UFRJ

UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO

Instituto de Matemática

Bayesian quantile regression analysis of complex survey data under informative sampling

Marcus Gerardus Lavagnole Nascimento

Advisor: Kelly Cristina Mota Gonçalves

2024

Bayesian quantile regression analysis of complex survey data under informative sampling

Marcus Gerardus Lavagnole Nascimento

Advisor: Kelly Cristina Mota Gonçalves

A dissertation submitted following the requirements of the degree of Doctor of Science in the Department of Statistical Methods.

Thesis Committee:

Prof. Kelly Cristina Mota Gonçalves, D.Sc.
Universidade Federal do Rio de Janeiro

Prof. Carlos Antonio Abanto-Valle, D.Sc.
Universidade Federal do Rio de Janeiro

Prof. Fernando Antonio da Silva Moura, Ph.D.
Universidade Federal do Rio de Janeiro

Prof. Bruno Santos, D.Sc.
Stockholm University

Prof. Pedro Luis do Nascimento Silva, Ph.D.
Escola Nacional de Ciências Estatísticas

Rio de Janeiro, RJ - Brazil
2024

Acknowledgements

Firstly, I would like to express my gratitude and appreciation for my supervisor, Professor Kelly Cristina Mota Gonçalves, who has encouraged and supported me throughout this journey. I am grateful for all the valuable comments and recommendations on this dissertation.

I would also like to thank Professors Rodrigo S. Targino and Leandro P. R. Pimentel for their energy, assistance, and comprehension at the beginning of this academic endeavor.

From the bottom of my heart, I would like to express all my love to my wife, Rebecca de Oliveira Souza, and to say a big thank you for all the unconditional support and patience.

To conclude, I cannot forget to thank my parents, Maria Christina Lavagnole and Leobaldo Silveira Nascimento, and my sisters, Hanna Lavagnole Nascimento and Rebecca Lavagnole Nascimento, who have always been by my side. I could not have done this without you.

Abstract

The interest in considering the relation among random variables in quantiles instead of the mean has emerged in various fields, and data collected from complex survey designs are of fundamental importance to different areas. The combination of both frameworks provides a powerful tool for supporting decisions and is useful for practitioners from diverse backgrounds. In this dissertation, we aim to advance in this literature by investigating new developments and extensions of Bayesian methods for quantile regression analysis of complex survey data under informative sampling. We not only focus on the absolutely continuous case as the previous works on the topic but also develop methods for count data and multiple outputs. Our methods are particularly appealing as they provide effective and easy-to-implement methodological tools.

Contents

List of Tables

List of Figures

1 Introduction

Over the last century, sampling has proved to be a fundamental tool to permit society to collect a large range of information about populations accurately. Complex survey designs, in particular, are essential to official statistical organizations and to the scientific advance in different areas: economics (Ferreira, Firpo, and Messina 2022), education (Binelli and Menezes-Filho 2019), public policy (Olson, Clark, and Reynolds 2019), health (Liu et al. 2018), to mention a few. It may be recognized as different from simple random sampling, in which each unit in the population is selected with equal probability. We refer to Valliant, Dever, and Kreuter (2018) for a detailed account of the various designs used in practice.

Three main motives for implementing complex survey designs instead of simple random sampling can be underlined. The first is for efficient estimation. Tillé and Wilhelm (2017) mentioned the "false intuition that a sample must be similar to a population" and pointed out how one can frequently estimate more efficiently by sampling units with unequal probabilities. The second is related to practical constraints introduced by the characteristics of the population target from which the sample must be drawn. Last but not least is the variable costs associated with data collection.

The sampling design formulation often establishes a correlation between the response variable of interest and sampling inclusion probabilities. Survey designs that instantiate this trait are denominated "informative" since the inclusion probabilities are informed by the response variable. An example is the proportion-to-size sampling design in the Current Employment Statistics (CES) survey employed by the U.S. Bureau of Labor Statistics. The CES is conducted to enable total employment statistics computation for area and industry-indexed do-

mains. Establishments with larger employment numbers are imputed higher unit inclusion probabilities.

Therefore, the information correspondence between the observed sample and the population differs in informative sampling designs. In this context, proposing a method on the observed sample may result in biased inference about the population, and accounting for the design might be required. Before this scenario, various methodologies were developed to account for informative sampling. In one class, the sampling design is parameterized into the model estimated on the sampled data (Little 2004). However, as highlighted by Leon-Novelo and Savitsky (2019), the sampling design is often a nuisance to the analyst, and marginalization over parameters indexing the sampling design distribution is needed. Further, it is also possible for the analyst not to know the sampling design to parameterize it.

Another class of methods utilizes sampling weights inversely proportional to the marginal inclusion probability to adjust the likelihood contribution for each unit in the observed sample. In this class, some approaches require a specific form for the likelihood, which does not permit the analyst to specify the population model of interest. Rao and Wu (2010), Dong, Elliott, and Raghunathan (2014), and Kunihama et al. (2016) imposed Dirichlet distribution priors for the mixture components, taking the hyperparameters as a function of the sampling weights. For designs where subgroups of sampled units have equal weights, Si, Pillai, and Gelman (2015) regressed the dependent variable on a Gaussian process function of the weights. The inferential focus of these approaches is the domain-level estimation of simple mean and total statistics.

Alternatively, some approaches are designed for inference about parameters that characterize an analyst-specified population model. For example, Pfeffermann, Krieger, and Rinott (1998) proposed a general formulation for a population model by applying the Bayes rule approach to derive a marginal sample probability density function to the variables of interest. Their approach allows discrete or continuous variables and the density function to depend on known values of concomitant variables, as in regression-type models. The approach was extended in different ways. Pfeffermann, Moura, and Silva (2006) and Silva and Moura (2022) introduced

univariate and multivariate multi-level models, respectively. Leon-Novelo and Savitsky (2019) adapted the approach to formulate a fully Bayesian method.

In addition, some formulations account for the informative sampling design by exponentiating each unit likelihood contribution under the analyst-specified model by its associated sampling weight. The resultant function, called pseudo-likelihood Chambers and Skinner (2003), possibly allows discrete or continuous variables and regression-type models. Following this, Savitsky and Toth (2016) built a sampling-weighted pseudo posterior distribution by convolving the pseudo-likelihood with the prior distributions for model parameters.

Although many methods that account for informative sampling enable regression-type models, only a few allow for quantile regression analysis. Quantile regression models comprehensively characterize the relationship between response and explanatory variables, provide a robust alternative to heteroscedasticity and outliers, and encompass a valuable tool for practitioners from myriad applied fields. Villarini et al. (2011) studied extreme rainfall by applying this tool to annual maximum daily rainfall records from 221 rain gages in the Midwest United States. Neelon et al. (2015) investigated emergency department-related medical expenditures and proposed a spatiotemporal quantile regression model due to a more pronounced variation in the extremes. Eide and Showalter (1999) estimated intergenerational earnings mobility models using quantile regression due to its less restrictive nature.

Concerning quantile regression models for complex survey data in a more general framework, Geraci (2016) offered guidance to analyze a continuous outcome when the variables of interest are partially observed. Under missing at random assumptions, imputation models are introduced. The motivating study regards the investigation of birthweight determinants in a UK-based cohort of children. (Zhao et al. 2020) in turn developed a Bayesian empirical likelihood approach based on estimating equations, which includes the quantile regression one. The authors demonstrated that their estimator is consistent, and the credible intervals are valid in the sense of producing asymptotically design-based frequentist properties.

When we look at methods that combine quantile regression and informative sampling, the work of Chen and Zhao (2019) is possibly the primary reference. The authors proposed

different weight-smoothing estimators that combine nonparametric methods for modeling the weight functions. Pseudo-population bootstrap methods are applied for variance estimation with associated confidence regions. Their methods are illustrated with the 1988 US National Maternal and Infant Health Survey. The paper of Wang, Kim, and Yang (2018) is also a relevant reference in the topic. They developed an approximate Bayesian approach that adopts the sampling distribution of a summary statistic to find the posterior distribution of the parameters of interest. Their general formulation includes quantile regression as a particular case.

In this dissertation, we aim to advance in this literature by investigating new developments and extensions of Bayesian methods for quantile regression analysis of complex survey data under informative sampling. We not only focus on the absolutely continuous case as the previous articles on the topic but also develop methods for count data and multiple outputs.

1.1 An overview for the thesis

The work is divided into three main chapters (Chapters 2, 3, and 4), structured as three independent papers.

As the primary reference combining quantile regression and informative sampling is Chen and Zhao (2019) and their work is restricted to a frequentist framework, in Chapter 2, we introduce different Bayesian methods relying on the survey-weighted estimator (Chen and Zhao 2019; Geraci 2016) and the estimating equations (Wang, Kim, and Yang 2018; Zhao et al. 2020). As Chen and Zhao (2019), we focus first on continuous response variables. We evaluate our methods in a model-based simulation study by considering different data-generating processes and informative and non-informative scenarios. We also propose a design-based simulation study. The content presented in Chapter 2 consists of a first preprint version of a paper entitled "Bayesian quantile regression models for complex survey data under informative sampling" already accepted for publication in the Journal of Survey Statistics and Methodology.

To the best of our knowledge, the previous works focused only on models for continuous response variables. With that in mind, in Chapter 3, we develop Bayesian quantile regression

models that appropriately deal with count outcome variables bounded by a known range. Our methods explore ideas related to the well-known quantile regression based on the asymmetric Laplace distribution and pseudo distributions. We evaluate the proposed methods in a design-based simulation study compared to a naive model fitting that ignores the informative sampling design under different scenarios. This chapter consists of the first preprint version of a paper entitled "Bayesian quantile regression models for bounded count data under informative sampling," which is under a submission process for possible publication.

In Chapter 4, we develop a Bayesian multiple-output quantile regression. We apply the ideas presented in Chapter 2 and those introduced by Guggisberg (2023). Concerned about computational efficiency, we provide an Expectation-Maximization approach in line with Zhao and Lian (2016). The methods are assessed through model and design-based simulation studies. This is an advanced work in progress that we intend to submit shortly.

Chapter 5 mainly briefly describes the ideas we plan to implement in future works.

All methods developed in this dissertation are implemented in R Core Team (2021), and the codes are available at `https://github.com/marcuslavagnole/BWQR_Informative_Sampling`.

2 Quantiles for absolutely continuous data under informative sampling

2.1 Introduction

Data collected from complex survey designs are of fundamental importance to the scientific advance in economics (Ferreira, Firpo, and Messina 2022), education (Binelli and Menezes-Filho 2019), public policy (Olson, Clark, and Reynolds 2019), and health (Liu et al. 2018), to mention a few. Complex survey designs are referred to as informative when the probabilities of selection for the units in the sample are associated with the variable of interest conditional on covariates (Fuller 2009). The statistical analysis of complex survey data under informative sampling that fails to account for this relation, ignoring design features, which includes stratification, clustering, and unequal weighting, possibly leads to biased results in the estimation of parameters that index the joint distribution assumed to have generated the population (Pfeffermann and Sverchkov 1999, 2003, 2009). For example, the Survey of Occupational Illnesses and Injuries (SOII) uses a stratified sampling design to capture work-related injuries or illnesses of workers who require medical care beyond first aid. The survey assigns higher sample inclusion probabilities to establishments that historically express higher injury rates, and predictions about illnesses and injuries among the population of establishments based on regression models will be biased.

Design-based methods using the design weights have been widely studied and applied to model the conditional mean values of the study variables by regression. Pfeffermann (2011), Scott and Wild (2011), among others, use the joint model of study variable and sampling indicator to explore a likelihood-based method that maximizes the conditional sample likelihood. Magee (1998), Beaumont (2008), and Kim and Skinner (2013), on the other hand, use predictions from a model for the conditional distribution of the design weights given the data to substitute the original design weights. Ultimately, Asparouhov (2006), Savitsky and Toth (2016), and Leon-Novelo and Savitsky (2019) replace the likelihood with the pseudo-likelihood.

However, the interest in considering the relation among random variables in quantiles instead of the mean has emerged in various fields. Quantile regression provides a more comprehensive description of the relationship between a response variable and covariates, and it is a robust alternative to heteroscedasticity and outliers. Villarini et al. (2011) studied extreme rainfall by applying this tool to annual maximum daily rainfall records from 221 rain gages in the Midwest United States. Neelon et al. (2015) investigated emergency department-related medical expenditures and proposed a spatiotemporal quantile regression model due to a more pronounced variation in the extremes. Eide and Showalter (1999) estimated intergenerational earnings mobility models using quantile regression due to its less restrictive nature.

Despite a myriad of papers developed and the extensive literature on survey data analysis and quantile regression models, only a few research papers explore quantile regression estimation accounting for complex survey sampling. In this regard, Li, Graubard, and Korn (2010) modified the double-kernel method and the bandwidth selection procedure introduced by Yu and Jones (1998) to include survey sample weights. Geraci (2016) described a general methodology for handling quantile regression analysis of complex survey data when the variables of interest are partially observed. Chen and Zhao (2019) focused on the quantile regression analysis of complex survey data under informative sampling designs and proposed several weight-smoothing estimators. Zhao et al. (2020) proposed a general formulation that includes quantile regression models to complex survey data analysis through a Bayesian empirical likelihood approach (Chen and Kim 2014; Kim 2009). From these papers, only Chen and Zhao (2019) focused on the estimation under informative sampling, and only Zhao et al.

(2020) explored a Bayesian estimation method.

In this chapter, we develop two new Bayesian methods where the quantile regression coefficients are defined at the super-population level, and their estimators are built upon the survey weights. First, proceeding from a survey-weighted estimator (Chen and Zhao 2019; Geraci 2016), we extend the well-known quantile regression model based on the asymmetric Laplace distribution to complex survey data under informative sampling following a similar argument of Yu and Moyeed (2001). From the ideas of Kozumi and Kobayashi (2011), we implement a simple Gibbs sampling algorithm for fitting the proposed model. Second, building on the estimating equations (Zhao et al. 2020), we extend the quantile regression model using a score likelihood (Wu and Narisetty 2021) to include sampling weights applying the results of Huang, Xu, and Tashnev (2015). We then carry out an adaptive Metropolis-Hastings (Shaby and Wells 2010) for sampling from the model parameters.

Our methods are compared with a naive model fitting, ignoring the informative sampling design in model-based and real-data-based simulations. In the first, four data-generating models for the population are specified. The models encompass symmetric and asymmetric cases, homoscedastic and heteroscedastic cases, and heavy and light-tailed cases. Consequently, we evaluate the performance of the proposed methods under different scenarios in terms of fitting and mean absolute error and variance of the estimators. In the second, we have a design-based study in which students in the ninth year from public elementary schools located in Campinas municipality in São Paulo State are fixed as the finite population, and their results in Prova Brasil are analyzed considering an informative sampling design. We compare the point estimates of the proposed methods and the naive model with estimates that consider the fully observed population and assess their uncertainty level.

Since modeling under informative sampling and quantile regression models are powerful tools for supporting decisions in different areas and are applied by practitioners from different backgrounds, our main contribution in this chapter is to provide two new and intuitive frameworks that are particularly appealing as they offer easy-to-implement methodological tools. For example, the asymmetric Laplace distribution is already widely applied in the quantile regression

literature, with extension to ordinal data (Rahman 2016), count data (Lee and Neocleous 2010), longitudinal data (Yuan and Yin 2010), state-space models (Gonçalves, Migon, and Bastos 2020), and spatial models (Lum and Gelfand 2012), among others.

2.2 Basic setup

Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, $\mathbf{x}_i \in \mathbb{R}^p$, denote a vector of explanatory variables, and y_i denote the study variable. Suppose that a finite population $\mathcal{F}_N = \{(y_i, \mathbf{x}_i), i = 1, \dots, N\}$ is generated from a super-population model $\mathcal{F}$ for which the true underlying distribution function is unknown. Nonetheless, through a linear quantile regression model, we can assume that for a fixed quantile level $\tau \in (0, 1)$ the finite population τ-th conditional quantile function of Y_i given $\mathbf{X}_i$ takes the parametric form

$$Q_{Y_i}(\tau|\mathbf{X}_i = \mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}(\tau), \;\; i = 1, \dots, N, \tag{1}$$

where $Q_{Y_i}(\tau|\mathbf{X}_i = \mathbf{x}_i) \equiv F^{-1}_{Y_i}(\tau; \mathbf{X}_i = \mathbf{x}_i)$, $F_{Y_i}(\tau; \mathbf{X}_i = \mathbf{x}_i)$ is the unknown cumulative distribution function of Y_i given $\mathbf{X}_i = \mathbf{x}_i$ evaluated at τ, and $\boldsymbol{\beta}(\tau) = (\beta_1(\tau), \dots, \beta_p(\tau))'$ is a p-dimensional vector of coefficients which depends on the quantile level of interest τ.

The true values of the unknown finite population parameters $\boldsymbol{\beta}_N(\tau)$ are defined as the solution to the optimization problem

$$\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{N} \rho_\tau\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right), \tag{2}$$

where $\rho_\tau(\ell) = \ell(\tau - \mathbb{1}(\ell < 0))$ is the asymmetric check loss function (Koenker and Bassett Jr. 1978). However, solving the expression (2) is not feasible in many real applications as we usually do not observe the entire population. Our problem, therefore, lies in making inference about the population parameters from a survey sample.

Suppose a survey sample $S = \{i_1, \dots, i_n\}$ with size $|S| = n$ in which the units are sampled with inclusion probability $\pi_i = \mathbb{P}(I_i = 1)$, where $I_i \in \{0, 1\}$ is a sampling indicator such that

$I_i = 1$ if unit i is selected and $I_i = 0$ otherwise. For cases in which S is obtained under Simple Random Sampling (SRS), meaning that π_i is the same for all i, the estimator for the finite population vector of parameters is

$$\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in S} \rho_\tau \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right).$$

When data is collected from designs more complex than the SRS, in particular, when the sampling design is informative, $\mathbb{P}(I_i = 1|x_i, y_i) \neq \mathbb{P}(I_i = 1|x_i)$, it is required to consider the design in the estimation process and weighting might be used to account for unequal inclusion probabilities assigned to sample observations. In the following two sections, we present new methods for quantile regression analysis of survey data under informative sampling based on survey-weighted estimators.

2.3 The asymmetric Laplace distribution as working likelihood

2.3.1 The setup

From the minimization problem in (2), Geraci (2016) and Chen and Zhao (2019) defined a survey-weighted estimator for making inference about the finite population parameters as

$$\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in S} w_i \rho_\tau \left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right). \tag{3}$$

As Savitsky and Toth (2016), we assume normalized sampling weights $w_i = n\tilde{w}_i / \sum_{i\in S} \tilde{w}_i$, where $\tilde{w}_i = 1/\pi_i$ are the sampling weights. From the previous expression, we build a similar argument to that of Yu and Moyeed (2001), who observed that the minimization of $\rho_\tau(\ell)$ is equivalent to the maximization of a likelihood function reached from independently distributed asymmetric Laplace densities.

Therefore, noticing that the minimization problem in (3) is equivalent to

$$\begin{aligned}\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\text{argmax}}\ \exp\left\{-\sum_{i\in S} w_i\rho_\tau\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\} &= \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\text{argmax}}\prod_{i\in S}\exp\left\{-w_i\rho_\tau\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\}\\ &= \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\text{argmax}}\prod_{i\in S} w_i\tau\left(1-\tau\right)\\ &\times\ \exp\left\{-w_i\rho_\tau\left(y_i-\mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\},\end{aligned}$$

which is the maximization of a likelihood function formed by combining independently distributed asymmetric Laplace densities with a scale equal to $1/w_i$, we propose a Bayesian quantile regression for complex survey data based on the asymmetric Laplace distribution (ALD). Instead of working with a fixed scale equals to $1/w_i$, we can also incorporate a scale parameter to obtain an ALD $(\eta_i(\tau), \tilde{\sigma}_i, \tau)$ with location parameter equals to $\eta_i(\tau) = \mathbf{x}_i'\boldsymbol{\beta}(\tau) \in \mathbb{R}$, scale parameter equals to $\tilde{\sigma}_i = \sigma/w_i \in (0, \infty)$, skewness parameter equals to τ, and likelihood function given by

$$\mathcal{L}(\boldsymbol{\beta}(\tau), \sigma; \mathbf{y}, \mathbf{x}, \mathbf{w}) = \prod_{i\in S} w_i \frac{\tau\left(1-\tau\right)}{\sigma}\ \exp\left\{-w_i\rho_\tau\left(\frac{y_i-\mathbf{x}_i'\boldsymbol{\beta}(\tau)}{\sigma}\right)\right\}, \tag{4}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a column vector, $\mathbf{x}$ is $p \times n$-dimensional matrix, and $\mathbf{w}$ is a n-dimensional column vector.

We, therefore, are not assuming the finite population is generated from a super-population model with an ALD likelihood function. Instead, it is used as a "working likelihood" for Bayesian quantile inference, deeming it as an instrument for efficient regression parameter estimation. More often than not, it is a misspecification of the true underlying likelihood.

2.3.2 Mixture representation

The mixture representation of the asymmetric Laplace distribution (Kotz, Kozubowski, and Podgórski 2001), which is based on data augmentation ideas (Tanner and Wong 1987), includes latent variables $\boldsymbol{\nu} = \{\nu_i : i \in S\}$ in the observed data $\{\mathbf{y}, \mathbf{x}, \mathbf{w}\}$ and, potentially, simplifies the posterior analysis. Kozumi and Kobayashi (2011), for example, exploit this to propose a Gibbs sampler to the linear quantile regression model. Following Kozumi and

Kobayashi (2011), we assume a mixture representation for the ALD, which can be represented by a location-scale mixture of normals as below:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}(\tau) + \tilde{\sigma}_i\theta\tilde{\nu}_i + \tilde{\sigma}_i\gamma\sqrt{\tilde{\nu}_i}\varepsilon_i, \tag{5}$$

in which $\theta = \frac{1-2\tau}{\tau(1-\tau)}$, $\gamma^2 = \frac{2}{\tau(1-\tau)}$, $\tilde{\nu}_i \sim \text{Exp}(1)$ and $\varepsilon_i \sim N(0,1)$ are mutually independent, and $\text{Exp}(\zeta)$ denotes an exponential distribution with mean ζ. A reparameterization of the expression (5) is given by

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}(\tau) + \theta\nu_i + \gamma\sqrt{\tilde{\sigma}_i\nu_i}\varepsilon_i, \tag{6}$$

where $\nu_i \sim \text{Exp}(\tilde{\sigma}_i)$.

As a consequence of the mixture representation in (6), we may express our quantile regression as the hierarchical model

$$\begin{aligned} y_i|\nu_i, \mathbf{x}_i, w_i, \boldsymbol{\beta}(\tau), \sigma &\sim N\left(\mathbf{x}_i'\boldsymbol{\beta}(\tau) + \theta\nu_i, \gamma^2\nu_i\sigma/w_i\right) \\ \nu_i|w_i, \sigma &\sim \text{Exp}(\sigma/w_i), \end{aligned} \tag{7}$$

and, from this hierarchical representation, the joint density function of the observed sample responses $\mathbf{y}$ and the latent variables $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ is written as

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\nu}|\mathbf{x}, \mathbf{w}, \boldsymbol{\beta}(\tau), \sigma) &= f(\mathbf{y}|\boldsymbol{\nu}, \mathbf{x}, \mathbf{w}, \boldsymbol{\beta}(\tau), \sigma) f(\boldsymbol{\nu}|\mathbf{w}, \sigma) \\ &= \prod_{i\in S} w_i^{1/2} \left(2\pi\sigma\nu_i\right)^{-1/2} \exp\left\{-\frac{w_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau) - \theta\nu_i\right)^2}{2\gamma^2\sigma\nu_i}\right\} \\ &\times \left(\sigma/w_i\right)^{-1} \exp\left\{-\frac{w_i\nu_i}{\sigma}\right\} \\ &= \left(\prod_{i\in S} w_i^{3/2}\nu_i^{-1/2}\right)(2\pi)^{-n/2}\sigma^{-3n/2} \\ &\times \exp\left\{-\sum_{i\in S}\left[\frac{w_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}(\tau) - \theta\nu_i\right)^2}{2\gamma^2\sigma\nu_i} + \frac{w_i\nu_i}{\sigma}\right]\right\}. \end{aligned}$$

2.3.3 Bayesian inference and computation

Our main interest lies in the posterior inference of $\boldsymbol{\lambda}(\tau) = (\boldsymbol{\beta}(\tau), \sigma)$ based on the observed data. However, sampling from $p(\boldsymbol{\lambda}(\tau)|\mathbf{y}, \mathbf{x}, \mathbf{w})$ is rather difficult since we do not have a closed form to this posterior density, and, in this case, Markov chain Monte Carlo (Gamerman and Lopes 2006, MCMC) algorithms to iteratively simulate observations for $\boldsymbol{\lambda}(\tau)$ from $p(\boldsymbol{\beta}(\tau)|\mathbf{y}, \mathbf{x}, \mathbf{w}, \sigma)$ and $p(\sigma|\mathbf{y}, \mathbf{x}, \mathbf{w}, \boldsymbol{\beta}(\tau))$ are a handy tool.

Proceeding with the Bayesian analysis, we exploit the mixture representation described in the previous section and assume prior independence between $\boldsymbol{\beta}(\tau)$ and σ, which means that $p(\boldsymbol{\lambda}(\tau)) = p(\boldsymbol{\beta}(\tau))p(\sigma)$. We then specify the prior distributions $\boldsymbol{\beta}(\tau) \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ and $\sigma \sim IG(a_0, b_0)$, where $IG(a, b)$ denotes the Inverse Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Consequently, the posterior density function of $\boldsymbol{\lambda}(\tau)$ and the latent variables conditional on the observed data is

$$
\begin{aligned}
p(\boldsymbol{\lambda}(\tau), \boldsymbol{\nu}|\mathbf{y}, \mathbf{x}, \mathbf{w}) &\propto f(\mathbf{y}|\boldsymbol{\nu}, \mathbf{x}, \mathbf{w}, \boldsymbol{\beta}(\tau), \sigma) f(\boldsymbol{\nu}|\mathbf{w}, \sigma) p(\boldsymbol{\beta}(\tau)) p(\sigma) \\
&\propto \sigma^{-3n/2} \exp\left\{ -\sum_{i \in S} \left[\frac{w_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau) - \theta \nu_i\right)^2}{2\gamma^2 \sigma \nu_i} + \frac{w_i \nu_i}{\sigma} \right] \right\} \\
&\times \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\beta}(\tau) - \boldsymbol{\mu}_0\right)' \Sigma_0^{-1} \left(\boldsymbol{\beta}(\tau) - \boldsymbol{\mu}_0\right) \right\} \\
&\times \sigma^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma} \right\}.
\end{aligned}
\tag{8}
$$

From (8), the resultant full conditional densities have closed form as follows

- $\boldsymbol{\beta}(\tau) \sim N(\boldsymbol{\mu}_1, \Sigma_1)$, where
  $\Sigma_1 = \left[\sum_{i \in S} \frac{w_i}{\gamma^2 \sigma \nu_i} \mathbf{x}_i \mathbf{x}_i' + \Sigma_0^{-1} \right]^{-1}$ and $\boldsymbol{\mu}_1 = \Sigma_1 \left[\sum_{i \in S} \frac{w_i (y_i - \theta \nu_i)}{\gamma^2 \sigma \nu_i} \mathbf{x}_i + \Sigma_0^{-1} \boldsymbol{\mu}_0 \right]$;
- $\sigma \sim IG\left(a_0 + \frac{3n}{2}, b_0 + \sum_{i \in S} \left[\frac{w_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau) - \theta \nu_i\right)^2}{2\gamma^2 \nu_i} + w_i \nu_i \right] \right)$;
- $\nu_i \sim GIG\left(\frac{1}{2}, \frac{w_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau)\right)^2}{\gamma^2 \sigma}, \frac{w_i \theta^2}{\gamma^2 \sigma} + \frac{2 w_i}{\sigma} \right)$ for $i \in S$.

As in Kozumi and Kobayashi (2011), the main advantage of this approach is a simple MCMC algorithm with Gibbs steps only.

2.4 Score based working likelihood

2.4.1 The setup

We can look at the population estimating equations instead of considering the minimization problem in (2). Following an argument similar to that of Koenker (2005), it is possible to demonstrate that solving the optimization problem in (2) is analogous to solving the following estimating equations for the coefficients $\boldsymbol{\beta}(\tau)$:

$$U_N\left(\boldsymbol{\beta}\left(\tau\right)\right) = \sum_{i=1}^{N} \mathbf{x}_i \psi_\tau\left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau)\right) = 0,$$

where $\psi_\tau(\ell) = \tau - \mathbb{1}(\ell < 0)$. Consequently, we can define survey-weighted estimating functions as

$$U_{\mathcal{S}}\left(\boldsymbol{\beta}\left(\tau\right)\right) = \sum_{i \in \mathcal{S}} w_i \mathbf{x}_i \psi_\tau\left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau)\right), \tag{9}$$

and obtain a survey-based estimator of the finite population parameters $\boldsymbol{\beta}_N(\tau)$ by solving $U_{\mathcal{S}}\left(\boldsymbol{\beta}\left(\tau\right)\right) = 0$. For further discussions on estimating-equation-based approaches in survey sampling, see Binder (1983) and Godambe and Thompson (1986).

Observing that the Equation (9) is a weighted sum of the score function for the quantile regression objective function in (2), we follow the ideas of Wu and Narisetty (2021) to propose a likelihood as below:

$$\mathcal{L}(\boldsymbol{\beta}(\tau); \mathbf{y}, \mathbf{x}, \mathbf{w}) = C \, \exp\left\{-\frac{1}{2n} U_{\mathcal{S}}\left(\boldsymbol{\beta}\left(\tau\right)\right)' \Omega U_{\mathcal{S}}\left(\boldsymbol{\beta}\left(\tau\right)\right)\right\}, \tag{10}$$

in which C is a constant free of $\boldsymbol{\beta}\left(\tau\right)$ and Ω is a $p \times p$ positive definite weight matrix. To define Ω, we adapt the weight matrix introduced by Wu and Narisetty (2021) to the results in Huang, Xu, and Tashnev (2015) to obtain

$$\Omega = \frac{n}{\tau(1-\tau)} \left(\sum_{i \in \mathcal{S}} w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)^{-1}.$$

As argued by Wu and Narisetty (2021), due to the quadratic form of the exponent in the posterior, there will be a correspondence between values of $\boldsymbol{\beta}(\tau)$ for which $U_{\mathcal{S}}(\boldsymbol{\beta}(\tau))$ is close to zero and the values of high posterior density. Furthermore, as the expected value of the score function is equal to zero only at the true value $\boldsymbol{\beta}_N(\tau)$, a posterior concentration around the truth is presumed.

Similarly to the previous approach, the likelihood in Equation (10) does not arise from a distributional specification for the super-population model, and, once again, we are using it as a "working likelihood" for Bayesian quantile inference, deeming it as an instrument for efficient regression parameter estimation.

2.4.2 Bayesian inference and computation

Proceeding with the Bayesian analysis, we specify a prior distribution $\boldsymbol{\beta}(\tau) \sim N(\boldsymbol{\mu}_0, \Sigma_0)$. Thus, the full conditional posterior distribution for the quantile regression coefficients $\boldsymbol{\beta}(\tau)$ is described by the function

$$\begin{aligned} p(\boldsymbol{\beta}(\tau)|\mathbf{y},\mathbf{x},\mathbf{w}) &\propto \mathcal{L}(\boldsymbol{\beta}(\tau);\mathbf{y},\mathbf{x},\mathbf{w})p(\boldsymbol{\beta}(\tau)) \\ &\propto \exp\left\{-\frac{1}{2n}U_{\mathcal{S}}(\boldsymbol{\beta}(\tau))'\,\Omega\,U_{\mathcal{S}}(\boldsymbol{\beta}(\tau))\right\} \\ &\times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}(\tau)-\boldsymbol{\mu}_0)'\,\Sigma_0^{-1}(\boldsymbol{\beta}(\tau)-\boldsymbol{\mu}_0)\right\}. \end{aligned} \tag{11}$$

Wu and Narisetty (2021) propose an importance sampling algorithm for computing the posterior estimates. Importance sampling is widely used as an alternative to MCMC, but its performance relies on a good proposal distribution (Robert and Casella 2013). Building a good proposal distribution for making inference about the true values of the unknown finite population parameters when we have complex survey data under informative sampling might be more complicated than when data is obtained under an SRS scheme. For that reason, we employ a Metropolis–Hastings algorithm to propose an MCMC procedure for sampling from the full posterior distribution in (11).

Let $\boldsymbol{\beta}^{(0)}$ be an initial value for the coefficients, and $\delta_m^{(0)}$ be an initial value for the tuning

quantity δ_m. Following the strategy of Shaby and Wells (2010), we employ an adaptive approach to update δ_m in an iterative manner. A sample of draws from the posterior density, therefore, is obtained by repeating the following three steps for $t = 0, \ldots, T$:

- Given the current state $\boldsymbol{\beta}^{(t)}$, generate $\boldsymbol{\beta}^*$ from a proposal distribution $\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)} \sim N(\boldsymbol{\beta}^{(t)}, \hat{\Sigma})$, where

$$\hat{\Sigma} = \delta^{(t)}\tau(1-\tau)\left(\frac{1}{n}\sum_{i\in\mathcal{S}} w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)^{-1};$$

- Generate $U \sim \text{Unif}(0,1)$, the uniform distribution on the interval $(0,1)$, and compute $\boldsymbol{\beta}^{(t+1)}$ applying the acceptance-rejection rule as follows:

$$\boldsymbol{\beta}^{(t+1)} = \begin{cases} \boldsymbol{\beta}^*, \text{ if } u \leq \alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) \\ \boldsymbol{\beta}^{(t)}, \text{ otherwise,} \end{cases}$$

where

$$\begin{aligned} \alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) &= \min\left\{1, \frac{p(\boldsymbol{\beta}^*|\mathbf{y},\mathbf{x},\mathbf{w})q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^{(t)}|\mathbf{y},\mathbf{x},\mathbf{w})q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})}\right\} \\ &= \min\left\{1, \frac{p(\boldsymbol{\beta}^*|\mathbf{y},\mathbf{x},\mathbf{w})}{p(\boldsymbol{\beta}^{(t)}|\mathbf{y},\mathbf{x},\mathbf{w})}\right\} \end{aligned}$$

- For every m steps, calculate $\hat{r}^{(t)} = \#\text{jumps}/m$ and $\kappa^{(t)} = k_0(1/t^{k_1})$, where $\#$jumps indicates the number of jumps, and set

$$\log(\delta_m^{(t+1)}) = \log(\delta_m^{(t)}) + \kappa^{(t)}(\hat{r}^{(t)} - 0.234).$$

The attenuation parameters $k_0 > 0$ and $k_1 \in (0,1]$ are the only quantities in the previous algorithms that are not entirely automatic. In practice, Shaby and Wells (2010) verified that their choices are not very influential and indicated to choose $k_0 = 1$ and $k_1 = 0.8$.

The idea behind the previous adaptive approach is quite intuitive. It takes a block of $m < T$ steps and estimates the acceptance rate for that block. When the acceptance rate exceeds

the optimal value of 0.234 (Roberts, Gelman, and Gilks 1997), it increases δ_m. On the other hand, if the acceptance rate is too low, it decreases δ_m. Moreover, adapting the logarithm of δ_m rather than δ_m itself is advantageous in two ways. Firstly, it guarantees that δ_m remains positive, and secondly, it permits multiplicative adjustments rather than additive.

2.5 Simulation study

In this section, we conduct a simulation study to assess the performance of our Bayesian weighted quantile regression models (BWQR) compared with an unweighted benchmark under different scenarios. We assume the well-known Bayesian quantile regression based on the asymmetric Laplace distribution (Kozumi and Kobayashi 2011; Yu and Moyeed 2001, BQR-AL) as the benchmark and denote the approach based on the asymmetric Laplace distribution as BWQR-AL and the approach based on the score likelihood as BWQR-SL.

We propose a model-based simulation study to analyze the ability of our models to characterize the entire population. Our models and the benchmark are fitted considering a sample from a finite population, and a comparison criterion based on the check loss function is calculated using all observations in the finite population. For that, we generate 200 finite populations with population size $N = 10,000$ from the following population models:

M1: $y_i = \beta_0(\tau) + \beta_1(\tau)x_{1i} + \epsilon_i$, where the error term ϵ_i is a standard normal variable independent of x_{1i};

M2: $y_i = \beta_0(\tau) + \beta_1(\tau)x_{1i} + \epsilon_i$, where $\epsilon_i \sim t_3$ - Student's t-distribution with three degrees of freedom - is independent of x_{1i};

M3: $y_i = \beta_0(\tau) + \beta_1(\tau)x_{1i} + (1 + 0.2x_{1i})\epsilon_i$, where ϵ_i is a standard normal variable independent of x_{1i};

M4: $y_i = \beta_0(\tau) + \beta_1(\tau)x_{1i} + \epsilon_i$, where $\epsilon_i \sim SN(-0.7740617, 1, 4)$ - Skew normal distribution (Azzalini 1985) with location, scale and skewness parameters equal to -0.7740617, 1 and 4 - is independent of x_{1i}. This particular choice for the location parameter implies a zero mean error.

For all models $\beta_0(\tau) = 2.0$, $\beta_1(\tau) = 1.5$, and x_{1i} is generated from a uniform variable on the interval $(0, 2)$. Models M1, M2, and M4 represent homoscedastic error models, but M2 and M4 represent heavy-tailed and asymmetric examples, respectively. Model M3 represents a heteroscedastic error model.

To obtain the final samples, we replicate a similar procedure to that of Chen and Zhao (2019), meaning that for each generated finite population of size N, a Poisson sample with an expected size equal to $n = 500$ was selected with inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$ where k_i is treated as two different cases. On both cases, we have $k_i = \{1 + \exp(z_i)\}^{-1}$, but we have $z_i \sim N(y_i, 0.25)$ for Case 1, and $z_i \sim N(3.5, 0.25)$ for Case 2. Thus, for all data-generating models, the first sampling procedure is informative since the inclusion probabilities depend on the response variable, while for the second, the same does not occur.

For each replica in our simulation, MCMC chains of length $25,000$ are drawn from the full conditional distributions described in Sections 2.3.3 and 2.4.2. A burn-in of $5,000$ and a thin of 20 are taken. Following Yang, Wang, and He (2016) and Benoit and den Poel (2017), $\sigma = 1$ is fixed for the BWQR-AL and the BQR-AL. Regarding the prior specification, we chose a non-informative prior for the regression parameters in which $\boldsymbol{\mu}_0 = \underline{0}_q$ and $\Sigma_0 = 1,000 I_q$, where $\underline{0}_q$ is a q-dimensional vector of zeros and I_q is an identity matrix with dimension $q \times q$.

Measuring and comparing the proposed models with the BQR-AL in terms of quantities based on a known fixed true value for the coefficients might not be appropriate in some of our scenarios since the linear conditional quantile function $Q_{Y_i}(\tau | X_{1i} = x_{1i}) = \beta_0(\tau) + \beta_1(\tau) x_{1i}$ is not comparable for all τ. Then, to evaluate the ability of our proposed methods to correctly describe the finite population in a wider range of scenarios, we calculate

$$\frac{1}{N} \sum_{i=1}^{N} \rho_\tau \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}(\tau) \right) \tag{12}$$

for all the $1,000$ samples generated in the inference procedures from the BWQR-AL, BWQR-SL, and BQR-AL, and take their medians. Lee, Noh, and Park (2014) and Barata, Prado, and Sansó (2022), for example, also applied criteria based on the check loss function to model selection and comparison, but not in a complex survey context. As with any error function,

the lower the quantity in 12, the better.

In Table 2.1, we summarize the results of our simulation study regarding both Case 1 and Case 2, all data-generating models from M1 to M4, and first, second, and third quartiles ($\tau = 0.25$, $\tau = 0.50$ and $\tau = 0.75$, respectively). Table 2.1 reports the median and the variances (in parenthesis) from the quantities generated using (12). Focusing firstly on Case 1, we observe that BWQR-AL and BWQR-SL have similar performance, with a slight advantage to the former. By our criteria, the BWQR-AL and the BWQR-SL present an evident gain if compared with our benchmark - the BQR-AL - in all scenarios, indicating that our methods outperform a naive model fitting, ignoring the informative sampling design.

Table 2.1: Median and variance ($\times 10^3$) in parentheses of the criteria considering models M1-M4, Cases 1 and 2, and the quartiles with expected sample sizes equal to 500.

τ	Method	Case 1			
		M1	M2	M3	M4
0.25	BWQR-AL	0.3206 (0.0155)	0.4672 (0.0647)	0.3849 (0.0187)	0.1791 (0.0042)
	BWQR-SL	0.3226 (0.0120)	0.4719 (0.0720)	0.3883 (0.0194)	0.1788 (0.0044)
	BQR-AL	0.4135 (0.1763)	0.8949 (2.3766)	0.5221 (0.2906)	0.1898 (0.0143)
0.50	BWQR-AL	0.4031 (0.0332)	0.5601 (0.1304)	0.4852 (0.0927)	0.2517 (0.0144)
	BWQR-SL	0.4052 (0.0299)	0.5621 (0.1272)	0.4879 (0.0876)	0.2517 (0.0160)
	BQR-AL	0.5493 (0.5047)	1.0168 (3.5542)	0.7145 (0.7194)	0.2763 (0.0397)
0.75	BWQR-AL	0.3242 (0.0930)	0.4775 (0.3676)	0.3928 (0.2584)	0.2225 (0.0484)
	BWQR-SL	0.3261 (0.0774)	0.4782 (0.3425)	0.3955 (0.2341)	0.2228 (0.0361)
	BQR-AL	0.4714 (0.7170)	0.8046 (2.6716)	0.6349 (1.1284)	0.2558 (0.1022)
		Case 2			
		M1	M2	M3	M4
0.25	BWQR-AL	0.3200 (0.0077)	0.4630 (0.0475)	0.3837 (0.0107)	0.1787 (0.0017)
	BWQR-SL	0.3197 (0.0078)	0.4629 (0.0473)	0.3834 (0.0107)	0.1777 (0.0018)
	BQR-AL	0.3200 (0.0077)	0.4629 (0.0478)	0.3838 (0.0102)	0.1786 (0.0017)
0.50	BWQR-AL	0.4013 (0.0104)	0.5528 (0.0499)	0.4811 (0.0134)	0.2505 (0.0036)
	BWQR-SL	0.4012 (0.0102)	0.5525 (0.0501)	0.4810 (0.0139)	0.2499 (0.0038)
	BQR-AL	0.4015 (0.0102)	0.5529 (0.0499)	0.4808 (0.0131)	0.2506 (0.0037)
0.75	BWQR-AL	0.3200 (0.0084)	0.4629 (0.0543)	0.3836 (0.0117)	0.2203 (0.0039)
	BWQR-SL	0.3196 (0.0084)	0.4629 (0.0554)	0.3835 (0.0116)	0.2196 (0.0041)
	BQR-AL	0.3200 (0.0082)	0.4628 (0.0542)	0.3836 (0.0113)	0.2201 (0.0039)

An issue of major concern pointed out by Pfeffermann, Moura, and Silva (2006) is understanding how including sample weights when it is not necessarily required impacts models' performance. Case 2, therefore, is designed to enable this analysis. We note that all models, BWQR-AL, BWQR-SL, and BQR-AL, have close results. From this analysis, we see that our models present a better fit under informative sampling and do not have a negative impact

when including sample weights is not required.

We also note that the medians of BWQR-AL and BWQR-SL are much less dispersed over the replicas when compared with the BQR-AL when Case 1 is considered. When Case 2 is analyzed, the results are closed instead.

Complementing the previous results, we proceed with an analysis focused on the coefficients. On both M1 and M2, the linear conditional quantile function takes the form $Q_{Y_i}(\tau|X_{1i} = x_{1i}) = \beta_0(\tau) + \beta_1(\tau)x_{1i}$ for any $\tau \in (0, 1)$, where $\beta_1(\tau) = 1.5$, and $\beta_0(\tau) = 2 + \Phi^{-1}(\tau)$ and $\beta_0(\tau) = 2 + F_{t_3}^{-1}(\tau)$ for M1 and M2, respectively. Regarding M3, the linear conditional quantile function $Q_{Y_i}(\tau|X_{1i} = x_{1i}) = \beta_0(\tau) + \beta_1(\tau)x_{1i}$ holds only at $\tau = 0.5$ with $\beta_0(\tau) = 2.0$ and $\beta_1(\tau) = 1.5$. Thus, for these cases, we can evaluate the capacity of our methods to recover the super-population coefficients.

Table 2.2 reports the absolute errors between the mean of the point estimates for the coefficients and the fixed true values. The variances of the point estimates are reported in parentheses. As point estimates for the coefficients, we used the mean of the MCMC draws. Concerning Case 1, both BWQR-AL and BWQR-SL perform substantially better than BQR-AL regarding the absolute error for all contemplated scenarios. The results indicate that our methods can directly perform posterior inference about the population using only quantities available for the observed sample when we have an informative sampling scenario. Regarding Case 2, we observe that our methods do not have a significant negative impact on the absolute error, strengthening the idea that our methods are competitive even when including sample weights is not required. In terms of variance, there are no discrepant differences among the models.

Table 2.2: Absolute error and variance in parentheses considering models M1-M3 at $\tau = 0.50$, models M1-M2 at $\tau = 0.25$ and $\tau = 0.75$, Cases 1 and 2, with expected sample sizes equal to 500.

τ	Model	Parameter	Case 1		
			M1	M2	M3
0.25	BWQR-AL	$\beta_0(\tau)$	0.0101 (0.0170)	0.0087 (0.0229)	-
		$\beta_1(\tau)$	0.0275 (0.0264)	0.0903 (0.0395)	-
	BWQR-SL	$\beta_0(\tau)$	0.0239 (0.0136)	0.0259 (0.0183)	-
		$\beta_1(\tau)$	0.0802 (0.0190)	0.2236 (0.0280)	-
	BQR-AL	$\beta_0(\tau)$	0.7152 (0.0075)	1.3441 (0.0696)	-
		$\beta_1(\tau)$	0.1569 (0.0131)	1.2615 (0.1406)	-
0.50	BWQR-AL	$\beta_0(\tau)$	0.0267 (0.0279)	0.0130 (0.0309)	0.0593 (0.0383)
		$\beta_1(\tau)$	0.0484 (0.0404)	0.1207 (0.0451)	0.1025 (0.0628)
	BWQR-SL	$\beta_0(\tau)$	0.0257 (0.0276)	0.0222 (0.0314)	0.0661 (0.0355)
		$\beta_1(\tau)$	0.0474 (0.0399)	0.1356 (0.0438)	0.1133 (0.0555)
	BQR-AL	$\beta_0(\tau)$	0.7649 (0.0057)	1.0522 (0.0249)	1.0158 (0.0082)
		$\beta_1(\tau)$	0.1296 (0.0111)	0.7290 (0.0596)	0.2198 (0.0127)
0.75	BWQR-AL	$\beta_0(\tau)$	0.0453 (0.0603)	0.0550 (0.0944)	-
		$\beta_1(\tau)$	0.0844 (0.0843)	0.2382 (0.0981)	-
	BWQR-SL	$\beta_0(\tau)$	0.0450 (0.0593)	0.0555 (0.1049)	-
		$\beta_1(\tau)$	0.0528 (0.0818)	0.1680 (0.1129)	-
	BQR-AL	$\beta_0(\tau)$	0.8025 (0.0065)	1.0531 (0.0131)	-
		$\beta_1(\tau)$	0.1099 (0.0122)	0.3854 (0.0267)	-
			Case 2		
			M1	M2	M3
0.25	BWQR-AL	$\beta_0(\tau)$	0.0142 (0.0143)	0.0079 (0.0217)	-
		$\beta_1(\tau)$	0.0151 (0.0108)	0.0023 (0.0155)	-
	BWQR-SL	$\beta_0(\tau)$	0.0093 (0.0142)	0.0163 (0.0209)	-
		$\beta_1(\tau)$	0.0149 (0.0107)	0.0036 (0.0150)	-
	BQR-AL	$\beta_0(\tau)$	0.0115 (0.0134)	0.0024 (0.0227)	-
		$\beta_1(\tau)$	0.0124 (0.0101)	0.0048 (0.0156)	-
0.50	BWQR-AL	$\beta_0(\tau)$	0.0113 (0.0127)	0.0035 (0.0126)	0.0013 (0.0186)
		$\beta_1(\tau)$	0.0069 (0.0102)	0.0014 (0.0104)	0.0013 (0.0146)
	BWQR-SL	$\beta_0(\tau)$	0.0132 (0.0121)	0.0025 (0.0127)	0.0020 (0.0179)
		$\beta_1(\tau)$	0.0089 (0.0099)	0.0007 (0.0106)	0.0001 (0.0141)
	BQR-AL	$\beta_0(\tau)$	0.0084 (0.0113)	0.0001 (0.0119)	0.0001 (0.0169)
		$\beta_1(\tau)$	0.0048 (0.0092)	0.0009 (0.0101)	0.0041 (0.0138)
0.75	BWQR-AL	$\beta_0(\tau)$	0.0024 (0.0139)	0.0004 (0.0220)	-
		$\beta_1(\tau)$	0.0014 (0.0100)	0.0037 (0.0176)	-
	BWQR-SL	$\beta_0(\tau)$	0.0055 (0.0136)	0.0098 (0.0214)	-
		$\beta_1(\tau)$	0.0027 (0.0098)	0.0027 (0.0170)	-
	BQR-AL	$\beta_0(\tau)$	0.0004 (0.0132)	0.0023 (0.0211)	-
		$\beta_1(\tau)$	0.0023 (0.0097)	0.0024 (0.0168)	-

2.6 Real-data-based simulation study

In this section, we propose a design-based simulation study that follows a similar path to Savitsky and Toth (2016) and Chen and Zhao (2019). For that, we use a database from Prova Brasil as Silva and Moura (2022), but here, we work with the edition of 2011 instead of 2009.

Prova Brasil is a large-scale proficiency test developed by the Brazilian National Institute of Education Research (INEP) for the Brazilian Ministry of Education. From standardized tests, Prova Brasil aims to evaluate the educational quality of the Brazilian public educational system. The tests are centered on two fields (Portuguese and Mathematics) and applied to students in the fifth and ninth years of public elementary schools. The Portuguese test is focused on reading, and the Mathematics test is focused on problem solutions. The proficiency scores stem from applying item response theory (IRT) models to the test results.

Therefore, Prova Brasil is an important tool for federal, state, and municipal educational agencies in designing, developing, implementing, and evaluating public policies oriented toward enhancing student performance and reducing educational inequalities. Educational agencies can identify frailties and correct distortions from the collected information, driving technical and financial resources to focal areas. In this regard, quantile regression models encompass a powerful tool to investigate distortions and inequalities in educational outcomes. Since several educational assessment data originate from complex surveys that might be informative, the methods proposed in this chapter can be very useful in the educational area. For example, Costanzo and Desimoni (2015) applied a quantile regression approach to address issues of inequality in education outcomes using INVALSI survey data, and Giambona and Porcu (2015) studied background determinants of reading achievement in Italy using the 2009 OECD-PISA survey.

In our analysis, we restrict the application to students in the ninth year from public elementary schools located in Campinas municipality in São Paulo State. A total of $11,004$ students in the ninth year, distributed among 128 schools, attended the exam with complete proficiency and

predictor variables. Although Prova Brasil is not a sample survey, it is subject to non-response due to school evasion or non-attendance on the day of the exam, and there is evidence from previous editions suggesting that students with low achievement are less likely to participate in the exam. Thus, it indicates an informative mechanism in the observed data collected from the application of Prova Brasil.

Based on the previous consideration, we propose a design-based simulation study where the $N = 11,004$ observations from students in the ninth year from public elementary schools located in Campinas are fixed as our finite population, and samples of expected size equals $n = 500$ are taken. For the study, we consider an indicator variable for nonwhite students (black, mixed, and indigenous) and an indicator variable for students lagging behind. The former is applied as a proxy for socioeconomic inequities as there is a discrepancy in poverty levels between whites and blacks in Brazil (Gradín 2009). The latter indicates students who have fallen behind other students in their cohorts. As a response variable, we consider the Mathematics proficiency scores.

Moreover, we generate 200 samples using a Poisson sample with inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$ where k_i is $k_i = 1 + \exp(z_i)$ and $z_i \sim N(y_i, 0.25)$. We have an informative single-stage sample design with unequal inclusion probabilities based on a proportional-to-size sample similar to the procedure described in Section 2.5. However, the sampling inclusion probability is directly proportional to the response variable as we intend to reproduce a scenario in which students with high achievement are more likely to participate.

To assess the performance of our BWQR models, we assume the point estimates generated from the BQR-AL considering the fully observed population as our best-case scenario. In addition, we assume the estimates from the BQR-AL considering the sampled data as the benchmark. Thus, the estimates from the BWQR-AL and the BWQR-SL should, ideally, get close to the results from the best-case scenario and present, at the same time, a better performance in comparison with the benchmark.

Similarly to the previous section, for each replica in our simulation, MCMC chains of length $25,000$ are drawn, a burn-in of $5,000$ and a thin of 20 are taken, and $\sigma = 1$ is fixed. Regarding

the prior specification, we chose a non-informative prior for the regression parameters in which $\boldsymbol{\mu}_0 = \underline{0}_3$ and $\Sigma_0 = 1,000 I_3$.

Figure 2.1 compares the point estimates generated from the 200 replicas for each of the three methods under analysis (BWQR-AL, BWQR-SL, and BQR-AL) with the point estimate generated from the best-case scenario. We analyze the estimates for three coefficients: intercept (first column of plot panels), an indicator variable for nonwhite students (second column of plot panels), and an indicator variable for students lagging behind (third column of plot panels). We also analyze three different quantiles, $\tau = 0.25$ (first row of plot panels), $\tau = 0.50$ (second row of plot panels), and $\tau = 0.75$ (third row of plot panels).

From Figure 2.1, we note that our methods outperform the benchmark, suggesting that the BWQR-AL and the BWQR-SL are effective in performing posterior inference about the population using only quantities available for the observed sample when we have an informative sampling scenario. The results in Figure 2.1 also seem to be in line with those presented in Table 2.2. Generally, our methods' estimates get closer to the best-case scenario estimates than the benchmark, supporting the smaller mean absolute errors in Table 2.2. We also observe, for example, a significant distortion in the estimates of the intercept when the naive model fitting ignoring the informative sampling design is considered.

Analyzing the coefficients, we notice that the indicator variables for nonwhite students and students lagging behind are negatively related to the proficiency score for all quantile levels. This negative association is well reported in the literature, especially if we look at the race variable (Botelho, Madeira, and Rangel 2015; Marteleto 2012), and the high correlation of both variables with socio-economic indicators explains it. The results reflect more the structural problems in Brazilian society than individual ability.

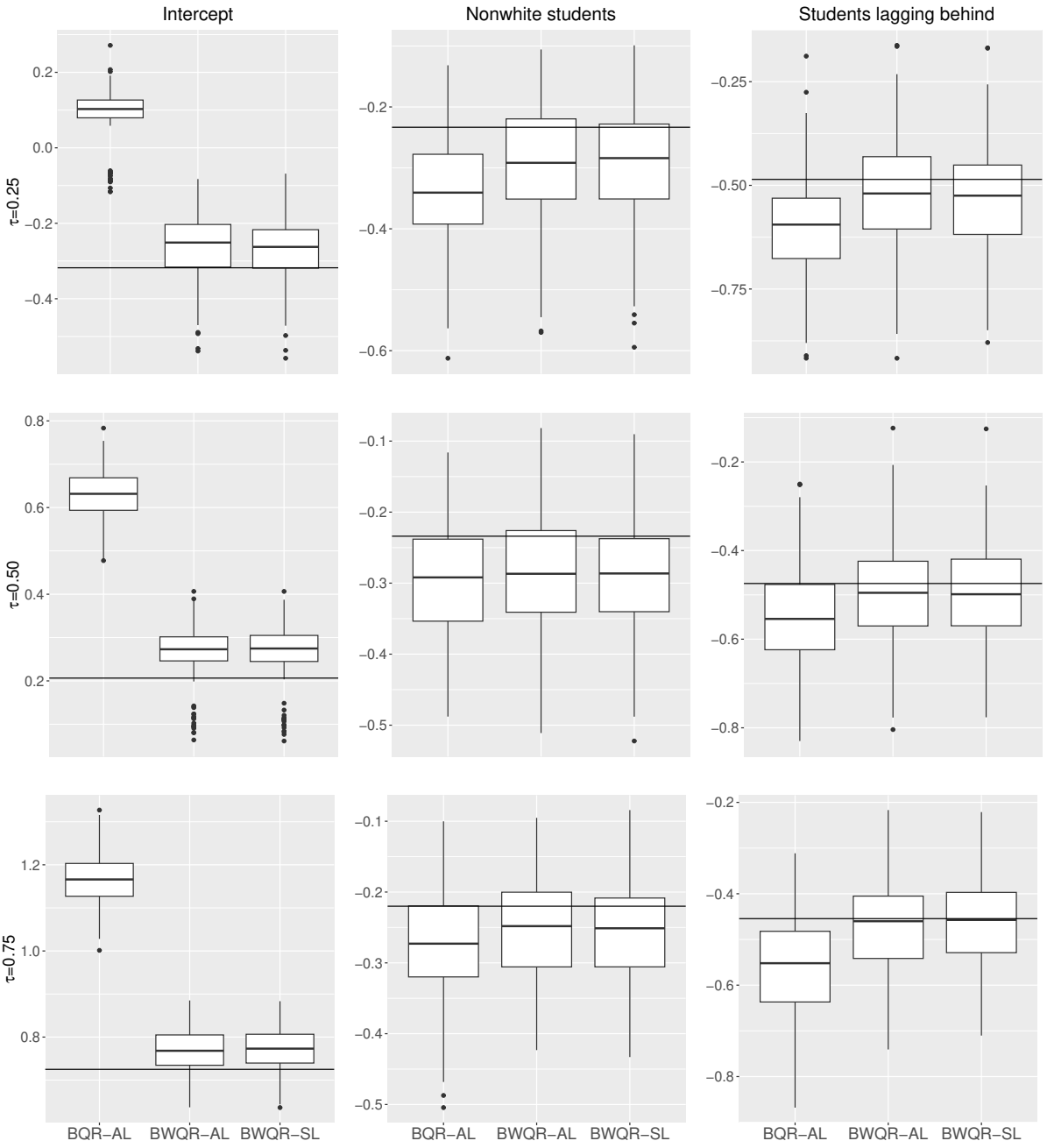

Figure 2.1: Comparison between our methods and the benchmark. The boxplots summarize the point estimates, and the solid line represents the best-case scenario.

Section 2.5 and the previous results mainly focus on analyzing and evaluating our methods' performance in point estimation. However, uncertainty is also a relevant issue. Figure 2.2 compares the lengths for the 95% Highest Posterior Density (HPD) intervals generated from the 200 replicas for each of the three methods under analysis (BWQR-AL, BWQR-SL, and BQR-AL). Similarly to the previous figure, we consider the interval estimates for all three

coefficients: intercept (first column of plot panels), an indicator variable for nonwhite students (second column of plot panels), and an indicator variable for students lagging behind (third column of plot panels). We also consider three different quantiles, $\tau = 0.25$ (first row of plot panels), $\tau = 0.50$ (second row of plot panels), and $\tau = 0.75$ (third row of plot panels).

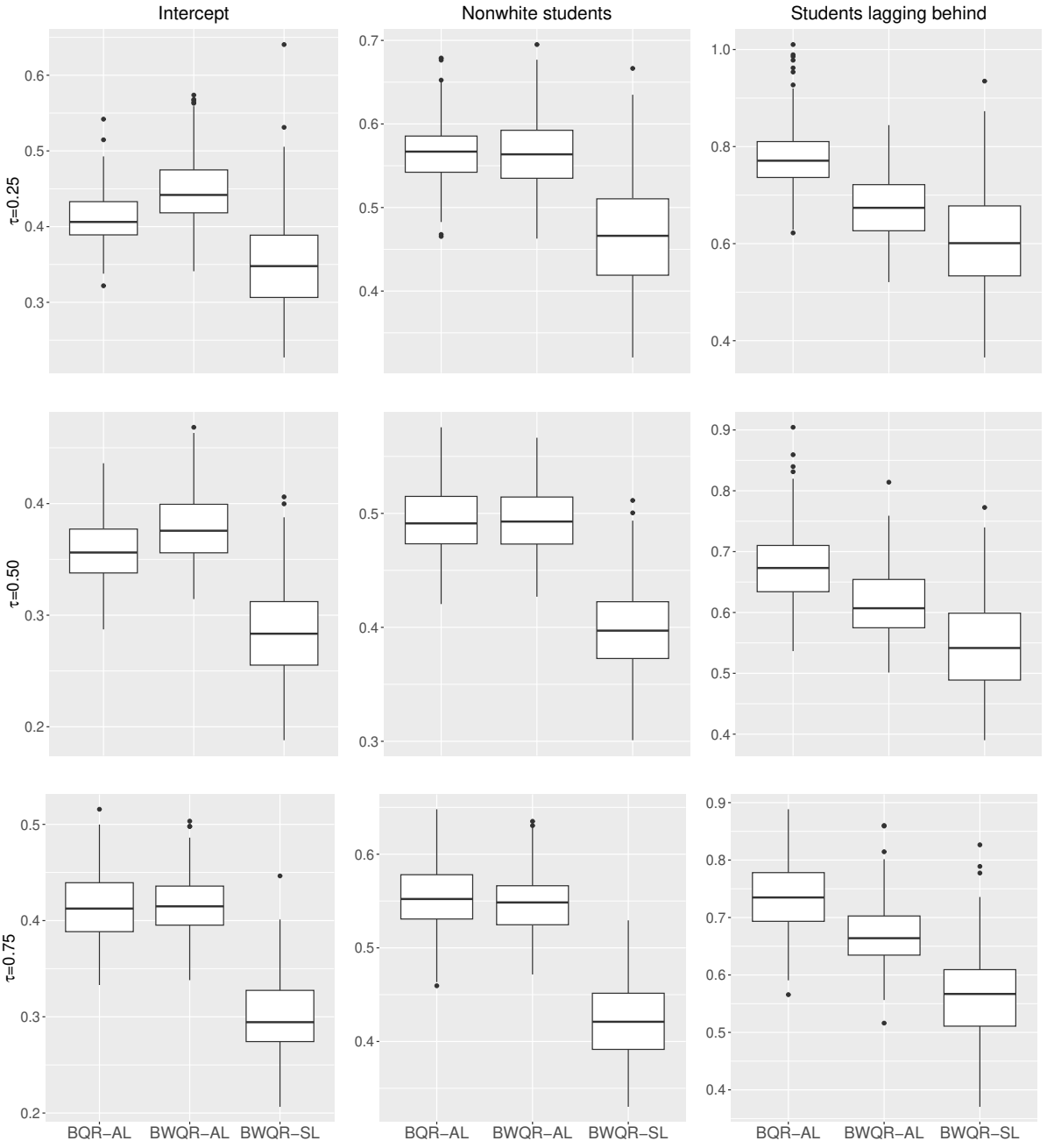

Figure 2.2: Comparison between our methods and the benchmark. The boxplots summarize the lengths of the interval estimates.

From Figure 2.2, the BWQR-SL seems to present a lower degree of uncertainty than the

BWQR-AL and BQR-AL for all coefficients and all quantiles under analysis. The BWQR-AL, in turn, seems to have a similar or lower degree of uncertainty than the BQR-AL, except for the intercept when the first and the second quartiles are considered.

Beyond the length, we also look at the percentage that the point estimate generated from the best-case scenario falls within the 95% HPD intervals generated from the 200 replicas for each of the three methods. This measure gives us a sense of coverage. Table 2.3 shows the results. We note that the percentages for the BWQR-SL tend to be smaller than the BWQR-AL and the BQR-AL and closer to the nominal level of 95% mainly when we consider $\tau = 0.25$ and $\tau = 0.50$. Jointly with the results presented in Figure 2.2, it indicates that the BWQR-SL better describes the uncertainties and that the BWQR-AL and the BQR-AL have excessively large intervals. We note that the percentages for the BQR-AL regarding the intercept are very close or equal to zero, indicating that not only the point estimates but also the interval estimates are precarious for this parameter.

Table 2.3: Percentage that the point estimate generated from the best-case scenario falls within the 95% HPD intervals generated from the 200 replicas for all the three methods.

τ	Method	Intercept	Nonwhite students	Students lagging behind
0.25	BWQR-AL	97.00	97.50	98.00
	BWQR-SL	92.50	95.50	95.50
	BQR-AL	1.50	99.00	98.50
0.50	BWQR-AL	99.50	98.50	99.50
	BWQR-SL	96.00	96.50	98.00
	BQR-AL	0.00	99.50	99.00
0.75	BWQR-AL	100.00	100.00	100.00
	BWQR-SL	98.50	99.50	99.50
	BQR-AL	0.00	99.00	99.50

2.7 Final remarks

In this chapter, we introduced two new and easy-to-implement methods for complex survey data under informative sampling. We obtained an approach based on the asymmetric Laplace likelihood from a design-weighted estimator (Chen and Zhao 2019; Geraci 2016) and following the argument of Yu and Moyeed (2001). From the ideas of Kozumi and Kobayashi (2011), we implemented a simple Markov chain Monte Carlo algorithm for fitting our quantile regression

model. Our second method extends the ideas of Wu and Narisetty (2021) to include sampling weights in their model, applying a result of Huang, Xu, and Tashnev (2015).

To evaluate the performance of our methods, we carried out two simulation studies and compared them with a naive model that ignores the informative sampling design. In the first simulation study, we considered a model-based simulation study in which four data-generating models for the population were specified, encompassing different possible characteristics for a data set. In the second simulation study, we considered a design-based study in which students in the ninth year from public elementary schools located in Campinas municipality in São Paulo State are fixed as the finite population, and the proficiency scores in Prova Brasil were taken as the response variable. Our methods outperformed the naive model fitting, ignoring the informative sampling design in all scenarios under analysis.

3 Quantiles for bounded count data under informative sampling

3.1 Introduction

Modeling count data under informative sampling is a substantive matter in survey statistics. Savitsky and Toth (2016) constructed a pseudo-posterior distribution utilizing sampling weights based on the marginal inclusion probabilities to exponentiate the likelihood contribution of each sampled unit to recover population information from the available sample. The authors illustrated their method on an application concerning the Bureau of Labor Statistics Job Openings and Labor Turnover Survey, specifying a count data model for the population based on the Poisson distribution. Parker, Holan, and Janicki (2020) combined the pseudo-likelihood approach and conjugate multivariate distribution theory to attain a unit-level model for count data that accounts for informative sampling designs. Their method was illustrated via an empirical simulation study using count data from the American Community Survey public use microdata sample.

At the same time, the interest in considering the relation among random variables in quantiles instead of the mean is also a relevant research topic that has emerged in various fields. Quantile regression provides a more comprehensive description of the relationship between a response variable and covariates, and it is a robust alternative to heteroscedasticity and outliers. However, direct estimation of quantile regression models for count data is not possible due to the association of the discreteness of the data and the nondifferentiability

of the sample objective function defining the estimator. To overcome this issue, Machado and Silva (2005) observed that quantiles of the randomly perturbed data have a one-to-one relation with the quantiles of the original data and proposed applying quantile regression to jittered count data as a mechanism to make inference about relevant aspects of the conditional quantiles of the counts.

From Machado and Silva (2005), Lee and Neocleous (2010) proposed a Bayesian quantile regression model for count data based on the asymmetric Laplace distribution and illustrated it with an application related to environmental epidemiology. The method from Machado and Silva (2005) has also been employed for further applied analysis, including fertility (Booth and Kee 2009; Miranda 2008), car accidents (Qin and Reyes 2011), frequency of doctor visits Moreira and Barros (2010); Winkelmann (2006), and students' performance (Grilli, Rampichini, and Varriale 2016).

Alternative approaches have also been developed from the solution introduced by Machado and Silva (2005) to artificially smooth the data by applying jittering. For example, Congdon (2017) proposed a Bayesian framework that combines the asymmetric Laplace distribution with the Poisson model. Padellini and Rue (2019) interpolated the quantile function using a continuous model-aware, allowing for proper quantile inference while retaining model interpretation. Tzavidis et al. (2015), in turn, proposed a semiparametric M-quantile method for counts. More recently, Frumento and Salvati (2021) applied the approach developed by Frumento and Bottai (2016) to describe the quantile regression coefficients.

Despite the methodologies previously mentioned, there is a lack of tools for the appropriate analysis of bounded count outcomes, even considering that it is often observed in many applications. Mullahy (2023), for example, underlined some health-related outcomes like the number of days per week on which adults reported engaging in vigorous or moderate exercise from the 2015 Health Survey of England (HSE) and the number of physically healthy days in the previous 30 days from the 2021 Behavioral Risk Factors Surveillance System (BRFSS). Bottai, Cai, and McKeown (2010) provided a simple approach using the logistic quantile regression to analyze bounded outcomes, but for the continuous case. To the best of our

knowledge, no method in the literature accommodates informative sampling and quantile regression analysis of count data, more specifically, bounded count data.

Therefore, this chapter aims to fulfill this lack by introducing effective and easy-to-implement frameworks and providing valuable tools for practitioners from different backgrounds. We propose two different Bayesian quantile regression models that accommodate bounded count data under informative sampling. From the side of the bounded count data, we apply the ideas of Machado and Silva (2005) and Lee and Neocleous (2010) to deal with quantiles for counts, adapting them by employing the methodology proposed by Bottai, Cai, and McKeown (2010). From the side of complex survey data under informative sampling, we propose two different methods based on the asymmetric Laplace distribution (Yu and Moyeed 2001); one is an extension of the approach introduced in Chapter 2, and the other is based on the pseudo posterior distribution (Savitsky and Toth 2016).

Since complex surveys comprise one of the main instruments for collecting information in the educational area and quantile regression models can characterize differences over the distribution of a data set, our framework can be particularly appealing, for example, in investigating distortions and inequalities in educational outcomes. In this context, Costanzo and Desimoni (2015) applied a quantile regression approach to address issues of inequality in education outcomes using INVALSI survey data, and Giambona and Porcu (2015) studied background determinants of reading achievement in Italy using the 2009 OECD-PISA survey.

With that in mind, we draw a design-based simulation study using data from Prova Brasil 2011. Prova Brasil is fundamental for Brazilian educational agencies in designing, developing, implementing, and evaluating public policies directed toward the learning process and reducing educational inequalities. The outcome is the number of correct answers among 26 questions in the Mathematics exam, and to assess the performance of our approaches, we compare them with a naive model fitting that ignores the informative sampling design under four different experiments. The first experiment analyzes different sampling designs, while the second has different degrees of informativeness. The third experiment covers different sample sizes. Lastly, we propose a prior sensitivity analysis that compares a non-informative prior with

two n-dependent priors (Narisetty and He 2014; Yang and He 2012).

3.2 Quantile regression for bounded count data

Suppose that Y is a variable of interest, and $\mathbf{X}$ is a p-dimensional vector of covariates. Let $Q_Y(\tau|\mathbf{x})$ stand for the 100τth of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$, meaning that $Q_Y(\tau|\mathbf{x}) \equiv F_Y^{-1}(\tau;\mathbf{x}) = \min\{\alpha|\mathbb{P}(Y \leq \alpha|\mathbf{x}) \geq \tau\}$, where F is the cumulative distribution function. Considering a linear quantile regression in which $Q_Y(\tau|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\beta}(\tau)$ is a p-dimensional vector of coefficients, Koenker and Bassett Jr. (1978) established sufficient conditions for asymptotically valid inference on the parameters of the model. One of these conditions states that the conditional probability density function $f(y|\mathbf{x})$ must be continuous and positive at $Q_Y(\tau|\mathbf{x})$.

In this chapter, we are interested in studying cases in which Y is a count random variable bounded from below and from above by two known constants a and b, where a and b are defined in the set of nonnegative integers ($\mathbb{N}_0$). As Y results from a bounded count, the sufficient conditions established by Koenker and Bassett Jr. (1978) are not satisfied. Regarding the case where Y is a count random variable with support in $\mathbb{N}_0$, Machado and Silva (2005) addressed this issue by adding a standard uniform random variable U, independent of Y and $\mathbf{X}$, to Y and working with $Z = Y + U$. This jittering procedure is a particular form of inserting smoothness (Pearson 1950; Stevens 1950), and implies a conditional quantile function that is continuous in τ. For some known monotone transformation $h_\tau(\cdot)$ which possibly depends on τ, Machado and Silva (2005), therefore, assumed that the τ-th conditional quantile function of Z given $\mathbf{X}$ takes the parametric form

$$Q_{h_\tau(Z)}(\tau|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}(\tau), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^p$ includes the intercept.

To ensure that $\mathbb{P}(h_\tau^{-1}(\mathbf{x}'\boldsymbol{\beta}(\tau)) \in \mathbb{N}_0) = 0$, and, consequently, for almost every realization of $\mathbf{X}$, the conditional density of the outcome at the quantile of interest will be continuous,

Machado and Silva (2005) also made some additional assumptions. The authors assumed that $\mathbf{x}$ can be partitioned as $(\mathbf{x}^{(D)}, \mathbf{x}^{(C)})$, where $\mathbf{x}^{(C)}$ corresponds to the continuous covariates and $\mathbf{x}^{(D)}$ to the discrete ones. The intercept is included on $\mathbf{x}^{(D)}$ and $\mathbf{x}^{(C)} \in \mathbb{R}^{p_C}$, $1 \leq p_C \leq p-1$, satisfying $\mathbb{P}(\mathbf{X}^{(C)} \in \mathcal{C}) = 0$ for any countable subset $\mathcal{C} \in \mathbb{R}^{p_C}$. Moreover, if $\boldsymbol{\beta}^{(C)}(\tau)$ denotes the components of $\boldsymbol{\beta}(\tau)$ corresponding to the continuous covariates $\mathbf{x}^{(C)}$, hence $\boldsymbol{\beta}^{(C)}(\tau) \neq 0$.

For a simple random sample $\{(y_i, x_i, u_i)\}_{i=1}^n$ of $(Y, \mathbf{X}, U)$, Machado and Silva (2005), therefore, defined an estimator for $\boldsymbol{\beta}(\tau)$ as any solution to the optimization problem

$$\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in S} \rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i' \boldsymbol{\beta}(\tau) \right),$$

where $\rho_\tau(\ell) = \ell(\tau - \mathbb{1}(\ell < 0))$ is the asymmetric check loss function (Koenker and Bassett Jr. 1978). And attentive that the object of ultimate interest does not lie at the estimator for $\boldsymbol{\beta}(\tau)$, but on the quantile function of the discrete data Y, Machado and Silva (2005) proved that under certain conditions $Q_Y(\tau|\mathbf{x})$ is consistently estimated by $\lceil h_\tau^{-1}(\mathbf{x}'\hat{\boldsymbol{\beta}}(\tau)) - 1 \rceil$, where $\lceil a \rceil$ denotes the ceiling function that returns the smallest integer greater than, or equal to, a, and $\hat{\boldsymbol{\beta}}(\tau)$ is the estimator of $\boldsymbol{\beta}(\tau)$.

On their paper, Machado and Silva (2005) specified $h_\tau^{-1}(\mathbf{x}'\boldsymbol{\beta}(\tau)) = \tau + \exp\{\mathbf{x}'\boldsymbol{\beta}(\tau)\}$ which means that

$$Q_Z(\tau|\mathbf{x}) = \tau + \exp\{\mathbf{x}'\boldsymbol{\beta}(\tau)\}$$

as quantile functions are equivariant to monotone transformations. The coefficients, hence, can be estimated by running a linear quantile regression to model the transformed response $Y^* = h_\tau(Z) = \log(Y + U - \tau)$.

Following the previous ideas, we apply the jittering procedure to the case we are interested in, obtaining $Z \in \{(\tilde{a}, \tilde{b})$, where $\tilde{a} = a$ and $\tilde{b} = b + 1$. As our outcome is defined in bounded support, we follow Bottai, Cai, and McKeown (2010) and specify

$$h_\tau^{-1}(\mathbf{x}'\boldsymbol{\beta}(\tau)) = \tau + \frac{\exp\{\mathbf{x}'\boldsymbol{\beta}(\tau)\}\tilde{b} + \tilde{a}}{\exp\{\mathbf{x}'\boldsymbol{\beta}(\tau)\} + 1}.$$

By doing this, we can estimate the coefficients by fitting a linear quantile regression to model the transformed response $Y^* = h_\tau(Z) = \log((Y + U - \tau - \tilde{a})/[\tilde{b} - (Y + U - \tau)])$.

3.3 Bayesian quantile regression for bounded count data under informative sampling

Suppose a population $\mathcal{F}_N = \{(y_i, \mathbf{x}_i), i = 1, \dots, N\}$ generated from a super-population model $\mathcal{F}$ where the underlying distribution function is unknown, and let $S = \{i_1, \dots, i_n\}$ be a sample with size $|S| = n$. When the sample S is a complex survey, and its design is informative, it is necessary to consider the design in the estimation process if we are interested in inferential problems about population quantities. A sampling design is informative when $\mathbb{P}(I_i = 1|x_i, y_i) \neq \mathbb{P}(I_i = 1|x_i)$, where $I_i \in \{0, 1\}$ is the sampling indicator such that $I_i = 1$ if unit i is selected and $I_i = 0$ otherwise, and $\pi_i = \mathbb{P}(I_i = 1)$ is the inclusion probability. With this in mind, we can define a survey-weighted estimator (Chen and Zhao 2019; Geraci 2016) for true values of the unknown population parameters $\boldsymbol{\beta}_0(\tau)$ as the solution to the optimization problem

$$\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in S} \tilde{w}_i \rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right), \tag{2}$$

where $\tilde{w}_i = 1/\pi_i$.

From the previous definition, we present two new methods for Bayesian quantile regression for bounded count data under informative sampling in the following two subsections.

3.3.1 The weighted scale approach

From Expression (2), it is straightforward to extend the Bayesian quantile regression model for count data proposed by Lee and Neocleous (2010) for bounded count data under informative sampling. From the content presented in Section 3.2 and Chapter 2, we can assume that $Y_i^* = h_\tau(Z_i) = \log((Y_i + U_i - \tau - \tilde{a})/[\tilde{b} - (Y_i + U_i - \tau)])$, $i \in \mathcal{S}$, follows an asymmetric Laplace distribution (ALD), $Y_i \sim \mathcal{AL}(\eta_i(\tau), \tilde{\sigma}_i, \tau)$, where the scale parameter is $\tilde{\sigma}_i = \sigma/w_i \in (0, \infty)$.

As Savitsky and Toth (2016), we take normalized sampling weights $w_i = n\pi_i^{-1} / \sum_{i \in S} \pi_i^{-1}$.

Following Kozumi and Kobayashi (2011) and taking the mixture representation for the asymmetric Laplace distribution, we rewrote the model as below:

$$y_i^* = \mathbf{x}_i' \boldsymbol{\beta}(\tau) + \theta \nu_i + \gamma \sqrt{\tilde{\sigma}_i \nu_i} \varepsilon_i, \tag{3}$$

where $\theta = \frac{1-2\tau}{\tau(1-\tau)}$, $\gamma^2 = \frac{2}{\tau(1-\tau)}$, $\nu_i \sim \text{Exp}(\tilde{\sigma}_i)$ and $\varepsilon_i \sim N(0,1)$ are mutually independent, and $\text{Exp}(\zeta)$ denotes an exponential distribution with mean ζ. As a consequence of the representation in Equation (3), we have the following hierarchical model

$$\begin{aligned} y_i^* | \nu_i, \mathbf{x}_i, w_i, \boldsymbol{\beta}(\tau), \sigma &\sim N\left(\mathbf{x}_i' \boldsymbol{\beta}(\tau) + \theta \nu_i, \gamma^2 \nu_i \sigma / w_i\right) \\ \nu_i | w_i, \sigma &\sim \text{Exp}(\sigma / w_i) \\ u_i &\sim \text{Unif}(0,1). \end{aligned} \tag{4}$$

The hierarchical representation in (4) is convenient as it proportionates closed form full conditional distributions when proceeding with the posterior inference of $\boldsymbol{\lambda}(\tau) = (\boldsymbol{\beta}(\tau), \sigma)$ based on the observed data. Assuming prior independence between $\boldsymbol{\beta}(\tau)$ and σ, and specifying the prior distributions $\boldsymbol{\beta}(\tau) \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ and $\sigma \sim IG(c_0, d_0)$, where $IG(c,d)$ denotes the Inverse Gamma distribution with shape parameter $c > 0$ and scale parameter $d > 0$, the resultant full conditional densities have closed form as follows

- $\boldsymbol{\beta}(\tau) | \sigma, \boldsymbol{\nu} \sim N(\boldsymbol{\mu}_1, \Sigma_1)$, where

$$\Sigma_1 = \left[\sum_{i \in S} \frac{w_i}{\gamma^2 \sigma \nu_i} \mathbf{x}_i \mathbf{x}_i' + \Sigma_0^{-1} \right]^{-1} \text{ and } \boldsymbol{\mu}_1 = \Sigma_1 \left[\sum_{i \in S} \frac{w_i \left(y_i^* - \theta \nu_i\right)}{\gamma^2 \sigma \nu_i} \mathbf{x}_i + \Sigma_0^{-1} \boldsymbol{\mu}_0 \right];$$

- $\sigma | \boldsymbol{\beta}(\tau), \boldsymbol{\nu} \sim IG\left(c_0 + \frac{3n}{2}, d_0 + \sum_{i \in S} \left[\frac{w_i \left(y_i^* - \mathbf{x}_i' \boldsymbol{\beta}(\tau) - \theta \nu_i\right)^2}{2\gamma^2 \nu_i} + w_i \nu_i \right] \right)$;
- $\nu_i | \boldsymbol{\beta}(\tau), \sigma \sim GIG\left(\frac{1}{2}, \frac{w_i \left(y_i^* - \mathbf{x}_i' \boldsymbol{\beta}(\tau)\right)^2}{\gamma^2 \sigma}, \frac{w_i \theta^2}{\gamma^2 \sigma} + \frac{2 w_i}{\sigma}\right)$.

Therefore, it results in a simple and efficient MCMC algorithm with Gibbs steps only where new values for U_i, $i \in \mathcal{S}$, are drawn from a standard uniform at each iteration of the procedure

(Lee and Neocleous 2010), and new Y_i^*'s are also built from these draws at each iteration of the procedure.

3.3.2 The pseudo-likelihood method

From Expression (2), but taking the normalized weights as in the previous subsection, we can also obtain the equivalence below:

$$\begin{aligned}
\underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in S} w_i \rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right) &\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \exp\left\{-\sum_{i\in S} w_i \rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\} \\
&\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i\in S} \exp\left\{-w_i \rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\} \\
&\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i\in S} \exp\left\{-\rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\}^{w_i} \\
&\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i\in S} \left[\tau(1-\tau)\right]^{w_i} \exp\left\{-\rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\}^{w_i} \quad (5) \\
&\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i\in S} \left[\tau(1-\tau) \exp\left\{-\rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\}\right]^{w_i} \\
&\equiv \underset{\boldsymbol{\beta}(\tau)\in\mathbb{R}^p}{\operatorname{argmax}} \prod_{i\in S} \left[f\left(h_\tau(z_i)|\mathbf{x}_i, \boldsymbol{\beta}(\tau)\right)\right]^{w_i},
\end{aligned}$$

where $f\left(h_\tau(z_i)|\mathbf{x}_i, \boldsymbol{\beta}(\tau)\right)$ is the probability density function of the asymmetric Laplace distribution with location equal to $\mathbf{x}_i'\boldsymbol{\beta}(\tau)$, scale equals 1, and skewness equals τ. As mentioned by Yu and Moyeed (2001), one can also include a scale parameter σ in this case. From the result in (5), we have a pseudo-likelihood approach (Chambers and Skinner 2003) with the ALD.

Following a Bayesian paradigm, our primary interest lies in the posterior inference of $\boldsymbol{\beta}(\tau)$ based on the observed data. For that, we assign a prior distribution for $\boldsymbol{\beta}(\tau)$, $\pi\left(\boldsymbol{\beta}(\tau)\right)$. Consequently, we can write a pseudo-posterior distribution as below:

$$\tilde{\pi}\left(\boldsymbol{\beta}(\tau)|\mathbf{y}, \mathbf{w}\right) \propto \prod_{i\in S} \left[\tau(1-\tau) \exp\left\{-\rho_\tau \left(h_\tau(z_i) - \mathbf{x}_i'\boldsymbol{\beta}(\tau)\right)\right\}\right]^{w_i} \pi\left(\boldsymbol{\beta}(\tau)\right). \quad (6)$$

The expression in (6) is straightly linked with the framework introduced by Savitsky and Toth (2016).

Proceeding with the Bayesian analysis and computation, we assume a prior distribution $\boldsymbol{\beta}(\tau) \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ for the parameter of interest. However, sampling from $\tilde{\pi}\left(\boldsymbol{\beta}(\tau)|\mathbf{y}, \mathbf{w}\right)$ is not straightforward, and Markov chain Monte Carlo (Gamerman and Lopes 2006, MCMC) algorithms are employed to simulate observations from our target distribution iteratively. In this chapter, we apply the adaptive Metropolis–Hastings algorithm proposed by Shaby and Wells (2010).

Assuming that $Y_i^* = h_\tau(Z_i) = \log((Y_i + U_i - \tau - \tilde{a})/[\tilde{b} - (Y_i + U_i - \tau)])$, $i \in \mathcal{S}$, suppose that $\boldsymbol{\beta}^{(0)}$ is an initial value for the coefficients, and $\delta_m^{(0)}$ is an initial value for the tuning quantity δ_m. Following the strategy of Shaby and Wells (2010), a sample of draws from (6) can be obtained by repeating the following three steps for $t = 0, \ldots, T$:

- Draw $U_i \sim \text{Unif}(0, 1)$, $i \in \mathcal{S}$;
- Provided the current state $\boldsymbol{\beta}^{(t)}$, generate $\boldsymbol{\beta}^*$ from a proposal distribution $\boldsymbol{\beta}|\boldsymbol{\beta}^{(t)} \sim N(\boldsymbol{\beta}^{(t)}, \hat{\Sigma})$, where

$$\hat{\Sigma} = \delta^{(t)}\tau(1-\tau)\left(\frac{1}{n}\sum_{i\in\mathcal{S}} w_i^2 \mathbf{x}_i \mathbf{x}_i'\right)^{-1}; \tag{7}$$

- Draw $\tilde{U} \sim \text{Unif}(0, 1)$, the standard uniform distribution, and compute $\boldsymbol{\beta}^{(t+1)}$ applying the acceptance-rejection rule as follows:

$$\boldsymbol{\beta}^{(t+1)} = \begin{cases} \boldsymbol{\beta}^*, \text{ if } \tilde{u} \leq \alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) \\ \boldsymbol{\beta}^{(t)}, \text{ otherwise,} \end{cases}$$

where

$$\begin{aligned} \alpha(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^*) &= \min\left\{1, \frac{p(\boldsymbol{\beta}^*|\mathbf{y}^*, \mathbf{x}, \mathbf{w}) q(\boldsymbol{\beta}^{(t)}|\boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^{(t)}|\mathbf{y}^*, \mathbf{x}, \mathbf{w}) q(\boldsymbol{\beta}^*|\boldsymbol{\beta}^{(t)})}\right\} \\ &= \min\left\{1, \frac{p(\boldsymbol{\beta}^*|\mathbf{y}^*, \mathbf{x}, \mathbf{w})}{p(\boldsymbol{\beta}^{(t)}|\mathbf{y}^*, \mathbf{x}, \mathbf{w})}\right\} \end{aligned}$$

- For every m steps, take $\hat{r}^{(t)} = \#\text{jumps}/m$ and $\kappa^{(t)} = k_0(1/t^{k_1})$, where #jumps indicates the number of jumps, and set

$$\log(\delta_m^{(t+1)}) = \log(\delta_m^{(t)}) + \kappa^{(t)}(\hat{r}^{(t)} - 0.234).$$

The attenuation parameters $k_0 > 0$ and $k_1 \in (0, 1]$ are the only discretionary quantities. However, Shaby and Wells (2010) checked that their choices are not very influential and suggested choosing $k_0 = 1$ and $k_1 = 0.8$. The proposal distribution in (7) is also implemented in Chapter 2. This approach tends to be less computationally costly for a fixed number of iterations T as we do not have to sample from the latent variables at each iteration.

3.4 Real-data-based simulation study

We further assess the performance of our method for bounded count data, proposing a simulation study based on a real data set that is similar to that analyzed in the previous chapter. The data was collected as a part of the 2011 Prova Brasil. Prova Brasil is a large-scale proficiency test developed by the Brazilian National Institute of Education Research (INEP) for the Brazilian Ministry of Education that aims to monitor the educational quality offered by the Brazilian public educational system through standardized tests. The tests are focused on Portuguese and Mathematics exams applied to students in the fifth and ninth years of public elementary schools.

There are indications from previous editions that students with low achievement are less likely to participate in the exam, suggesting that Prova Brasil is subject to a non-response bias caused by school evasion or non-attendance on the day of the exam. Hence, although it does not consist of a sample survey, there is an informative mechanism in the observed data, and based on it, we propose a simulation study that intends to reproduce this effect. For that, we fix the $N = 10,941$ observations from students in the ninth year from public elementary schools located in Campinas municipality in São Paulo State as our finite population. Moreover, we consider the number of correct answers in the Mathematics exam as the response variable

and the proficiency score in the Portuguese exam (x_1) as the covariate. This proficiency score stems from applying item response theory (IRT) models to the test results. Regarding the response variable, its domain is bounded on $\{0, 1, 2, \dots, 25, 26\}$, where 26 is the total number of questions in the exam.

To evaluate the effectiveness of our methodologies, we compare the proposed methods, the method where the scale parameter of the ALD is proportional to the weight (BWQR-AL) and the pseudo-likelihood method (BWQR-PL), with a best-case scenario and a benchmark. As the best-case scenario, we take the estimates from the unweighted Bayesian quantile regression for bounded count data (BQR-AL) considering the fully observed population. As the benchmark, we take the results from the BQR-AL considering the sampled data. Thus, the BWQR-AL and the BWQR-PL should be close to the best-case scenario and present, at the same time, a better performance in comparison with the benchmark. The comparisons are made under different sampling designs, different degrees of informativeness, and different prior specifications.

In the following subsections, for each of the scenarios, we generate $M = 500$ replicas, and analyze the quantiles $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$. For each replica, MCMC chains of length $25,000$ are drawn from the full conditional distributions described in Section 3.3, and a burn-in of $5,000$ and a thin of 5 are taken. Following Yang, Wang, and He (2016) and Benoit and den Poel (2017), $\sigma = 1$ is fixed. We also denote the intercept as $\beta_0(\tau)$ and the coefficient for the explanatory variable as $\beta_1(\tau)$

3.4.1 Experiment 1: Analysis under different sampling designs

We evaluate and compare our methods to the benchmark in this experiment through different sampling designs. We consider three different probability proportional to size (PPS) processes to obtain the final samples: Poisson, stratified, and systemic. In all cases, samples with size equals to $n = 500$ are selected with inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$, where $k_i = \{1 + \exp(2.5 - 0.5z_i)\}^{-1}$ and $z_i \sim N(1 + 0.5y_i + x_{1i}, 0.25)$. These particular choices for k_i and z_i are similar to those assumed by Chen and Zhao (2019). By doing this, we

have an informative single-stage sample design based on a PPS sample in which the sampling inclusion probability is directly proportional to the response variable. We are setting, therefore, a scenario in which students with high achievement have a higher probability of participating.

Regarding the prior specification, we chose a non-informative prior for the regression parameters in which $\boldsymbol{\mu}_0 = \underline{0}_2$ and $\Sigma_0 = 1,000 I_2$, where $\underline{0}_2$ is a 2-dimensional vector of zeros and I_2 is an identity matrix with dimension 2×2.

Table 3.1 reports the variance of the estimators and a measure analogous to bias in which we replace the true value with the estimate from the best-case scenario. Hereafter, we refer to the latter as bias. In terms of bias, we observe that the BWQR-AL and the BWQR-PL outperform the BQR-AL regardless of the quantile, the sampling scheme, and the coefficient under analysis. We note that the estimates of the proposed methods are close to the best-case scenario, irrespective of the scenario. Comparing the BWQR-AL with the BWQR-PL and considering the quantiles $\tau \in \{0.10, 0.25, 0.75\}$, there is not an evident advantage between them as they alternate the lowest bias values among the coefficients. However, for the quantiles $\tau \in \{0.50, 0.90\}$, there is a preference for the BWQR-AL.

Concerning the variance, we note that for all sampling schemes and $\tau \in \{0.10, 0.25, 0.50\}$, the BQR-AL has a better performance in general; the only exception is the Stratified case at $\tau = 0.50$ where the BWQR-PL has a lower variance. At the same time, our methods perform better for all sampling schemes and for $\tau \in \{0.75, 0.90\}$, except for $\beta_1(\tau)$ considering the Systematic scheme and $\tau = 0.75$. Comparing the BWQR-AL with the BWQR-PL, we note that the latter slightly outperforms the former in all scenarios under analysis.

Hence, our methods seem to present a better performance regarding the variance when we consider the quantiles where the sampling inclusion probabilities are higher. The remaining results about the variance of the estimators might be reflecting, in some sense, the increase in variance often observed from using the weights in a fully parametric context.

Table 3.1: Bias ($\times 10^2$) and variance ($\times 10^2$) in parenthesis under different sampling designs. The lowest bias values are indicated in bold.

τ	Method	Poisson		Stratified		Systematic	
		$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$
0.10	BQR-AL	1.38887	0.27542	0.58377	0.09724	1.43594	0.25236
		(0.11896)	(0.24278)	(0.15620)	(0.21631)	(0.11126)	(0.28696)
	BWQR-AL	**0.06030**	0.00140	**0.04787**	0.00547	**0.06090**	0.00096
		(0.27902)	(0.61907)	(0.21556)	(0.31598)	(0.27404)	(0.63133)
	BWQR-PL	0.12632	**0.00076**	0.10376	**0.00396**	0.12369	**0.00042**
		(0.26736)	(0.59832)	(0.20663)	(0.30726)	(0.26200)	(0.60536)
0.25	BQR-AL	0.99223	0.27350	0.68118	0.00534	1.00606	0.26758
		(0.06394)	(0.15783)	(0.08025)	(0.13386)	(0.06312)	(0.19325)
	BWQR-AL	**0.05348**	0.00683	**0.01713**	0.00438	**0.04079**	0.00506
		(0.10127)	(0.31372)	(0.10436)	(0.16057)	(0.09658)	(0.30887)
	BWQR-PL	0.07824	**0.00443**	0.03283	**0.00250**	0.06253	**0.00291**
		(0.09926)	(0.30257)	(0.09961)	(0.15755)	(0.09372)	(0.30202)
0.50	BQR-AL	0.90993	0.41661	0.81339	0.05017	0.93445	0.44667
		(0.08605)	(0.17783)	(0.07329)	(0.13932)	(0.08516)	(0.17511)
	BWQR-AL	**0.02430**	**0.00017**	**0.00266**	**0.00284**	**0.02162**	**0.00170**
		(0.09640)	(0.21332)	(0.07970)	(0.14094)	(0.09190)	(0.21083)
	BWQR-PL	0.02781	0.00204	0.00399	0.00665	0.02546	0.00529
		(0.09146)	(0.20621)	(0.07317)	(0.13566)	(0.08766)	(0.20471)
0.75	BQR-AL	0.78852	0.29109	0.7726	0.10844	0.77831	0.31300
		(0.12231)	(0.18683)	(0.09234)	(0.16221)	(0.11965)	(0.17894)
	BWQR-AL	0.00158	0.00087	**0.00117**	**0.00006**	0.00146	**0.00000**
		(0.11498)	(0.17830)	(0.08539)	(0.14499)	(0.11756)	(0.18957)
	BWQR-PL	**0.00001**	**0.00009**	0.00559	0.00082	**0.00000**	0.00049
		(0.10858)	(0.17332)	(0.07944)	(0.14097)	(0.11050)	(0.18782)
0.90	BQR-AL	0.78845	0.26960	0.84475	0.14099	0.82855	0.28535
		(0.20187)	(0.34418)	(0.18887)	(0.32948)	(0.19635)	(0.31997)
	BWQR-AL	**0.00479**	**0.00256**	**0.02359**	**0.00482**	**0.00904**	**0.00623**
		(0.18614)	(0.30839)	(0.16621)	(0.27776)	(0.17263)	(0.30814)
	BWQR-PL	0.02496	0.00457	0.05964	0.00742	0.03299	0.00846
		(0.17509)	(0.29383)	(0.15402)	(0.26437)	(0.16222)	(0.29555)

3.4.2 Experiment 2: Analysis under different degrees of informativeness

In this experiment, we compare the performances of the proposed methods and the BQR-AL under different degrees of informativeness. For that, we take Poisson samples with expected size equal to $n = 500$ with the following inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$, where $k_i = \{1 + \exp(2.5 - 0.5z_i)\}^{-1}$ and z_i assumes three different specifications: 1) $z_i \sim N(1 + 0.5y_i + x_{1i}, 0.25)$ (Informative I); 2) $z_i \sim N(1 + 1.5y_i + x_{1i}, 0.25)$ (Informative II); 3) $z_i \sim N(1 + x_{1i}, 0.25)$ (Non-informative). The first two specifications are informative since both the inclusion probabilities depend on the response variable. However, the first represents a case of a higher degree of information. The third case represents a non-informative scenario since

the inclusion probabilities do not depend on the response variable. This scenario is important to address how including sample weights when it is not necessarily required impacts models' performance (Pfeffermann, Moura, and Silva 2006).

To illustrate the different degrees of informativeness we proposed, we use a similar procedure to Chen and Zhao (2019) and calculate the partial correlation between the response variables and weights given the covariate to measure the degrees of informativeness. For that, we fit a generalized linear model to the response variable and a linear model to the weights in the logarithmic scale. From the estimated residuals, we compute the correlations. Figure 3.1 summarizes the results.

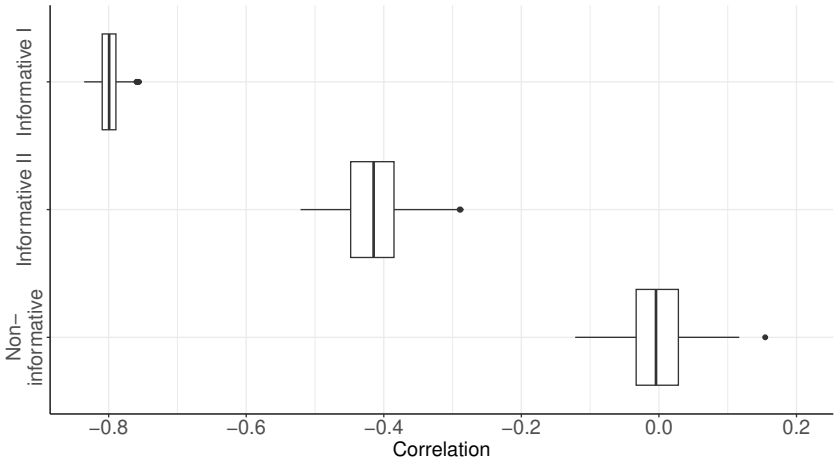

Figure 3.1: Boxplots of the correlations from each replica for the different degrees of informativeness.

Table 3.2 reports the variance of the estimators and the bias, as explained in the previous subsection. Note that we are making the same choice as the prior experiment for the first specification; we then take the results from the Poisson scheme in Table 3.1 as the scenario with a higher degree of informativeness and also report them in Table 3.2 to facilitate the visualization and the comparisons. Therefore, considering the case with a high degree of informativeness (Informative I), we observe that the proposed methods are advantageous in terms of bias, and the BWQR-AL and the BWQR-PL alternate, which is better depending on the quantile and the coefficient. Regarding the variance, our methods perform better in the quantiles with higher sampling inclusion probabilities.

Table 3.2: Bias ($\times 10^2$) and variance ($\times 10^2$) in parenthesis under different degrees of informativeness. The lowest bias values are indicated in bold.

τ	Method	Informative I		Informative II		Non-informative	
		$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$
0.10	BQR-AL	1.38887	0.27542	**0.00139**	0.04787	0.04194	0.11661
		(0.11896)	(0.24278)	(0.15508)	(0.24669)	(0.21014)	(0.26802)
	BWQR-AL	**0.06030**	0.00140	0.03167	0.01080	**0.03100**	**0.00000**
		(0.27902)	(0.61907)	(0.18531)	(0.36154)	(0.22400)	(0.35247)
	BWQR-PL	0.12632	**0.00076**	0.08123	**0.00818**	0.08048	0.00013
		(0.26736)	(0.59832)	(0.17748)	(0.34841)	(0.21291)	(0.34001)
0.25	BQR-AL	0.99223	0.27350	**0.00147**	0.01456	0.01534	0.12181
		(0.06394)	(0.15783)	(0.10293)	(0.15034)	(0.12677)	(0.18346)
	BWQR-AL	**0.05348**	0.00683	0.00784	0.00622	**0.00968**	0.00256
		(0.10127)	(0.31372)	(0.11213)	(0.18050)	(0.13185)	(0.21969)
	BWQR-PL	0.07824	**0.00443**	0.01943	**0.00357**	0.02196	**0.00098**
		(0.09926)	(0.30257)	(0.10785)	(0.17312)	(0.12777)	(0.21334)
0.50	BQR-AL	0.90993	0.41661	0.00370	0.02620	0.00289	0.06114
		(0.08605)	(0.17783)	(0.11975)	(0.15553)	(0.11769)	(0.19390)
	BWQR-AL	**0.02430**	**0.00017**	**0.00049**	**0.00096**	**0.00166**	**0.00231**
		(0.09640)	(0.21332)	(0.12430)	(0.16114)	(0.12444)	(0.23098)
	BWQR-PL	0.02781	0.00204	0.00114	0.00398	0.00250	0.00631
		(0.09146)	(0.20621)	(0.11769)	(0.15559)	(0.11827)	(0.22104)
0.75	BQR-AL	0.78852	0.29109	0.00915	0.00406	**0.00002**	0.08907
		(0.12231)	(0.18683)	(0.13333)	(0.16429)	(0.12483)	(0.19430)
	BWQR-AL	0.00158	0.00087	**0.00167**	0.00034	0.00070	**0.00011**
		(0.11498)	(0.17830)	(0.13408)	(0.16878)	(0.13171)	(0.23936)
	BWQR-PL	**0.00001**	**0.00009**	0.00650	**0.00000**	0.00422	0.00013
		(0.10858)	(0.17332)	(0.12816)	(0.16519)	(0.12501)	(0.23366)
0.90	BQR-AL	0.78845	0.26960	0.03889	0.00841	0.01908	0.10438
		(0.20187)	(0.34418)	(0.24955)	(0.32129)	(0.25054)	(0.37479)
	BWQR-AL	**0.00479**	**0.00256**	**0.02526**	**0.00087**	**0.01477**	**0.00011**
		(0.18614)	(0.30839)	(0.25022)	(0.32582)	(0.26485)	(0.42957)
	BWQR-PL	0.02496	0.00457	0.06062	0.00209	0.04577	0.00039
		(0.17509)	(0.29383)	(0.23386)	(0.31098)	(0.24632)	(0.40948)

Analyzing the case Informative II, the BWQR-AL performs better in terms of bias when $\tau \in \{0.50, 0.75, 0.90\}$, except for $\beta_1(\tau)$ at $\tau = 0.75$. For $\tau \in \{0.10, 0.25\}$, the benchmark outperforms the proposed methods regarding the intercept, and the BWQR-PL has the best performance for $\beta_1(\tau)$. For the Non-informative case, we observe that our models are competitive compared to the benchmark. This allows us to say that including sample weights when not needed does not imply a marked decrease in performance, mainly for the BWQR-AL.

3.4.3 Experiment 3: Analysis under different sample sizes

Our third experiment evaluates the performances under different sample sizes, particularly $n \in \{250, 500, 1000\}$. We take Poisson samples with inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$, where $k_i = \{1 + \exp(2.5 - 0.5z_i)\}^{-1}$ and $z_i \sim N(1 + 0.5y_i + x_{1i}, 0.25)$. Note that the case in

which $n = 500$ is employed as a reference case is identical to the Poisson scheme in the first experiment and the Informative I scenario in the second experiment.

Table 3.3 shows the results regarding bias and variance. We first compare the results among the sample sizes. As expected, the variances of the estimators decrease as the sample size grows aside from the quantile, the coefficient, or the model under analysis. The same conjoint movement about bias is not observed, so we interpret the results individually for each model. Considering the BWQR-AL, we see that the bias of $\beta_0(\tau)$ declines as n increases; the only exception occurs for $n = 500$ at $\tau = 0.75$. We see a similar tendency for $\beta_1(\tau)$, the exceptions occur for $n = 500$ at $\tau = 0.90$ and for $n = 1,000$ at $\tau = 0.10$ and $\tau = 0.25$. Considering the BWQR-PL, we also note a general decrease as n increases for the bias of $\beta_0(\tau)$; the exceptions arise for $n = 1,000$ at $\tau = 0.75$ and $\tau = 0.90$. For $\beta_1(\tau)$, the bias increments are observed for $n = 500$ at $\tau = 0.50$ and $\tau = 0.90$, and for $n = 1,000$ at $\tau = 0.10$ and $\tau = 0.25$. These declines as n increases are not observed for the BQR-AL; for example, there are bias increments in almost all scenarios for the intercept. The same happens for $\beta_1(\tau)$ mainly when $n = 500$.

Comparing the results from the proposed methods to those from the BQR-AL, we note that ours performs better regardless of the scenario under analysis. Comparing the BWQR-AL with the BWQR-PL and considering the quantiles $\tau \in \{0.10, 0.25\}$, the former seems to perform better for the intercept, while the latter performs better for $\beta_1(\tau)$. For the quantiles $\tau \in \{0.50, 0.90\}$, there is a preference for the BWQR-AL. For $\tau = 0.75$, on the other hand, there is a preference for the BWQR-PL.

Considering the variance, the BQR-AL has lower variances at $\tau \in \{0.10, 0.25, 0.50\}$, while the BWQR-AL and the BWQR-PL present a similar or lower variance at $\tau \in \{0.75, 0.90\}$. Regarding the proposed methods, we observe that the BWQR-PL slightly outperforms the BWQR-AL in all scenarios under analysis. These findings are similar to those observed in the first experiment.

Table 3.3: Bias ($\times 10^2$) and variance ($\times 10^2$) in parenthesis under different sample sizes. The lowest bias values are indicated in bold.

τ	Method	$n = 250$		$n = 500$		$n = 1,000$	
		$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$	$\beta_0(\tau)$	$\beta_1(\tau)$
0.10	BQR-AL	0.69275	0.16753	1.38887	0.27542	1.77049	0.28889
		(0.22124)	(0.49142)	(0.11896)	(0.24278)	(0.06342)	(0.14027)
	BWQR-AL	**0.33717**	0.00370	**0.06030**	0.00140	**0.01911**	0.00382
		(0.46643)	(1.09898)	(0.27902)	(0.61907)	(0.13913)	(0.35933)
	BWQR-PL	0.45884	**0.00275**	0.12632	**0.00076**	0.06660	**0.00244**
		(0.45927)	(1.07609)	(0.26736)	(0.59832)	(0.13282)	(0.34240)
0.25	BQR-AL	0.66673	0.18930	0.99223	0.27350	1.23214	0.26865
		(0.12079)	(0.35159)	(0.06394)	(0.15783)	(0.03618)	(0.08395)
	BWQR-AL	**0.20340**	0.02744	**0.05348**	0.00683	**0.01334**	0.00730
		(0.20565)	(0.58951)	(0.10127)	(0.31372)	(0.06085)	(0.15946)
	BWQR-PL	0.23953	**0.02255**	0.07824	**0.00443**	0.02903	**0.00452**
		(0.20006)	(0.57837)	(0.09926)	(0.30257)	(0.05812)	(0.15397)
0.50	BQR-AL	0.77323	0.38314	0.90993	0.41661	1.07629	0.40594
		(0.14901)	(0.37493)	(0.08605)	(0.17783)	(0.04178)	(0.08256)
	BWQR-AL	**0.07424**	0.00098	**0.02430**	**0.00017**	**0.00423**	**0.00001**
		(0.16165)	(0.44334)	(0.09640)	(0.21332)	(0.05505)	(0.10772)
	BWQR-PL	0.07838	**0.00003**	0.02781	0.00204	0.00608	0.00111
		(0.15386)	(0.43007)	(0.09146)	(0.20621)	(0.04959)	(0.10211)
0.75	BQR-AL	0.82102	0.27300	0.78852	0.29109	0.82854	0.30374
		(0.21411)	(0.42460)	(0.12231)	(0.18683)	(0.04877)	(0.08273)
	BWQR-AL	0.00149	0.00362	0.00158	0.00087	**0.00012**	0.00014
		(0.20440)	(0.43250)	(0.11498)	(0.17830)	(0.04998)	(0.08536)
	BWQR-PL	**0.00011**	**0.00156**	**0.00001**	**0.00009**	0.00116	**0.00003**
		(0.19610)	(0.42610)	(0.10858)	(0.17332)	(0.04663)	(0.08394)
0.90	BQR-AL	1.06773	0.25030	0.78845	0.26960	0.76144	0.20078
		(0.46875)	(0.64844)	(0.20187)	(0.34418)	(0.11829)	(0.19243)
	BWQR-AL	**0.04284**	**0.00112**	**0.00479**	**0.00256**	**0.00426**	**0.00020**
		(0.39512)	(0.59147)	(0.18614)	(0.30839)	(0.10837)	(0.17281)
	BWQR-PL	0.08034	0.00181	0.02496	0.00457	0.02634	0.00123
		(0.38141)	(0.57837)	(0.17509)	(0.29383)	(0.09611)	(0.15766)

3.4.4 Experiment 4: Prior sensitivity analysis

Once we have validated that the proposed methodologies are comparable to the best-case scenario and outperform the benchmark, we carry out a prior sensitivity analysis in this section to assess the effect of specifying different prior distributions in the BWQR-AL. In a similar context, Zhao et al. (2020) considered n-dependent priors (Narisetty and He 2014; Yang and He 2012), justifying that these priors naturally stem in practice, for example, when a previous survey or a pilot survey is available.

In their simulation study, Zhao et al. (2020) centered the prior distributions on the parameter's true value; however, as the authors pointed out, it is unfeasible and cannot be implemented in practical analysis. To overcome this limitation, instead of centering the distributions on the

true value of the parameter, we draw a pilot survey with expected size equals $n_0 = 100$ for each of the 500 Monte Carlo samples and center the priors on the point estimates resultant from fitting the BWQR-AL.

We assume the following prior specifications: a non-informative prior where $\boldsymbol{\mu}_0 = \underline{0}_2$ and $\Sigma_0 = 1,000 I_2$ (Prior I); a n-dependent prior in which $\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\beta}}(\tau)$ and $\Sigma_0 = n_0^{-1/2} I_2$ (Prior II); and a n-dependent prior in which $\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\beta}}(\tau)$ and $\Sigma_0 = n_0^{-1/4} I_2$ (Prior III). We denote the point estimates computed from the pilot surveys by $\tilde{\boldsymbol{\beta}}(\tau)$. Note that both Prior II and Prior III yield valid and strong prior information, but the former is even more informative as the variance is of order $n_0^{-1/2}$.

Ultimately, we generate final samples of expected size equal $n = 500$ independently of the pilot samples. For both pilot and final samples, we use a Poisson sampling with inclusion probabilities $\pi_i = n_0 k_i / \sum_{j=1}^{N} k_j$ and $\pi_i = n k_i / \sum_{j=1}^{N} k_j$, respectively, where $k_i = \{1+\exp(2.5-0.5z_i)\}^{-1}$ and $z_i \sim N(1 + 0.5y_i + x_{1i}, 0.25)$.

Figure 3.2 summarizes the point estimates from all replicas resulting from the BWQR-AL and the BWQR-PL regarding the different prior specifications (Prior I, Prior II, and Prior III) and compares them with the point estimate generated from the best-case scenario. We observe no discrepancy among the results for both proposed models, indicating that the prior specification is not markedly influential in their ability to produce estimates close to the best-case scenario estimates.

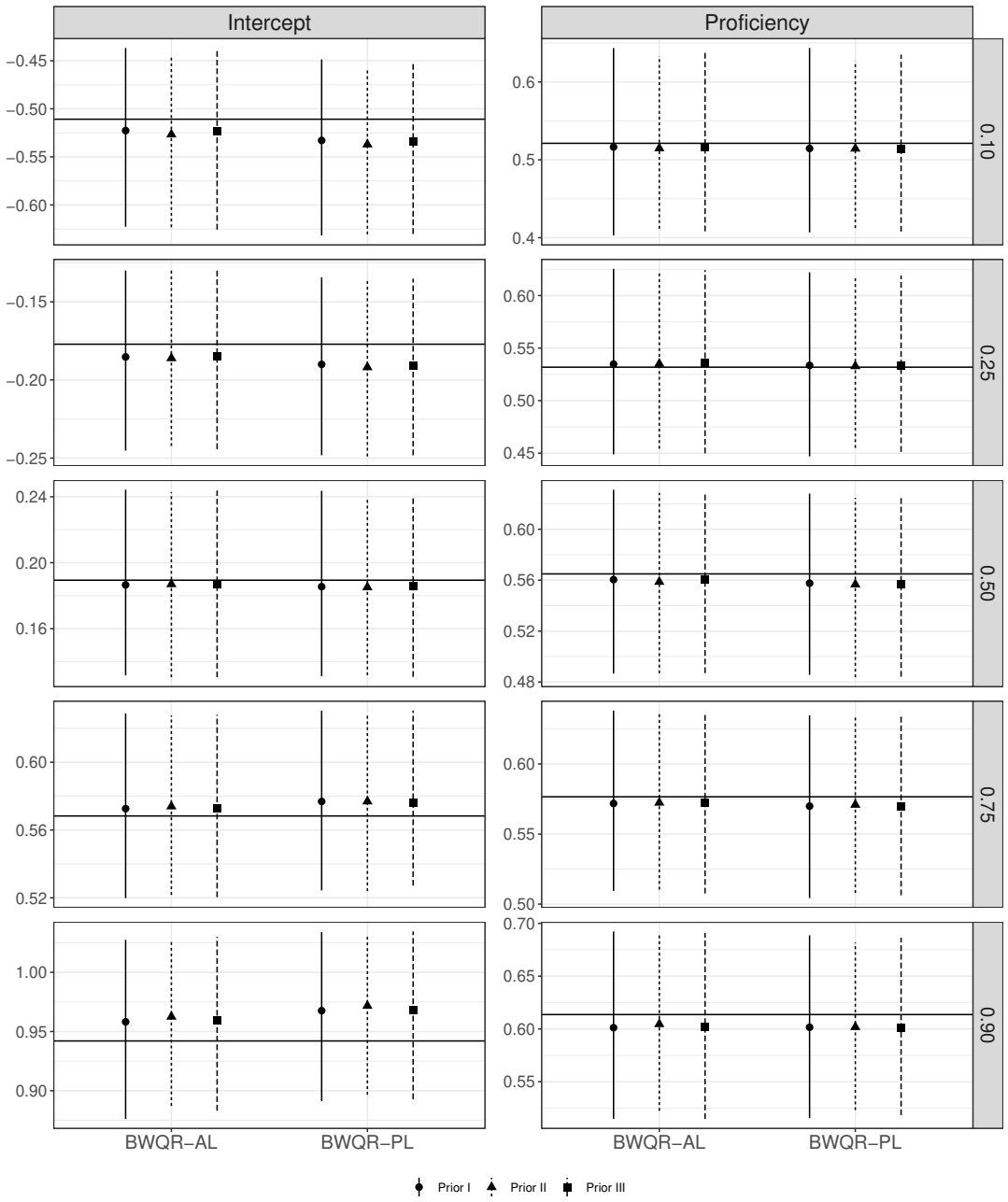

Figure 3.2: Comparison among the different prior specifications. The dots represent the median, and the lines represent ranges from quantile 5% to quantile 95%. The horizontal solid lines represent the best-case scenario.

3.5 Final remarks

This chapter introduced two different Bayesian quantile regression frameworks that accommodate bounded count data under informative sampling. From the side of the bounded count

data, we took the ideas of Machado and Silva (2005) and Lee and Neocleous (2010) to deal with quantiles for counts, adapting them by applying the methodology proposed by Bottai, Cai, and McKeown (2010) to take on values within a known range. From the side of complex survey data under informative sampling, we extended the approach introduced in Chapter 2 and derived an alternative method finding an equivalence between the survey-weighted estimator Chen and Zhao (2019); Geraci (2016) and the method based on the pseudo posterior (Savitsky and Toth 2016).

The asymmetric Laplace distribution was employed as a "working likelihood" for Bayesian quantile inference in both methods. Thus, we did not assume the finite population is generated from a super-population model with an ALD likelihood function but used the distribution as an instrument for an efficient regression parameter estimation. As a result, it was possible to obtain a simple and efficient Gibbs sampling algorithm that is very similar to the one described by Kozumi and Kobayashi (2011) and a simple Metropolis-Hastings algorithm. By doing this, we provided valuable tools that are easy to implement and can be applied by practitioners from different backgrounds.

Although both frameworks presented in this article employed the asymmetric Laplace distribution, how the function is applied differs. In one method, the sampling weights are incorporated into the likelihood through the scale of the ALD. In the other, a sampling-weighted pseudo posterior distribution is constructed by exponentiating each unit likelihood contribution. In summary, setting off from the survey-weighted estimator, we built two different approaches based on the ALD.

To evaluate the performance of our methodologies, we conducted a design-based simulation study using data from Prova Brasil 2011. This simulation study encompassed four different experiments. The first experiment explored different sampling designs, while the second had different degrees of informativeness. The third experiment covered different sample sizes. We observed no clear preference between the proposed models; depending on the scenario, one can perform better than the other. In all three experiments, we compared the methodologies we introduced with a naive model fitting that ignores the informative sampling design. Our

methods outperformed the naive model and, at the same time, produced estimates close to a best-case scenario that considered the entire finite population in the estimation process. By doing this, we could verify that our methods can effectively describe the underlying finite population. Lastly, we developed a prior sensitivity analysis that compared a non-informative prior with two n-dependent priors. The findings indicated no discrepant results among the specifications.

4 Quantiles for multiple-output under informative sampling

4.1 Introduction

Complex surveys are a fundamental tool for data collection in various areas. In education, they are widely applied for student assessment, enabling government agencies and policymakers to measure the educational system's effectiveness. For example, the Organisation for Economic Co-operation and Development (OECD) promotes and implements the Programme for International Student Assessment (PISA) in different countries worldwide. The PISA test evaluates 15-year-olds in applying their reading, mathematics, and science knowledge and skills to solve real-life challenges. In economics, it is vastly employed by official statistics offices to characterize the labor market. In the United Kingdom, the Labour Force Survey (LFS), for example, studies the employment circumstances of the population.

In this context, understanding the relation of multiple response variables with potential covariates might be not only an interesting research topic but also a matter with practical implications in society. However, when data is accessed through a complex sampling scheme, the standard multivariate approaches may lead to biased estimates, and incorporating the sampling design into the model is required. Veiga, Smith, and Brown (2014) extended the probability-weighted iterative generalized least squares estimation method (Pfeffermann et al. 1998, PWIGLS) and applied multivariate multilevel models to investigate how income dynamics differ between the formal and informal sectors in Brazil using the Brazilian Labour Force

Survey.

Motivated by an application in the educational area, Silva and Moura (2022) extended the models developed by Pfeffermann, Moura, and Silva (2006) and developed a multivariate "sample model" (Pfeffermann and Sverchkov 2003) that potentially accommodates complex survey data under an informative design. The authors jointly modeled the proficiency scores in Portuguese and Mathematics obtained from a large-scale standardized proficiency test implemented in Brazil. Considering informative sampling designs, Savitsky and Toth (2016) built a pseudo-posterior distribution utilizing sampling weights and illustrated their method through an application concerning the Bureau of Labor Statistics Job Openings and Labor Turnover Survey (JOLTS). The authors defined a multivariate response variable as the number of job hires and the total separations.

Although computing the relation among multiple responses and explanatory variables in the mean is valuable, interest in establishing this relation among quantiles has recently increased. Multiple-output (i.e., multivariate) quantile regression may supply more comprehensive information on the relationship between response variables and covariates. One can define multiple-output quantiles in different manners (Carlier, Chernozhukov, and Galichon 2016; Chaudhuri 1996; Hallin, Paindaveine, and Siman 2010; Kong and Mizera 2012; Serfling 2002; Small 1990; Wei 2008) with no general agreement about the most appropriate. Developments for Bayesian multiple-output quantiles are less abundant. Bhattacharya and Ghosal (2021) used a geometric definition for a multiple-output quantile location, and Guggisberg (2023) proposed a Bayesian framework for multiple-output quantiles defined parametrically in Hallin, Paindaveine, and Siman (2010). Santos and Kneib (2020) introduced a method for noncrossing quantile contours and structured additive predictors from the methodology of Guggisberg (2023).

In this chapter, we develop a Bayesian multiple-output quantile regression for complex survey data under informative sampling. To the best of our knowledge, advancements in accounting for informative sampling have not yet been proposed in the multiple-output quantile regression literature. For that purpose, we extend Guggisberg (2023) following the ideas of Chapter 2 to

propose a method that relies on the asymmetric Laplace distribution. This distribution is vastly applied for single-output Bayesian quantile regression models (Kozumi and Kobayashi 2011; Yu and Moyeed 2001). From the location-scale mixture representation of this distribution, we introduce an efficient Expectation–Maximization algorithm that propitiates substantial computational savings compared to the commonly used Markov Chain Monte Carlo algorithm. By doing this, our contribution also includes introducing a fast computation method to Bayesian multiple-output quantile regression models.

We assess our method performance through a model-based simulation study where the finite populations were drawn following a data-generating process similar to the one applied by Guggisberg (2023). The study consists of two experiments. First, the samples are generated under different sampling schemes and informative and non-informative designs. In the second experiment, we propose a prior sensitivity analysis that compares a non-informative prior with three n-dependent priors (Narisetty and He 2014; Yang and He 2012). We also propose a design-based simulation study using data from Prova Brasil 2011, fixing the students from the municipality of Campinas who took part in the exam as a finite population. Our method is compared to a naive model that ignores the sampling weights in both simulation studies.

4.2 Multiple-Output Quantile Regression

In this section, we introduce the main ideas for multiple-output quantiles defined parametrically, as in Hallin, Paindaveine, and Siman (2010) and, posteriorly, in Guggisberg (2023). Both papers followed a notion of directional quantiles that coincide with Tukey depth, also known as half-space depth, and that can be accessed with standard quantile regression techniques.

For that, we first delineate some concepts and notations. Suppose that $\mathbf{Y} = (Y_1, \dots, Y_q)'$ is a q-dimensional random vector of outputs, and let $\mathbf{u} = (u_1, \dots, u_q)'$ be a direction such that $\mathbf{u} \in \mathcal{S}^{q-1} = \{\mathbf{v} \in \mathbb{R}^q : \|\mathbf{v}\|_2 = 1\}$, and $\tau \in (0, 1)$ be a magnitude. By $\|\cdot\|_2$, we denote the L_2 norm. The directional quantile is given by the vector $\boldsymbol{\tau} = \tau\mathbf{u}$, and its direction and magnitude are given by $\boldsymbol{\tau} \in \mathcal{B}^q = \{\mathbf{v} \in \mathbb{R}^q : 0 < \|\mathbf{v}\|_2 < 1\}$, in which $\mathcal{B}^q$ is a q-dimensional unit ball centered at $\mathbf{0}$ (with center removed). Additionally, let $\boldsymbol{\Gamma}_u$ be a $q \times (q-1)$ matrix for which

$(\mathbf{u} : \mathbf{\Gamma}_u)$ is an orthogonal basis of $\mathbb{R}^q$.

Denoting $Y_u = \mathbf{u}'\mathbf{Y}$ and $\mathbf{Y}_u^{\perp} = \mathbf{\Gamma}_u'\mathbf{Y}$, Hallin, Paindaveine, and Siman (2010) defined the $\boldsymbol{\tau}$-th quantile of $\mathbf{Y}$ as any element of the set Λ_τ of hyperplanes $\lambda_\tau = \{\mathbf{y} \in \mathbb{R}^q : y_u = \alpha_\tau + \boldsymbol{\beta}_{\tau y}'\mathbf{y}_u^{\perp}\}$ such that

$$(\alpha_\tau, \boldsymbol{\beta}_{\tau y}') \in \underset{(a,\mathbf{b}_y')\in\mathbb{R}^q}{\operatorname{argmin}} \mathrm{E}\left[\rho_\tau\left(Y_u - a - \mathbf{b}_y'\mathbf{Y}_u^{\perp}\right)\right], \tag{1}$$

where $\rho_\tau(k) = k(\tau - \mathbb{I}(k < 0))$ is the asymmetric check loss function, and $\mathbb{I}$ is the indicator function. Supposing that $\mathbf{X} = (X_1, \ldots, X_p)'$ is a random vector of covariates and taking $\lambda_\tau = \{\mathbf{y} \in \mathbb{R}^q : y_u = \alpha_\tau + \boldsymbol{\beta}_{\tau y}'\mathbf{y}_u^{\perp} + \boldsymbol{\beta}_{\tau x}'\mathbf{x}\}$, one can include predictor variables in expression (1) by rewriting the minimization problem as

$$(\alpha_\tau, \boldsymbol{\beta}_{\tau y}', \boldsymbol{\beta}_{\tau x}') \in \underset{(a,\mathbf{b}_y',\mathbf{b}_x')\in\mathbb{R}^{q+p}}{\operatorname{argmin}} \mathrm{E}\left[\rho_\tau\left(Y_u - a - \mathbf{b}_y'\mathbf{Y}_u^{\perp} - \mathbf{b}_x'\mathbf{X}\right)\right]. \tag{2}$$

Hallin et al. (2015) referred to this model as the "unconditional" model and Guggisberg (2023) as the parametric model.

The notion of directional quantiles comes from a quantile contour obtained from the boundary of the intersection of the closed upper half-spaces λ_τ for a fixed τ and all $\mathbf{u}$. Each element $(\alpha_\tau, \boldsymbol{\beta}_{\tau y}', \boldsymbol{\beta}_{\tau x}')$ as defined in (2) delineates an upper (closed) quantile half-space

$$H_\tau^+ = H_\tau^+(\alpha_\tau, \boldsymbol{\beta}_{\tau y}', \boldsymbol{\beta}_{\tau x}') := \{\mathbf{y} \in \mathbb{R}^q : y_u \geq \alpha_\tau + \boldsymbol{\beta}_{\tau y}'\mathbf{y}_u^{\perp} + \boldsymbol{\beta}_{\tau x}'\mathbf{x}\}.$$

From the upper quantile half-spaces, it is possible to build a quantile region $R(\tau)$ for a fixed τ as

$$R(\tau) = \bigcap_{\mathbf{u}\in\mathcal{S}^{q-1}} H_{\tau\mathbf{u}}^+.$$

Hallin, Paindaveine, and Siman (2010) proved that these quantile regions are related to the Tukey depth under certain conditions. As a result of the directional quantile approach, one can get the Tukey depth region through the analysis of $R(\tau)$.

Considering a simple random sample (SRS) $\{(\mathbf{y}_i, \mathbf{x}_i)\}_{i=1}^n$ of $(\mathbf{Y}, \mathbf{X})$, it is possible to define an empirical version to the directional multivariate quantile regression as any element of the collection Λ_τ^* of hyperplanes $\lambda_\tau^* = \{\mathbf{y} \in \mathbb{R}^q : y_u = \hat{\alpha}_\tau + \hat{\boldsymbol{\beta}}'_{\tau y}\mathbf{y}_u^\perp + \hat{\boldsymbol{\beta}}'_{\tau x}\mathbf{x}\}$ such that

$$(\hat{\alpha}_\tau, \hat{\boldsymbol{\beta}}'_{\tau y}, \hat{\boldsymbol{\beta}}'_{\tau x}) \in \underset{(a, \mathbf{b}'_y, \mathbf{b}'_x) \in \mathbb{R}^{q+p}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau \left(y_{ui} - a - \mathbf{b}'_y \mathbf{y}_{ui}^\perp - \mathbf{b}'_x \mathbf{x}_i\right). \tag{3}$$

Following the ideas of Yu and Moyeed (2001), instead of defining estimators as solutions for the minimization problem in (3), Guggisberg (2023) proposed a Bayesian approach that assumes that Y_u follows an asymmetric Laplace distribution (ALD), $Y_u|\mathbf{Y}_u^\perp, \mathbf{X}, \alpha_\tau, \boldsymbol{\beta}_{\tau y}, \boldsymbol{\beta}_{\tau x}, \sigma_\tau \sim ALD(\alpha_\tau + \boldsymbol{\beta}'_{\tau y}\mathbf{y}_u^\perp + \boldsymbol{\beta}'_{\tau x}\mathbf{x}, \sigma_\tau, \tau)$, which implies that the likelihood is

$$\mathcal{L}(\alpha_\tau, \boldsymbol{\beta}'_{\tau y}, \boldsymbol{\beta}'_{\tau x}) = \prod_{i=1}^{n} \frac{\tau(1-\tau)}{\sigma_\tau} \times \exp\left\{-\frac{1}{\sigma_\tau}\rho_\tau \left(y_{ui} - \alpha_\tau - \boldsymbol{\beta}'_{\tau y}\mathbf{y}_{ui}^\perp - \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i\right)\right\},$$

where the location $\eta_{\tau i} = \alpha_\tau + \boldsymbol{\beta}'_{\tau y}\mathbf{y}_{ui}^\perp + \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i$, the scale σ_τ, and the skewness τ are parameters of the ALD. It is worth mentioning that ALD is being used here as a "working likelihood" for Bayesian inference, and more often than not, it is a misspecification of the true underlying likelihood.

4.3 A Bayesian approach for complex survey data under informative sampling

4.3.1 The setup

Let $\mathcal{F}_N = \{(\mathbf{y}_i, \mathbf{x}_i), i = 1, \ldots, N\}$ be a population resultant from a super-population model $\mathcal{F}$ where the underlying distribution function is unknown, and suppose a sample $S = \{i_1, \ldots, i_n\}$ with size $|S| = n$. The empirical versions of the directional multivariate quantile regression described in Section 4.2 particularly work for SRS sampling designs. When the sample S is drawn from a complex survey under an informative design, it is required to incorporate this design somehow in the estimation process. By an informative sampling design we refer to cases in which $\mathbb{P}(I_i = 1|x_i, y_i) \neq \mathbb{P}(I_i = 1|x_i)$, where $I_i \in \{0, 1\}$ is the sampling indicator

such that $I_i = 1$ if unit i is selected and $I_i = 0$ otherwise, and $\pi_i = \mathbb{P}(I_i = 1)$ is the inclusion probability.

To deal with this issue, we define an empirical survey-weighted version of the directional multivariate quantile regression as any element of the collection Λ_τ^* of hyperplanes $\lambda_\tau^* = \{\mathbf{y} \in \mathbb{R}^q : y_u = \hat{\alpha}_\tau + \hat{\boldsymbol{\beta}}'_{\tau y}\mathbf{y}_u^\perp + \hat{\boldsymbol{\beta}}'_{\tau x}\mathbf{x}\}$ such that

$$(\hat{\alpha}_\tau, \hat{\boldsymbol{\beta}}'_{\tau y}, \hat{\boldsymbol{\beta}}'_{\tau x}) \in \underset{(a,\mathbf{b}'_y,\mathbf{b}'_x)\in\mathbb{R}^{q+p}}{\operatorname{argmin}} \sum_{i\in S} w_i \rho_\tau \left(y_{ui} - a - \mathbf{b}'_y\mathbf{y}_{ui}^\perp - \mathbf{b}'_x\mathbf{x}_i\right),$$

where w_i is the normalized sampling weight $w_i = n\pi_i^{-1}/\sum_{i\in S}\pi_i^{-1}$ (Savitsky and Toth 2016). From this definition, and following Guggisberg (2023) and Chapter 2, we can assume that Y_{ui} follows an asymmetric Laplace distribution (ALD), but now $Y_{ui}|\mathbf{Y}_{ui}^\perp, \mathbf{X}_i, w_i, \alpha_\tau, \boldsymbol{\beta}_{\tau y}, \boldsymbol{\beta}_{\tau x}, \sigma_\tau \sim ALD(\eta_{\tau i}, \tilde{\sigma}_{\tau i}, \tau)$, where $\tilde{\sigma}_{\tau i} = \sigma_\tau / w_i$. By doing this, we have the likelihood function is

$$\mathcal{L}(\alpha_\tau, \boldsymbol{\beta}'_{\tau y}, \boldsymbol{\beta}'_{\tau x}) = \prod_{i=1}^{n} w_i \frac{\tau(1-\tau)}{\sigma_\tau} \times \exp\left\{-\frac{w_i}{\sigma_\tau}\rho_\tau\left(y_{ui} - \alpha_\tau - \boldsymbol{\beta}'_{\tau y}\mathbf{y}_{ui}^\perp - \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i\right)\right\}.$$

4.3.2 Mixture representation

Based on data augmentation techniques (Tanner and Wong 1987), the mixture representation of the asymmetric Laplace distribution (Kotz, Kozubowski, and Podgórski 2001) includes latent variables $\boldsymbol{\nu} = \{\nu_i : i \in S\}$ in the observed data and, potentially, facilitates inference procedures. For example, Kozumi and Kobayashi (2011) assumed a location-scale mixture representation for the ALD and introduced a Gibbs sampler to the linear quantile regression model. As Guggisberg (2023), we exploit the ideas of Kozumi and Kobayashi (2011) and consider that

$$y_{ui} = \alpha_\tau + \boldsymbol{\beta}'_{\tau y}\mathbf{y}_{ui}^\perp + \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i + \tilde{\sigma}_{\tau i}\theta\tilde{\nu}_i + \tilde{\sigma}_{\tau i}\gamma\sqrt{\tilde{\nu}_i}\varepsilon_i, \tag{4}$$

where $\theta = \frac{1-2\tau}{\tau(1-\tau)}$, $\gamma^2 = \frac{2}{\tau(1-\tau)}$, and $\tilde{\nu}_i \sim \text{Exp}(1)$ and $\varepsilon_i \sim N(0,1)$ are mutually independent. By $\text{Exp}(\zeta)$, we denote an exponential distribution with mean ζ. The the expression in (4) can

be reparameterized as below:

$$y_{ui} = \alpha_\tau + \boldsymbol{\beta}'_{\tau y}\mathbf{y}^{\perp}_{ui} + \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i + \theta\nu_i + \gamma\sqrt{\tilde{\sigma}_{\tau i}\nu_i}\varepsilon_i, \tag{5}$$

where $\nu_i \sim \text{Exp}(\tilde{\sigma}_{\tau i})$.

From the mixture representation in (5), our model can be represented by the following hierarchical structure:

$$\begin{aligned} Y_{ui}|\mathbf{Y}^{\perp}_{ui}, \mathbf{X}_i, w_i, \nu_i, \alpha_\tau, \boldsymbol{\beta}_{\tau y}, \boldsymbol{\beta}_{\tau x}, \sigma_\tau &\sim N\left(\alpha_\tau + \boldsymbol{\beta}'_{\tau y}\mathbf{y}^{\perp}_{ui} + \boldsymbol{\beta}'_{\tau x}\mathbf{x}_i + \theta\nu_i, \gamma^2\nu_i\tilde{\sigma}_{\tau i}\right) \\ \nu_i|w_i, \sigma_\tau &\sim \text{Exp}(\tilde{\sigma}_{\tau i}). \end{aligned} \tag{6}$$

Thus, the joint density function of the observed data and the latent variables $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)$ is given by

$$\begin{aligned} f(\mathbf{y}_u, \boldsymbol{\nu}|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}_\tau, \sigma_\tau) &= f(\mathbf{y}_u|\boldsymbol{\nu}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta}_\tau, \sigma_\tau) f(\boldsymbol{\nu}|\mathbf{w}, \sigma_\tau) \\ &= \prod_{i\in S} w_i^{1/2} \left(2\pi\sigma_\tau\nu_i\right)^{-1/2} \exp\left\{-\frac{w_i\left(y_{ui} - \mathbf{z}'_i\boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2\sigma_\tau\nu_i}\right\} \\ &\times \left(\sigma_\tau/w_i\right)^{-1} \exp\left\{-\frac{w_i\nu_i}{\sigma_\tau}\right\} \\ &= \left(\prod_{i\in S} w_i^{3/2}\nu_i^{-1/2}\right)\left(2\pi\right)^{-n/2}\sigma_\tau^{-3n/2} \\ &\times \exp\left\{-\sum_{i\in S}\left[\frac{w_i\left(y_{ui} - \mathbf{z}'_i\boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2\sigma_\tau\nu_i} + \frac{w_i\nu_i}{\sigma_\tau}\right]\right\}, \end{aligned}$$

where $\boldsymbol{\beta}_\tau = (\alpha_\tau, \boldsymbol{\beta}'_{\tau y}, \boldsymbol{\beta}'_{\tau x})'$ and $\mathbf{z}_i = (1, \mathbf{y}^{\perp\prime}_{ui}, \mathbf{x}'_i)'$.

4.3.3 Gibbs sampler

Our primary interest lies in the posterior inference of $(\alpha_\tau, \boldsymbol{\beta}'_{\tau y}, \boldsymbol{\beta}'_{\tau x}, \sigma_\tau)$. By exploiting the hierarchical representation described above, the resultant full conditional densities may have closed, and a simple Markov chain Monte Carlo (Gamerman and Lopes 2006, MCMC) algorithm with Gibbs steps only can be implemented to simulate observations for the parameter vector of interest iteratively.

For that, we first need to specify conjugate priors. We assume prior independence between $\boldsymbol{\beta}_\tau$ and σ_τ, which means that $p(\boldsymbol{\beta}_\tau, \sigma_\tau) = p(\boldsymbol{\beta}_\tau)p(\sigma_\tau)$, and consider $\boldsymbol{\beta}_\tau \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ as Guggisberg (2023) and $\sigma_\tau \sim IG(a_0, b_0)$. Here, $IG(a, b)$ denotes the Inverse Gamma distribution with shape parameter $a > 0$ and scale parameter $b > 0$. Hence, the posterior density function of $\boldsymbol{\beta}_\tau$, σ_τ and the latent variables conditional on the observed data is

$$\begin{aligned} p(\boldsymbol{\beta}_\tau, \sigma_\tau, \boldsymbol{\nu}|\mathbf{y}_u, \mathbf{z}, \mathbf{w}) &\propto f(\mathbf{y}_u|\boldsymbol{\nu}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta}_\tau, \sigma_\tau) f(\boldsymbol{\nu}|\mathbf{w}, \sigma_\tau) p(\boldsymbol{\beta}(\tau)) p(\sigma_\tau) \\ &\propto \sigma_\tau^{-3n/2} \exp\left\{ -\sum_{i \in S} \left[\frac{w_i \left(y_{ui} - \mathbf{z}_i' \boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2 \sigma_\tau \nu_i} + \frac{w_i \nu_i}{\sigma_\tau} \right] \right\} \\ &\times \left(\prod_{i \in S} \nu_i^{-1/2} \right) \exp\left\{ -\frac{1}{2} \left(\boldsymbol{\beta}_\tau - \boldsymbol{\mu}_0\right)' \Sigma_0^{-1} \left(\boldsymbol{\beta}_\tau - \boldsymbol{\mu}_0\right) \right\} \\ &\times \sigma_\tau^{-a_0 - 1} \exp\left\{ -\frac{b_0}{\sigma_\tau} \right\}. \end{aligned} \tag{7}$$

From (7), we have that the full conditional densities are as follows

- $\boldsymbol{\beta}_\tau|\mathbf{y}_u, \mathbf{z}, \mathbf{w}, \sigma_\tau, \boldsymbol{\nu} \sim N(\boldsymbol{\mu}_1, \Sigma_1)$, where
  $\Sigma_1 = \left[\sum_{i \in S} \frac{w_i}{\gamma^2 \sigma_\tau \nu_i} \mathbf{z}_i \mathbf{z}_i' + \Sigma_0^{-1} \right]^{-1}$ and $\boldsymbol{\mu}_1 = \Sigma_1 \left[\sum_{i \in S} \frac{w_i (y_{ui} - \theta\nu_i)}{\gamma^2 \sigma_\tau \nu_i} \mathbf{z}_i + \Sigma_0^{-1} \boldsymbol{\mu}_0 \right]$;
- $\sigma_\tau|\mathbf{y}_u, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta}_\tau, \boldsymbol{\nu} \sim IG\left(a_0 + \frac{3n}{2}, b_0 + \sum_{i \in S} \left[\frac{w_i \left(y_{ui} - \mathbf{z}_i' \boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2 \nu_i} + w_i \nu_i \right] \right)$;
- $\nu_i|y_{ui}, \mathbf{z}_i, w_i, \boldsymbol{\beta}_\tau, \sigma_\tau \sim GIG\left(\frac{1}{2}, \frac{w_i \left(y_{ui} - \mathbf{z}_i' \boldsymbol{\beta}_\tau\right)^2}{\gamma^2 \sigma_\tau}, \frac{w_i \theta^2}{\gamma^2 \sigma_\tau} + \frac{2 w_i}{\sigma_\tau} \right)$ for $i \in S$.

4.4 The Expectation-Maximization algorithm

Gibbs samplers similar to the one presented in Subsection 4.3.3 are widely applied for Bayesian quantile regression (Guggisberg 2023; Kozumi and Kobayashi 2011) since they are easily implemented. However, in the context of multiple-output quantile regression, it is required to repeat the procedure in several directions to construe a quantile region for a fixed quantile τ. Hammer, Yazidi, and Rue (2022), for example, gave some examples using fifty directions. Thus, running a Gibbs sampler for different quantiles and directions might be computationally costly. Bearing this in mind, in this section, we introduce an Expectation-Maximization (EM) algorithm as an alternative to overcome this issue. To our knowledge, there is no paper

introducing the EM algorithm for multiple-output quantile regression models.

Unlike the frequently used MCMC algorithms that estimate the entire posterior distribution or variational Bayes methods (Blei, Kucukelbir, and McAuliffe 2017; Waldmann and Kneib 2015; Wand et al. 2011), the EM algorithm iteratively calculates posterior modes of parameters and latent variables, potentially being much less computationally costly. The location-scale mixture representation of the ALD presented in Subsection 4.3.2 plays an essential role in this fast implementation for estimating quantiles.

To develop our algorithm, we follow Zhao and Lian (2016). The expectation (E) step consists of replacing the latent variables $\boldsymbol{\nu}$ with their expectations conditional on the observed data and the parameters, whereas the maximization (M) step consists of maximizing the expected log-posterior (objective function) resultant from the E step. In the M step, each parameter maximization occurs conditionally in the other parameters. This approach is known as the Expectation Conditional Maximization (Meng and Rubin 1993, ECM) algorithm. Iterating both steps, a sequence of values is generated, and it monotonically converges a local maximum of the posterior distribution.

In our approach, the objective function is given by

$$
\begin{aligned}
Q(\boldsymbol{\beta}_\tau, \sigma_\tau) &= \mathrm{E}_{\boldsymbol{\nu}}\left[\log\left(p(\boldsymbol{\beta}_\tau, \sigma_\tau, \boldsymbol{\nu}|\mathbf{y}_u, \mathbf{z}, \mathbf{w})\right)\right] \\
&= C + \mathrm{E}_{\boldsymbol{\nu}}\left[-\frac{3n + a_0 + 1}{2}\log\sigma_\tau - \sum_{i\in S}\frac{w_i\left(y_{ui} - \mathbf{z}_i'\boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2\sigma_\tau\nu_i} - \sum_{i\in S}\frac{w_i\nu_i}{\sigma_\tau}\right. \\
&\left. - \frac{b_0}{\sigma_\tau} - \frac{1}{2}\boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\beta}_\tau - \boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\mu}_0\right] \\
&= C - \frac{3n + a_0 + 1}{2}\log(\sigma_\tau) - \mathrm{E}_{\boldsymbol{\nu}}\left[\sum_{i\in S}\frac{w_i\left(y_{ui} - \mathbf{z}_i'\boldsymbol{\beta}_\tau - \theta\nu_i\right)^2}{2\gamma^2\sigma_\tau\nu_i}\right] - \mathrm{E}_{\boldsymbol{\nu}}\left[\sum_{i\in S}\frac{w_i\nu_i}{\sigma_\tau}\right] \quad (8) \\
&- \frac{b_0}{\sigma_\tau} - \frac{1}{2}\boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\beta}_\tau - \boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\mu}_0 \\
&= C - \frac{3n + a_0 + 1}{2}\log(\sigma_\tau) - \sum_{i\in S}\frac{w_i\left(y_{ui} - \mathbf{z}_i'\boldsymbol{\beta}_\tau - \theta/\mathrm{E}[\nu_i^{-1}]\right)^2}{2\gamma^2\sigma_\tau/\mathrm{E}[\nu_i^{-1}]} \\
&- \sum_{i\in S}\frac{w_i\theta^2\left(\mathrm{E}[\nu_i] - \mathrm{E}[\nu_i^{-1}]^{-1}\right)}{2\gamma^2\sigma_\tau} - \sum_{i\in S}\frac{w_i\mathrm{E}[\nu_i]}{\sigma_\tau} - \frac{b_0}{\sigma_\tau} - \frac{1}{2}\boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\beta}_\tau - \boldsymbol{\beta}_\tau'\Sigma_0^{-1}\boldsymbol{\mu}_0,
\end{aligned}
$$

where C is a is a constant. Observe that the term $\mathrm{E}_{\boldsymbol{\nu}}\left[-\sum_{i\in S}\log(\nu_i/2)\right]$ that comes from the

conditional density $f(\mathbf{y}_u|\boldsymbol{\nu}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta}_\tau, \sigma_\tau)$ was absorbed into the constant C as it only depends on the current parameter estimates and does not influence the subsequent M step. In the third equality in (8), we used

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\nu}}\left[\sum_{i\in S}\frac{w_i\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau-\theta\nu_i\right)^2}{2\gamma^2\sigma_\tau\nu_i}\right] &= \mathsf{E}_{\boldsymbol{\nu}}\left[\sum_{i\in S} w_i\left(\frac{\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)^2}{2\gamma^2\sigma_\tau\nu_i}-\frac{\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)\theta}{\gamma^2\sigma_\tau}+\frac{\theta^2\nu_i}{2\gamma^2\sigma_\tau}\right)\right] \\
&= \sum_{i\in S} w_i\left(\frac{\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)^2}{2\gamma^2\sigma_\tau/\mathsf{E}[\nu_i^{-1}]}-\frac{\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)\theta}{\gamma^2\sigma_\tau}+\frac{\theta^2\mathsf{E}[\nu_i]}{2\gamma^2\sigma_\tau}\right) \\
&= \sum_{i\in S} w_i\left(\frac{\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)^2-2\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)\theta/\mathsf{E}[\nu_i^{-1}]+\theta^2/\mathsf{E}^2[\nu_i^{-1}]}{2\gamma^2\sigma_\tau/\mathsf{E}[\nu_i^{-1}]}\right) \\
&+ \sum_{i\in S} w_i\left(\frac{\theta^2\mathsf{E}[\nu_i]}{2\gamma^2\sigma_\tau}-\frac{\theta^2/\mathsf{E}[\nu_i^{-1}]}{2\gamma^2\sigma_\tau}\right) \\
&= \sum_{i\in S}\frac{w_i\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau-\theta/\mathsf{E}[\nu_i^{-1}]\right)^2}{2\gamma^2\sigma_\tau/\mathsf{E}[\nu_i^{-1}]} \\
&+ \sum_{i\in S}\frac{w_i\theta^2\left(\mathsf{E}[\nu_i]-\mathsf{E}[\nu_i^{-1}]^{-1}\right)}{2\gamma^2\sigma_\tau}.
\end{aligned}$$

Finally, from the objective function in (8), we can write the algorithm as follows:

1. Initialize $\hat{\boldsymbol{\beta}}_\tau$ and $\hat{\sigma}_\tau$.

2. E-step:

   As can be noted by $Q(\boldsymbol{\beta}_\tau, \sigma_\tau)$ in (8), we are left to evaluate $\mathsf{E}[\nu_i]$ and $\mathsf{E}[\nu_i^{-1}]$. Given that we have seen that the full conditional distribution of ν_i follows generalized inverse Gaussian (GIG) distribution, for $a_1 = \sqrt{\frac{w_i\left(y_{ui}-\mathbf{z}_i'\boldsymbol{\beta}_\tau\right)^2}{\gamma^2\sigma_\tau}}$ and $b_1 = \sqrt{\frac{w_i\theta^2}{\gamma^2\sigma_\tau}+\frac{2w_i}{\sigma_\tau}}$, we have that

   $$\mathsf{E}[\nu_i^{-1}] = \frac{a_1}{b_1} \text{ and } \mathsf{E}[\nu_i] = \frac{b_1}{a_1}\times\frac{K_{3/2}(a_1b_1)}{K_{1/2}(a_1b_1)},$$

   where $K_\ell(\cdot)$ is the modified Bessel function of third kind. The equalities follow from the fact that $\mathsf{E}[\nu_i^r] = \left(\frac{a_1}{b_1}\right)^r\frac{K_{1/2+r}(a_1b_1)}{K_{1/2}(a_1b_1)}$ (Karlis 2002) and $K_{1/2}(\cdot) = K_{-1/2}(\cdot)$ (Abramowitz and Stegun 1968).

3. M-step:

To maximize the expected log-posterior, we take the first derivatives of $Q(\boldsymbol{\beta}_\tau, \sigma_\tau)$ with respect to the parameters $\boldsymbol{\beta}_\tau$ and σ_τ, and equal to 0. By doing this, we have that

$$\hat{\boldsymbol{\beta}}_\tau \leftarrow \left(\mathbf{z}'\Sigma_w^{-1}\mathbf{z}' + \Sigma_0^{-1}\right)^{-1}\left(\mathbf{z}'\Sigma_w^{-1}\mathbf{y}^* + \Sigma_0^{-1}\boldsymbol{\beta}_0\right),$$

where $\mathbf{z}$ is a $n \times (q+p)$ matrix, $\Sigma_w = \gamma^2\hat{\sigma}_\tau \text{diag}((w_{i_1}\text{E}[\nu_{i_1}^{-1}])^{-1}, \ldots, (w_{i_n}\text{E}[\nu_{i_n}^{-1}])^{-1})$, and $\mathbf{y}_* = \mathbf{y}_u - \theta/\text{E}[\boldsymbol{\nu}^{-1}]$. And

$$\hat{\sigma}_\tau \leftarrow \left[2b_0 + \sum_{i\in S} w_i \text{E}[\nu_i^{-1}]\left(y_{ui} - \mathbf{z}_i'\boldsymbol{\beta}_\tau - \theta/\text{E}[\nu_i^{-1}]\right)^2 + \sum_{i\in S} w_i\theta^2(\text{E}[\nu_i] - \text{E}[\nu_i^{-1}]^{-1}) + 2\gamma^2\sum_{i\in S} w_i\text{E}[\nu_i]\right] \times \left[(3n + a_0 + 1)\gamma^2\right]^{-1}.$$

4. Repeat steps 2 and 3 until the increment in the objective function is negligible.

4.5 Simulation study

In this section, we propose a model-based simulation study to assess the performance of our Multiple-output Bayesian quantile regression for complex survey data under informative sampling (BWQR-MO) using the ECM algorithm introduced in Section 4.4. For that, we consider experiments like those explored by Nascimento and Gonçalves (to appear).

We follow a data generating process (DGP) similar to Guggisberg (2023) as below:

$$\begin{aligned} &\mathbf{Y} = \tilde{\mathbf{X}} + \begin{bmatrix}0\\ X\end{bmatrix}, \text{ where } \begin{bmatrix}X\\ \tilde{\mathbf{X}}\end{bmatrix} \sim N\left(\begin{bmatrix}\mu_X\\ \boldsymbol{\mu}_{\tilde{\mathbf{X}}}\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma'_{X\tilde{\mathbf{X}}}\\ \Sigma_{X\tilde{\mathbf{X}}} & \Sigma_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}}\end{bmatrix}\right), \\ &\mu_X = 0,\ \boldsymbol{\mu}_{\tilde{\mathbf{X}}} = \begin{bmatrix}0\\ 0\end{bmatrix},\ \Sigma_{XX} = 4,\ \Sigma'_{X\tilde{\mathbf{X}}} = \begin{bmatrix}0\\ 2\end{bmatrix}, \text{ and } \Sigma_{\tilde{\mathbf{X}}\tilde{\mathbf{X}}} = \begin{bmatrix}1 & 1.5\\ 1.5 & 9\end{bmatrix}. \end{aligned} \tag{9}$$

As Guggisberg (2023), we analyze two directions, $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$ and $\mathbf{u} = (0, 1)$, the orthogonal directions $\boldsymbol{\Gamma}_u = (1/\sqrt{2}, -1/\sqrt{2})$ and $\boldsymbol{\Gamma}_u = (1, 0)$, and the depth $\tau = 0.2$. Since the population parameters are not known for the DGP described in (9), we use the values found by Guggisberg (2023) through the numerical minimization of the objective function. The author calculated the expectation of the objective function with a Monte Carlo simulation

sample of 10^6. Table 4.1 presents the true underlying values for the parameters.

Table 4.1: Population parameters.

u	Parameters		
	α_τ	$\beta_{\tau y}$	$\beta_{\tau x}$
$(0, 1)$	-2.02	1.50	1.50
$(1/\sqrt{2}, 1/\sqrt{2})$	-1.16	-1.17	-0.18

4.5.1 Experiment 1

Our first experiment explores three sampling schemes - Poisson, stratified, systematic - and informative and non-informative scenarios. We comprise the non-informative scenario to evaluate how including sampling weights unnecessarily impacts models' performance (Pfeffermann, Moura, and Silva 2006). As a benchmark, we also fit a naive model ignoring the sampling weights design (BQR-MO).

For each scenario, we draw $1,000$ finite populations with population size $N = 10,000$ from the DGP in (9). From the generated populations, samples with size equals to $n = 500$ are selected with inclusion probabilities $\pi_i = nk_i / \sum_{j=1}^{N} k_j$, where $k_i = \{1 + \exp(2.5 - 0.5v_i)\}^{-1}$. For the informative case, $v_i \sim N(1 + 0.25y_{1i} + 0.25y_{2i}, 0.25)$, whereas for the non-informative case, $v_i \sim N(1, 0.25)$.

For each replica in our simulation, the BWQR-MO and the BQR-MO are fitted using the ECM algorithm presented in Section 4.4. Regarding the prior specification, we follow Guggisberg (2023), and specify a non-informative prior for location parameters in which $\boldsymbol{\mu}_0 = \underline{0}_3$ and $\Sigma_0 = 1,000 I_3$, where $\underline{0}_3$ is a 3-dimensional vector of zeros and I_3 is an identity matrix with dimension 3×3. For the scale parameter, we follow Santos and Kneib (2020), and also specify non-informative prior in which $a_0 = b_0 = 0.001$.

Table 4.2 reports the estimators' bias and variance. Considering the informative case and looking at bias first, we observe that for α_τ, the proposed method performs considerably better than the benchmark for the three sampling schemes and both directions **u** under analysis. Considering $\mathbf{u} = (0, 1)$, although the difference is smaller, the proposed method also performs better for $\beta_{\tau y}$ and $\beta_{\tau x}$. For $\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$, we see that the proposed method

performs better for $\beta_{\tau y}$, whereas it is slightly worse for $\beta_{\tau x}$ when we look at the Poisson and the stratified schemes.

Considering the non-informative case, we note no expressive differences regarding bias and variance between the BWQR-MO and the BQR-MO. From that, we can say that the BWQR-MO is effective when applied to complex survey data under informative sampling, and it is not harmful when unnecessarily applied to a non-informative sample.

Focusing on the variance, the BQR-MO has a better general performance, with more apparent differences in the Informative case. This result probably reflects the increase in variance often observed from using the weights in a fully parametric context.

Table 4.2: Bias ($\times 10^2$) and variance ($\times 10^2$) in parenthesis under different sampling designs.

Parameter	Method	Informative			Non-informative		
		Poisson	Stratified	Systematic	Poisson	Stratified	Systematic
				$\mathbf{u} = (0, 1)$			
α_τ	BQR-MO	41.0671	25.3745	41.7847	0.0268	0.0431	0.0186
		(2.6223)	(2.5231)	(2.6809)	(2.2757)	(2.2398)	(2.3231)
	BWQR-MO	0.0231	0.0351	0.0428	0.0353	0.0387	0.0190
		(4.0991)	(3.1080)	(3.8844)	(2.4615)	(2.3586)	(2.4829)
$\beta_{\tau y}$	BQR-MO	0.0514	0.2347	0.0810	0.0002	0.0001	0.0023
		(2.3599)	(2.2137)	(2.3537)	(2.3934)	(2.4648)	(2.3857)
	BWQR-MO	0.0089	0.0007	0.0068	0.0020	0.0013	0.0015
		(4.5719)	(3.2623)	(4.7272)	(2.4949)	(2.5717)	(2.5155)
$\beta_{\tau x}$	BQR-MO	0.0332	0.0756	0.0347	0.0002	0.0012	0.0003
		(0.5929)	(0.5876)	(0.6028)	(0.5958)	(0.5422)	(0.5676)
	BWQR-MO	0.0075	0.0002	0.0015	0.0000	0.0012	0.0000
		(1.1602)	(0.8090)	(1.1368)	(0.6271)	(0.5667)	(0.5762)
				$\mathbf{u} = (1/\sqrt{2}, 1/\sqrt{2})$			
α_τ	BQR-MO	9.5091	5.4925	8.9595	0.0039	0.0050	0.0108
		(0.9662)	(0.9399)	(1.0537)	(0.8501)	(0.8800)	(0.8055)
	BWQR-MO	0.0141	0.0023	0.0051	0.0027	0.0068	0.0087
		(1.3503)	(1.2495)	(1.3290)	(0.8730)	(0.9890)	(0.8387)
$\beta_{\tau y}$	BQR-MO	0.0128	0.0084	0.0072	0.0018	0.0012	0.0015
		(0.2567)	(0.2479)	(0.2690)	(0.2685)	(0.2501)	(0.2592)
	BWQR-MO	0.0023	0.0017	0.0018	0.0015	0.0015	0.0017
		(0.4788)	(0.3409)	(0.4722)	(0.2816)	(0.2601)	(0.2702)
$\beta_{\tau x}$	BQR-MO	0.0000	0.0000	0.0001	0.0047	0.0017	0.0019
		(0.4748)	(0.4860)	(0.4859)	(0.4955)	(0.4692)	(0.4976)
	BWQR-MO	0.0002	0.0016	0.0000	0.0047	0.0024	0.0023
		(0.8772)	(0.6992)	(0.7945)	(0.5123)	(0.4920)	(0.5271)

4.5.2 Experiment 2

The second experiment encompasses a prior sensitivity analysis. We aim to assess how specifying different prior distributions influences the proposed model estimates. For that, we propose

a similar study to Zhao et al. (2020) and consider n-dependent priors (Narisetty and He 2014; Yang and He 2012) as they can naturally arise in practice in cases where a previous survey or a pilot survey is available, for example. However, instead of only centering the prior distributions on the parameter's true value like Zhao et al. (2020), we also simulate the scenario in which a pilot survey is available. It enables a more realistic experiment since centering on the actual value is not feasible and cannot be implemented in practical analysis, as pointed out by Zhao et al. (2020).

Therefore, our study consists of drawing $1,000$ finite populations with population size $N = 10,000$ from the DGP in (9). For each generated population, we extract a pilot sample with a size equal to $n_0 = 150$ and a sample with a size equal to $n = 500$. The samples are selected according to a Poisson sampling process with inclusion probabilities $\pi_i = n_0 k_i / \sum_{j=1}^{N} k_j$ and $\pi_i = n k_i / \sum_{j=1}^{N} k_j$, respectively, where $k_i = \{1 + \exp(2.5 - 0.5 v_i)\}^{-1}$ and $v_i \sim N(1 + 0.25 y_{1i} + 0.25 y_{2i}, 0.25)$.

Since σ_τ is a nuisance scale parameter, we are interested in investigating the prior sensitivity related to the regression coefficients in this study. Due to it, σ_τ is fixed at 1 (Guggisberg 2023). Regarding the prior specification, we take four examples: a non-informative prior similar to the one specified in the previous experiment, i. e., $\boldsymbol{\mu}_0 = \underline{0}_3$ and $\Sigma_0 = 1,000 I_3$ (Prior I); a n-dependent prior where $\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\beta}}_\tau$ and $\Sigma_0 = n_0^{-1/2} I_3$ (Prior II); a n-dependent prior where $\boldsymbol{\mu}_0 = \tilde{\boldsymbol{\beta}}_\tau$ and $\Sigma_0 = n_0^{-1/4} I_3$ (Prior III); and a n-dependent prior where $\boldsymbol{\mu}_0$ is equals to the actual values and $\Sigma_0 = n_0^{-1/2} I_3$ (Prior IV). The point estimates computed from the pilot surveys are denoted by $\tilde{\boldsymbol{\beta}}_\tau$. Prior II, Prior III, and Prior IV yield valid and highly informative priors, but Prior II and Prior IV are even more informative as they are more concentrated around the mean.

Table 4.3 reports the estimators' bias and variance for the prior specifications under analysis. Prior IV and Prior II present lower variances as their specifications correspond to a more concentrated distribution around the mean. On the other hand, Prior I corresponds to a higher variance. However, the differences are not significant.

Except for α_τ regarding Prior IV, the bias are not markedly different among the priors under

analysis. Interestingly, the performances of the more realistic priors (Prior I, Prior II, and Prior III) are not significantly worse than an ideal case where the mean is centered in the real value and the distribution is highly concentrated around it. Then, it seems that the prior specification does not considerably influence estimators' properties like bias and variance.

Table 4.3: Bias ($\times 10^2$) and variance ($\times 10^2$) in parenthesis for the prior specifications under analysis.

u	Parameter	Prior I	Prior II	Prior III	Prior IV
(-1,0)	α_τ	0.0210 (4.0948)	0.0203 (3.5604)	0.0205 (3.8881)	0.0125 (3.3528)
	$\beta_{\tau y}$	0.0054 (4.3788)	0.0048 (3.6965)	0.0039 (4.1727)	0.0030 (3.6449)
	$\beta_{\tau x}$	0.0001 (1.1425)	0.0005 (1.1152)	0.0002 (1.1296)	0.0001 (1.0828)
$\left(\frac{3}{\sqrt{10}}, \frac{1}{\sqrt{10}}\right)$	α_τ	0.0058 (1.3883)	0.0067 (1.3276)	0.0051 (1.2991)	0.0058 (1.2752)
	$\beta_{\tau y}$	0.0063 (0.5107)	0.0049 (0.4738)	0.0062 (0.4967)	0.0057 (0.4680)
	$\beta_{\tau x}$	0.0041 (0.8415)	0.0030 (0.7650)	0.0040 (0.8123)	0.0039 (0.7572)

4.6 Real-data-based simulation study

Complex surveys are commonly employed to collect data and information in the educational area. A good example is the test implemented by the Programme for International Student Assessment (PISA) from the Organisation for Economic Co-operation and Development (OECD). PISA test measures 15-year-olds' competence in applying their reading, mathematics, and science knowledge and skills to solve real-life challenges. It is implemented in different countries worldwide, and its default sampling design is a two-stage stratified sample.

In Brazil, the National Institute of Education Research (INEP) for the Brazilian Ministry of Education assesses student performance and collects data through large-scale standardized proficiency tests like Prova Brasil. Prova Brasil is a crucial tool for monitoring the educational quality of the Brazilian public educational system. The proficiency test encompasses Mathematics and Portuguese exams and is applied to students in the fifth and ninth years of public

elementary schools.

In this chapter, similarly to Silva and Moura (2022), we analyze Mathematics (Y_1) and Portuguese (Y_2) proficiency scores, but in our case, we use the edition from 2011 instead of the edition from 2009. The scores come from applying item response theory (IRT) models to the test results. Quantile regression models like ours can be particularly appealing in the educational area as they can characterize differences over the distribution of a data set. For example, Costanzo and Desimoni (2015) addressed issues of educational inequality, and Giambona and Porcu (2015) investigated background determinants of reading achievements.

Prova Brasil is subject to non-response provoked by school evasion or non-attendance on the day of the exam. Evidence from previous editions indicates that students with low achievement are less likely to take part in the exam, suggesting an informative mechanism in the observed data. From this consideration, we propose a design-based simulation study where we fix the $N = 11,004$ observations from students in the ninth year from public elementary schools located in Campinas, a municipality in São Paulo State, as our finite population. As covariates, We consider an indicator variable for nonwhite students (black, mixed, and indigenous) and an indicator variable for students lagging behind. The first is taken as a proxy for socioeconomic inequities (Gradín 2009), and the second denotes students who have fallen behind other students in their cohorts.

From the fixed finite population, we draw $1,000$ samples with expected size equal $n = 500$ through a Poisson sampling process, in which the inclusion probabilities are $\pi_i = nk_i / \sum_{j=1}^{N} k_j$, where $k_i = \{1 + \exp(2.5 - 0.5v_i)\}^{-1}$ and $v_i \sim N(1 + 0.25y_{1i} + 0.25y_{2i}, 0.25)$. We then fit the BQR-MO and the BWQR-MO considering $\mathbf{u} = (3/\sqrt{10}, 1/\sqrt{10})$, $\mathbf{\Gamma}_u = (1/\sqrt{10}, -3/\sqrt{10})$, $\tau \in \{0.10, 0.25, 0.50, 0.75, 0.90\}$, and the same non-informative priors from Subsection 4.5.1. The BQR-MO and the BWQR-MO estimates are compared to a best-case scenario. As a best-case scenario, we mean the estimates from the BQR-MO considering the fully observed population. Figure 4.1 summarizes the results. The coefficients $\beta_{1\tau x}$ and $\beta_{2\tau x}$ refer to the indicators of nonwhite students and students lagging behind, respectively.

From Figure 4.1, we can observe similar results to those obtained in Subsection 4.5.1 as we see

that the major differences in terms of bias occur for α_τ. Here, we note that it is a more general result since we analyze a more comprehensive number of quantiles. In Subsection 4.5.1, the BQR-MO presented lower variances. However, from the range lengths, we notice that it seems true mainly in the quantiles where the observations have lower inclusion probabilities ($\tau \in \{0.10, 0.25, 0.75\}$).

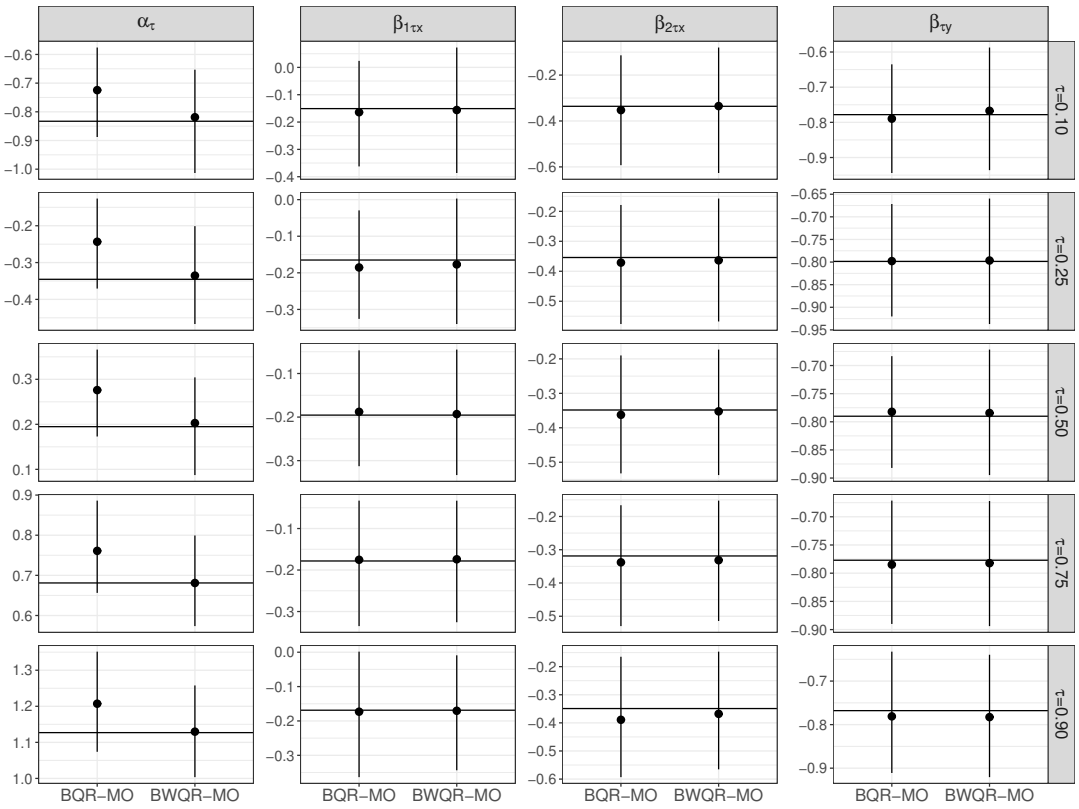

Figure 4.1: Comparison among different quantiles. The ranges summarize the point estimates. The dots represent the median, and the lines represent ranges from quantile 5% to quantile 95%. The horizontal solid lines represent the best-case scenario.

4.7 Final remarks

In this chapter, we introduced a multiple-output Bayesian quantile regression model that accommodates complex survey data under informative sampling. For that, we extended the model proposed by Guggisberg (2023) using the ideas of Chapter 2 to construe a framework that can characterize population parameters through a set of observed data drawn from an informative sampling design. Moreover, due to the computational cost that applying the usual MCMC algorithms in our context may imply, we adapted the EM algorithm described in Zhao

and Lian (2016) to the developed methodology model. By doing this, we provided a tool that can be applied by practitioners from different backgrounds with a low computational cost.

To evaluate the performance of our methods, we proposed a model-based simulation study where the finite populations were drawn following a data-generating process similar to the one applied by Guggisberg (2023). Our study was divided into two experiments. First, the samples were generated under different sampling schemes (Poisson, stratified, and systematic) and informative and non-informative designs. As a result, we observed that the proposed method is advantageous compared to a benchmark that ignores the sampling weights for the informative scenario regardless of the sampling scheme. Considering the non-informative case, we saw that the proposed model performed close to the benchmark. In the second experiment, we developed a prior sensitivity analysis that compared a non-informative prior with two n-dependent priors where the distributions were centered on estimates from artificial pilot samples and a n-dependent prior where the distribution was centered on the true underlying values. The findings pointed to no discrepant results among the specifications.

In addition, we also proposed a design-based simulation study using data from Prova Brasil 2011, fixing the students from the municipality of Campinas who took part in the exam as a finite population. This study was focused on analyzing a more comprehensive range of quantiles. As in the model-based simulation study, we noticed that the main harm in not including the survey design in the estimation process relies on the intercept.

5 Conclusion and future works

This dissertation consisted of three main chapters. In the first, Chapter 2, we presented three alternatives for the Bayesian quantile regression analysis of continuous data under informative sampling. From the survey-weighted estimator (Chen and Zhao 2019; Geraci 2016), we extended the well-known quantile regression model based on the asymmetric Laplace distribution to complex survey data under informative sampling following a similar argument of Yu and Moyeed (2001). In addition, building on the estimating equations (Zhao et al. 2020), we extended the quantile regression model using a score likelihood (Wu and Narisetty 2021) to include sampling weights and derive a quantile regression as a particular case of the approximate approach proposed by Wang, Kim, and Yang (2018). For the first, we implemented a simple Gibbs sampling algorithm following the ideas of Kozumi and Kobayashi (2011). For the second, we employed an adaptative Metropolis-Hastings algorithm. The chapter consisted of a preliminary version of the paper Nascimento and Gonçalves (to appear).

In Chapter 3, we proposed two different methods for analyzing bounded count data under informative sampling. From the side of the bounded count data, we applied the ideas of Machado and Silva (2005) and Lee and Neocleous (2010) to deal with quantiles for counts, adapting them by employing the methodology proposed by Bottai, Cai, and McKeown (2010). From the side of complex survey data under informative sampling, we introduced two different methods based on the asymmetric Laplace distribution (Yu and Moyeed 2001); one is an extension of the approach introduced in Chapter 2, and the other is based on the pseudo posterior distribution (Savitsky and Toth 2016).

Lastly, in Chapter 4, we developed a framework for analyzing multiple-output under informative

sampling. For that purpose, we extended Guggisberg (2023) following the ideas of Chapter 2 to propose a method that relies on the asymmetric Laplace distribution. This distribution is vastly applied for single-output Bayesian quantile regression models Kozumi and Kobayashi (2011); Yu and Moyeed (2001). From the location-scale mixture representation of this distribution, we introduced an efficient Expectation–Maximization algorithm that propitiates substantial computational savings compared to the commonly used Markov Chain Monte Carlo algorithm.

A concern that possibly emerges is the asymptotic validity of the posterior credible intervals. For example, in the case of the usual Bayesian quantile regression with asymmetric Laplace likelihood, Sriram (2015) and Yang, Wang, and He (2016) studied the stationary distribution for the resulting posterior and proposed a simple adjustment to the covariance matrix of the posterior chain. Wu and Narisetty (2021) also investigated the resulting posterior stationary distribution for the model based on the score likelihood.

The main challenge is that the standard assumptions in the quantile regression literature (He and Shao 1996; Koenker 2005) used by Sriram (2015), Yang, Wang, and He (2016) and Wu and Narisetty (2021) are not straightforwardly applied when the survey weights are included in the models. The regularity conditions concerning the conditional quantile function linearity and the conditional distribution and density function continuities could remain the same. However, the conditions concerning the distribution of the covariates and the existence of asymptotic limits need to be adapted. We could apply the ideas of Huang, Xu, and Tashnev (2015), but for that, we would need to consider given survey weights or $w_i = w_i(\mathbf{x}_i, \tau)$ as the authors made these assumptions to obtain their theoretical results.

Since quantile regression models and the analysis of complex survey data under informative sampling are of great interest to practitioners from different areas and different backgrounds, and one of our primary purposes in this dissertation was to provide effective and easy-to-implement methodological tools, we intend to develop a package containing our methods as a future work. Possibly, including new alternatives for analyzing ordinal data by extending the work of Rahman (2016).

Another idea we intend to pursue is to allow the Expectation-Maximization approach pre-

sented in Chapter 4 for account to noncrossing quantile. Our objective is to propose a fast computation alternative to the methodology presented by Santos and Kneib (2020). In summary, we aim to apply the solution from Bondell, Reich, and Wang (2010) in the maximization step of the coefficients. Bondell, Reich, and Wang (2010) introduced a simple constrained version of quantile regression to avoid the crossing problem that can be incorporated into our algorithm.

Bibliography

Abramowitz, Milton, and Irene A Stegun. 1968. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Vol. 55. US Government printing office.

Asparouhov, Tihomir. 2006. "General Multi-Level Modeling with Sampling Weights." *Communications in Statistics - Theory and Methods* 35 (3): 439–460.

Azzalini, Adelchi. 1985. "A Class of Distributions Which Includes the Normal Ones." *Scandinavian Journal of Statistics* 12 (2): 171–178.

Barata, Raquel, Raquel Prado, and Bruno Sansó. 2022. "Fast inference for time-varying quantiles via flexible dynamic models with application to the characterization of atmospheric rivers." *Annals of Applied Statistics* 16 (1): 247–271.

Beaumont, Jean-François. 2008. "A new approach to weighting and inference in sample surveys." *Biometrika* 95 (3): 539–553.

Benoit, Dries F., and Dirk Van den Poel. 2017. "bayesQR: A Bayesian Approach to Quantile Regression." *Journal of Statistical Software* 76 (7): 1–32.

Bhattacharya, Indrabati, and Subhashis Ghosal. 2021. "Bayesian multivariate quantile regression using Dependent Dirichlet Process prior." *Journal of Multivariate Analysis* 185: 104763.

Binder, David A. 1983. "On the Variances of Asymptotically Normal Estimators from Complex Surveys." *International Statistical Review* 51 (3): 279–292.

Binelli, Chiara, and Naercio Menezes-Filho. 2019. "Why Brazil fell behind in college education?" *Economics of Education Review* 72: 80–106.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* 112 (518): 859–877.

Bondell, Howard D., Brian J. Reich, and Huixia Wang. 2010. "Noncrossing quantile regression curve estimation." *Biometrika* 97 (4): 825–838.

Booth, Alison L., and Hiau Joo Kee. 2009. "Intergenerational Transmission of Fertility Patterns." *Oxford Bulletin of Economics and Statistics* 71 (2): 183–208.

Botelho, Fernando, Ricardo A. Madeira, and Marcos A. Rangel. 2015. "Racial Discrimination in Grading: Evidence from Brazil." *American Economic Journal: Applied Economics* 7 (4): 37–52.

Bottai, Matteo, Bo Cai, and Robert E. McKeown. 2010. "Logistic quantile regression for bounded outcomes." *Statistics in Medicine* 29 (2): 309–317.

Carlier, Guillaume, Victor Chernozhukov, and Alfred Galichon. 2016. "Vector quantile regression: An optimal transport approach." *Annals of Statistics* 44 (3): 1165–1192.

Chambers, R. L., and Chris J. Skinner. 2003. *Analysis of Survey Data*. 1st ed., Wiley Series in Survey Methodology,. New Jersey: Wiley.

Chaudhuri, Probal. 1996. "On a Geometric Notion of Quantiles for Multivariate Data." *Journal of the American Statistical Association* 91 (434): 862–872.

Chen, Sixia, and Jae Kwang Kim. 2014. "Population empirical likelihood for nonparametric inference in survey sampling." *Statistica Sinica* 24 (1): 335–355.

Chen, Sixia, and Yan Daniel Zhao. 2019. "Quantile Regression Analysis of Survey Data Under Informative Sampling." *Journal of Survey Statistics and Methodology* 7 (2): 157–174.

Congdon, Peter. 2017. "Quantile Regression for Area Disease Counts: Bayesian Estimation using Generalized Poisson Regression." *International journal of statistics in medical research* 6: 92–103.

Costanzo, Antonella, and Marta Desimoni. 2015. "Beyond the mean estimate: a quantile regression analysis of inequalities in educational outcomes using INVALSI survey data." *Large-scale Assessments in Education* 5: 14.

Dong, Qi, Michael R. Elliott, and Trivellore E. Raghunathan. 2014. "A nonparametric method to generate synthetic populations to adjust for complex sampling design features." *Survey Methodology* 40 (1): 29–46.

Eide, Eric R., and Mark H. Showalter. 1999. "Factors Affecting the Transmission of Earnings across Generations: A Quantile Regression Approach." *Journal of Human Resources* 34 (2): 253–267.

Ferreira, Francisco H. G., Sergio P. Firpo, and Julián Messina. 2022. "Labor Market Experience and Falling Earnings Inequality in Brazil: 1995–2012." *World Bank Economic Review* 36 (1): 37–67.

Frumento, Paolo, and Matteo Bottai. 2016. "Parametric modeling of quantile regression coefficient functions." *Biometrics* 72 (1): 74–84.

Frumento, Paolo, and Nicola Salvati. 2021. "Parametric modeling of quantile regression coefficient functions with count data." *Statistical Methods & Applications* 30: 1237–1258.

Fuller, Wayne A. 2009. *Sampling Statistics*. 1st ed. New Jersey: John Wiley & Sons, Ltd.

Gamerman, Dani, and Hedibert F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. London: Chapman & Hall.

Geraci, Marco. 2016. "Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants." *Statistical Methods in Medical Research* 25 (4): 1393–1421.

Giambona, Francesca, and Mariano Porcu. 2015. "Student background determinants of reading achievement in Italy. A quantile regression analysis." *International Journal of Educational Development* 44: 95–107.

Godambe, Vidyadhar P., and Mary E. Thompson. 1986. "Parameters of Superpopulation and Survey Population: Their Relationships and Estimation." *International Statistical Review* 54 (2): 127–138.

Gonçalves, Kelly C. M., Hélio S. Migon, and Leonardo S. Bastos. 2020. "Dynamic Quantile Linear Models: A Bayesian Approach." *Bayesian Analysis* 15 (2): 335–362.

Gradín, Carlos. 2009. "Why is Poverty So High Among Afro-Brazilians? A Decomposition Analysis of the Racial Poverty Gap." *Journal of Development Studies* 45 (9): 1426–1452.

Grilli, Leonardo, Carla Rampichini, and Roberta Varriale. 2016. "Statistical modelling of gained university credits to evaluate the role of pre-enrolment assessment tests: An approach based on quantile regression for counts." *Statistical Modelling* 16 (1): 47–66.

Guggisberg, Michael. 2023. "A Bayesian Approach to Multiple-Output Quantile Regression." *Journal of the American Statistical Association* 118 (544): 2736–2745.

Hallin, Marc, Zudi Lu, Davy Paindaveine, and Miroslav Šiman. 2015. "Local bilinear multiple-output quantile/depth regression." *Bernoulli* 21 (3): 1435–1466.

Hallin, Marc, Davy Paindaveine, and Miroslav Siman. 2010. "Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth." *Annals of Statistics* 38 (2): 635–669.

Hammer, Hugo L., Anis Yazidi, and Håvard Rue. 2022. "Estimating Tukey depth using incremental quantile estimators." *Pattern Recognition* 122: 108339.

He, Xuming, and Qi-Man Shao. 1996. "A general Bahadur representation of M-estimators and its application to linear regression with nonstochastic designs." *Annals of Statistics* 24 (6): 2608–2630.

Huang, Mei Ling, Xiaojian Xu, and Dmitry Tashnev. 2015. "A weighted linear quantile regression." *Journal of Statistical Computation and Simulation* 85 (13): 2596–2618.

Karlis, Dimitris. 2002. "An EM type algorithm for maximum likelihood estimation of the normal–inverse Gaussian distribution." *Statistics & Probability Letters* 57 (1): 43–52.

Kim, Jae Kwang. 2009. "Calibration estimation using empirical likelihood in survey sampling." *Statistica Sinica* 19 (1): 145–157.

Kim, Jae Kwang, and Chris J. Skinner. 2013. "Weighting in survey analysis under informative sampling." *Biometrika* 100 (2): 385–398.

Koenker, Roger. 2005. *Quantile Regression*. 1st ed. Cambridge: Cambridge University Press.

Koenker, Roger, and Gilbert Bassett Jr. 1978. "Regression Quantiles." *Econometrica* 46 (1): 33–50.

Kong, Linglong, and Ivan Mizera. 2012. "Quantile tomography: using quantiles with multivariate data." *Statistica Sinica* 22 (4): 1589–1610.

Kotz, Samuel, Tomaz J. Kozubowski, and Krzysztof Podgórski. 2001. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. 1st ed. Boston: Birkhauser.

Kozumi, Hideo, and Genya Kobayashi. 2011. "Gibbs sampling methods for Bayesian quantile regression." *Journal of Statistical Computation and Simulation* 81 (11): 1565–1578.

Kunihama, T., A.H. Herring, C.T. Halpern, and D.B. Dunson. 2016. "Nonparametric Bayes modeling with sample survey weights." *Statistics & Probability Letters* 113: 41–48.

Lee, Duncan, and Tereza Neocleous. 2010. "Bayesian quantile regression for count data with application to environmental epidemiology." *Applied Statistics* 59 (5): 905–920.

Lee, Eun Ryung, Hohsuk Noh, and Byeong U. Park. 2014. "Model Selection via Bayesian Information Criterion for Quantile Regression Models." *Journal of the American Statistical Association* 109 (505): 216–229.

Leon-Novelo, Luis G., and Terrance D. Savitsky. 2019. "Fully Bayesian estimation under informative sampling." *Electronic Journal of Statistics* 13 (1): 1608–1645.

Li, Yan, Barry I. Graubard, and Edward L. Korn. 2010. "Application of Nonparametric Quantile Regression to Body Mass Index Percentile Curves from Survey Data." *Statistics in Medicine* 29 (5): 558–572.

Little, Roderick J. 2004. "To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99 (466): 546–556.

Liu, Zuyun, Pei-Lun Kuo, Steve Horvath, Eileen Crimmins, Luigi Ferrucci, and Morgan Levine. 2018. "A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: A cohort study." *PLoS Medicine* 15 (12): e1002718.

Lum, Kristian, and Alan E. Gelfand. 2012. "Spatial Quantile Multiple Regression Using the Asymmetric Laplace Process." *Bayesian Analysis* 7 (2): 235–258.

Machado, José A. F, and J. M. C. Santos Silva. 2005. "Quantiles for Counts." *Journal of the American Statistical Association* 100 (472): 1226–1237.

Magee, Lonnie. 1998. "Improving survey-weighted least squares regression." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 60 (1): 115–126.

Marteleto, Leticia J. 2012. "Educational Inequality by Race in Brazil, 1982–2007: Structural Changes and Shifts in Racial Classification." *Demography* 49 (1): 337–358.

Meng, Xiao-Li, and Donald B. Rubin. 1993. "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework." *Biometrika* 80 (2): 267–278.

Miranda, Alfonso. 2008. "Planned fertility and family background: a quantile regression for counts analysis." *Journal of Population Economics* 21: 67–81.

Moreira, Sara, and Pedro Pita Barros. 2010. "Double health insurance coverage and health care utilisation: evidence from quantile regression." *Health Economics* 19 (9): 1075–1092.

Mullahy, John. 2023. *Analyzing Bounded Count Data*. Working Paper 31814. National Bureau of Economic Research.

Narisetty, Naveen Naidu, and Xuming He. 2014. "Bayesian variable selection with shrinking and diffusing priors." *Annals of Statistics* 42 (2): 789–817.

Nascimento, Marcus L., and Kelly M. Gonçalves. to appear. "Bayesian quantile regression models for complex survey data under informative sampling." *Journal of Survey Statistics and Methodology* .

Neelon, Brian, Fan Li, Lane F. Burgette, and Sara E. Benjamin Neelon. 2015. "A spatiotemporal quantile regression model for emergency department expenditures." *Statistics in Medicine* 34 (17): 2559–2575.

Olson, Zachary, Rachel Gardner Clark, and Sarah Anne Reynolds. 2019. "Can a conditional cash transfer reduce teen fertility? The case of Brazil's Bolsa Familia." *Journal of Health Economics* 63: 128–144.

Padellini, Tullia, and Haavard Rue. 2019. "Model-aware Quantile Regression for Discrete Data." .

Parker, Paul A., Scott H. Holan, and Ryan Janicki. 2020. "Conjugate Bayesian unit-level modelling of count data under informative sampling designs." *Stat* 9 (1): e267.

Pearson, E. S. 1950. "On Questions Raised by the Combination of Tests Based on Discontinuous Distributions." *Biometrika* 37 (3–4): 383–398.

Pfeffermann, D., C. J. Skinner, D. J. Holmes, H. Goldstein, and J. Rasbash. 1998. "Weighting for Unequal Selection Probabilities in Multilevel Models." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 60 (1): 23–40.

Pfeffermann, Danny. 2011. "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?" *Survey Methodology* 37 (2): 115–136.

Pfeffermann, Danny, Abba M. Krieger, and Yosef Rinott. 1998. "Parametric distributions of complex survey data under informative probability sampling." *Statistica Sinica* 8: 1087–1114.

Pfeffermann, Danny, and Michail Sverchkov. 1999. "Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data." *Sankhya: The Indian Journal of Statistics, Series B* 61 (1): 166–186.

Pfeffermann, Danny, and Michail Sverchkov. 2003. "Fitting Generalized Linear Models under Informative Sampling." In *Analysis of Survey Data*, edited by Raymond Chambers and Chris J. Skinner, Chap. 12, 175–195. New Jersey: John Wiley & Sons, Ltd.

Pfeffermann, Danny, and Michail Sverchkov. 2009. "Inference under Informative Sampling." In *Handbook of Statistics*, edited by Danny Pfeffermann and Calyampudi Radhakrishna Rao, Chap. 29, 455–487. Amsterdam: Elsevier.

Pfeffermann, Danny Pfeffermann, Fernando Antonio da Silva Moura, and Pedro Luis do Nascimento Silva. 2006. "Multi-level modelling under informative sampling." *Biometrika* 93 (4): 943–959.

Qin, Xiao, and Perla E. Reyes. 2011. "Conditional Quantile Analysis for Crash Count Data." *Journal of Transportation Engineering* 137 (9): 601–607.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Rahman, Mohammad Arshad. 2016. "Bayesian Quantile Regression for Ordinal Models." *Bayesian Analysis* 11 (1): 1–24.

Rao, J. N. K., and Changbao Wu. 2010. "Bayesian pseudo-empirical-likelihood intervals for complex surveys." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 72 (4): 533–544.

Robert, Christian P., and George Casella. 2013. *Monte Carlo Statistical Methods*. 2nd ed. New York: Springer.

Roberts, Gareth O., Andrew Gelman, and Walter R. Gilks. 1997. "Weak convergence and optimal scaling of random walk Metropolis algorithms." *Annals of Applied Probability* 7 (1): 110–120.

Santos, Bruno, and Thomas Kneib. 2020. "Noncrossing structured additive multiple-output Bayesian quantile regression models." *Statistics and Computing* 30: 855–869.

Savitsky, Terrance D., and Daniell Toth. 2016. "Bayesian estimation under informative sampling." *Electronic Journal of Statistics* 10 (1): 1677–1708.

Scott, Alastair J., and Chris J. Wild. 2011. "Fitting regression models with response-biased samples." *Canadian Journal of Statistics* 39 (3): 519–536.

Serfling, Robert. 2002. "Quantile functions for multivariate analysis: approaches and applications." *Statistica Neerlandica* 56 (2): 214–232.

Shaby, Benjamin, and Martin T. Wells. 2010. *Exploring an Adaptive Metropolis Algorithm*. Technical Report.

Si, Yajuan, Natesh S. Pillai, and Andrew Gelman. 2015. "Bayesian Nonparametric Weighted Sampling Inference." *Bayesian Analysis* 10 (3): 605–625.

Silva, Pedro Luis do N., and Fernando Antônio da S. Moura. 2022. "Fitting Multivariate Multilevel Models under Informative Sampling." *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 185 (4): 1663–1678.

Small, Christopher G. 1990. "A Survey of Multidimensional Medians." *International Statistical Review* 58 (3): 263–277.

Sriram, Karthik. 2015. "A sandwich likelihood correction for Bayesian quantile regression based on the misspecified asymmetric Laplace density." *Statistics & Probability Letters* 107: 18–26.

Stevens, W. L. 1950. "Fiducial Limits of the Parameter of a Discontinuous Distribution." *Biometrika* 37 (1–2): 117–129.

Tanner, Martin A., and Wing Hung Wong. 1987. "The Calculation of Posterior Distributions by Data Augmentation." *Journal of the American Statistical Association* 82 (398): 528–540.

Tillé, Yves, and Matthieu Wilhelm. 2017. "Probability Sampling Designs: Principles for Choice of Design and Balancing." *Statistical Science* 32 (2): 176–189.

Tzavidis, Nikos, M. Giovanna Ranalli, Nicola Salvati, Emanuela Dreassi, and Ray Chambers. 2015. "Robust small area prediction for counts." *Statistical Methods in Medical Research* 24 (3): 373–395.

Valliant, Richard, Jill A. Dever, and Frauke Kreuter. 2018. *Practical Tools for Designing and Weighting Survey Samples*. 2nd ed. Cham: Springer.

Veiga, Alinne, Peter W. F. Smith, and James J. Brown. 2014. "The Use of Sample Weights in Multivariate Multilevel Models with an Application to Income Data Collected by Using a Rotating Panel Survey." *Journal of the Royal Statistical Society, Series C (Applied Statistics)* 63 (1): 65–84.

Villarini, Gabriele, James A. Smith, Mary Lynn Baeck, Renato Vitolo, David B. Stephenson, and Witold F.Krajewski. 2011. "On the frequency of heavy rainfall for the Midwest of the United States." *Journal of Hydrology* 400 (1–2): 103–120.

Waldmann, Elisabeth, and Thomas Kneib. 2015. "Variational approximations in geoadditive latent Gaussian regression: mean and quantile regression." *Statistics and Computing* 25: 1247–1263.

Wand, Matthew P., John T. Ormerod, Simone A. Padoan, and Rudolf Frühwirth. 2011. "Mean Field Variational Bayes for Elaborate Distributions." *Bayesian Analysis* 6 (4): 847–900.

Wang, Zhonglei, Jae Kwang Kim, and Shu Yang. 2018. "Approximate Bayesian inference under informative sampling." *Biometrika* 105 (1): 91–102.

Wei, Ying. 2008. "An Approach to Multivariate Covariate-Dependent Quantile Contours With Application to Bivariate Conditional Growth Charts." *Journal of the American Statistical Association* 103 (481): 397–409.

Winkelmann, Rainer. 2006. "Reforming health care: Evidence from quantile regressions for counts." *Journal of Health Economics* 25 (1): 131–145.

Wu, Teng, and Naveen N. Narisetty. 2021. "Bayesian Multiple Quantile Regression for Linear Models Using a Score Likelihood." *Bayesian Analysis* 16 (3): 875–903.

Yang, Yunwen, and Xuming He. 2012. "Bayesian empirical likelihood for quantile regression." *Annals of Statistics* 40 (2): 1102–1131.

Yang, Yunwen, Huixia Judy Wang, and Xuming He. 2016. "Posterior Inference in Bayesian Quantile Regression with Asymmetric Laplace Likelihood." *International Statistical Review* 84 (3): 327–344.

Yu, Keming, and Chris Jones. 1998. "Local Linear Quantile Regression." *Journal of the American Statistical Association* 93 (441): 228–237.

Yu, Keming, and Rana A. Moyeed. 2001. "Bayesian quantile regression." *Statistics & Probability Letters* 55 (4): 437–447.

Yuan, Ying, and Guosheng Yin. 2010. "Bayesian Quantile Regression for Longitudinal Studies with Nonignorable Missing Data." *Biometrics* 66 (1): 105–114.

Zhao, Kaifeng, and Heng Lian. 2016. "The Expectation–Maximization approach for Bayesian quantile regression." *Computational Statistics and Data Analysis* 96: 1–11.

Zhao, Puying, Malay Ghosh, J. N. K. Rao, and Changbao Wu. 2020. "Bayesian empirical likelihood inference with complex survey data." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 82 (1): 155–174.