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

Survey Methodology

Semiparametric quantile regression imputation for a complex survey with application to the Conservation Effects Assessment Project

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005

VOLUME 31

NUMBER 1

Canada

by Emily Berg and Cindy Yu

Release date: June 27, 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.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 249-270
Statistics Canada, Catalogue No. 12-001-X

Semiparametric quantile regression imputation for a complex survey with application to the Conservation Effects Assessment Project

Emily Berg and Cindy Yu[1]

Abstract

Development of imputation procedures appropriate for data with extreme values or nonlinear relationships to covariates is a significant challenge in large scale surveys. We develop an imputation procedure for complex surveys based on semiparametric quantile regression. We apply the method to the Conservation Effects Assessment Project (CEAP), a large-scale survey that collects data used in quantifying soil loss from crop fields. In the imputation procedure, we first generate imputed values from a semiparametric model for the quantiles of the conditional distribution of the response given a covariate. Then, we estimate the parameters of interest using the generalized method of moments (GMM). We derive the asymptotic distribution of the GMM estimators for a general class of complex survey designs. In simulations meant to represent the CEAP data, we evaluate variance estimators based on the asymptotic distribution and compare the semiparametric quantile regression imputation (QRI) method to fully parametric and nonparametric alternatives. The QRI procedure is more efficient than nonparametric and fully parametric alternatives, and empirical coverages of confidence intervals are within 1% of the nominal 95% level. An application to estimation of mean erosion indicates that QRI may be a viable option for CEAP.

Key Words: Informative sample design; B-spline; Erosion.

1 Introduction

Missing data have important implications for analyses of survey data. Missing data can arise because sampled units refuse to participate in the survey, are difficult to locate, do not respond to sensitive questions, or drop out of longitudinal studies. If the missing values are related to the variable of interest, an analysis of the complete data with no modification for missing values, is biased. Weighting and imputation are two broad classes of missing data adjustments.

Two types of weighting adjustments are calibration (D’arrigo and Skinner, 2010 and Kott, 2006) and propensity score estimation (Kim and Riddles, 2012). In calibration, the weights for the respondents are adjusted so that the weighted sum of an auxiliary variable for the respondents is equal to the corresponding mean for the full sample or a population mean. In propensity score estimation, the sampling weight is multiplied by the inverse of an estimated response probability.

Imputation completes the data set, replacing missing response variables with imputed values. Imputation can simplify analyses in the presence of item nonresponse and improve consistency in results across users. We consider imputation of a response y, which may be missing, using an auxiliary variable x that is observed for the full sample. To allow flexibility in the model assumptions, we use a semiparametric quantile regression model to describe the relationship between x and y.

[1] Emily Berg, Department of Statistics, Iowa State University. E-mail: emilyb@iastate.edu; Cindy Yu, Department of Statistics, Iowa State University.

A diverse range of imputation procedures exists (Kim and Shao, 2013). Parametric fractional imputation (Kim, 2011) and parametric multiple imputation (Rubin, 2004) generate imputed values from an estimate of a fully parametric model for the conditional distribution of the response given covariates. Hot deck imputation (i.e., Andridge and Little, 2010), in contrast, includes, a class of nonparametric procedures in which imputed values are selected from respondents. In some hot deck procedures, weights are assigned according to a proximity measure, defined by imputation classes (Brick and Kalton, 1996) or a metric (Rubin, 2004; Little, 1988) such as a kernel distance (Wang and Chen, 2009). Nonparametric imputation is more robust to model misspecification than fully parametric methods, but estimators based on nonparametric procedures can have poor efficiency in small samples. Semiparametric quantile regression imputation (QRI) is a compromise between nonparametric and fully parametric imputation procedures. In QRI, the imputed values for a single missing value are the estimated quantiles of the distribution of the missing observation conditional on a function of auxiliary variables. Because a semiparametric model for the quantile function is used, QRI is robust to model misspecification, and because values are imputed from estimated quantiles, QRI is resistant to extreme values. Chen and Yu (2016) develop QRI for simple random sampling from an infinite population. We extend Chen and Yu (2016) to allow unequal selection probabilities.

Many imputation procedures rely on a missing at random (MAR) assumption (Rubin, 1976). A common assumption is that the response variable $(y,$ which may be missing) is conditionally independent of the missing indicator (1 if a response is provided and 0 otherwise) given the observed data. A direct application of this MAR definition to a complex survey specifies independence of the response variable and missing indicator variable conditional on the auxiliary variable and the sample inclusion indicators (Little, 1982; Pfeffermann, 2011). Berg, Kim and Skinner (2016) call the missing at random assumption that is defined conditional on the sample inclusion indicators sample missing at random. An alternative assumption, called population missing at random (Berg et al., 2016), is that the response variable is conditionally independent of the missing indicator given the auxiliary variable in the superpopulation, unconditional on the sample inclusion indicators. Berg et al. (2016) show that these two assumptions are not equivalent. We discuss these MAR concepts precisely in Section 2 and develop our procedure to be sufficiently flexible to accommodate either condition.

Our interest in semiparametric quantile regression for a complex survey is motivated in part by the Conservation Effects Assessment Project (CEAP), a complex survey intended to quantify soil and nutrient loss from crop fields. Because distributions of the response variables are highly skewed and contain extreme values, specification of an adequate fully parametric imputation model is difficult, and hot deck imputation procedures may have large variances. We investigate the use of QRI to address these issues in imputation for CEAP.

We demonstrate the theoretical validity and applicability of semiparametric quantile regression imputation in the context of a complex survey. Section 2 and Section 3, respectively, present the imputation algorithm and asymptotic properties. Section 4 and Section 5 demonstrate the properties of QRI through the

CEAP application and simulations, respectively. Section 6 concludes with a summary and a discussion of areas for future research.

2 Quantile regression imputation for complex survey data

Consider a conceptual framework in which samples are drawn from a finite population generated from a superpopulation model (Fuller, 2009b, Chapter 6). Let x_i and y_i have joint distribution $f(x_i, y_i)$ in the superpopulation. We define the conditional distribution of y_i given x_i through the conditional quantile function. Let $q_\tau(x_i)$ denote the τ^{th} quantile of the conditional distribution of y_i given x_i in the superpopulation, where $q_\tau(x_i)$ is defined by

$$P(y_i \le q_\tau(x_i) \mid x_i) = \tau. \tag{2.1}$$

We specify a model for the quantiles because quantile regression models can describe a wide variety of distributions, as illustrated in Figure 2.1. The left panel of Figure 2.1 depicts a linear quantile regression model in which each conditional quantile function is represented with a different intercept and a different slope. The use of a different slope allows describing data with nonconstant variances. The right panel of Figure 2.1 illustrates a generalization to semiparamtric quantile regression, where the τ^{th} quantile of the conditional distribution of y_i is represented as a continuous function of x_i. In the imputation procedure, we assume $q_\tau(\cdot)$ is a function with $p+1$ continuous derivatives. We approximate $q_\tau(x_i)$ with a B-spline (de Boor, 2001; Chen and Yu, 2016; Yoshida, 2013; Hastie, Tibshirani and Friedman, 2009), as we explain in more detail in Section 2.2. To enable the use of the B-spline, we assume x_i has compact support but do not require further distributional assumptions for x_i.

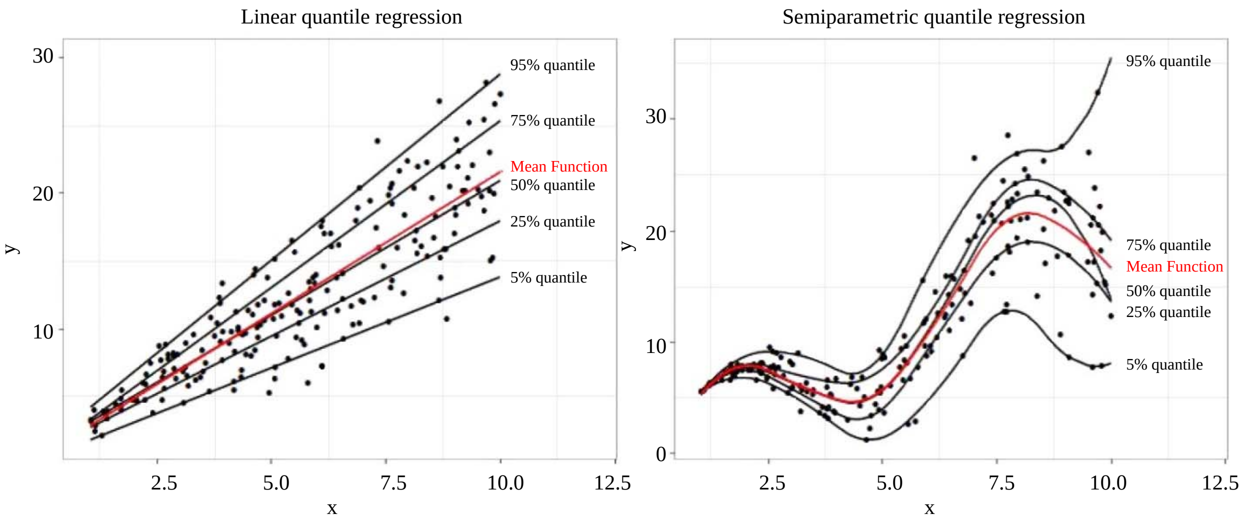

Figure 2.1 Illustration of linear quantile regression (left) and semiparametric quantile regression (right).

We consider estimation of parameters that are defined in terms of the superpopulation model relating y_i to x_i, rather than finite population parameters. The true parameter of interest, $\boldsymbol{\theta}_o$, is a $d-$dimensional vector satisfying,

$$E\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)] = \mathbf{0}, \tag{2.2}$$

where $\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)$ is an $r-$dimensional function with two continuous derivatives, and $r \geq d$. The expectation operator $E\,[\cdot]$ denotes expectation with respect to the superpopulation model. Note that $E\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)] = E\left[E_{y|x}\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)]\right]$, where

$$\begin{aligned} E_{y|x}\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)] &= \int_{-\infty}^{\infty} \mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)\, f_{y|x}\,(y_i \mid x_i)\, dy_i \\ &= \int_0^1 \mathbf{g}\left(F_{y|x}^{-1}\,(\tau),\, x_i;\, \boldsymbol{\theta}_o\right) \frac{f_{y|x}\left(F_{y|x}^{-1}\,(\tau) \mid x_i\right)}{f_{y|x}\left(F_{y|x}^{-1}\,(\tau) \mid x_i\right)}\, d\tau = \int_0^1 \mathbf{g}\,(q_\tau\,(x_i),\, x_i;\, \boldsymbol{\theta}_o)\, d\tau, \end{aligned} \tag{2.3}$$

and $F_{y|x}\,(y_i \mid x_i)$ and $f_{y|x}\,(y_i \mid x_i)$, respectively, denote the cumulative distribution function (cdf) and probability density function (pdf) of the conditional distribution of y_i given x_i. The second equality in (2.3) follows from the probability integral transform and a change of variables from y_i to the uniformly distributed τ with pdf $f\,(\tau) = I\,[\tau \in (0,1)]$, where $I\,[\cdot]$ is the indicator variable that takes the value 1 if the argument is true and 0 otherwise. The relationship defined by the third equality in (2.3) plays an important role in the imputation procedure. For each missing y_i, we construct J imputed values defined $\left\{\hat{q}_{\tau_1}\,(x_i),\, \ldots,\, \hat{q}_{\tau_J}\,(x_i)\right\}$, where $\hat{q}_{\tau_j}\,(x_i)$ estimates $q_{\tau_j}\,(x_i)$, and $\tau_1,\, \ldots,\, \tau_J$ form a fine grid on the interval $[0,\, 1]$. We then estimate $E_{y|x}\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)]$ by approximating the integral in the last expression of (2.3) with an average of the J imputed values.

The imputation procedure consists of two main steps. We first construct the imputed values, estimating $q_\tau\,(x_i)$ using a linear combination of B-spline basis functions. We then estimate $\boldsymbol{\theta}_o$ using the generalized method of moments (GMM), replacing missing y_i with the estimate of $E_{y|x}\,[\mathbf{g}\,(y_i,\, x_i;\, \boldsymbol{\theta}_o)]$ based on the imputed values and the relationship (2.3). To formalize the procedure, we require specific assumptions about the design and the response mechanism, which we specify in Section 2.1. Section 2.2 explains estimation of the quantile function, and Section 2.3 describes the generalized method of moments. Software for implementing the procedures is available from the authors.

2.1 Assumptions on design and response mechanism

Let I_i be the sample membership indicator, defined by $I_i = 1$ if unit i is selected. Let π_i and π_{ij} be the first and second order inclusion probabilities, respectively, defined by

$$\left[\pi_i,\, \pi_{ij}\right] = \left[P\,(I_i = 1 \mid y_i,\, x_i),\, P\left(I_i = 1,\, I_j = 1 \mid y_i,\, x_i,\, y_j,\, x_j\right)\right]. \tag{2.4}$$

Dependence of π_i on y_i in (2.4) represents a possible correlation between y_i and π_i that can cause the sample design to be informative for the quantile regression model (2.1). We denote the selected sample by A, where $A = \{i : I_i = 1\}$.

We assume x_i is observed for all i in A, while y_i may be missing. Let δ_i be the response indicator, defined by $\delta_i = 1$ if y_i is observed, and $\delta_i = 0$ if y_i is missing. Assume $\delta_i \sim \text{Bernoulli}(p_i)$, where the response probability p_i is defined as

$$p_i = P(\delta_i = 1 | y_i, x_i, I_i). \tag{2.5}$$

To define an approximately unbiased imputation procedure, we require an assumption about the relationship between δ_i and y_i. A common approach in missing data analysis is to assume that the response variable, y_i, is independent of the missing indicator, δ_i, conditional on the observed values (Little, 1982 and Pfeffermann, 2011). This assumption is a widely used interpretation of the missing at random (MAR) definition given in Rubin (1976) and clarified in Mealli and Rubin (2015). For a complex survey, the relationship between the inclusion probabilities, the response probabilities, and y can be complex if the response indicators and the sample inclusion indicators depend on a variable that is not included in the imputation model.

We follow the approach of Berg et al. (2016) and consider two assumptions about the relationship between δ_i and y_i. We define sample missing at random (SMAR) to mean

$$P(\delta_i = 1 | x_i, y_i, I_i) = P(\delta_i = 1 | x_i, I_i). \tag{2.6}$$

In contrast, we define population missing at random (PMAR) to mean

$$P(\delta_i = 1 | x_i, y_i) = P(\delta_i = 1 | x_i). \tag{2.7}$$

Berg et al. (2016) discuss situations in which the PMAR assumption may be viewed as reasonable and provide examples where PMAR holds while SMAR fails. If the response probabilities and the sample inclusion probabilities depend on a variable that is not included in the imputation model, then the PMAR may hold while SMAR does not. One example of a variable that may be excluded from the imputation model is a design variable. The analyst may omit a design variable from the imputation model if the design variable is unavailable at the imputation stage or because the imputation model is a subject-matter model relating y_i to x_i. We develop the QRI procedure to be flexible enough to accommodate either PMAR or SMAR. In practice, the analyst can decide whether PMAR or SMAR is more realistic for a particular application. In Section 2.2, we explain precisely how the nature of the missing at random assumption can impact the use of sampling weights in the estimation procedure. In the theory of Section 3, we focus on the situation in which assumption (2.7) holds.

2.2 Quantile regression with penalized B-Splines

We approximate the quantile function defining the relationship between y_i and x_i in the superpopulation with a linear combination of B-spline basis functions. A B-spline basis of order p spans the linear space of piecewise polynomials of degree $p-1$ with continuous derivatives up to order $p-2$.

B-splines allow improvements in computational efficiency over direct use of polynomial splines (Hastie, Tibshirani and Friedman, 2009).

To define the B-spline, we borrow terminology from Hastie, Tibshirani, and Friedman (2009) and Chen and Yu (2016). Assume x_i has compact support on the interval $[M_1, M_2]$. Define $K_n - 1$ interior knots, spaced at equidistant locations in the interval $[M_1, M_2]$ by, $\kappa_i = M_1 + [M_2 - M_1][K_n]^{-1} i$, for $i = 1, \ldots, K_n - 1$. Define p boundary knots at M_1 by κ_k for $k = -p+1, \ldots, 0$, and denote the p boundary knots at M_2 by κ_k for $k = K_n, \ldots, K_n + p - 1$. The $p^{\text{th}}-$degree B-spline basis functions for the knot sequence $\kappa_{-p+1}, \ldots, \kappa_{K_n+p-1}$ are the elements of the $K_n + p-$dimensional vector,

$$\mathbf{B}(x) = \left(B_{-p+1}^{[p]}(x), \ldots, B_{K_n}^{[p]}(x)\right)', \tag{2.8}$$

where $B_i^{[s]}(x)$ $(s = 1, \ldots, p)$ is defined recursively through divided differences. Specifically,

$$B_i^{[1]}(x) = I[\kappa_i \le x \le \kappa_{i+1}], \quad \text{for} \quad i = -p+1, \ldots, K_n + p - 2, \tag{2.9}$$

and

$$B_i^{[s]}(x) = \frac{x - \kappa_i}{\kappa_{i+s-1} - \kappa_i} B_i^{[s-1]}(x) + \frac{\kappa_{i+s} - x}{\kappa_{i+s} - \kappa_{i+1}} B_{i+1}^{[s-1]}(x), \tag{2.10}$$

for $i = -p+1, \ldots, K_n + p - 1 - s$ and $s = 2, \ldots, p$.

The estimator of the quantile regression function is defined by

$$\hat{q}_\tau(x) = \mathbf{B}(x)' \hat{\boldsymbol{\beta}}_\tau, \tag{2.11}$$

where the estimator $\hat{\boldsymbol{\beta}}_\tau$ is obtained by minimizing the quadratic form,

$$Q_\tau(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i w_i b_i \rho_\tau\left(y_i - \mathbf{B}(x_i)' \boldsymbol{\beta}\right) + \frac{\lambda_n}{2} \boldsymbol{\beta}' \mathbf{D}_m' \mathbf{D}_m \boldsymbol{\beta}, \tag{2.12}$$

where $w_i = \pi_i^{-1}\left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1}$, λ_n is a specified smoothing parameter, and $\rho_\tau(\cdot)$, b_i, and $\mathbf{D}_m$ are defined as follows. The function $\rho_\tau(u)$ in the first term of (2.12), is the check function of Koenker and Bassett (1978) defined by

$$\rho_\tau(u) = u(\tau - I[u < 0]). \tag{2.13}$$

Koenker's check function (2.13) is a standard optimization criterion for quantile regression because $q_\tau(x)$ minimizes the function $R(a) = E[\rho_\tau(y - a) | x]$ across a. The second term of (2.12) imposes a roughness penalty on the estimated quantile regression function. The matrix $\mathbf{D}_m$ is the m^{th} difference matrix with (i, j) element, $d_{ij} = (-1)^{j-i} C(m, j-i) I[0 \le j - i \le m] + (1 - I[0 \le j - i \le m])$, where $C(a, b)$ is the choose function. When $m = 2$, $\mathbf{D}_m$ has an interpretation related to the integral of the square of the second derivative of the function defined by the B-spline. Because the second derivative of a straight line is zero, the use of $\mathbf{D}_m$ for $m = 2$ shrinks the estimated quantile regression function toward a straight line. The appropriate choice of b_i in the first term of (2.12) depends on the assumptions about the nonresponse mechanism. If (2.6) holds, then one may set $b_i = w_i^{-1},$ which leads to the unweighted estimating equation of Chen and Yu (2016). If (2.6) is not satisfied, the unweighted estimator may lead to bias, and setting $b_i = 1$ is one way to attain an approximately unbiased estimator (Berg et al., 2016). We focus on the conservative choice of $b_i = 1,$ which leads to consistent estimators under (2.7) without requiring (2.6).

Remark 1. For simplicity, we consider a univariate x_i with support on a closed interval. Chen and Yu (2016) show that the procedure extends directly to a $h-$dimensional vector $\mathbf{x}_i,$ each element of which has support on a closed interval. To extend the procedure to a vector $\mathbf{x}_i,$ Chen and Yu (2016) define $\mathbf{B}(\mathbf{x}_i) = \left(\mathbf{B}(x_{1i})', \mathbf{B}(x_{2i})', \ldots, \mathbf{B}(x_{hi})'\right),$ where $x_{\tilde{h}i}$ is the $\tilde{h}^{\text{th}}$ element of $\mathbf{x}_i,$ for $\tilde{h} = 1, \ldots, h.$

2.3 GMM estimation based on quantile regression imputation

Recall that the population parameter of interest is defined by the estimating equation in (2.2). We define a full sample estimator of $\boldsymbol{\theta}_o$ by

$$\hat{\boldsymbol{\theta}}_A = \operatorname{argmin}_{\theta} \mathbf{G}_{n,A}(\boldsymbol{\theta})' \mathbf{G}_{n,A}(\boldsymbol{\theta}), \tag{2.14}$$

where

$$\mathbf{G}_{n,A}(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}), \tag{2.15}$$

w_i is defined following (2.12), and $i = 1, \ldots, n$ index the elements in $A.$ The estimator defined by (2.15) is a a generalized method of moments estimator, where each element of $\mathbf{G}_{n,A}$ defines a deviation between a sample moment and the corresponding population parameter. For instance, if $\boldsymbol{\theta}_o = E[y_i],$ then $\mathbf{g}_i(y_i; \boldsymbol{\theta}_o) = (y_i - \boldsymbol{\theta}_o).$ Additional examples are provided in the simulation study of Section 5. Because y_i is unobserved for nonrespondents, $\hat{\boldsymbol{\theta}}_A$ is unattainable.

An imputed version of (2.15) is defined by replacing $\mathbf{g}(y_i, x_i, \boldsymbol{\theta})$ for an unobserved unit i by an estimator of the expected value. From (2.3), an estimator of $E_{y|x}[\mathbf{g}(y_i, x_i, \boldsymbol{\theta})]$ is $\int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau,$ where $\hat{q}_{\tau i} = \hat{q}_{\tau}(x_i) = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_{\tau}.$ We then define the estimator $\hat{\boldsymbol{\theta}}$ by,

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{\mathbf{G}_n(\boldsymbol{\theta})' \mathbf{G}_n(\boldsymbol{\theta})\right\}, \tag{2.16}$$

where

$$\mathbf{G}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} w_i \left\{\delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1-\delta_i) \int_0^1 \mathbf{g}(\hat{q}_{\tau i}, x_i, \boldsymbol{\theta})\, d\tau\right\}. \tag{2.17}$$

For specific $\mathbf{g}_i$ the minimizer of (2.16) has a closed form expression. For the case in which $\boldsymbol{\theta}_o = E[y_i],$ and $\hat{\boldsymbol{\theta}}$ is the Hájek estimator defined by

$$\hat{\boldsymbol{\theta}} = \sum_{i=1}^{n} w_i \left\{\delta_i y_i + (1-\delta_i) \int_0^1 \hat{q}_{\tau}(x_i)\, d\tau\right\}.$$

In other situations, a closed form expression may not exist and standard numerical procedures, such as Newton-Raphson, can be used to minimize (2.17). In deriving the asymptotic results of Section 3, we assume that $\boldsymbol{\theta}_o$ is the unique value such that $E[\mathbf{g}_i(y_i, \boldsymbol{\theta}_o)] = \mathbf{0}$, which relates to the existence of a unique minimum of (2.16). See Fuller (1996, page 252) for a similar condition and a discussion of the theory of estimators that minimize a quadratic form.

In practice, an approximation for the integral is required. We use a midpoint approximation (i.e., Nusser, Carriquiry, Dodd and Fuller, 1996). Let the fixed sequence $0 < \tau_1 \leq \tau_2 \cdots \leq \tau_J < 1$ be the mid-points of J evenly-spaced sub-intervals of $[0, 1]$. For non-respondent i, construct J imputed values,

$$y_{ij}^* = \mathbf{B}(x_i)' \hat{\boldsymbol{\beta}}_{\tau_j}, \quad j = 1, \ldots, J, \tag{2.18}$$

where $\hat{\boldsymbol{\beta}}_{\tau_j}$ is obtained by minimizing $Q_{\tau_j}(\boldsymbol{\beta})$ in (2.12). We define the estimator $\hat{\boldsymbol{\theta}}_J$ to satisfy

$$\hat{\boldsymbol{\theta}}_J = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ \mathbf{G}_{n,J}(\boldsymbol{\theta})' \mathbf{G}_{n,J}(\boldsymbol{\theta}) \right\}, \tag{2.19}$$

where

$$\mathbf{G}_{n,J}(\boldsymbol{\theta}) := \mathbf{G}_n\left(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}\right) = \sum_{i=1}^{n} w_i \left\{ \delta_i \mathbf{g}(y_i, x_i, \boldsymbol{\theta}) + (1 - \delta_i) J^{-1} \sum_{j=1}^{J} \mathbf{g}\left(y_{ij}^*, x_i, \boldsymbol{\theta}\right) \right\}, \tag{2.20}$$

w_i is defined following (2.12), and $\hat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\beta}}_{\tau_1}, \ldots, \hat{\boldsymbol{\beta}}_{\tau_J}\right)'$. The imputation procedure above differs from Chen and Yu (2016) in that the midpoint approximation for the integral is used instead of Monte Carlo integration. Both the midpoint approximation and Monte Carlo integration are justified by the probability integral transform, which relates the expectation to the conditional quantile function, as explained in (2.3). For functions with bounded second derivatives, the error in the midpoint approximation is $O(J^{-2})$. We also prefer the midpoint approximation because in simulations, it reduces the variance of the estimator and reduces instability in the variance estimator due to extreme quantiles relative to Monte Carlo simulation. Jang and Wang (2015) discuss the potential problem of unstable estimators for extreme quantiles from unstructured quantile regression models.

3 Asymptotic distributions and variance estimation

We derive an asymptotic normal distribution for the QRI estimator $\hat{\boldsymbol{\theta}}$ defined in (2.16), although the estimator $\hat{\boldsymbol{\theta}}_J$, defined in (2.19), with a finite number of (J) imputations is necessary in practice. This approach of developing theory under an assumption of an infinite number of imputed values has been used previously. See, for example, Clayton, Spiegelhalter, Dunn and Pickles (1998) and Robins and Wang (2000). The simulations in Section 5 demonstrate that the asymptotic normal distribution derived for $J = \infty$ is a reasonable approximation for the distribution of the estimator constructed with finite J. We outline the main concepts underlying the proofs of lemma 1, lemma 2, and Theorem 1, deferring details to Section B of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

The derivation of the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ proceeds in three main steps. Lemma 1 gives the asymptotic distribution of the estimators of the quantile regression coefficients. Lemma 2 presents the asymptotic distribution of the estimating equation (2.17). These two lemmas are analogous to lemma 1 and lemma 2 of Chen and Yu (2016). Theorem 1 then provides the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

3.1 Asymptotic normality of $\hat{\boldsymbol{\theta}}$

We consider a sequence of samples and finite populations indexed by N, where the sample size $n \to \infty$ as $N \to \infty$. To define the regularity conditions, we introduce the notation $\mathcal{F}_N$ to represent an element of the sequence of finite populations with size N and use the notation "$\mid \mathcal{F}_N$" to indicate that the reference distribution is the distribution based on repeated sampling conditional on the finite population of size N. For example, $E\left[\hat{Y} \mid \mathcal{F}_N\right]$ and $V\left\{\hat{Y} \mid \mathcal{F}_N\right\}$, respectively, denote the conditional expectation and variance of the outcome $\hat{Y}$ with respect to the randomization distribution generated from repeated sampling from $\mathcal{F}_N$. Similarly, $\hat{Y} \xrightarrow{d} Y \mid \mathcal{F}_N$ a.s., means that $\hat{Y}$ converges in distribution to Y almost surely with respect to the process of repeated sampling from the sequence of finite populations as $N \to \infty$. The convergence is with probability 1 because $\mathcal{F}_N$ is a random realization from the superpopulation model (2.1).

The regularity conditions on the sample design and tuning parameters for the estimator of the B-spline model are as follows:

1. Any variable v_i such that $E\left[\,|v_i|^{2+\delta}\right] < \infty$, where $\delta > 0$, satisfies,

$$\sqrt{n}\,(\bar{v}_{\mathrm{HT}} - \bar{v}_N) \mid \mathcal{F}_N \xrightarrow{d} N(0, V_\infty) \quad \text{a.s.,} \tag{3.1}$$

   where $(\bar{v}_{\mathrm{HT}}, \bar{v}_N) = N^{-1}\sum_{i=1}^{N}(\pi_i^{-1} v_i I_i,\, v_i),\, V_\infty = \lim_{N\to\infty} V_N$, and $V_N = nV\{\bar{v}_{\mathrm{HT}} \mid \mathcal{F}_N\}$ is the conditional variance of the Horvitz-Thompson mean, $\bar{v}_{\mathrm{HT}}$, given $\mathcal{F}_N$.

2. $nn_B^{-1} \to 1$ and $n_B N^{-1} \to f_\infty \in [0, 1]$, where n_B is the expected sample size.
3. There exist constants C_1, C_2, and C_3 such that $0 < C_1 \le n_B N^{-1}\pi_i^{-1} \le C_2 < \infty$, and

$$\left| n_B \left(\pi_{ij} - \pi_i \pi_j\right) \pi_i^{-1} \pi_j^{-1} \right| \le C_3 < \infty \quad \text{a.s.} \tag{3.2}$$

4. The value determining the number of interior knots $K_n = O\left(n_B^{\frac{1}{2p+3}}\right)$.
5. $\lambda_n = O(n_B^{\nu})$ for $\nu \le (2p+3)^{-1}(p+m+1)$.

Condition 3 is also used in Fuller (2009a). Condition 3 holds for simple random sampling, where $(\pi_{ij} - \pi_i \pi_j)\,\pi_i^{-1}\pi_j^{-1} = n^{-1}(n-1)\,(N-1)^{-1}N - 1$, and for Poisson sampling, where $(\pi_{ij} - \pi_i \pi_j)\,\pi_i^{-1}\pi_j^{-1} = 0$. Fuller (2009a) explains that condition 3 holds for many stratified designs and that the designer has the control to ensure condition 3.

Under assumptions 4-5, Barrow and Smith (1978) show that a $\boldsymbol{\beta}_\tau^*$ exists that satisfies,

$$\sup_{x\in[M_1, M_2]} \left| q_\tau(x) - b_\tau^a(x) - \mathbf{B}(x)'\, \boldsymbol{\beta}_\tau^* \right| = o\left(K_n^{-(p+1)}\right), \tag{3.3}$$

where $\mathbf{B}(x)'\boldsymbol{\beta}_\tau^*$ is the best L_∞ approximation for $q_\tau(x)$, and $b_\tau^{(a)}(x)$ is a bias of the B-spline approximation for the true quantile function, satisfying, $b_\tau^a(x) = O\left(K_n^{-(p+1)}\right)$. For details of the form of the bias term, see Chen and Yu (2016) and Yoshida (2013). The property (3.3) is used extensively in the derivation of lemma 1.

The proofs of both lemma 1 and lemma 2 use a result given in Theorem 1.3.6 of Fuller (2009b). Because of the importance of this theorem to the results of this section, we state this theorem as Fact 1:

Fact 1. (Theorem 1.3.6 of Fuller (2009b)): Suppose

$$\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N(0, V_{11}) \quad \text{a.s.,} \quad \text{and} \quad \theta_N - \theta_o \xrightarrow{d} N(0, V_{22}). \tag{3.4}$$

Then, $\left(\hat{\theta} - \theta_o\right) \xrightarrow{d} N(0, V_{11} + V_{22})$.

Note that V_{11} in Fact 1 is a fixed limit and not a design variance because the design variance is a random function of the finite population in this framework. The condition $\left(\hat{\theta} - \theta_N\right) | \mathcal{F}_N \xrightarrow{d} N(0, V_{11})$ a.s., holds for a broad class of designs, such as those discussed in Isaki and Fuller (1982).

Lemma 1. Under assumptions 1-5 and for fixed $x_i \in [M_1, M_2]$ and $\tau \in (0, 1)$,

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau(x_i) - \mathbf{B}(x_i)'\boldsymbol{\beta}_\tau^* + b_\tau^\lambda(x_i)\right) \xrightarrow{d} N\left(0, \mathbf{B}(x_i)'\boldsymbol{\Sigma}_\infty(\tau)\mathbf{B}(x_i)\right), \tag{3.5}$$

and

$$\sqrt{\frac{n}{K_n}}\left(\hat{q}_\tau(x_i) - q_\tau(x_i) + b_\tau^a(x_i) + b_\tau^\lambda(x_i)\right) \xrightarrow{d} N\left(0, \mathbf{B}(x_i)'\boldsymbol{\Sigma}_\infty(\tau)\mathbf{B}(x_i)\right), \tag{3.6}$$

where

$$\begin{aligned}
b_\tau^\lambda(x_i) &= \lim_{N\to\infty} \frac{\tilde{\lambda}_n}{n}\mathbf{B}(x_i)'\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{D}_m'\mathbf{D}_m\boldsymbol{\beta}_\tau^*, \qquad (3.7)\\
\boldsymbol{\Omega}_n(\tau) &= \mathbf{H}(\tau) + \frac{\tilde{\lambda}_n}{n}\mathbf{D}_m'\mathbf{D}_m,\\
\boldsymbol{\Sigma}_\infty(\tau) &= \lim_{N\to\infty}\frac{1}{K_n}\boldsymbol{\Omega}_n(\tau)^{-1}\left(\mathbf{V}_{1,\infty}(\tau) + f_\infty\tau(1-\tau)\boldsymbol{\Phi}\right)\boldsymbol{\Omega}_n(\tau)^{-1},\\
\mathbf{H}(\tau) &= E\left[p_i\mathbf{B}(x_i)f_{y|x,i}(q_{\tau i})\mathbf{B}(x_i)'\right],\\
\boldsymbol{\Phi} &= E\left[p_i\mathbf{B}(x_i)\mathbf{B}(x_i)'\right],\\
\mathbf{V}_{1,\infty}(\tau) &= \lim_{N\to\infty}\frac{n}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\delta_i\delta_j\mathbf{B}(x_i)\psi_\tau(u_{i\tau})\mathbf{B}\left(x_j\right)'\psi_\tau\left(u_{j\tau}\right),
\end{aligned}$$

$u_{i\tau} = y_i - \mathbf{B}(x_i)' \boldsymbol{\beta}_\tau^*$, $\psi_\tau(u) = \tau - I[u < 0]$, $\tilde{\lambda}_n = n\hat{N}N^{-1}\lambda_n$, $\hat{N} = \sum_{i=1}^n \pi_i^{-1}$, and $f_{y|x,i}(q)$ is the pdf of y_i given x_i evaluated at q.

The main idea of the proof of lemma 1 is to show that the estimator of the quantile regression coefficient has a Bahadur representation given in corollary 1 below:

Corollary 1: By the proof of lemma 1, the estimator of the quantile regression coefficient has the following Bahadur representation:

$$\sqrt{\frac{n}{K_n}}\left(\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau^* + \frac{\tilde{\lambda}_n}{n}\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{D}_m'\mathbf{D}_m\boldsymbol{\beta}_\tau^*\right) = \sqrt{\frac{n}{K_n}}\boldsymbol{\Omega}_n(\tau)^{-1}\frac{1}{N}\sum_{i=1}^n \pi_i^{-1}\delta_i\mathbf{B}(x_i)\psi_\tau(u_{i\tau}) + o_p(1). \tag{3.8}$$

The derivation of the Bahadur representation follows the basic approach of Koenker (2005) and Yoshida (2013). To account for the complex sample design, condition (3.2) is used to bound sums of covariances induced by nontrivial second order inclusion probabilities. For independent random variables from an infinite population (as in Chen and Yu (2016), Yoshida (2013) and Koenker (2005)), the corresponding covariances are zero. Given the Bahadur representation (3.8), lemma 1 follows from an application of the regularity condition in (3.1) and Fact 1 to the elements of the Horvitz-Thompson mean in (3.8). The $V_{1,\infty}$ in $\boldsymbol{\Sigma}_\infty(\tau)$ essentially plays the role of V_{11} in Fact 1 and is the limit of the design variance of the Horvitz-Thompson mean. The second term in $\boldsymbol{\Sigma}_\infty(\tau)$ is the asymptotic variance of the design-expectation of the Horvitz-Thompson mean and plays the role of V_{22} in Fact 1.

Lemma 2 and Theorem 1 require additional regularity conditions about the estimating equation. The regularity conditions on the estimation are similar to those in Chen and Yu (2016) and are therefore deferred to Section A of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

Lemma 2. Under the assumptions of lemma 1 and the regularity conditions provided in Section A of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016),

$$\sqrt{n}\,\mathbf{G}_n(\boldsymbol{\theta}_o) \overset{d}{\to} N(\mathbf{0}, \mathbf{V}_G(\boldsymbol{\theta}_o)), \tag{3.9}$$

where

$$\begin{aligned}
\mathbf{V}_G(\boldsymbol{\theta}_o) &= f_\infty V\{\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\} + \lim_{N\to\infty}\mathbf{V}_{\xi,N}(\boldsymbol{\theta}_o), \qquad (3.10)\\
\mathbf{V}_{\xi,N} &= nN^{-2}\sum_{i=1}^N\sum_{j=1}^N \frac{\pi_{ij} - \pi_i\pi_j}{\pi_i\pi_j}\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)\boldsymbol{\xi}_j(\boldsymbol{\theta}_o)',\\
\boldsymbol{\xi}_i(\boldsymbol{\theta}_o) &= \delta_i\mathbf{g}_i(y_i;\boldsymbol{\theta}_o) + (1-\delta_i)\int_0^1 \mathbf{g}_i(q_\tau(x_i);\boldsymbol{\theta}_o)\,d\tau + \delta_i\mathbf{h}_{ni}(\boldsymbol{\theta}_o),\\
\mathbf{h}_{n_i}(\boldsymbol{\theta}_o) &= \int_0^1 E\left[\left(1-p_j\right)\dot{\mathbf{g}}_{j,y}\left(q_\tau\left(x_j\right);\boldsymbol{\theta}_o\right)\mathbf{B}\left(x_j\right)'\right]\boldsymbol{\Omega}_n(\tau)^{-1}\mathbf{B}(x_i)\psi_\tau(u_{i\tau})\,d\tau,
\end{aligned}$$

and $\dot{\mathbf{g}}_{i,y}(y_i;\, \boldsymbol{\theta}_o)$ is the partial derivative of $\mathbf{g}_i(a;\, \boldsymbol{\theta})$ with respect to a evaluated at y_i.

The proof of lemma 2 centers on the Taylor expansion given by

$$\begin{aligned}\mathbf{g}_i(\hat{q}_\tau(x_i);\, \boldsymbol{\theta}_o) &= \mathbf{g}_i(q_\tau(x_i);\, \boldsymbol{\theta}_o) + \dot{\mathbf{g}}_{i,y}(q_\tau(x_i);\, \boldsymbol{\theta}_o)(\hat{q}_\tau(x_i) - q_\tau(x_i)) \\ &\quad + \ddot{\mathbf{g}}_{i,y}(q_\tau(x_i);\, \boldsymbol{\theta}_o)(\tilde{q}_\tau(x_i) - q_\tau(x_i))^2,\end{aligned} \tag{3.11}$$

where $\tilde{q}_\tau(x_i)$ is between $\hat{q}_\tau(x_i)$ and $q_\tau(x_i)$, and $\ddot{\mathbf{g}}_{i,y}(q_\tau(x_i);\, \boldsymbol{\theta}_o)$ denotes the vector of partial derivatives of the elements of $\dot{\mathbf{g}}_{i,y}(a,\, \boldsymbol{\theta}_o)$ with respect to a evaluated at $q_\tau(x_i)$. By arguments similar to those of Chen and Yu (2016), $n \left\| \ddot{g}_{i,y}(q_\tau(x_i);\, \boldsymbol{\theta}_o)(\tilde{q}_\tau(x_i) - q_\tau(x_i))^2 \right\| = O(1)$. Lemma 2 then follows from the linear approximation for $\hat{q}_\tau(x_i) - q_\tau(x_i)$ in lemma 1.

Theorem 1. Under the assumptions of lemmas 1 and 2, the QRI estimator $\hat{\boldsymbol{\theta}}$ defined in (2.16), constructed with $J = \infty$, satisfies, $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$, where

$$\boldsymbol{\Sigma}_\theta = \left[\boldsymbol{\Gamma}(\boldsymbol{\theta}_o)'\, \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)\right]^{-1} \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)'\, \mathbf{V}_G(\boldsymbol{\theta}_o)\, \boldsymbol{\Gamma}(\boldsymbol{\theta}_o) \left[\boldsymbol{\Gamma}(\boldsymbol{\theta}_o)'\, \boldsymbol{\Gamma}(\boldsymbol{\theta}_o)\right]^{-1}, \tag{3.12}$$

$\mathbf{G}(\boldsymbol{\theta}) = E[\mathbf{G}_N(\boldsymbol{\theta},\, \mathbf{y})]$, $\mathbf{G}_N(\boldsymbol{\theta},\, \mathbf{y}) = N^{-1}\sum_{i=1}^{N} \delta_i g(y_i,\, x_i)$, and $\boldsymbol{\Gamma}(\boldsymbol{\theta}_o) = E[\partial/\partial\boldsymbol{\theta}\, \mathbf{G}_N(\boldsymbol{\theta})]$.

By Pakes and Pollard (1989), Theorem 1 is satisfied if the following hold:

1. $\sup_\theta |\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})| = o_p(1)$,
2. For $\zeta_n \to 0$, $\sup_{|\boldsymbol{\theta} - \boldsymbol{\theta}_o| < \zeta_n} |\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta}) - \mathbf{G}_n(\boldsymbol{\theta}_o)| = o_p(n_B^{-0.5})$,

where ζ_n is arbitrarily small. Because of the complex sample design, the proof that these conditions hold proceeds in two steps, considering first the deviation $|\mathbf{G}_n(\boldsymbol{\theta}) - \mathbf{G}_N(\boldsymbol{\theta})|$ and then the deviation $|\mathbf{G}_N(\boldsymbol{\theta}) - \mathbf{G}(\boldsymbol{\theta})|$. The result then follows from the triangle inequality.

3.2 Variance estimation

We estimate the variance of $\hat{\boldsymbol{\theta}}_J$ using the linearization method (Fuller, 2009b, page 64). We use the asymptotic covariance matrix in (3.12) to estimate the variance of $\hat{\boldsymbol{\theta}}_J$, the estimator of $\boldsymbol{\theta}_o$ defined in (2.19), constructed with a finite number of imputed values. To estimate $\mathbf{V}_G(\boldsymbol{\theta}_o)$, a design-consistent variance estimator is applied to an estimator of the mean of an estimator of $\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)$ defined in (3.10). The estimator of $\boldsymbol{\xi}_i(\boldsymbol{\theta}_o)$ is obtained by replacing $\boldsymbol{\theta}_o$ and $\boldsymbol{\beta}_\tau^*$ with estimators $\hat{\boldsymbol{\theta}}_J$ and $\hat{\boldsymbol{\beta}}_\tau$, respectively.

The estimator of variance is defined,

$$\hat{\boldsymbol{\Sigma}}_\theta = \left[\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)'\, \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)\right]^{-1} \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)' \left[\hat{\mathbf{V}}_{G,\infty}\left(\hat{\boldsymbol{\theta}}_J\right)\right] \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right) \left[\hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)'\, \hat{\boldsymbol{\Gamma}}\left(\hat{\boldsymbol{\theta}}_J\right)\right]^{-1}, \tag{3.13}$$

where $\hat{\mathbf{V}}_{G,\infty}\left(\hat{\boldsymbol{\theta}}_J\right) = \hat{f}_\infty \hat{V}\left\{\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right\} + \hat{\mathbf{V}}_{\xi,N}\left(\hat{\boldsymbol{\theta}}_J\right)$,

$$
\begin{aligned}
\hat{V}\left\{\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right\} &= \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)' - \frac{1}{\hat{N}\left(\hat{N}-1\right)}\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)\left(\sum_{i=1}^{n}\pi_i^{-1}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\right)', \qquad (3.14)\\
\hat{\mathbf{V}}_{\xi,N}\left(\hat{\boldsymbol{\theta}}_J\right) &= \frac{n}{\hat{N}^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right)\hat{\boldsymbol{\xi}}_j\left(\hat{\boldsymbol{\theta}}_J\right)',\\
\hat{\boldsymbol{\xi}}_i\left(\hat{\boldsymbol{\theta}}_J\right) &= \delta_i\mathbf{g}_i\left(y_i;\hat{\boldsymbol{\theta}}_J\right) + (1-\delta_i)\,J^{-1}\sum_{j=1}^{J}\mathbf{g}_i\left(\mathbf{B}(x_i)'\,\hat{\boldsymbol{\beta}}_{\tau_j};\hat{\boldsymbol{\theta}}_J\right) + \delta_i\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right),\\
\hat{\mathbf{h}}_{ni}\left(\hat{\boldsymbol{\theta}}_J\right) &= J^{-1}\sum_{j=1}^{J}N^{-1}\sum_{k=1}^{n}\pi_k^{-1}(1-\delta_k)\,\dot{\mathbf{g}}_{k,y}\left(\mathbf{B}(x_k)'\,\hat{\boldsymbol{\beta}}_{\tau_j};\hat{\boldsymbol{\theta}}_J\right)\mathbf{B}(x_k)'\,\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right)^{-1}\mathbf{B}(x_i)\,\psi_\tau\left(\hat{u}_{i\tau_j}\right),\\
\hat{\boldsymbol{\Omega}}_n\left(\tau_j\right) &= \hat{\mathbf{H}}\left(\tau_j\right) + \frac{\hat{f}_\infty\tilde{\lambda}_n}{n}\mathbf{D}_m'\mathbf{D}_m,\\
\hat{\mathbf{H}}(\tau) &= \frac{1}{\hat{N}}\sum_{i=1}^{n}\pi_i^{-1}\delta_i\mathbf{B}(x_i)\,\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)\mathbf{B}(x_i)',
\end{aligned}
$$

$\hat{f}_\infty = n\hat{N}^{-1}$, $\hat{N} = \sum_{i=1}^{n}\pi_i^{-1}$, and $\hat{u}_{i\tau_j} = y_i - \mathbf{B}(x_i)'\,\hat{\boldsymbol{\beta}}_{\tau_j}$. An estimator of $\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right)$ is the inverse of an estimator of the derivative of the quantile function and is defined by

$$
\hat{f}_{y|x,i}\left(\hat{q}_\tau(x_i)\right) = \max\left\{\frac{2a_{n,\tau}}{\mathbf{B}(x_i)'\left(\hat{\boldsymbol{\beta}}_{\tau+a_{n,\tau}} - \hat{\boldsymbol{\beta}}_{\tau-a_{n,\tau}}\right)}, 0\right\}, \qquad (3.15)
$$

where the bandwidth $a_{n,\tau}$ is given by

$$
a_{n,\tau} = n^{-0.2}\left[\frac{4.5\phi\left(\Phi^{-1}(\tau)\right)^4}{\left(2\Phi^{-1}(\tau)^2+1\right)^2}\right], \qquad (3.16)
$$

with $\phi(\cdot)$ and $\Phi(\cdot)$, respectively, the pdf and cdf of a standard normal distribution. See Wei, Ma and Carroll (2012) and Koenker (2005) for discussions of (3.15) and (3.16), respectively.

4 Application to Conservation Effects Assessment Project

The cropland component of the Conservation Effects Assessment Project (CEAP) consists of a series of surveys meant to measure soil and nutrient loss from crop fields. The first cropland assessment was a national survey conducted over the period 2003-2006. Data collection for a second national survey, planned for 2015-2016, was on-going at the time of writing this paper. Each of the time periods 2003-2006 and 2015-2016 is considered one time point for estimation. Data are collected over multiple years (i.e., 2003-2006 or 2015-2016) for operational reasons, and no unit is in the sample for two years in the same time period. Temporal changes of interest are changes between the two time periods, rather than changes between two years in the same time period. The temporal structure leads to unbalanced data because some units respond in both time periods, some units never respond, and some units respond in only one of the two time periods. Providing the data user with a complete, imputed data set with a single set of weights simplifies analyses involving more than one time point.

We investigate the feasibility of imputation for CEAP using a subset of the data collected during 2003-2005. We omit the data collected in 2006 because the sample design changed, and we do not have the information required to compute sampling weights for the 2006 survey. The data from the 2015-2016 survey are not yet collected. This analysis is considered an investigation of the feasibility of using QRI to impute missing data in CEAP in the direction of addressing the broader problem of estimation of change over time.

An understanding of the CEAP sample design requires an understanding of the design of the National Resources Inventory (NRI). The NRI monitors status and trends in land use, land cover, and erosion, with emphasis on characteristics related to natural resources and agriculture. Primary sampling units in the NRI are land areas called segments, which are approximately 160 acres. Each segment contains approximately three secondary sampling units, which are randomly selected locations called points. From 1982-1997, the same sample of approximately 300,000 segments, referred to as the foundation sample, was revisited every five years. The foundation sample is a stratified sample of segments, with a typical sampling rate of approximately 4%. See Nusser and Goebel (1997) for details of the design of the NRI foundation sample. In 2000, the NRI transitioned to annual sample design. Because revisiting every sampled segment in the foundation sample on an annual basis is infeasible, a rotating panel design is used. A subsample of the foundation segments, called the core panel, is revisited annually. The core panel is supplemented with a rotation panel, which changes each year. In essence, the core and rotation panels are stratified samples of the foundation sample. The strata, called sample classes, depend on the characteristics of the NRI segment observed in 1982-1997, such as presence of wetlands, cropland, and forest. See Nusser (2006) and Breidt and Fuller (1999) for further detail on the NRI annual samples.

For the Conservation Effects Assessment Project (CEAP), data collectors visit a subset of the NRI points that are located in sampled crop fields and collect more detailed information on crop choices and conservation practices. The sample for the 2003-2005 CEAP survey essentially consists of segments in the NRI core panel, 2002 rotation panel, and 2003 rotation panel that contain at least one cropland point. For segments containing more than one cropland point, one cropland point was selected randomly. The selection of one point per segment is an effort to improve geographic spread and reduce the number of instances in which a farm operator associated with multiple sampled points is selected into the sample, thereby reducing the respondent burden.

Because the first phase sampling rate for the NRI is small $(\approx 4\%)$, we approximate the CEAP sample as a probability proportional to size with replacement sample. The selection probabilities for CEAP largely reflect the sample design for the NRI. Details of construction of first and second order selection probabilities for CEAP are provided in Section C of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

Data collection for crop fields sampled for the CEAP survey consists of multiple components. An important component is a farmer interview survey that collects detailed information on farming managements and conservation practices. Nonresponse can occur in CEAP if a farmer refuses to participate in the interview.

Response variables in CEAP are measurements of different types of soil and nutrient loss, obtained from a physical process model called the Agricultural Policy Environmental Extender (APEX). The APEX model converts data from the farmer surveys as well as information from administrative sources and the NRI to numerical measures of erosion. For this study, we consider a measure of soil loss due to sheet and rill erosion called RUSLE2, discussed further in Section 4.1.

The NRI survey provides a convenient source of auxiliary information for imputing CEAP response variables. Because the NRI survey data are collected through aerial photographs of sampled segments, nonresponse due to refusals does not occur in the NRI. As a consequence, NRI data are available for all sampled points in CEAP. Furthermore, the NRI collects data related to land use, conservation practices and erosion – characteristics that are expected to be correlated with outputs of the APEX model. As an auxiliary variable, we use USLE, a measure of sheet and rill erosion collected in the NRI.

Domains of interest in CEAP are ten “CEAP production regions”. We focus on estimation of mean RUSLE2 for seven states (Iowa, Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin) that comprise the majority of the CEAP production region called the Corn Belt. We use semiparametric quantile regression to impute missing values for RUSLE2 using USLE as an auxiliary variable for each of these seven states in the Corn Belt region.

4.1 Imputation model and procedures

The variable of interest, RUSLE2, is a measure of sheet and rill erosion obtained from the APEX model. Because interest is in mean erosion on a per acre basis, the parameter of interest $\theta,$ the mean RUSLE2 erosion in the state, is defined as a ratio by,

$$\theta = \frac{E\left[\sum_{i=1}^{m_{ek}} R_{ik} D_k m_k^{-1}\right]}{E\left[m_{ek} D_k m_k^{-1}\right]}, \tag{4.1}$$

where R_{ik} is the RUSLE2 erosion for point i in segment k sampled in the period 2003-2005, D_k is the area of segment k, m_k is the total number of points in segment k, and m_{ek} is the number of points in segment k that are eligible for the CEAP survey. As discussed above, the period 2003-2005 is considered one time point, and no point is sampled more than once in this collection of years. Therefore, each sampled unit has one value R_{ik} for this set of years, and R_{ik} does not need a subscript of t for year.

The RUSLE2 erosion is an advancement of a simpler measure of sheet and rill erosion called USLE. The USLE is a product of five numerical indexes associated with slope steepness and length, rainfall, soil erodibility, conservation practices, and crop managements. While RUSLE2 is only observed for respondents to the CEAP survey, USLE is available from the main NRI sample for all points in the CEAP sample. We use the average USLE across years 2003-2005 as the covariate in the imputation model. Specifically, for point i in segment k, we define, $U_{ik} = 3^{-1}\sum_{t=2003}^{2005} U_{tik}$, where U_{tik} is the USLE soil loss in the NRI for point i in segment k for year t.

Because the RUSLE2 and USLE are highly skewed, the quantile regression model is applied after transforming both R_{ik} and U_{ik} by a power of 0.2. The quantile regression model postulated for the superpopulation can be expressed as, $P(y_{ik} \leq \tau \mid x_{ik}) = q_\tau(x_{ik})$, where $y_{ik} = R_{ik}^{0.2}$, and $x_{ik} = U_{ik}^{0.2}$. The unknown function $q_\tau(x_{ik})$ is approximated by a linear combination of B-spline basis functions generated from x_{ik}. To define the penalized B-spline, we set $p = 3$, $m = 2$, $K_n = 16$, and $\lambda = 0.004$.

Because the quantity of interest is erosion on a per acre basis, the estimator $\hat{\theta}$ of θ defined in (4.1) is a ratio of two estimators. That is, $\hat{\theta} = \hat{\theta}_2^{-1}\hat{\theta}_1$, where $\hat{\theta}_1$ is an estimator of $\theta_1 = E[D_k U_{ik}]$, and $\theta_2 = E[D_k]$. The estimator of θ_2 is the Hájek estimator, $\hat{\theta}_2 = \left(\sum_{k=1}^{n} \pi_{ik}^{-1} D_k\right)\left(\sum_{k=1}^{n} \pi_{ik}^{-1}\right)^{-1}$, where π_{ik} is the probability of selecting point i in segment k into the CEAP sample. The estimator $\hat{\theta}_1$ of θ_1 is obtained from GMM with $g(y, \theta_1) = (D_k y^5 - \theta_1)$.

4.2 Estimates and variance estimates

Table 4.1 contains estimates of average RUSLE2 soil loss based on QRI, along with estimated standard errors for seven states in the Corn Belt CEAP region. For comparison, the complete case estimator $(\bar{R}_{cc})$ and corresponding estimated standard error is also provided in Table 4.1. The complete case estimator is the ratio of Hájek estimators constructed using only the units that provide a usable response for RUSLE2.

For each of the seven states, the complete case estimator is larger than the estimator based on the imputed data. The imputation procedure reduces the estimator of θ, relative to the complete case estimator, because the weighted mean of U_{ik} among sampled units is smaller than the mean of U_{ik} among respondents, as shown in the last two rows of Table 4.1.

As expected, the estimated standard error for $\hat{\theta}$ is smaller than the estimated standard error for the complete case estimator. The ratios of the estimated variances for the complete case estimator to the estimated variances of $\hat{\theta}$ range from 1.103 for MN to 1.252 for IN. This comparison demonstrates the potential for efficiency gain due to the use of imputation. The reduction in estimated standard deviation occurs because the imputation procedure uses U_{ik} for the full sample, while the complete case estimator is based only on R_{ik} for the subset of respondents.

Table 4.1
Complete-case estimator $(\bar{R}_{cc})$ and QRI-GMM estimator $(\hat{\theta})$ of mean RUSLE2 soil loss (θ), corresponding standard errors, sample sizes (n), number of respondents (n_r), and weighted covariate means for sampled units $(\hat{U}_s)$ and weighted covariate means among respondents $(\hat{U}_r)$ for seven states in the Corn Belt

	IL	IN	IA	MI	MN	OH	WI
$\bar{R}_{cc}$	0.3301	0.2994	0.3464	0.3214	0.1741	0.3700	0.5226
$\mathrm{SE}(\bar{R}_{cc})$	0.0112	0.0179	0.0144	0.0209	0.0068	0.0213	0.0354
$\hat{\theta}$	0.3281	0.2901	0.3408	0.3145	0.1646	0.3636	0.4977
$\mathrm{SE}(\hat{\theta})$	0.0106	0.0160	0.0134	0.0189	0.0063	0.0201	0.0337
n	1,823	1,151	1,492	935	1,649	1,053	662
n_r	1,275	751	1,011	585	1,008	698	414
$\hat{U}_r$	4.0775	3.7781	5.2046	1.6029	2.1063	2.1071	4.7586
$\hat{U}_s$	4.0909	3.6107	5.0385	1.5776	1.8973	2.0761	4.2232

5 Simulations

We construct a simulation study to represent properties of the CEAP data and design. An extended set of simulations using the simulation models of Chen and Yu (2016) yields similar results and is not presented here for brevity. The objectives of the simulations are to evaluate the variance estimator and to compare QRI to nonparametric and fully parametric alternatives.

The fully parametric imputation procedure is parametric fractional imputation (Kim, 2011). The imputation model specified for parametric fractional imputation (PFI) is $y_i = \gamma_0 + \gamma_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. The imputed values for PFI are generated as, $y_{ij}^* \sim N(\hat{\gamma}_0 + \hat{\gamma}_1 x_i, \hat{\sigma}_\epsilon^2)$, where $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}_\epsilon^2)'$ satisfies $\mathbf{S}_w(\hat{\boldsymbol{\gamma}}) = \mathbf{0}$,

$$\mathbf{S}_w(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \pi_i^{-1} \delta_i \mathbf{d}_i, \tag{5.1}$$

and $\mathbf{d}_i = \left(y_i - \gamma_0 - \gamma_1 x_i, (y_i - \gamma_0 - \gamma_1 x_i) x_i, (y_i - \gamma_0 - \gamma_1 x_i)^2 / \sigma_\epsilon^2 - 1\right)'$. By incorporating π_i^{-1} in the score function (5.1), the estimator is consistent if the population model is a linear model with iid normally distributed errors and either the MAR assumption in (2.7) or (2.6) holds.

The non-parametric imputation (NPI) procedure is based on Wang and Chen (2009). For NPI, the j^{th} imputed value for nonrespondent i, y_{ij}^*, is generated from a multinomial distribution with sample space $\{y_s : I_s = \delta_s = 1\}$. Specifically,

$$P\left(y_{ij}^* = y_s\right) = \frac{\pi_i^{-1} K\{(x_i - x_s)/h\}}{\sum_{j=1}^{N} I_j \delta_j \pi_j^{-1} K\{(x_i - x_j)/h\}}, \tag{5.2}$$

where $K(\cdot)$ is a normal kernel with bandwidth h selected by applying the method of Sheather and Jones (1991), as implemented in the R function $dpik$, to $\{x_i : I_s = \delta_s = 1\}$.

The QRI procedure is implemented as described in Sections 2-3. To define the penalized B-spline, we set $p = 3$, $m = 2$, $K_n = 16$, and $\lambda = 0.004$. The value of $\lambda = 0.004$ is the median of the values selected using the R function “cobbs” across 1,000 samples of a preliminary simulation. To select λ using “cobbs”, we first use the R function “cobbs” to obtain λ_{τ_j} for $\tau_1, \ldots, \tau_J$. The selected λ is the minimum of the $\left\{\lambda_{\tau_j} : j = 1, \ldots, J\right\}$, which introduces the least amount of smoothing from among the selected λ_{τ_j}.

In simulations not presented here, we also consider multiple imputation. Modifications to standard multiple imputation procedures are needed to produce unbiased estimators for a situation in the sample missing at random assumption (2.6) does not hold (Berg et al., 2016; Reiter, Raghunathan and Kinney, 2006). Because an exploration of the modifications to multiple imputation needed to ensure consistent estimation is beyond the scope of this study, we restrict attention to PFI, NPI, and QRI.

For all three imputation procedures, GMM based on the imputed values is used to estimate the parameters. Note that this differs from Wang and Chen (2009), which uses empirical likelihood instead of

GMM. The number of imputations for the simulation is $J = 50$. The Monte Carlo (MC) sample size is 1,000.

We consider estimation of several parameters: $\theta_1 = E[y_i]$, $\theta_2 = V\{y_i\}$, $\theta_3 = \text{Cor}\{y_i, x_i\}$, $\theta_4 = E[E[y_i \mid x_i \le 0.65]]$, and $\theta_5 = P(y_i \le 8)$. With the exception of θ_5, GMM estimators of these parameters satisfy the assumptions required for the theory of Section 3. In particular, the function $\mathbf{g}_i(\cdot;\boldsymbol{\theta})$ defining the estimator of $(\theta_1, \theta_2, \theta_3, \theta_4)$ has two continuous derivatives. The estimator of θ_5 does not fall in the framework of Section 3 because $I[a \le 8]$ is a non-smooth function of a; however, we evaluate the empirical properties of $\hat{\theta}_5$ defined as

$$\hat{\theta}_5 = \left(\sum_{i=1}^{n} \pi_i^{-1}\right)^{-1} \sum_{i=1}^{n} \pi_i^{-1} \left\{\delta_i I[y_i \le 8] + (1-\delta_i) J^{-1} \sum_{j=1}^{J} I\left[y_{ij}^* \le 8\right]\right\}. \tag{5.3}$$

For details on the function $\mathbf{g}_i(\cdot;\boldsymbol{\theta})$ defining the estimators for the simulation, see Section D of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016).

5.1 Superpopulation model and design for simulations

The superpopulation model represents four aspects of the CEAP data and survey: (1) the shape of the expectation function, (2) the inclusion of a mean-variance relationship, (3) the use of probability proportional to size (PPS) with-replacement sampling, and (4) the sample sizes and response rates. The specific model for the simulation is $y_i = m(x_i) + e_i$, where $e_i \sim N\left(0, \sigma_e^2 m(x_i)^2\right)$, $m(x_i) = 2 + 10(1 + 8\exp(-5x_i))^{-\frac{5}{4}}$, and $x_i \sim \text{Trunc. Norm.}(0.5, 0.3)$. The sample design is PPS with replacement, where the probability of selecting unit i on a single draw is $\left(\sum_{i=1}^{N} \tilde{\psi}_i\right)^{-1} \tilde{\psi}_i$, $\text{logit}(\tilde{\psi}_i) = -3 - 0.33 z_i + 0.1 y_i$, $z_i \sim \text{Trunc. Norm.}(0.5, 0.3)$, and $N = 50{,}000$. The number of draws is $n = 1{,}500$, leading to a median sample size of 1,477, where the sample size is the number of unique units in the sample. The first and second order selection probabilities corresponding to $\tilde{\psi}_i$ are, $\pi_i = 1 - (1 - \tilde{\psi}_i)^n$, and $\pi_{ij} = 1 - (1 - \tilde{\psi}_i)^n - \left(1 - \tilde{\psi}_j\right)^n + \left(1 - \tilde{\psi}_j - \tilde{\psi}_i\right)^n$. The response indicator $\delta_i \sim \text{Bernoulli}(p_i)$, where $\text{logit}(p_i) = 0.5x_i + 1.5z_i$, which yields a median response rate of 0.631.

By the model for y_i given x_i, the assumption of population missing at random (2.7) holds for this simulation. Incorporating z_i in the models for p_i and π_i is the approach used in Berg et al. (2016) that causes the sample missing at random assumption (2.6) to fail. The variable z_i can be interpreted a design variable that is omitted from the imputation model.

5.2 Results

Table 5.1 contains three measures for comparing the QRI estimator to the PFI and NPI estimators. The percent relative MC MSE for estimator k ($k = \text{PFI, NPI}$) is defined,

$$\text{Pct. Rel. MSE}(k) = 100 \frac{\text{MSE}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{MSE}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}{\text{MSE}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}, \tag{5.4}$$

where $\hat{\theta}(k)$ is the estimator based on imputation procedure k. The percent relative variance for estimator k is defined,

$$\text{Pct. Rel. Var}(k) = 100\frac{\text{Var}_{\text{MC}}\left(\hat{\theta}(k)\right) - \text{Var}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}{\text{Var}_{\text{MC}}\left(\hat{\theta}(\text{QRI})\right)}, \tag{5.5}$$

for $k = \text{NPI, PFI}$. The percent of mean squared error due to squared bias is defined by

$$\text{Pct. Bias}(k) = 100\frac{\left(E_{\text{MC}}\left(\hat{\theta}(k)\right) - \theta\right)^2}{\text{MSE}_{\text{MC}}\left(\hat{\theta}(k)\right)}, \tag{5.6}$$

where $k = \text{NPI, PFI, QRI}$. The MSE of the QRI estimator is smaller than the MSE of the NPI and PFI estimators for all parameters. The PFI estimator is biased because the model underlying the PFI procedure does not account for the nonlinearity in the quantile curves or the nonconstant variances. The NPI procedure has a relatively large variance for sample sizes such as those obtained in the CEAP survey. The squared MC bias of the QRI procedure is less than 0.5% of MC MSE for all parameters.

The last two columns of Table 5.1 contain the relative bias of the variance estimator and the empirical coverage of normal theory 95% confidence intervals. The relative bias of the variance estimator defined as

$$\text{Rel. Bias} = \frac{E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right] - V_{\text{MC}}\left(\hat{\theta}\right)}{V_{\text{MC}},\left(\hat{\theta}\right)}, \tag{5.7}$$

where $E_{\text{MC}}\left[\hat{V}\left(\hat{\theta}\right)\right]$ is the MC mean of the variance estimators and $V_{\text{MC}}\left(\hat{\theta}\right)$ is the MC variance of the QRI estimator. The MC relative bias of the variance estimator for the QRI estimator is between -6% and -1%. Empirical coverages of normal theory confidence intervals are within 1% of the nominal 95% level.

Table 5.1
MC properties of estimators and variance estimators for simulation with PPS with replacement sample design. Pct. Rel. MSE (5.4): Difference between the MC variance of the PFI or NPI estimator and the MC MSE of the QRI estimator, relative to the MC MSE of the QRI estimator. Pct. Rel. Var. (5.5): Difference between the MC variance of the PFI or NPI estimator and the MC MSE of the QRI estimator, relative to the MC MSE of the QRI estimator. Pct. Bias (5.6): percent of MC MSE of PFI, NPI, and QRI estimators due to squared MC bias. Rel. Bias = MC relative bias of variance estimator defined in (5.7). Coverage = MC coverage of 95% confidence intervals

	Pct. Rel. MSE		Pct. Rel. Var.		Pct. Bias			Rel. Bias	Coverage
	NPI	PFI	NPI	PFI	NPI	PFI	QRI	QRI	QRI
θ_1	0.509	1.624	0.211	1.589	0.304	0.041	0.006	-2.386	0.945
θ_2	3.308	1.882	1.011	-0.151	2.225	1.998	0.002	-1.113	0.951
θ_3	1.518	5.449	0.979	2.605	0.840	2.999	0.311	-5.772	0.943
θ_4	515.980	26.752	10.501	12.415	82.101	11.508	0.222	-3.182	0.952
θ_5	5.879	61.416	5.659	-2.345	0.223	39.510	0.015	–	–

6 Discussion

QRI is developed for a complex survey setting. Alternative choices of weights are discussed, and a closed form variance estimator is provided based on a linear approximation. Consistency and asymptotic normality of the estimators are demonstrated under the framework of an infinite number of imputed values. In simulations designed to represent the CEAP data, the variance estimator based on the asymptotic distribution has a relative bias less than 6% in absolute value and leads to confidence intervals with coverage close to the nominal level for finite J. Further, the estimator based on QRI is more efficient than an estimator based on PFI or NPI because QRI provides a reasonable compromise between bias and variance.

The quantile regression imputation procedure is applied to estimate mean erosion in seven states in the midwestern United States using data from the Conservation Effects Assessment Project. The analysis demonstrates that QRI presents a viable alternative to weighting adjustments currently used to account for nonresponse in CEAP.

Areas for improvement to QRI include the choice of τ_j, the choice of b_i, refinements to estimation of the quantile curves, and variance estimation for non-differentiable $\mathbf{g}(\cdot)$ functions. Development of automated methods to select the nuisance parameters, appropriate for selection of multiple quantiles in a complex survey setting, is an area for future research. Estimation of the quantile curves subject to a restriction that the estimated curves are non-overlapping, has potential to improve estimation of the derivatives needed for the variance estimator. Section E of the online supplement https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf, (Berg and Yu, 2016) provides further discussion of areas for improvement.

Acknowledgements

This work was supported by Cooperative Agreement No. 68-3A75-4-122 between the USDA Natural Resources Conservation Service and the Center for Survey Statistics and Methodology at Iowa State University.

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78, 40-64.

Barrow, D.L., and Smith, P.W. (1978). Asymptotic properties of the best L_2 [0, 1] approximation by Splines with variable knots. *Quaterly of Applied Mathematics*, 33, 293-304.

Berg, E.J., and Yu, C. (2016). Supplement to "Semiparametric quantile regression imputation for a complex survey with application to the conservation effects assessment project". Available at: https://github.com/emilyjb/Semiparametric-QRI-Supplement/blob/master/SupplementToQRI.pdf.

Berg, E.J., Kim, J.K. and Skinner, C. (2016). Imputation under informative sampling. *Journal of Survey Statistics and Methodology*, 4, 436-462.

Breidt, F.J., and Fuller, W.A. (1999). Design of supplemented panel surveys with application to the National Resources Inventory. *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 391-403.

Brick, J.M., and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Chen, S., and Yu, C. (2016). Parameter estimation through semiparametric quantile regression imputation. *Electronical Journal of Statistics*, 10, 3621-3647.

Clayton, D., Spiegelhalter, D., Dunn, G. and Pickles, A. (1998). Analysis of longitudinal binary data from multiphase sampling. *Journal of the Royal Statistical Society, Series B*, 60, 71-87.

D'Arrigo, J., and Skinner, C. (2010). Linearization variance estimation for generalized raking estimators in the presence of nonresponse. *Survey Methodology*, 36, 2, 181-192. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2010002/article/11380-eng.pdf.

De Boor, C. (2001). *A Practical Guide to Splines* (Revised Edition), New York: Springer-Verlag.

Fuller, W.A. (1996). *Introduction to Statistical Time Series: Second Edition*. New York: John Wiley & Sons, Inc.

Fuller, W.A. (2009a). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Fuller, W.A. (2009b). *Sampling Statistics*. New York: John Wiley & Sons, Inc. Vol. 560.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer. Vol. 2, No. 1.

Isaki, T.C., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Jang, W., and Wang, J.H. (2015). A semiparametric Bayesian approach for joint-quantile regression with clustered data. *Computational Statistics and Data Analysis*, 84, 99-115.

Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78, 21-39.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis. *Biometrika*, 98(1), 119-132.

Kim, J.K., and Riddles, M.K. (2012). Some theory for propensity-score-adjustment estimators in survey sampling. *Survey Methodology*, 38, 2, 157-165. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11754-eng.pdf.

Kim, J.K., and Shao, J. (2013). *Statistical Methods for Handling Incomplete Data*, Chapman and Hall/CRC, Boca Raton.

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

Koenker, R. (2005). *Quantile Regression*. Cambridge university press. No. 38.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 2, 133-142. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9547-eng.pdf.

Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data missing values. *Applied Statistics*, 37, 23-38.

Mealli, F., and Rubin, D. (2015). Clarifying missing at random and related definitions, and implications when coupled with exchangeability. *Biometrika*, 102, 995-1000.

Nusser, S.M., and Goebel, J.J. (1997). The National Resources Inventory: A long-term multi-resource monitoring programme. *Environmental and Ecological Statistics*, 4(3), 181-204.

Nusser, S.M. (2006). National Resources Inventory (NRI), US. *Encyclopedia of Environmetrics Second Edition*, 1-3.

Nusser, S.M., Carriquiry, A.L., Dodd, K.W. and Fuller, W.A. (1996). A semiparametric transformation approach to estimating usual daily intake distributions. *Journal of the American Statistical Association*, 91(436), 1440-1449.

Pakes, A., and Pollard, D. (1989). Simulation and the asymptotic of optimization estimators. *Econometrica*, 57(4), 1027-1057.

Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37, 2, 115-136. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2011002/article/11602-eng.pdf.

Reiter, J.P., Raghunathan, T.E. and Kinney, S.K. (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology*, 32, 2, 143-149. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2006002/article/9548-eng.pdf.

Robins, J.M., and Wang, N. (2000). Inference for imputation estimators. *Biometrika*. 87, 113-124.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc. Vol. 81.

Sheather, S.J., and Jones, M.C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

Wang, D., and Chen, S.X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 490-517.

Wei, Y., Ma, Y. and Carroll, R.J. (2012). Multiple imputation in quantile regression. *Biometrika*, 99, 423-438.

Yoshida, T. (2013). Asymptotics for penalized spline estimators in quantile regression. *Communications in Statistics - Theory and Methods*, DOI 10.1080/03610926.2013.765477.