_i, z_i), i = 1, 2, \ldots, N\}$ is generated from a super-population model $\mathcal{F}$, where x_i is a $p \times 1$ vector of covariates, y_i is the study variable, and z_i is the design variable, which may not be observed. We assume y_i

given x_i follow the following QR model:

$$y_i = x_i'\beta_\tau + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

where β_τ is $p \times 1$ unknown coefficient vector, and ϵ_i is an error term such that $\Pr(\epsilon_i \leq 0 | x_i) = \tau$ for the τ-th quantile ($0 < \tau < 1$). A complex survey sample S is drawn with sampling indicator I_i such that $I_i = 1$ if unit i is selected and 0 otherwise, $i = 1, \ldots, N$. The first- and second-order inclusion probabilities are denoted as $\pi_i = E(I_i)$ for selecting unit i and $\pi_{ij} = E(I_i I_j)$ for unit i and unit j. Consequently, the corresponding design weight for unit i is $d_i = \pi_i^{-1}$, which is known for units in S.

Following Kim and Skinner (2013), the sampling design is assumed to be informative in the sense that the design weights are functions of the covariates and the design variable, i.e., $\pi_i = \pi_i(X, Z)$ with $X = (x_1, x_2, \ldots, x_N)$ and $Z = (z_1, z_2, \ldots, z_N)$. In other words, the sampling will only be informative if y and z are related, conditional on x. Under the informative sampling, we have $\Pr(I_i = 1 | x_i, y_i) \neq \Pr(I_i = 1 | x_i)$, i.e., the selection of unit i depends not only on the covariates, but also on the study variable. The design-weighted estimator of the QR coefficents is

$$\hat{\beta}_{\tau,d} = \arg\min_{\beta} \sum_{i \in S} d_i \rho_\tau (y_i - x_i'\beta), \tag{2}$$

where $\rho_\tau(u) = u \{\tau - I(u < 0)\}$. By using an argument similar to that of Koenker (2005) and after some algebra, it can be shown that $\hat{\beta}_{\tau,d}$ is the solution of the following estimating equation

$$\hat{U}_d(\beta) = \sum_{i=1}^{N} I_i d_i \{\tau - I(y_i - x_i'\beta < 0)\} x_i = 0. \tag{3}$$

The estimator $\hat{\beta}_{\tau,d}$ is consistent for estimating β_τ, by an argument similar to that of Wang and Opsomer (2011). Its efficiency may be further improved by modeling the design weights as described in the next section.

3 Proposed methods

In this section, we propose five new weight smoothing estimators of QR coefficient β_τ in generalized linear models. Specifically, we consider estimators that satisfy the

following estimating equation by replacing the original design weight in Equation (3) with a modified weight w_i

$$\hat{U}_w(\beta) = \sum_{i=1}^{N} I_i w_i \left\{\tau - I(y_i - x_i'\beta < 0)\right\} x_i = 0, \tag{4}$$

where w_i are new weights chosen to improve the efficiency of the new estimators. All of the weight smoothing methods were initially developed for regression analysis of mean values of the study variable. We adapted these methods to our QR problem.

3.1 Smoothed Design-Weight (SDW) Estimator

Beaumont (2008) used a smoothing weight $E(d_i|y_i, I_i = 1)$ to estimate the population mean of y. Kim and Skinner (2013) extended the idea by using $w_i = \tilde{d}_{i,x,y} = E(d_i|x_i, y_i, I_i = 1)$ in the context of linear regression to obtain regression coefficient estimates. For our QR analysis, we use the same weights in Equation (4) and denote the corresponding estimator as $\hat{\beta}_{\tau,SDW}$. As in Kim and Skinner (2013), we can show that $\hat{\beta}_{\tau,SDW}$ is consistent and $V(\hat{\beta}_{\tau,SDW}) < V(\hat{\beta}_{\tau,d})$ if the conditional expectation $E(d_i|x_i, y_i, I_i = 1)$ is correctly modeled.

In general, $\tilde{d}_{i,x,y}$ in Equation (4) is unknown. To estimate $\tilde{d}_{i,x,y}$, one can use a parametric model, such as linear or non-linear regression method, or a non-parametric model, such as splines. However, the parametric model approach is vulnerable to model misspecification, and the non-parametric model approach is subject to the well-known curse of dimensionality if the dimension of covariate x is large. Instead, we fit the following generalized additive model (GAM) to estimate $\tilde{d}_{i,x,y}$ (Hastie and Tibshirani, 1990).

$$\log(d_i - 1) = g_0 + \sum_{t=1}^{p} g_t(x_{it}) + g_{p+1}(y_i) + e_i, \quad i \in S, \tag{5}$$

where x_{it} is the t-th variable in x_i, g_0 is an unknown parameter, $g_t, t = 1, \ldots, p+1$, are unknown functions that satisfy certain regularity conditions, and e_i is assumed to have normal distribution with mean 0 and variance σ^2. Model (5) is quite general and can be easily extended to more general cases with unequal variance and non-Gaussian exponential family distributions. For simplicity, we only consider models

(5) with lower-order spline functions and Gaussian errors with constant variance. After obtaining estimators $\hat{g}_0$, $\hat{g}_t, t = 1, \ldots, p+1$ and $\hat{\sigma}^2$, we estimate $\tilde{d}_{i,x,y}$ by using $\hat{\tilde{d}}_{i,x,y} = 1 + \exp(\hat{g}_0 + \sum_{t=1}^{p} \hat{g}_t(x_{it}) + \hat{g}_{p+1}(y_i) + \hat{\sigma}^2/2)$.

3.2 Unsmoothed Pfeffermann-Sverchkov (PS) Estimator and Smoothed Pfeffermann-Sverchkov (SPS) Estimator

Pfeffermann-Sverchkov (1999) proposed weights $w_i = d_i \hat{\tilde{d}}_{i,x}^{-1}$ where $\hat{\tilde{d}}_{i,x} = \hat{E}(d_i|x_i, I_i = 1)$ to produce efficient and consistent estimates of linear regression coefficients. We propose to obtain $\hat{\tilde{d}}_{i,x}$ by using a similar technique to that used to obtain $\hat{\tilde{d}}_{i,x,y}$ in section 3.1. The extension to quantile regression is trivial, and we denote the corresponding estimator as $\hat{\beta}_{\tau,PS}$. Note that $\hat{\beta}_{\tau,PS}$ is consistent even if the model $E(d_i|x_i, I_i = 1)$ is misspecified (see the justification for the consistency of UOPT estimator).

To further improve efficiency, Pfeffermann-Sverchkov (1999) proposed weights $w_i = \hat{\tilde{d}}_{i,x,y} \hat{\tilde{d}}_{i,x}^{-1}$, which yield a consistent and a more efficient estimator if the weight model $E(d_i|x_i, y_i, I_i = 1)$ is correctly specified. We denote this smoothed Pfeffermann-Sverchkov (SPS) estimator as $\hat{\beta}_{\tau,SPS}$.

Remark 3.1 *The SPS estimator minimizes the following prediction distance function*

$$Q(\beta) = \int \rho_\tau(y - x'\beta) f(y|x) dy. \tag{6}$$

Because

$$f(y|x, I = 1) = f(y|x) \frac{\Pr(I = 1|x, y)}{\Pr(I = 1|x)}$$

and

$$E(d|x, y, I = 1) = \frac{1}{E(\pi|x, y)}, \quad E(d|x, I = 1) = \frac{1}{E(\pi|x)},$$

a consistent estimator can be obtained by solving (4) with $w_i = E(d_i|x_i, y_i, I_i = 1) \{E(d_i|x_i, I_i = 1)\}^{-1}$.

3.3 Unsmoothed and Smoothed Optimal (UOPT and SOPT) Estimators

In this section, we propose two novel optimal weight modification estimators. Under the correct weight models, one will be more efficient than $\hat{\beta}_{\tau,PS}$, and the other will

be more efficient than $\hat{\beta}_{\tau,SDW}$ and $\hat{\beta}_{\tau,SPS}$. We assume $\pi_i = \pi(x_i, z_i)$ and the sampling design is Poisson; Kim and Skinner (2013) also made this assumption to derive the optimal weight in linear regression models.

Consider a class of estimators that solves (4) with $w_i = d_i q(x_i)$. The UOPT estimator is obtained by choosing $q_i = q(x_i)$ to minimize the variance of the following class of estimators

$$\hat{\beta}_{\tau,q} = \arg\min_{\beta} \sum_{i \in S} d_i q_i \rho_\tau(y_i - x_i'\beta), \tag{7}$$

or equivalently as the solution of the following estimating equations

$$\hat{U}_{dq}(\beta) = \sum_{i=1}^{N} I_i d_i q_i \left\{\tau - I(y_i - x_i'\beta < 0)\right\} x_i = 0. \tag{8}$$

According to Koenker (2005), we have $E\left\{\tau - I(y_i - x_i'\beta_\tau < 0)|x_i\right\} = 0$, so

$$\begin{aligned} E\left\{\hat{U}_{dq}(\beta_\tau)\right\} &= E\left[\sum_{i=1}^{N} I_i d_i q_i \left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\} x_i\right] \\ &= E\left[\sum_{i=1}^{N} q_i \left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\} x_i\right] \\ &= E\left[\sum_{i=1}^{N} q_i x_i E\left\{\tau - I(y_i - x_i'\beta_\tau < 0)|x_i\right\}\right] \\ &= 0. \end{aligned} \tag{9}$$

By the argument in Van der Vaart (1998, Ch. 5) and according to (9), it can be shown that $\hat{\beta}_{\tau,q}$ is consistent for β_τ for arbitrary $q_i = q(x_i)$, under mild regularity conditions. After some algebra, it can be shown that $\hat{\beta}_{\tau,q}$ has the following asymptotic expansion

$$\hat{\beta}_{\tau,q} = \beta_{\tau,q} + \left\{\sum_{i=1}^{N} q_i x_i x_i' f_{y|x}(x_i'\beta_\tau)\right\}^{-1} \hat{U}_{dq}(\beta_\tau) + o_p(n^{-1/2}) \tag{10}$$

and the corresponding asymptotic conditional variance can be written as

$$\left\{\sum_{i=1}^{N} q_i x_i x_i' f_{y|x}(x_i'\beta_\tau)\right\}^{-1} \sum_{i=1}^{N} E(d_i e_i^2|x_i) q_i^2 x_i x_i' \left\{\sum_{i=1}^{N} q_i x_i x_i' f_{y|x}(x_i'\beta_\tau)\right\}^{-1}, \tag{11}$$

where $f_{y|x}(x_i'\beta_\tau)$ is the conditional density of y given x evaluated at $x_i'\beta_\tau$ and $e_i = \tau - I(y_i - x_i'\beta_\tau < 0)$. Thus, $q_{i,1}^* = v_{i,1}^{-1} f_{y|x}(x_i'\beta_\tau)$ with $v_{i,1} = E(d_i e_i^2|x_i)$ minimizes the

variance defined in (11). Specifically, we have

$$\begin{aligned} v_{i,1} &= E(d_i e_i^2|x_i) \\ &= (\tau-1)^2 E\left(d_i|x_i; y_i < x_i'\beta_\tau\right) \Pr(y_i < x_i'\beta_\tau|x_i) \\ &+ \tau^2 E\left(d_i|x_i; y_i \geq x_i'\beta_\tau\right) \Pr(y_i \geq x_i'\beta_\tau|x_i). \end{aligned} \tag{12}$$

The estimator $\hat{q}_{i,1}^*$ is discussed in Section 4. Denote the estimator by using $w_i = d_i\hat{q}_{i,1}^*$ as $\hat{\beta}_{\tau,UOPT}$. It is easy to see that estimators $\hat{\beta}_{\tau,B}$ and $\hat{\beta}_{\tau,PS}$ belong to this class of estimators, so the UOPT estimator is more efficient.

For a more efficient estimator than $\hat{\beta}_{\tau,SDW}$ and $\hat{\beta}_{\tau,SPS}$, the SOPT estimator is obtained by minimizing variance for a class of estimators defined by $w_i = \tilde{d}_{i,x,y}q(x_i)$. By arguments similar to those for UOPT, the corresponding estimators are consistent, since

$$\begin{aligned} E\left\{\hat{U}_w(\beta_\tau)\right\} &= E\left[\sum_{i=1}^N I_i\tilde{d}_{i,x,y}q(x_i)\left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\}x_i\right] \\ &= E\left[E\left[\sum_{i=1}^N I_i d_i q(x_i)\left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\}x_i|x,y\right]\right] \\ &= E\left[\sum_{i=1}^N I_i d_i q(x_i)\left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\}x_i\right] \\ &= E\left[\sum_{i=1}^N q(x_i)\left\{\tau - I(y_i - x_i'\beta_\tau < 0)\right\}x_i\right] \\ &= E\left[\sum_{i=1}^N q_i x_i E\left\{\tau - I(y_i - x_i'\beta_\tau < 0)|x_i\right\}\right] \\ &= 0. \end{aligned} \tag{13}$$

Under the correct weight models, SOPT is even more efficient than UOPT, as seen in our simulation studies. By using similar techniques to those used for the UOPT estimator, it can be shown that the optimal choice of q_i is $q_{i,2}^* = \tilde{v}_{i,2}^{-1}f_{y|x}(x_i'\beta_\tau)$ with $\tilde{v}_{i,2} = E(\tilde{d}_i e_i^2|x_i)$. Specifically, we have

$$\begin{aligned} \tilde{v}_{i,2} &= E(\tilde{d}_i e_i^2|x_i) \\ &= (\tau-1)^2 E\left(\tilde{d}_i|x_i; y_i < x_i'\beta_\tau\right) \Pr(y_i < x_i'\beta_\tau|x_i) \\ &+ \tau^2 E\left(\tilde{d}_i|x_i; y_i \geq x_i'\beta_\tau\right) \Pr(y_i \geq x_i'\beta_\tau|x_i). \end{aligned} \tag{14}$$

We discuss how to obtain the estimator $\hat{q}^*_{i,2}$ of $q^*_{i,2}$ in Section 4. We denote the estimator by using $w_i = \hat{\tilde{d}}_{i,x,y}\hat{q}^*_{i,2}$ as $\hat{\beta}_{\tau,SOPT}$.

4 Algorithms for Computing the UOPT and SOPT Estimators

In this section, we discuss algorithms for computing the UOPT and SOPT estimators by the generalized additive model (GAM) approach. The UOPT estimator $q^*_{i,1}$ can be estimated by the following steps:

1. Set $\hat{\beta}^{(0)}_\tau = \hat{\beta}_{\tau,d}$, the estimator in equation (3).

2. Estimate $\hat{f}_{y|x}$ by using the GAM approach and assuming a normal distribution of y_i, where $\hat{f}_{y|x}$ is the estimated conditional density of y given x. The conditional expectation $E(y|x)$ is assumed to be additive in each component of x_i and their lower order interactions, including second- and third- order terms.

3. Estimate $E(d_i|x_i; y_i < x_i'\beta_\tau)$ by

$$\hat{E}(d_i|x_i; y_i < x_i'\beta_\tau) = 1 + \exp\left\{\hat{g}_{01} + \sum_{t=1}^{p}\hat{g}_{t1}(x_{it}) + \hat{\sigma}_1^2/2\right\},$$

   by assuming the following generalized additive model (GAM):

$$\log(d_i - 1) = g_{01} + \sum_{t=1}^{p} g_{t1}(x_{it}) + e_{1i}, \quad i \in S_1, \tag{15}$$

   where S_1 denotes the units in S such that $y_i < x_i'\hat{\beta}^{(t)}_\tau$ and using similar techniques as the estimation of $\tilde{d}_{i,x,y}$ described in section 3.1. Estimate $\hat{E}(d_i|x_i; y_i \geq x_i'\beta_\tau)$ by similar techniques. Estimate $\hat{\Pr}(y_i < x_i'\beta_\tau|x_i)$ and $\hat{\Pr}(y_i \geq x_i'\beta_\tau|x_i)$ by substituting the estimated density $\hat{f}_{y|x}$ in Step 2. Then, according to (12),

$$\begin{aligned}\hat{v}^{(t)}_{i,1} &= (\tau-1)^2\hat{E}\left(d_i|x_i; y_i < x_i'\beta_\tau\right)\hat{\Pr}(y_i < x_i'\beta_\tau|x_i) \\ &+ \tau^2\hat{E}\left(d_i|x_i; y_i \geq x_i'\beta_\tau\right)\hat{\Pr}(y_i \geq x_i'\beta_\tau|x_i).\end{aligned} \tag{16}$$

4. Estimate $\hat{q}^{*(t)}_{i,1} = \hat{v}^{(t)-1}_{i,1}\hat{f}_{y|x}(x_i'\hat{\beta}^{(t)}_\tau)$ and the corresponding optimal estimator $\hat{\beta}^{(t+1)}_\tau$ by solving equation (8) with $\hat{q}^{*(t)}_{i,1}$.

5. Repeat Step 3 to Step 4 with updated estimator $\hat{\beta}_\tau^{(t+1)}$ until convergence.

The SOPT estimator $q_{i,2}^*$ is obtained as follows:

1. Same as Step 1 for UOPT estimator.
2. Estimate $\hat{\tilde{d}}_{i,x,y}$ of $\tilde{d}_{i,x,y}$ by using the GAM approach described in section 3.1.
3. Same as Step 2 for the UOPT estimator.
4. Same as Step 3 for the UOPT estimator, with d_i replaced by $\hat{\tilde{d}}_{i,x,y}$ in the model.
5. Estimate $\hat{q}_{i,2}^{*(t)} = \hat{v}_{i,2}^{(t)-1}\hat{f}_{y|x}(x_i'\hat{\beta}_\tau^{(t)})$ and the corresponding optimal estimator $\hat{\beta}_\tau^{(t+1)}$ by solving equation (4) with $w_i = \hat{\tilde{d}}_{i,x,y}\hat{q}_{i,2}^{*(t)}$.
6. Repeat Steps 4 and 5 until convergence.

5 Variance estimation

We now describe bootstrap estimates of variance for our proposed estimators, with associated confidence regions. The Taylor linearization approach to variance estimation involves tedious technical derivation, especially when our proposed estimation procedure includes semi-parametric methods.

We apply pseudo-population bootstrap methods (Gross, 1980, Booth et al. 1994, Conti et al. 2017), which are simple and practical and have been shown to work effectively under high entropy designs (Conti et al. 2017), such as Rao-Sampford (Rao, 1965; Sampford, 1967) and randomized proportional-to-size systematic sampling. Our proposed bootstrap method can be described as follows:

1. For $k = 1, \ldots, N$, choose a unit i from the original sample S independently with probability $\pi_i^{-1}/\sum_{j\in S}\pi_j^{-1}$. If at trial k the unit $i \in S$ is selected, define $(x_k^*, y_k^*, z_k^*) = (x_i, y_i, z_i)$.
2. The pseudo-bootstrap population is then $\mathcal{F}_N^* = \{(x_k^*, y_k^*, z_k^*), k = 1, \ldots, N\}$. Draw a bootstrap sample S^* from $\mathcal{F}_N^*$ by using the same design as the original design with first-order inclusion probabilities $nz_k^*/\sum_{i=1}^N z_i^*$. If z_i is unknown,

then one can use π_k^*, which is the corresponding original inclusion probability for the k-th element in the pseudo bootstrap population.

3. Obtain the bootstrap sample estimator $\hat{\beta}_\tau^*$ from S^* by using our proposed method.

Generate B bootstrap samples by above procedure, with corresponding estimators $\hat{\beta}_\tau^{*(b)}$ for $b = 1, \ldots, B$. Then, the bootstrap variance estimator is:

$$\hat{V}^* = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_\tau^{*(b)} - \bar{\hat{\beta}}_\tau^*)(\hat{\beta}_\tau^{*(b)} - \bar{\hat{\beta}}_\tau^*)', \tag{17}$$

where $\bar{\hat{\beta}}_\tau^* = B^{-1}\sum_{b=1}^{B}\hat{\beta}_\tau^{*(b)}$. The $(1-\alpha)100\%$ confidence region of β_τ can be written as

$$(\hat{\beta}_\tau - \beta_\tau)\hat{V}^{*-1}(\hat{\beta}_\tau - \beta_\tau)' < \chi^2_{p,1-\alpha}, \tag{18}$$

where $\chi^2_{q,1-\alpha}$ is the $(1-\alpha)100$-th percentile of a chi-square distribution with degrees of freedom q, the dimension of β_τ. Alternatively, one can use bootstrap percentiles of statistics $(\hat{\beta}_\tau^{*(b)} - \beta_\tau^{*(b)})\hat{V}^{*-1}(\hat{\beta}_\tau^{*(b)} - \beta_\tau^{*(b)})'$ to obtain the confidence region. For inference with individual parameter $\beta_{\tau,a}$ defined in β_τ where $a = 1, \ldots, p$, one can use the following normal-based confidence interval:

$$\left(\hat{\beta}_{\tau,a} - z_{1-\alpha/2}\sqrt{\hat{V}^*_{aa}}, \quad \hat{\beta}_{\tau,a} + z_{1-\alpha/2}\sqrt{\hat{V}^*_{aa}}\right), \tag{19}$$

where $\hat{\beta}_{\tau,a}$ is the a-th component of $\hat{\beta}_\tau$ and $\hat{V}^*_{aa}$ is the corresponding estimated variance of $\hat{\beta}_{\tau,a}$.

6 Simulation study

We now compare the performance of all six estimators in a simulation study. We generated $M = 1,000$ finite populations with population size $N = 10,000$ from the following population model: $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + (1 + \psi_1 x_{1i} + \psi_1 x_{2i})\epsilon_i$, where $(\beta_0, \beta_1, \beta_2) = (1, -1, -0.5)$, covariates (x_{1i}, x_{2i}) were independently and identically distributed (iid) with a normal distribution with means $E(x_{1i}) = E(x_{2i}) = 0$ and variances $V(x_{1i}) = V(x_{2i}) = 1$, and ϵ_i were iid with a standard normal distribution.

The parameter ψ_1 was set to zero (homoscedastic variance) or 0.2 (heteroscedastic variance).

For each generated finite population of size N, a Poisson sample was then selected with inclusion probabilities $\pi_i = nk_i/(\sum_{j=1}^{N} k_j)$, where $n = 400$ was the expected sample size and k_i was the size variable such that $k_i = \{1 + \exp(2.5 - 0.5z_i)\}^{-1}$ and $z_i \sim N(1 + y_i, 0.5^2)$. Note that this sampling was informative, because the inclusion probabilities depended on the outcome variable y. Specifically, the correlation between ϵ and π was about 0.6. Sample sizes varied, but were all close to 400.

For the population model, it can be shown that the τ-th conditional quantile of y_i is $Q_\tau(y_i|x_i) = \beta_{0\tau} + x_{1i}\beta_{1\tau} + x_{2i}\beta_{2\tau}$, where $\beta_{0\tau} = \beta_0 + Q_\tau(\epsilon_i)$, $\beta_{1\tau} = \beta_1 + \psi_1 Q_\tau(\epsilon_i)$, $\beta_{2\tau} = \beta_2 + \psi_1 Q_\tau(\epsilon_i)$ with $Q_\tau(\epsilon_i)$ as the τ-th quantile of ϵ_i, which could be readily calculated. Our parameters of interest were the QR regression coefficients $\beta_\tau = (\beta_{1\tau}, \beta_{2\tau})$. In the simulation study, $\tau = 0.4$ and 0.6 were considered.

We compared the six estimators described above in terms of MC relative bias (RBias), MC relative standard error (RSE), MC relative root mean squared error (RRMSE), and MC coverage properties, including MC coverage probability (CP), standard error relative bias (SERBias), and relative average confidence interval length (RCILen). The formulas for those quantities are as follows:

$$\text{RBias} = \frac{\hat{\beta} - \beta}{|\beta|},$$

$$\text{RSE} = \frac{\left\{(M-1)^{-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2\right\}^{1/2}}{|\beta|},$$

$$\text{RRMSE} = \frac{\left\{\left(\hat{\beta} - \beta\right)^2 + (M-1)^{-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2\right\}^{1/2}}{|\beta|},$$

$$\text{CP} = \frac{1}{M}\sum_{m=1}^{M} I(LB^{(m)} < \beta < UB^{(m)}),$$

$$\text{SERBias} = \frac{M^{-1}\sum_{m=1}^{M}\left\{\hat{V}^{(m)}\right\}^{1/2} - \left\{(M-1)^{-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2\right\}^{1/2}}{\left\{(M-1)^{-1}\sum_{m=1}^{M}\left(\hat{\beta}^{(m)} - \hat{\beta}\right)^2\right\}^{1/2}},$$

$$\text{RCILen} = \frac{M^{-1}\sum_{m=1}^{M}(UB^{(m)} - LB^{(m)})}{|\beta|},$$

where β represents the true value for parameters $\beta_{1\tau}$ or $\beta_{2\tau}$, $\hat{\beta}^{(m)}$ represents the estimator based on the m-th Monte Carlo sample for β, $\hat{\beta} = M^{-1}\sum_{m=1}^{M}\hat{\beta}^{(m)}$, $LB^{(m)}$ and $UB^{(m)}$ represent the lower and upper 95% confidence interval bounds for β based on the formula (19) in section 5, and $\hat{V}^{(m)}$ represents our proposed bootstrap variance estimator based on the m-th Monte Carlo sample. We selected 200 bootstrap samples for variance estimation for each MC sample.

The point estimation results are presented in Table 1 for $\psi_1 = 0$ and Table 2 for $\psi_1 = 0.2$. Under the homoscedastic scenario in Table 1, all estimators had small RBias, which was consistent with the underlying theorem. The DW and SDW estimators had the largest RRMSE, since the DW estimator did not use any smoothing technique to reduce variance, and the single smoothing model in the SDW estimator was not efficient. The UOPT, PS, SPS, and SOPT estimators had comparable RSE and RRMSE. To test the sensitivity of model specification, we assumed equal variance structure under the heteroscedastic scenario; the results were comparable with assuming the correct heteroscedastic variance structure. As shown in Table 2, all estimators had small bias for most of the cases. The UOPT and SOPT estimators had significantly smaller RSE and RRMSE than did other estimators, for all cases. For simplicity, we only presented the results for coverage properties for the scenario in which $\psi_1 = 0.2$ and $\tau = 0.6$ (Table 3). Other scenarios had similar results. The UOPT and SOPT estimators had better or comparable coverage to other estimators for most of the cases, and their CP were close to the nominal level of 95%. The SERBias for all estimators based on our proposed bootstrap methods were less than 8.8%, which verified our proposed variance estimation approach. The DW and SDW estimators had larger RCILen than did other estimators. The UOPT and SOPT estimators had RCILen smaller than that of the PS and SPS estimators. We also considered the scenario where the correlation between ϵ and π is about 0.3, and the results were similar (results not presented here).

Table 1: The Monte Carlo relative bias (RBias) ($\times 10^3$), relative standard error (RSE) ($\times 10^3$), and relative root mean squared error (RRMSE) ($\times 10^3$) for six different methods with $\psi_1 = 0$.

Tau	Par	Method	RBias	RSE	RRMSE
0.4	$\beta_{1\tau}$	DW	-1	94	94
		UOPT	5	79	80
		SDW	2	90	90
		PS	2	79	79
		SPS	5	77	77
		SOPT	8	78	78
	$\beta_{2\tau}$	DW	7	169	169
		UOPT	4	149	149
		SDW	10	162	162
		PS	2	148	148
		SPS	6	143	143
		SOPT	8	145	145
0.6	$\beta_{1\tau}$	DW	-1	81	81
		UOPT	6	68	68
		SDW	2	78	78
		PS	4	68	68
		SPS	8	67	67
		SOPT	9	68	68
	$\beta_{2\tau}$	DW	-1	150	150
		UOPT	5	133	133
		SDW	4	145	145
		PS	3	131	131
		SPS	8	127	127
		SOPT	9	129	129

Table 2: The Monte Carlo relative bias (RBias) ($\times 10^3$), relative standard error (RSE) ($\times 10^3$), and relative root mean squared error (RRMSE) ($\times 10^3$) for six different methods with $\psi_1 = 0.2$.

Tau	Par	Method	RBias	RSE	RRMSE
0.4	$\beta_{1\tau}$	DW	1	91	91
		UOPT	9	54	55
		SDW	5	89	89
		PS	6	60	60
		SPS	10	59	60
		SOPT	11	55	56
	$\beta_{2\tau}$	DW	3	155	155
		UOPT	9	104	104
		SDW	9	152	152
		PS	6	114	114
		SPS	12	110	111
		SOPT	12	102	102
0.6	$\beta_{1\tau}$	DW	-1	86	86
		UOPT	7	53	54
		SDW	4	83	83
		PS	5	56	56
		SPS	8	55	56
		SOPT	10	54	55
	$\beta_{2\tau}$	DW	-3	155	155
		UOPT	9	112	113
		SDW	3	152	152
		PS	5	116	116
		SPS	10	113	114
		SOPT	13	110	111

Table 3: The coverage probability (CP) ($\times 10^3$), standard error relative bias (SERBias) ($\times 10^3$), and relative average confidence interval length (RCILen) ($\times 10^3$) for six different methods with $\psi_1 = 0.2$ and $\tau = 0.6$.

Par	Method	CP	SERBias	RCILen
$\beta_{1\tau}$	DW	948	-6	337
	UOPT	953	88	228
	SDW	939	14	330
	PS	954	64	234
	SPS	952	65	230
	SOPT	949	66	225
$\beta_{2\tau}$	DW	960	61	646
	UOPT	949	65	469
	SDW	959	60	631
	PS	945	60	483
	SPS	950	67	474
	SOPT	952	74	462

7 Real-Data-Based Simulation Study

We further compare our estimators on a real data set previously analyzed by Korn and Graubard (1995) and Pfeffermann and Sverchkov (1999). The data were collected as part of the 1988 U.S. National Maternal and Infant Health Survey, which used a stratified random sample of vital records corresponding to live births, late fetal deaths, and infant deaths in the United States. The strata were constructed using the mother's race and child's birth weight, and the sampling fractions varied according to strata.

Pfeffermann and Sverchkov (1999) treated birth weight (measured in grams) as the study variable Y and gestational age (measured in weeks) as the predictor X. After deleting 506 observations with missing values, the finite population size was reduced to $9,447$. One can fit the following linear regression model using the finite population and obtain the estimated model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, 9447, \tag{20}$$

with $\beta_0 = -2695.27$ and $\beta_1 = 149.04$. The p-values for all regression coefficients were highly significant ($p < .0001$). The R^2 value was about 0.6. The original design

was informative because the strata were determined using the study variable birth weight. The correlation between d_0 and $\hat{\epsilon}$ was 0.32, where d_0 was the original design weight in the survey and $\hat{\epsilon}$ was the estimated residuals obtained from model (20). In other words, even after adjusting for predictor variable gestational age, a correlation remained between the design weights and the study variable.

Rather than the mean model described in (20), we estimated the quantile regression of Y on X, with parameters of interest the τth quantile regression coefficients $\beta_{0\tau}$ and $\beta_{1\tau}$. Before conducting the simulation, we first fit the mean regression model, as well as quantile regression models with $\tau = 0.2$, 0.4, 0.6, and 0.8. The results are presented in Figure 1. From Figure 1, it is clear that the quantile regression fitted lines are not parallel, unlike the conventional homoscedastic mean regression model. This result suggests the skewness of distribution for Birth Weight and shows that quantile regression provides a more comprehensive analysis.

For the simulation, we chose $\tau = 0.8$ for illustration. We conducted $1,000$ Monte Carlo simulations to compare the six quantile regression coefficient estimators. In each simulation, one sample was generated from the finite population with an expected sample size of 400 by using the Poisson sampling design with inclusion probability $\pi_i = 400 d_{0,i}^{-1} / \sum_{j=1}^{N} d_{0,j}^{-1}, i = 1, \ldots, N$, where $d_{0,j}$ was the design weight for the jth subject in the finite population. The bootstrap size was set to 200 for variance estimation for all six estimators, and 95% confidence intervals were constructed.

Before comparing the performance of our proposed estimators for quantile regression coefficients, we first compared the performance of the design-based estimator of median regression coefficients and mean regression coefficients, using only the linear term for the purpose of illustration. The purpose of this comparison was to show that there is an advantage in using quantile regression instead of mean regression for data with certain features. The results in Table 4 show that the estimators of median regression coefficients have smaller relative bias, relative standard error and relative root mean squared error than do the estimators of mean regression. This occurs because the distribution of residual terms displays some skewness and the Kolmogorov-Smirnov test rejects the normality assumption ($p < 0.05$). Furthermore,

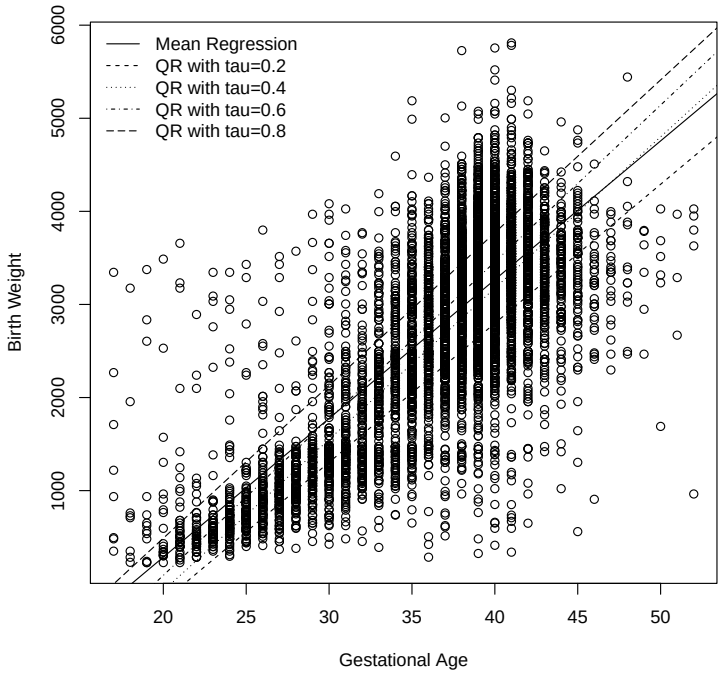

Figure 1: Mean regression and quantile regression models

there was some heteroscedastic trend in variance.

Table 5 summarizes the simulation results, comparing the performance of all six estimators. For point estimation, the DW and SDW estimators had larger RBias than did the other estimators. The DW estimator had the largest RSE and RRMSE for all cases as expected, because the efficiency of the DW estimator is improved through weight smoothing. The SOPT estimator had the smallest RSE and RRMSE. The SPS estimator was the second-best estimator in terms of RSE and RRMSE. The UOPT estimator had similar RRMSE to that of the PS estimator, as in Kim and Skinner (2013). All confidence coverages were close to the nominal rate of 95%.

Table 4: The Monte Carlo relative bias (RBias), relative standard error (RSE), and relative root mean squared error (RRMSE) for comparing mean regression with median regression.

Parameters	Method	RBias	RSE	RRMSE
β_0	Mean	-0.013	0.194	0.195
$\beta_{0\tau}$	Median	0.000	0.116	0.116
β_1	Mean	0.005	0.097	0.097
$\beta_{1\tau}$	Median	0.001	0.071	0.071

Table 5: The Monte Carlo relative bias (RBias), relative standard error (RSE), and relative root mean squared error (RRMSE) for six different methods.

Parameters	Method	RBias	RSE	RRMSE
$\beta_{0\tau}$	DW	0.063	0.285	0.292
	UOPT	-0.021	0.223	0.224
	SDW	0.081	0.228	0.242
	PS	-0.043	0.213	0.218
	SPS	-0.001	0.199	0.199
	SOPT	0.001	0.182	0.182
$\beta_{1\tau}$	DW	-0.027	0.128	0.131
	UOPT	0.005	0.100	0.100
	SDW	-0.035	0.102	0.108
	PS	0.018	0.097	0.098
	SPS	-0.002	0.089	0.089
	SOPT	-0.002	0.082	0.082

8 Discussion

In this paper, we proposed several weight smoothing estimators for estimating quantile regression coefficients in complex surveys under informative sampling design. Our proposed estimators were compared in terms of point estimation and variance estimation by using both simulated data and a Real-Data-Based Simulation Study. All proposed estimators have smaller standard errors than the original design-based estimator. Unsmoothed and smoothed optimal estimators showed a better balance of variance, bias, and coverage rate, compared with other estimators. Smoothed estimators, based on nonparametric weight smoothing models, outperformed unsmoothed estimators. All related R codes, as well as an example data file, are posted at the following website: https://github.com/yandzhao/Quantile-Regression-of-Survey-Data. For future research, we will consider estimating quantile regression coefficients with a clustered informative sampling design.

Acknowledgement

The authors sincerely thank Professor Danny Pfeffermann and Dr. Michael Sverchkov for sharing the 1988 US National Maternal and Infant Health Survey data with us. This work was supported partially by the funding provided by National Institutes of Health, National Institute of General Medical Sciences (Grant 1 U54GM104938), an IDeA-CTR to the University of Oklahoma Health Sciences Center.

References

Antal, E., and Y. Tille (2011), "A direct bootstrap method for complex sampling designs from a finite population," *Journal of the American Statistical Association*, **106**, 534-543.

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

Booth, J. G., Butler, R. W., and Hall, P. (1994), "Bootstrap methods for finite populations," *Journal of the American Statistical Association*, **89**, 1282-1289.

Cameron, A. C. and Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chambers, R. L., and Skinner, C. J. (2003), *Analysis of Survey Data*. Chichester: Wiley.

Chambers, R. L. (2003), Introduction to part A. In *Analysis of Survey Data*, Ed. R. L. Chambers and C. J. Skinner. Chichester: Wiley.

Conti, P. L., Marelia, D., and Mecatti, F. (2017), "Recovering sampling distributions of statistics of finite populations via resampling: a predictive approach," *submitted*.

Deaton, A. (1997), *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Johns Hopkins University Press.

Fuller, W. (2009), *Sampling Statistics*, Hoboken: Wiley.

Geraci, M. (2016), "Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants," *Statistical Methods in Medical Research*, available online.

Gross, S. (1980), "Median estimation in sample surveys," In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 181-184.

Harrington, D. M., Barreira, T. V., Staiano, A. E., and Katzmarzyk, P. T. (2014), "The descriptive epidemiology of sitting among US adults, NHANES 2009/2010," *Journal of Science Medicine in Sport*, **17**, 371-375.

Hastie, T., and Tibshirani. R. (1990), *Generalized Additive Models*. New York: Chapman and Hall.

Heeringa, S. G., West, B. T, and Berglund, P. A. (2010), *Applied Survey Data Analysis*. Boca Raton, FL: Taylor and Francis Group.

He, X., and Shao, Q. (1996), "A general Bahadur representation of M -estimators and its application to linear regression with nonstochastic designs," *The Annals of Statistics*, **24**, 2608-2630.

Kim, J. K., and Skinner, C. J. (2013), "Weighting in survey analysis under informative sampling," *Biometrika*, **100**, 385-398.

Knight, K. (1998), "Limiting distribution for L1 regression estimators under general conditions," *The Annals of Statistics*, **26**, 755-770.

Koenker, R., and Bassett, G. (1978), "Regression quantiles," *Econometrica*, **46**, 33-50.

Koenker, R. (2005), *Quantile Regression*. Cambridge.

Korn, E. L., and Graubard, B. I. (1995), "Examples of differing weighted and unweighted estimates from a sample survey," *The American Statistician*, **49**, 291-295.

Li, Y., Graubard, B. I., and Korn, E. L. (2010), "Application of nonparametric quantile regression to body mass percentile curves from survey data," *Statistics in Medicine*, **29**, 558-572.

Magee, L. (1998), "Improving survey-weighted least squares regression," *Journal of Royal Statistical Society, Series B*, **60**, 115-126.

Mu, Y., and He, X. (2007), "Power transformation toward a linear regression quantile," *Journal of the American Statistical Association*, **102**, 269-279.

Nelson, D. E., Powell-Griner, E., Town, M., and Kovar, M. G. (2003), "A comparison of national estimates from the National Health Interview Survey and the Behavioral Risk Factor Surveillance System," *American Journal of Public Health*, **93**, 1335-1341.

Pfeffermann, D., and Sverchkov, M. Y. (1999), "Parametric and semi-parametric estimation of regression models fitted to survey data," *Sankhya B*, **61**, 166-186.

Pfeffermann, D., and Sverchkov, M. Y. (2003), "Fitting generalized linear models under informative sampling," In *Analysis of Survey Data*, Ed. R. L. Chambers and C. J. Skinner. Chichester: Wiley.

Pfeffermann, D., and Sverchkov, M. Y. (2009), "Inference under informative sampling," In *Handbook of Statistics 29B; Sample surveys: Inference and Analysis*, Ed. D. Pfeffermann and C. R. Rao. Amsterdam: North Holland.

Pfeffermann, D. (2011), "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?" *Survey Methodology*, **37**, 115-136.

Rao, J. N. K. (1965), "On two simple schemes of unequal probability sampling without replacement," *Journal of the Indian Statistical Association*, **3**, 173-180.

Rao, J. N. K., and Wu, C. F. J. (1988), "Resampling inference with complex survey data," *Journal of the American Statistical Association*, **83**, 231241.

Sampford, M. R. (1967), "On sampling without replacement with unequal probabilities of selection," *Biometrika*, **54(3-4)**, 499-513.

Scott, A., and Wild, C. (2011), "Fitting regression models with response-biased samples," *Canadian Journal of Statistics*, **39**, 519-536.

Shao, J., and Tu, D. (1995), *The jackknife and bootstrap*, New York: Springer-Verlag.

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Wang, J. Q., and Opsomer, J. D. (2011), "On asymptotic normality and variance estimation for nondifferentiable survey estimators," *Biometrika*, **98**, 91-106.

Wolter, K. M. (2007), *Introduction to variance estimation*, New York: Springer-Verlag.