Computational Statistics (2024) 39:583–603
https://doi.org/10.1007/s00180-022-01308-2

ORIGINAL PAPER

Quantile regression with interval-censored data in questionnaire-based studies

Angel G. Angelov[1,2] · **Magnus Ekström**[1,3] · **Klarizze Puzon**[4] · **Agustin Arcenas**[5] · **Bengt Kriström**[6]

Received: 6 September 2021 / Accepted: 25 November 2022 / Published online: 8 December 2022

Abstract
Interval-censored data can arise in questionnaire-based studies when the respondent gives an answer in the form of an interval without having pre-specified ranges. Such data are called self-selected interval data. In this case, the assumption of independent censoring is not fulfilled, and therefore the ordinary methods for interval-censored data are not suitable. This paper explores a quantile regression model for self-selected interval data and suggests an estimator based on estimating equations. The consistency of the estimator is shown. Bootstrap procedures for constructing confidence intervals are considered. A simulation study indicates satisfactory performance of the proposed methods. An application to data concerning price estimates is presented.

Keywords Interval-censored data · Dependent censoring · Self-selected interval · Quantile regression · Estimating equation

✉ Angel G. Angelov
agangelov@gmail.com

✉ Magnus Ekström
magnus.ekstrom@umu.se

1 Department of Statistics,Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

2 Department of Probability, Operations Research and Statistics, Faculty of Mathematics and Informatics, Sofia University St. Kliment Ohridski, Sofia, Bulgaria

3 Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden

4 United Nations University World Institute for Development Economics Research, Helsinki, Finland

5 School of Economics, University of the Philippines, Diliman, Quezon City, Philippines

6 Department of Forest Economics, Swedish University of Agricultural Sciences, Umeå, Sweden

1 Introduction

Quantile regression is a flexible approach to analyzing relationships between a response variable and a set of covariates. While the classical least-squares regression methods capture the central tendency of the data, quantile regression methods allow estimating the full range of conditional quantile functions and thus can provide a more complete analysis. Other attractive properties of quantile regression are equivariance to monotone transformations, robustness to outlying observations, and flexibility to distributional assumptions (Koenker 2005).

In many studies, the response variable of interest is observed to lie within an interval instead of being observed exactly. Such observations are called interval-censored and they often arise when the variable of interest is the time to some event (Kalbfleisch and Prentice 2002; Sun 2006; Bogaerts et al. 2017). Interval-censored data may also occur in questionnaire-based studies when the respondent is requested to give an answer in the form of an interval without having a list of ranges to choose from. This type of data is referred to as self-selected interval data (Belyaev and Kriström 2010, 2012, 2015). Similar question formats have been explored by Press and Tanur (2004a, 2004b), Håkansson (2008), and Mahieu et al. (2017). Such formats are appropriate for asking questions which are hard to answer with an exact amount and for sensitive questions because they allow partial information to be elicited from respondents who are unable or unwilling to provide exact values.

Estimation procedures for quantile regression with interval-censored data have been suggested by Kim et al. (2010), Shen (2013), Zhou et al. (2017), Li et al. (2020), and Frumento (2022). These methods rely on the assumption of independent censoring, i.e., the observation process that generates the censoring is independent of the variable of interest, conditional on the covariates included in the model (Sun 2006). However, for self-selected interval data this is not a reasonable assumption because the respondent is the one who chooses the interval. Not accounting for the dependent censoring in self-selected interval data can lead to bias in the estimation (Angelov and Ekström 2017, 2019).

Building upon the ideas of McKeague et al. (2001), Shen (2013), and Angelov and Ekström (2017), we suggest an estimator for quantile regression where the response variable is of self-selected interval data type and the covariates are discrete. In questionnaire-based studies, most often the covariates are discrete, such as gender, level of education, employment status, and answers to Likert-scale questions, or ones that are discretized such as age, personal income, and monthly expenses. In Sect. 2, we outline the sampling scheme for self-selected interval data. Section 3 describes the model and the suggested estimation procedure. A simulation study is reported in Sect. 4. In Sect. 5, the methods are applied to data from a study where the respondents provided estimates of the prices of rice and two types of fish. In the Appendix are given proofs and auxiliary results.

2 Data collection scheme

We consider a two-stage scheme for collecting data. The motivation behind this scheme is that more information is needed than a single interval from each respondent in order to consistently estimate the underlying distribution function or related parameters. Therefore the respondent is asked to select a sub-interval of the interval that he/she stated. The problem of deciding where to split the stated interval into sub-intervals can be resolved using some previously collected data (in a pilot stage or an earlier survey) or based on other knowledge about the quantity of interest. Another possibility is to include a predetermined degree of rounding in the instruction for the respondents, e.g., to state intervals with endpoints rounded to a multiple of 10, and then the points of split will be chosen among the multiples of 10.

In the *pilot stage*, a random sample of individuals is selected and each individual is requested to give an answer in the form of an interval containing his/her value of the quantity of interest. It is assumed that the endpoints of the intervals are rounded (e.g., to the nearest multiple of 10) and that they are bounded from above by some large number. Let $\{d_j^\star\}$ be the set of endpoints of all observed intervals. The pilot-stage data are used only for obtaining the set $\{d_j^\star\}$.

In the *main stage*, a new random sample of n individuals is selected and each individual is asked to state an interval containing his/her value of the quantity of interest. We refer to this question as Qu1. Then, follow-up questions are asked according to one of the following designs.

Design A. The interval stated at Qu1 is split into two or three sub-intervals and the respondent is asked to select one of these sub-intervals. The points of split are chosen in some random fashion among the points $d_j^\star$ that are within the stated interval, e.g., equally likely. We refer to this question as Qu2.

Design B. The interval stated at Qu1 is split into two sub-intervals and the respondent is asked to select one of these sub-intervals. The point of split is the $d_j^\star$ that is the closest to the middle of the interval; if there are two points that are equally close to the middle, one of them is taken at random. We refer to this question as Qu2a. The interval selected at Qu2a is thereafter split similarly into two sub-intervals and the respondent is asked to select one of them. We refer to this question as Qu2b.

The respondent may refuse to answer Qu2 (Qu2a and Qu2b); we assume that the respondent chooses not to answer independently of his/her true value. If there are no points $d_j^\star$ within the interval stated at Qu1 or Qu2a, the respective follow-up question is not asked. We assume that if a respondent has answered Qu2 (Qu2a), he/she has chosen the interval containing his/her true value, independent of how the interval stated at Qu1 was split. An analogous assumption is made about the response to Qu2b.

In Design B, if we know the intervals stated at Qu1 and Qu2b, we can find out the answer to Qu2a. Thus, if Qu2b is answered, the data from Qu2a can be omitted. Let

Qu2Δ denote the last follow-up question that was answered by the respondent. If the respondent did not answer Qu2a (Qu2 in Design A), we say that there is no answer at Qu2Δ. Designs A and B are studied in Angelov and Ekström (2019), where they are referred to as schemes A and B.

Let $d_0 < d_1 < \ldots < d_{J-1} < d_J$ be the endpoints of all intervals observed at the main stage. The assumptions that the endpoints are rounded and bounded from above imply that J remains fixed for large sample sizes. Let us define a set of intervals $\mathcal{V} = \{\mathbf{v}_j\}$, where $\mathbf{v}_j = (d_{j-1}, d_j]$, $j = 1, \ldots, J$, and let $\mathcal{U} = \{\mathbf{u}_h\}$ be the set of all intervals that can be expressed as a union of intervals from $\mathcal{V}$, i.e., $\mathcal{U} = \{(d_l, d_r] : \ d_l < d_r, \ l, r = 0, \ldots, J\}$. We denote $\mathcal{J}_h$ to be the set of indices of intervals from $\mathcal{V}$ contained in $\mathbf{u}_h$, i.e., $\mathcal{J}_h = \{j : \ \mathbf{v}_j \subseteq \mathbf{u}_h\}$. For example, if $\mathcal{V} = \{(0,2], (2,5], (5,10]\}$, then $\mathcal{U} = \{(0,2], (2,5], (5,10], (0,5], (2,10], (0,10]\}$. Also, $\mathbf{u}_4 = (0,5] = \mathbf{v}_1 \cup \mathbf{v}_2$, hence $\mathcal{J}_4 = \{1,2\}$.

3 Model and methods

Let us denote the observations $\mathbf{dat}_i = (l_{1i}, r_{1i}, l_{2i}, r_{2i}, \mathbf{x}_i)$, $i = 1, \ldots, n$, where $(l_{1i}, r_{1i}]$ is the interval stated at Qu1, $(l_{2i}, r_{2i}]$ is the interval stated at Qu2Δ, and $\mathbf{x}_i = (1, x_{1i}, \ldots, x_{di})$ is a covariate vector. Each data point $(l_{1i}, r_{1i}, l_{2i}, r_{2i}, \mathbf{x}_i)$ is an observed value of random vector $(L_{1i}, R_{1i}, L_{2i}, R_{2i}, \mathbf{X}_i)$, $i = 1, \ldots, n$, $\mathbf{X}_i = (1, X_{1i}, \ldots, X_{di})$. The unobservable values $y_1, \ldots, y_n$ of the quantity of interest are values of independent random variables $Y_1, \ldots, Y_n$ and $L_{1i} \le L_{2i} < Y_i \le R_{2i} \le R_{1i}$. The distribution of Y_i depends on the value of $\mathbf{X}_i$. It is assumed that $\mathbf{X}_i$ takes finitely many values.

Let $Q_\tau(\mathbf{x}_i)$ be the τ-th quantile of Y_i conditional on $\mathbf{X}_i = \mathbf{x}_i$,

$$Q_\tau(\mathbf{x}_i) = \inf\{y : \mathbb{P}(Y_i \le y \mid \mathbf{x}_i) \ge \tau\}.$$

We assume that

$$Q_\tau(\mathbf{x}_i) = \boldsymbol{\beta}_\tau \mathbf{x}_i^\intercal = \beta_{0\tau} + \beta_{1\tau} x_{1i} + \ldots + \beta_{d\tau} x_{di},$$

where $\boldsymbol{\beta}_\tau \in \boldsymbol{\Theta} \subseteq \mathbb{R}^{d+1}$ is a parameter vector (a vector of regression coefficients).

For uncensored data, an estimate of $\boldsymbol{\beta}_\tau$ can be obtained by solving the estimating equation

$$\sum_{i=1}^{n} \Big(\mathbb{1}\{y_i \ge \boldsymbol{\beta}_\tau \mathbf{x}_i^\intercal\} - (1-\tau) \Big) \mathbf{x}_i = 0. \tag{1}$$

Following the ideas of McKeague et al. (2001) and Shen (2013), we replace the unobservable $\mathbb{1}\{y_i \ge \boldsymbol{\beta}_\tau \mathbf{x}_i^\intercal\}$ in (1) by an estimate of the conditional probability that $Y_i \ge \boldsymbol{\beta}_\tau \mathbf{x}_i^\intercal$ given $\mathbf{dat}_i$. Thus we arrive at the following estimating equation:

$$\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}_\tau) = \sum_{i=1}^{n} \Big(\widetilde{G}_i(\boldsymbol{\beta}_\tau \mathbf{x}_i^\intercal \mid \mathbf{dat}_i) - (1-\tau) \Big) \mathbf{x}_i = 0, \tag{2}$$

where $\widetilde{G}_i(\boldsymbol{\beta}_\tau \mathbf{x}_i^\top \mid \mathbf{dat}_i)$ is an estimate of the probability $G_i(\boldsymbol{\beta}_\tau \mathbf{x}_i^\top \mid \mathbf{dat}_i) = \mathbb{P}(Y_i \geq \boldsymbol{\beta}_\tau \mathbf{x}_i^\top \mid \mathbf{dat}_i)$. We define $\widehat{\boldsymbol{\beta}}_\tau$ to be the root of estimating equation (2).

Unless otherwise stated, hereafter we focus on the case $\tau = 0.5$ which corresponds to a median regression model and we omit the subscript τ in $\boldsymbol{\beta}_\tau$ and $\boldsymbol{\Psi}_\tau$. However, the suggested estimation procedure is applicable to an arbitrary $\tau \in (0, 1)$.

The set of combinations of possible values of $\mathbf{X}_i$ is denoted by $\{\boldsymbol{\xi}_k\}$, $k = 1, \dots, K$, i.e., there are K combinations in total. Let $c(h) = |\mathcal{J}_h|$; thus we can write $\mathcal{J}_h = \{j_{1(h)}, \dots, j_{c(h)}\}$, where $j_{1(h)} < j_{2(h)} < \dots < j_{c(h)}$ and $d_{j_{1(h)}} < d_{j_{2(h)}} < \dots < d_{j_{c(h)}}$.

Let us define

$$p_{j|h,k} = \mathbb{P}(Y_i \in \mathbf{v}_j \mid (L_{1i}, R_{1i}] = \mathbf{u}_h, \mathbf{X}_i = \boldsymbol{\xi}_k),$$
$$p_{j|h*s,k} = \mathbb{P}(Y_i \in \mathbf{v}_j \mid (L_{1i}, R_{1i}] = \mathbf{u}_h, (L_{2i}, R_{2i}] = \mathbf{u}_s, \mathbf{X}_i = \boldsymbol{\xi}_k),$$

where $\mathbf{u}_s \subset \mathbf{u}_h$. The following relation between $p_{j|h,k}$ and $p_{j|h*s,k}$ is fulfilled:

$$p_{j|h*s,k} = \frac{p_{j|h,k}}{\sum_{j \in \mathcal{J}_s} p_{j|h,k}}. \tag{3}$$

We need to estimate $p_{j|h,k}$ and $p_{j|h*s,k}$ in order to find an estimate $\widetilde{G}_i$, which is needed in (2). The conditional probabilities $p_{j|h,k}$ reflect the relative position of Y_i within the stated interval $(L_{1i}, R_{1i}]$. These probabilities are estimated using the data from Qu2Δ, where the respondent selects a sub-interval of $(L_{1i}, R_{1i}]$. The estimate $\widetilde{p}_{j|h,k}$ is obtained by applying the procedure proposed in Angelov and Ekström (2017) to the subset of data corresponding to $\mathbf{X}_i = \boldsymbol{\xi}_k$, namely, $\widetilde{p}_{j|h,k}$, $j \in \mathcal{J}_h$, is the maximizer of the log-likelihood

$$\sum_j n_{hjk} \log p_{j|h,k} + \sum_s n_{h*s,k} \log\left(\sum_{j \in \mathcal{J}_s} p_{j|h,k} \right),$$

where n_{hjk} is the number of respondents who stated $\mathbf{u}_h$ at Qu1, $\mathbf{v}_j$ at Qu2Δ ($\mathbf{v}_j \subseteq \mathbf{u}_h$) and have covariate value $\boldsymbol{\xi}_k$, while $n_{h*s,k}$ is the number of respondents who stated $\mathbf{u}_h$ at Qu1, $\mathbf{u}_s$ at Qu2Δ ($\mathbf{u}_s$ is a union of at least two intervals from $\mathcal{V}$, $\mathbf{u}_s \subset \mathbf{u}_h$) and have covariate value $\boldsymbol{\xi}_k$.

The estimate $\widetilde{p}_{j|h*s,k}$ is computed using the relation (3), i.e.,

$$\widetilde{p}_{j|h*s,k} = \frac{\widetilde{p}_{j|h,k}}{\sum_{j \in \mathcal{J}_s} \widetilde{p}_{j|h,k}}.$$

If independent censoring is assumed and the survival function of Y_i is close to linear over $(L_{1i}, R_{1i}]$, then the distribution of the relative position of Y_i within the interval $(L_{1i}, R_{1i}]$ will be close to uniform. This will not be realistic if the respondents exhibit some specific behavior when choosing the intervals, e.g., if they tend

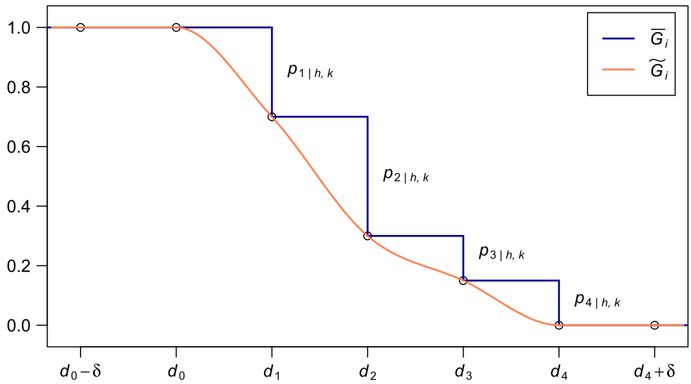

Fig. 1 An illustration of $\overline{G}_i$ and $\widetilde{G}_i$ for some i, where $(L_{1i}, R_{1i}] = \mathbf{u}_h$, $(L_{2i}, R_{2i}] =$ NA, $\mathbf{X}_i = \boldsymbol{\xi}_k$, and $\mathbf{u}_h = \mathbf{v}_1 \cup \mathbf{v}_2 \cup \mathbf{v}_3 \cup \mathbf{v}_4 = (d_0, d_4]$

to choose an interval such that the true value is located in the right half of the interval. Therefore, assuming independent censoring in such cases may lead to bias in the estimation of $\boldsymbol{\beta}$.

If $(L_{1i}, R_{1i}] = \mathbf{u}_h$, $(L_{2i}, R_{2i}] =$ NA (no answer), and $\mathbf{X}_i = \boldsymbol{\xi}_k$, then an estimate, $\overline{G}_i(y \mid \mathbf{dat}_i)$, of $G_i(y \mid \mathbf{dat}_i)$ can be derived as follows:

$$\overline{G}_i(y \mid \mathbf{dat}_i) = \begin{cases} 1 & \text{if } y < d_{j_{1(h)}}; \\ 1 - \sum_{j=j_{1(h)}}^{j_{1(h)}} \widetilde{p}_{j|h,k} & \text{if } y \in [d_{j_{1(h)}}, d_{j_{2(h)}}); \\ 1 - \sum_{j=j_{1(h)}}^{j_{2(h)}} \widetilde{p}_{j|h,k} & \text{if } y \in [d_{j_{2(h)}}, d_{j_{3(h)}}); \\ \dots & \\ 1 - \sum_{j=j_{1(h)}}^{j_{c(h)-1}} \widetilde{p}_{j|h,k} & \text{if } y \in [d_{j_{c(h)-1}}, d_{j_{c(h)}}); \\ 0 & \text{if } y \geq d_{j_{c(h)}}. \end{cases}$$

Thus, $\overline{G}_i$ is a step function with jumps at the points $d_{j_{1(h)}}, \dots, d_{j_{c(h)}}$. However, it will be more convenient to use a smoothed version of $\overline{G}_i$ and we employ spline interpolation for that purpose. The procedure for obtaining the smooth version $\widetilde{G}_i$ is described below. Figure 1 visualizes the functions $\overline{G}_i$ and $\widetilde{G}_i$ in an artificial example. Let δ be a positive constant.

Case 1 Suppose that $(L_{1i}, R_{1i}] = \mathbf{u}_h$, $(L_{2i}, R_{2i}] =$ NA, and $\mathbf{X}_i = \boldsymbol{\xi}_k$. Then $\widetilde{G}_i$ is the monotone cubic spline (see Fritsch and Carlson 1980) through the points:

First coordinate	Second coordinate
$d_{j_{1(h)}-1}-\delta$	1
$d_{j_{1(h)}-1}$	1
$d_{j_{1(h)}}$	$1-\sum_{j=j_{1(h)}}^{j_{1(h)}} \widetilde{p}_{j\|h,k}$
…	…
$d_{j_{c(h)-1}}$	$1-\sum_{j=j_{1(h)}}^{j_{c(h)-1}} \widetilde{p}_{j\|h,k}$
$d_{j_{c(h)}}$	0
$d_{j_{c(h)}}+\delta$	0

By adding the points $(d_{j_{1(h)}-1}-\delta, 1)$ and $(d_{j_{c(h)}}+\delta, 0)$, we get a spline $\widetilde{G}_i(y \mid \mathbf{dat}_i)$ such that $\widetilde{G}_i(y \mid \mathbf{dat}_i)=1$ if $y \leq d_{j_{1(h)}-1}$ and $\widetilde{G}_i(y \mid \mathbf{dat}_i)=0$ if $y \geq d_{j_{c(h)}}$. The constant δ can be chosen, e.g., as $\delta=\min_j |d_j-d_{j+1}|$; although any positive constant should work.

Case 2 Suppose that $(L_{1i}, R_{1i}]=\mathbf{u}_h$, $(L_{2i}, R_{2i}]=\mathbf{u}_s$, and $\mathbf{X}_i=\boldsymbol{\xi}_k$. Then $\widetilde{G}_i$ is the monotone cubic spline through the points:

First coordinate	Second coordinate
$d_{j_{1(s)}-1}-\delta$	1
$d_{j_{1(s)}-1}$	1
$d_{j_{1(s)}}$	$1-\sum_{j=j_{1(s)}}^{j_{1(s)}} \widetilde{p}_{j\|h*s,k}$
…	…
$d_{j_{c(s)-1}}$	$1-\sum_{j=j_{1(s)}}^{j_{c(s)-1}} \widetilde{p}_{j\|h*s,k}$
$d_{j_{c(s)}}$	0
$d_{j_{c(s)}}+\delta$	0

Case 3 Suppose that $(L_{2i}, R_{2i}]=\mathbf{v}_j$. Then $\widetilde{G}_i$ is the monotone cubic spline through the points:

First coordinate	Second coordinate
$d_{j-1}-\delta$	1
d_{j-1}	1
d_j	0
$d_j+\delta$	0

Let $\boldsymbol{\Psi}^\bullet(\boldsymbol{\beta})$ be an estimating function based on the true G_i rather than on $\widetilde{G}_i$, i.e.,

$$\boldsymbol{\Psi}^\bullet(\boldsymbol{\beta})=\sum_{i=1}^{n}\left(G_i(\boldsymbol{\beta}\,\mathbf{x}_i^\intercal \mid \mathbf{dat}_i)-\frac{1}{2}\right)\mathbf{x}_i.$$

Let $D(\boldsymbol{\beta})=n^{-1}\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\Psi}^\bullet(\boldsymbol{\beta})$. Let $\boldsymbol{\beta}^0$ be the true value of $\boldsymbol{\beta}$, i.e., the median of Y_i conditional on $\mathbf{X}_i=\mathbf{x}_i$ is given by $\boldsymbol{\beta}^0\mathbf{x}_i^\intercal$.

Assumption 1 $D(\boldsymbol{\beta}^0)\xrightarrow{\text{a.s.}} A$, where A is negative definite.

Table 1 Average computation time (in seconds)

Sample size	Method	One covariate	Two covariates
$n = 100$	NM	3.6	7.6
	BFGS	21.8	22.8
$n = 500$	NM	16.0	36.3
	BFGS	109.6	123.6
$n = 1000$	NM	31.8	65.0
	BFGS	217.0	268.4

The results are based on 30 replications in each case

Assumption 2 If the probabilities $\mathbb{P}(Y_i \geq d_j \,|\, \mathbf{dat}_i)$ are known for all possibly observed points d_j, then the survival function $G_i(y \,|\, \mathbf{dat}_i) = \mathbb{P}(Y_i \geq y \,|\, \mathbf{dat}_i)$ is the monotone cubic spline through the points $(d_j, \mathbb{P}(Y_i \geq d_j \,|\, \mathbf{dat}_i))$.

Assumption 3 $\sum_j n_{hjk}/(\sum_j n_{hjk} + \sum_s n_{h*s,k}) \overset{\text{a.s.}}{\longrightarrow} \gamma_{h,k} > 0$ as $n \longrightarrow \infty$.

We can regard Assumption 2 as a sensible approximation of the true underlying survival function. The very nature of a distributional model is a simplified and idealized representation of the underlying survival function, and thus there is no 'true' model that perfectly describes the survival function and how it depends on the covariates.

Assumption 3 ensures the strong consistency of $\widetilde{p}_{j|h,k}$, see Angelov and Ekström (2017).

The almost sure convergence of $\widehat{\boldsymbol{\beta}}$ is established in the following theorem.

Theorem 1 *Suppose that Assumptions 1–3 are satisfied. Then* $\widehat{\boldsymbol{\beta}} \overset{\text{a.s.}}{\longrightarrow} \boldsymbol{\beta}^0$ *as* $n \longrightarrow \infty$.

For $b = 1, \ldots, B$, let $\mathbf{dat}^*_{1,b}, \ldots, \mathbf{dat}^*_{n,b}$ be a random sample with replacement from the data $\mathbf{dat}_1, \ldots, \mathbf{dat}_n$. We say that $\mathbf{dat}^*_{1,b}, \ldots, \mathbf{dat}^*_{n,b}$ is the b-th bootstrap sample. Let $\widehat{\boldsymbol{\beta}}^*_b = (\widehat{\beta}^*_{0,b}, \ldots, \widehat{\beta}^*_{d,b})$ be the estimate of $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_d)$ from the bootstrap sample $\mathbf{dat}^*_{1,b}, \ldots, \mathbf{dat}^*_{n,b}$. Let $\widehat{\beta}^{\text{boot}}_r(\alpha)$ be the sample α quantile of $\widehat{\beta}^*_{r,1}, \ldots, \widehat{\beta}^*_{r,B}$ and let $\widehat{s}^{\text{boot}}_r$ be the sample standard deviation of $\widehat{\beta}^*_{r,1}, \ldots, \widehat{\beta}^*_{r,B}$, i.e.,

$$\widehat{s}^{\text{boot}}_r = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} \left(\widehat{\beta}^*_{r,b} - \frac{1}{B} \sum_{t=1}^{B} \widehat{\beta}^*_{r,t} \right)^2}.$$

Let $z_{1-\alpha}$ denote the $(1-\alpha)$ quantile of the standard normal distribution, i.e., for $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z < z_{1-\alpha}) = 1 - \alpha$.

We will explore the following confidence intervals for β_r with nominal level $1 - \alpha$:

- Bootstrap percentile confidence interval

$$\left[\widehat{\beta}_r^{\text{boot}}(\alpha/2), \quad \widehat{\beta}_r^{\text{boot}}(1-\alpha/2)\right], \tag{4}$$

- Wald-type confidence interval with bootstrap standard error

$$\left[\widehat{\beta}_r - z_{1-\alpha/2}\,\widehat{s}_r^{\text{boot}}, \quad \widehat{\beta}_r + z_{1-\alpha/2}\,\widehat{s}_r^{\text{boot}}\right]. \tag{5}$$

For monotone cubic spline interpolation, we use the R function `splinefun` with the option `method="monoH.FC"`, which corresponds to the method of Fritsch and Carlson (1980). The estimate $\widehat{\boldsymbol{\beta}}_\tau$ is obtained as a minimizer of $\|\boldsymbol{\Psi}_\tau(\boldsymbol{\beta}_\tau)\|$, where $\|\cdot\|$ is the Euclidean norm. For this task, the Nelder–Mead (NM) algorithm is used (the R function `optim` with the option `method="Nelder-Mead"`). The Broyden–Fletcher–Goldfarb–Shanno (BFGS) method can also be used (the R function `optim` with `method="BFGS"`); however, our experiments suggested that it is much slower than the Nelder–Mead algorithm for this particular optimization problem. Table 1 displays the average computation time for the suggested estimation procedure (using the NM algorithm and the BFGS algorithm) under different settings on a laptop computer with Intel(R) Pentium(R) CPU 2117U 1.8 GHz, RAM 4.0 GB.

4 Simulation study

4.1 Setup

Let $Y_1, \ldots, Y_n$ be independent random variables that have a Weibull distribution,

$$\mathbb{P}\,(Y_i > y \,|\, \mathbf{x}_i) = \exp\left(-\left(\frac{y}{\lambda_i}\right)^{\nu}\right),$$
$$\lambda_i = \frac{\boldsymbol{\beta}\,\mathbf{x}_i^{\mathsf{T}}}{\left(\log\frac{1}{1-\tau}\right)^{1/\nu}}.$$

Then, the τ-th quantile of Y_i is $\boldsymbol{\beta}\,\mathbf{x}_i^{\mathsf{T}}$.

We generate $Y_1, \ldots, Y_n$ according to the above definition with $\nu = 1.5$ and consider two cases for the covariates: (i) one covariate x_{1i} taking values 1, 2, or 3; (ii) two covariates x_{1i} and x_{2i}, where x_{1i} takes values 2 or 3 and x_{2i} takes values 0 or 1.

Let $U_1^{\mathrm{L}}, \ldots, U_n^{\mathrm{L}}$ and $U_1^{\mathrm{R}}, \ldots, U_n^{\mathrm{R}}$ be sequences of independent random variables:

$$\begin{aligned} U_i^{\mathrm{L}} &= M_i\,U_i^{(1)} + (1-M_i)\,U_i^{(2)}, \\ U_i^{\mathrm{R}} &= M_i\,U_i^{(2)} + (1-M_i)\,U_i^{(1)}, \end{aligned} \tag{6}$$

where $M_i \sim \text{Bernoulli}(p_{\mathrm{M}})$, $U_i^{(1)}$ and $U_i^{(2)}$ are random variables defined later. Let $(L_{1i}, R_{1i}]$ be the interval stated by the i-th respondent at question Qu1. The left endpoints are generated as $L_{1i} = (Y_i - U_i^{\mathrm{L}})\,\mathbb{1}\{Y_i - U_i^{\mathrm{L}} > 0\}$ rounded downwards to the nearest multiple of 10. The right endpoints are generated as $R_{1i} = Y_i + U_i^{\mathrm{R}}$ rounded

Table 2 Simulation settings

Setting	$U_i^{(1)}$	$U_i^{(2)}$	Covariates	p_{M}
S11	Unif(0, 20)	Unif(20, 40)	$x_{1i} \in \{1, 2, 3\}$	$0.2x_{1i} - 0.1$
S21	Unif(0, 12)	Unif(12, 24)	$x_{1i} \in \{1, 2, 3\}$	$0.2x_{1i} - 0.1$
S22	Unif(0, 12)	Unif(12, 24)	$x_{1i} \in \{2, 3\}$, $x_{2i} \in \{0, 1\}$	$0.2(x_{1i} + x_{2i}) - 0.3$

upwards to the nearest multiple of 10. We consider two settings for the random variables $U_i^{(1)}$ and $U_i^{(2)}$ in (6), see Table 2. In setting S11, the median length of the interval at Qu1 is 50, while in settings S21 and S22 the median length is 30. The data for the follow-up question are generated according to Design A; the interval $(L_{1i}, R_{1i}]$ is split into two sub-intervals, the point of split is chosen equally likely from all the possible points $d_j^\star$ that are within the interval. The probability that a respondent gives no answer to Qu2Δ is $p_{\mathrm{NA}} = 1/4$. The parameter p_{M} of the Bernoulli random variables M_i is considered to be a function of the covariates (see Table 2). For example, in setting S11, $p_{\mathrm{M}} = 0.2x_{1i} - 0.1$, which leads to tree possible values, $p_{\mathrm{M}} = 0.1, 0.3, 0.5$. Figure 2 illustrates the relative position of Y_i in the interval $(L_{1i}, R_{1i}]$, i.e., $(Y_i - L_{1i})/(R_{1i} - L_{1i})$, for the different values of p_{M} under setting S11. Instead of simulating pilot-stage data, a pre-determined set of points $\{d_j^\star\} = \{0, 10, 20, \ldots, 450\}$ is used (cf. Angelov and Ekström 2019).

All computations were performed with R (see R Core Team 2019). The R code can be obtained from the corresponding author upon request.

4.2 Results

We conducted simulations for a range of sample sizes where we compare the proposed estimator with the estimator of Shen (2013), which assumes independent censoring. Our estimator can be seen as an extension of Shen's estimator to the case of dependent censoring. With such comparison we can see the benefit of using an estimator that accounts for dependent censoring. Shen's estimator is applied to the dataset where each data point includes only the last interval stated by the respondent. Relative bias is defined as the bias divided by the true value of the parameter. Tables 3, 4, and 5 display the results based on 10000 simulated datasets (replications). We see that in most cases the root mean square error is smaller for our estimator. The bias of our estimator is considerably lower than the bias of Shen's estimator (with some exceptions for $n = 100$ under setting S22). Moreover, the bias of our estimator gets closer to zero as the sample size increases, while the bias of the other estimator does not change noticeably when increasing the sample size. The bias of our estimator for smaller sample sizes might be explained by the not large number of observations for each combination of h and k which may lead to poor estimates of some of the probabilities $p_{j|h,k}$.

Simulations concerning the bootstrap confidence intervals (4) and (5) are reported in Table 6. The results are based on 1000 simulated samples of sizes $n = 100$ and $n = 1500$. One bootstrap confidence interval is calculated using 1000 bootstrap samples. For the bootstrap percentile confidence intervals, the coverage is fairly close to

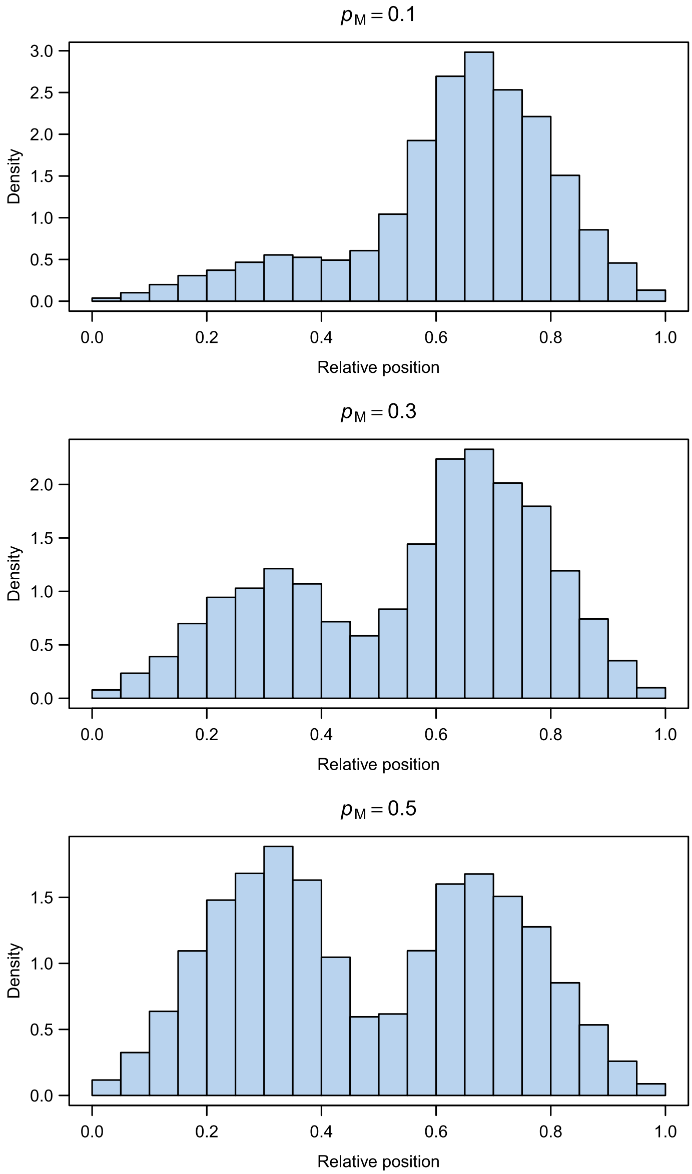

Fig. 2 Relative position of Y_i in the interval $(L_{1i}, R_{1i}]$, i.e., $(Y_i - L_{1i})/(R_{1i} - L_{1i})$ for three different values of p_M corresponding to $x_i = 1, 2, 3$. The histograms are based on a generated dataset of size $n = 50000$ under setting S11

the nominal level of 0.95. The bootstrap percentile method has previously shown good performance in the context of quantile regression (see, e.g., Wang and Wang 2009; De Backer et al. 2019). The Wald-type confidence intervals with bootstrap standard

Table 3 Simulation results under setting S11. Mean, relative bias (RB), and root mean square error (RMSE) based on 10000 replications. Comparison of our estimator (New) and the estimator of Shen (2013). The true value of the parameter is $\boldsymbol{\beta}^0 = (50, 12)$, $\tau = 0.5$

Estimator	n	$\widehat{\beta}_0$			$\widehat{\beta}_1$		
		Mean	RB	RMSE	Mean	RB	RMSE
New	100	46.219	−0.076	16.822	13.515	0.126	8.530
New	500	48.655	−0.027	8.287	12.521	0.043	4.186
New	1000	49.431	−0.011	6.109	12.225	0.019	3.077
New	1500	49.917	−0.002	5.183	12.053	0.004	2.583
Shen (2013)	100	43.650	−0.127	19.983	14.209	0.184	9.935
Shen (2013)	500	43.678	−0.126	10.716	14.094	0.175	4.881
Shen (2013)	1000	43.601	−0.128	8.845	14.106	0.175	3.758
Shen (2013)	1500	43.679	−0.126	8.017	14.077	0.173	3.271

Table 4 Simulation results under setting S21. Mean, relative bias (RB), and root mean square error (RMSE) based on 10000 replications. Comparison of our estimator (New) and the estimator of Shen (2013)

Estimator	n	$\widehat{\beta}_0$			$\widehat{\beta}_1$		
		Mean	RB	RMSE	Mean	RB	RMSE
$\tau = 0.25$, $\boldsymbol{\beta}^0 = (27, 7)$							
New	100	25.549	−0.054	12.906	7.647	0.092	6.585
New	500	26.854	−0.005	6.271	7.091	0.013	3.203
New	1000	27.095	0.004	4.374	6.982	−0.003	2.296
Shen (2013)	100	24.278	−0.101	13.828	8.076	0.154	7.051
Shen (2013)	500	23.827	−0.118	6.960	8.094	0.156	3.379
Shen (2013)	1000	23.752	−0.120	5.435	8.109	0.158	2.500
$\tau = 0.5$, $\boldsymbol{\beta}^0 = (50, 12)$							
New	100	48.691	−0.026	16.833	12.584	0.049	8.585
New	500	49.726	−0.005	8.045	12.155	0.013	4.106
New	1000	50.017	0.000	5.788	12.029	0.002	2.949
Shen (2013)	100	47.235	−0.055	17.594	13.006	0.084	8.985
Shen (2013)	500	46.605	−0.068	8.694	13.138	0.095	4.253
Shen (2013)	1000	46.690	−0.066	6.445	13.116	0.093	3.050
$\tau = 0.75$, $\boldsymbol{\beta}^0 = (80, 19)$							
New	100	78.946	−0.013	22.585	19.322	0.017	11.602
New	500	79.619	−0.005	10.789	19.154	0.008	5.472
New	1000	79.691	−0.004	7.795	19.143	0.008	3.949
Shen (2013)	100	77.320	−0.034	22.654	19.864	0.045	11.787
Shen (2013)	500	76.648	−0.042	11.442	20.120	0.059	5.724
Shen (2013)	1000	76.577	−0.043	8.475	20.144	0.060	4.135

Table 5 Simulation results under setting S22. Mean, relative bias (RB), and root mean square error (RMSE) based on 10000 replications. Comparison of our estimator (New) and the estimator of Shen (2013)

Estimator	n	$\widehat{\beta}_0$			$\widehat{\beta}_1$			$\widehat{\beta}_2$		
		Mean	RB	RMSE	Mean	RB	RMSE	Mean	RB	RMSE
$\tau = 0.25, \quad \boldsymbol{\beta}^0 = (18, 9, 5)$										
New	100	15.201	−0.156	27.948	9.984	0.109	11.286	5.887	0.177	11.188
New	500	17.356	−0.036	13.603	9.196	0.022	5.470	5.204	0.041	5.417
New	1000	17.716	−0.016	9.747	9.097	0.011	3.920	5.098	0.020	3.953
Shen (2013)	100	15.016	−0.166	26.082	9.931	0.103	10.505	5.975	0.195	11.505
Shen (2013)	500	13.945	−0.225	14.072	10.034	0.115	5.511	6.177	0.235	5.520
Shen (2013)	1000	13.788	−0.234	10.480	10.059	0.118	4.009	6.101	0.220	3.963
$\tau = 0.5, \quad \boldsymbol{\beta}^0 = (35, 15, 10)$										
New	100	33.034	−0.056	35.691	15.557	0.037	14.380	10.923	0.092	14.574
New	500	34.495	−0.014	17.100	15.144	0.010	6.869	10.160	0.016	6.912
New	1000	34.707	−0.008	12.557	15.116	0.008	5.071	10.019	0.002	4.918
Shen (2013)	100	32.568	−0.069	31.940	15.500	0.033	12.854	11.133	0.113	14.492
Shen (2013)	500	30.790	−0.120	17.774	16.083	0.072	7.022	11.074	0.107	6.890
Shen (2013)	1000	30.318	−0.134	13.010	16.227	0.082	5.024	11.097	0.110	5.007
$\tau = 0.75, \quad \boldsymbol{\beta}^0 = (55, 24, 16)$										
New	100	53.291	−0.031	47.834	24.448	0.019	19.212	16.815	0.051	19.847
New	500	54.271	−0.013	23.114	24.135	0.006	9.291	16.294	0.018	9.054
New	1000	54.513	−0.009	16.674	24.126	0.005	6.694	16.223	0.014	6.693
Shen (2013)	100	54.077	−0.017	36.366	23.661	−0.014	14.755	17.292	0.081	17.601
Shen (2013)	500	51.020	−0.072	23.805	24.936	0.039	9.508	17.123	0.070	9.440
Shen (2013)	1000	50.670	−0.079	17.028	25.067	0.044	6.694	17.131	0.071	6.759

Table 6 Confidence intervals: coverage proportion (CP) and average length (AL) based on 1000 replications and 1000 bootstrap samples under setting S11. The nominal level is 0.95. The true value of the parameter is $\boldsymbol{\beta}^0 = (50, 12), \ \tau = 0.5$

Method	n	β_0		β_1	
		CP	AL	CP	AL
Bootstrap percentile	100	0.941	63.727	0.956	32.981
Wald with BootSE	100	0.922	64.285	0.948	33.244
Bootstrap percentile	1500	0.954	18.780	0.943	9.639
Wald with BootSE	1500	0.930	19.102	0.927	9.706

error (Wald with BootSE) are on average longer and their coverage is in some cases too low. Therefore, the bootstrap percentile confidence intervals are recommended.

5 Application

We apply the proposed methods to data concerning price estimates from a study conducted in Aklan, a province in the Philippines. The focus of the sampling process was the capital city, Kalibo. The administrative divisions, barangays, of Kalibo were classified into either coastal or inland communities. Two coastal barangays (Pook and Old Buswang) and two inland barangays (Tigayon and Estancia) were randomly selected. In each barangay, a number of households were randomly chosen. With their consent, a member of a sampled household (preferably, the head) was asked to participate in a survey. They were told to answer as honest as possible, and that their identity and personal data gathered will be kept confidential. The questionnaire was written in English, but trained enumerators explained questions in the local language Tagalog.

The participants were asked to provide estimates of the prices of rice and two types of fish (galunggong and bangus). They answered by means of self-selected intervals. As a follow-up question, the respondents were asked whether the price is more likely to be in the left or in the right half of the interval. Price estimates were given for two time periods: April 2019 (summer/fishing season) and September 2019 (typhoon/non-fishing season); thus the dataset contains six price estimates:

(RA) Price of 1 kg of rice in April 2019;

(RS) Price of 1 kg of rice in September 2019;

(GA) Price of 1 kg of galunggong in April 2019;

(GS) Price of 1 kg of galunggong in September 2019;

(BA) Price of 1 kg of bangus in April 2019;

(BS) Price of 1 kg of bangus in September 2019.

Table 7 Observed market prices per kilogram

	Product	Period	Price (in pesos)
(RA)	Rice	April 2019	38.25
(RS)	Rice	September 2019	38.00
(GA)	Galunggong	April 2019	110.00
(GS)	Galunggong	September 2019	130.00
(BA)	Bangus	April 2019	160.00
(BS)	Bangus	September 2019	160.00

The data for rice are from the Philippine Statistics Authority. The data for galunggong and bangus are from the Bureau of Fisheries and Aquatic Resources and the Municipal Economic Enterprise Development Office, Municipality of Kalibo

Data collection took place in August 2019, therefore the price estimate for April 2019 is a recall, while the price estimate for September 2019 is a forecast. The observed market prices for the given periods can be found in Table 7.

First, we investigated how the 0.25-quantile, the median, and the 0.75-quantile of the price depend on the level of education of the respondent. Consider the following models:

$$\texttt{Qnt025(Price)} = \beta_0 + \beta_1\,\texttt{Education}, \tag{7}$$

$$\texttt{Median(Price)} = \beta_0 + \beta_1\,\texttt{Education}, \tag{8}$$

$$\texttt{Qnt075(Price)} = \beta_0 + \beta_1\,\texttt{Education}, \tag{9}$$

where `Education` is a variable with values 1 = 'Lower than college level' and 2 = 'College level or higher'. In the first model, the parameter β_1 shows how the 0.25-quantile of the price differs between respondents with college education compared to those with lower education. In the second model, the parameter β_1 shows how the median price differs between respondents with college education compared to those with lower education. In the third model, the interpretation is similar.

Point estimates and confidence intervals for the parameter β_1 based on the collected data ($n = 178$) are presented in Fig. 3. The results indicate that people with college education tend to give higher price estimates. However, for each of the six prices, the confidence intervals are quite long and contain zero, which implies that the hypothesis that $\beta_1 = 0$ can not be rejected at the 5% significance level.

Point estimates for the 0.25-quantile, the median, and the 0.75-quantile of the prices together with confidence intervals are shown in Fig. 4. For rice and galunggong (cheaper fish), respondents tend to overestimate the prices (observed market price is

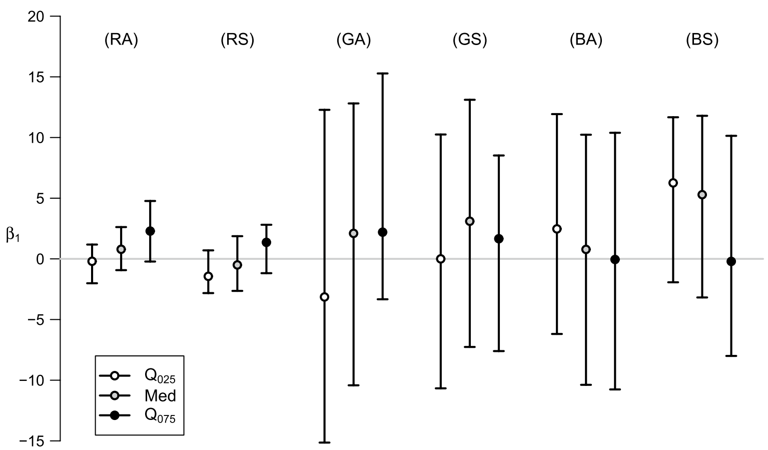

Fig. 3 Estimates and bootstrap percentile confidence intervals for the parameter β_1 in the models with one covariate (7, 8 and 9). The confidence intervals are based on 50000 bootstrap samples. The confidence level is 0.95

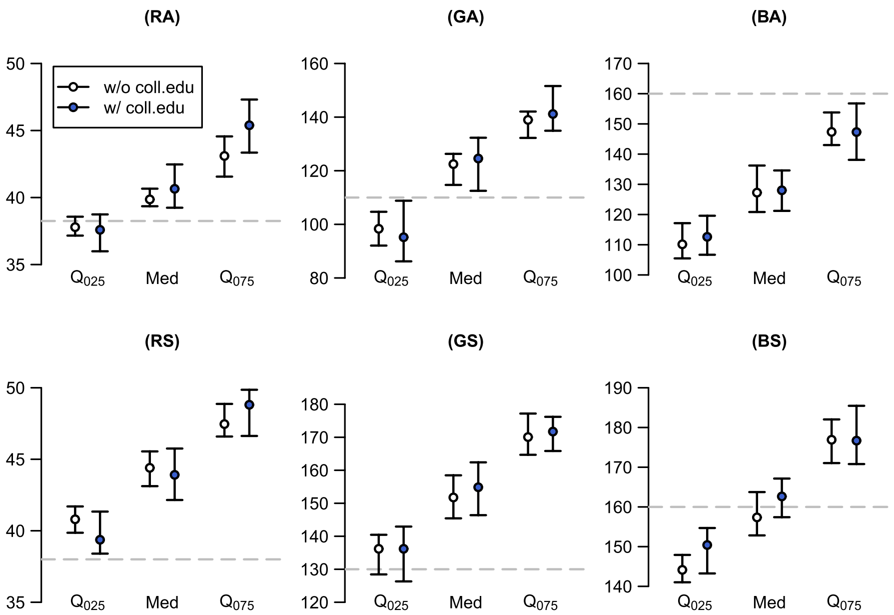

Fig. 4 Estimates and bootstrap percentile confidence intervals for the 0.25-quantile, the median, and the 0.75-quantile of the prices using the models with one covariate (7, 8 and 9). The confidence intervals are based on 50000 bootstrap samples. The confidence level is 0.95. In each plot, the observed market price (see Table 7) is displayed with a horizontal dashed line

below the lower bound of the confidence intervals for the medians). For bangus (luxury fish), respondents underestimated the price in April (observed market price is above the upper bound of the confidence intervals for the medians and the 0.75-quantiles). However, they gave more accurate estimates for the price of bangus in September (observed market price is within the confidence intervals for the medians).

Respondents expected prices to be higher in the typhoon season compared to the non-typhoon season, which in reality happened only with the price of galunggong, while the prices of rice and bangus remained stable.

We also considered models with two covariates:

$$\texttt{Qnt025(Price)} = \beta_0 + \beta_1\,\texttt{Education} + \beta_2\,\texttt{HouseholdHead}, \tag{10}$$

$$\texttt{Median(Price)} = \beta_0 + \beta_1\,\texttt{Education} + \beta_2\,\texttt{HouseholdHead}, \tag{11}$$

$$\texttt{Qnt075(Price)} = \beta_0 + \beta_1\,\texttt{Education} + \beta_2\,\texttt{HouseholdHead}, \tag{12}$$

where `HouseholdHead` is a variable which takes value 1, if the respondent is head of the household, and 0 otherwise.

Point estimates and confidence intervals for the parameters β_1 and β_2 are presented in Figs. 5 and 6. The results indicate that people with college education tend to give higher price estimates compared to those without college education. Heads

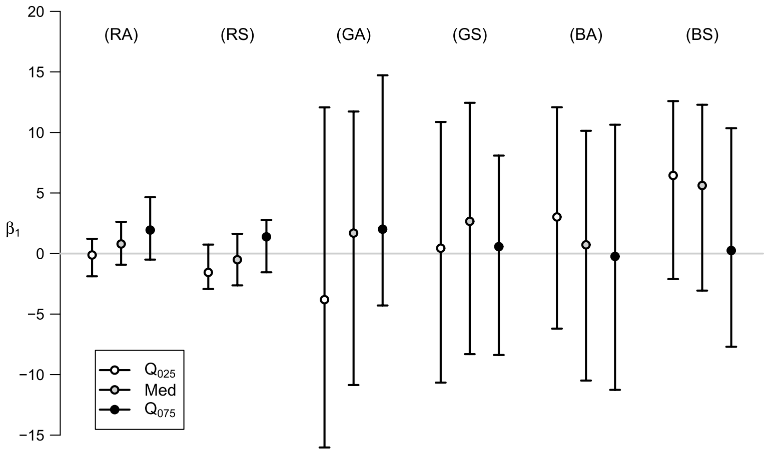

Fig. 5 Estimates and bootstrap percentile confidence intervals for the parameter β_1 in the models with two covariates (10, 11 and 12). The confidence intervals are based on 50000 bootstrap samples. The confidence level is 0.95

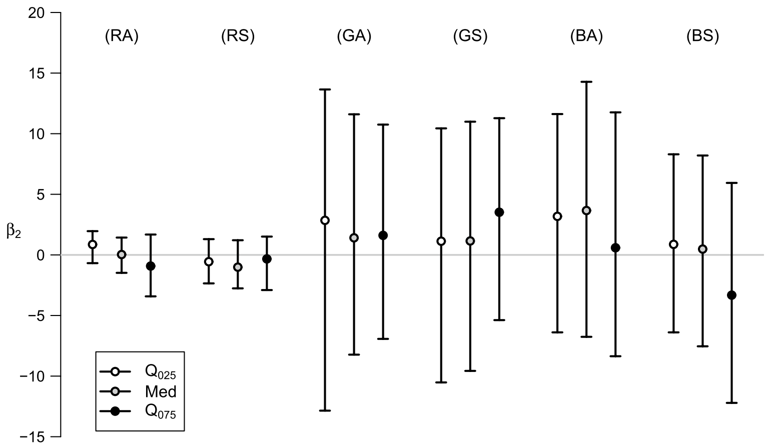

Fig. 6 Estimates and bootstrap percentile confidence intervals for the parameter β_2 in the models with two covariates (10, 11 and 12). The confidence intervals are based on 50000 bootstrap samples. The confidence level is 0.95

of households tend to give higher price estimates for galunggong and bangus compared to people who are not heads of households. However, all the confidence intervals for the parameters β_1 and β_2 contain zero. Therefore, in each case the hypotheses $\beta_1 = 0$ and $\beta_2 = 0$ can not be rejected at the 5% significance level.

6 Concluding remarks

We suggested an estimator for quantile regression for self-selected interval data with discrete covariates. We proved the strong consistency of the estimator. Our simulation study indicated that the proposed estimator performs better than an existing estimator which assumes independent censoring. A simple bootstrap procedure for constructing confidence intervals (the bootstrap percentile) showed satisfactory performance in the simulations.

A Appendix

A.1 Continuity of splines

Here we show the continuity of monotone cubic splines (see Fritsch and Carlson 1980) with respect to the data points. The notation in this section is independent of that in the rest of the paper.

Suppose that we have data points (x_i, y_i), $i = 1, \dots, m$, where $x_1 < x_2 < \dots < x_m$ and $y_1 \geq y_2 \geq \dots \geq y_m$. Let $g(x)$ be a monotone piecewise cubic function such that $g(x_i) = y_i$, $i = 1, \dots, m$. In each interval $[x_i, x_{i+1}]$, $g(x)$ is a cubic polynomial:

$$g(x) = y_i H_1(x) + y_{i+1} H_2(x) + a_i H_3(x) + a_{i+1} H_4(x),$$

where

$$\begin{aligned}
H_1(x) &= \varphi_1((x_{i+1} - x)/(x_{i+1} - x_i)),\\
H_2(x) &= \varphi_1((x - x_i)/(x_{i+1} - x_i)),\\
H_3(x) &= -(x_{i+1} - x_i)\,\varphi_2((x_{i+1} - x)/(x_{i+1} - x_i)),\\
H_4(x) &= (x_{i+1} - x_i)\,\varphi_2((x - x_i)/(x_{i+1} - x_i)),\\
\varphi_1(t) &= 3t^2 - 2t^3,\\
\varphi_2(t) &= t^3 - t^2.
\end{aligned}$$

We use the following procedure for calculating a_i, $i = 1, \dots, m$.

Step 1. If $y_{i+1} = y_i$, set $a_i^{[0]} = a_{i+1}^{[0]} = 0$. Else,

$$\begin{aligned}
a_i^{[0]} &= \frac{1}{2}\left(\frac{y_i - y_{i-1}}{x_i - x_{i-1}} + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right), \quad i = 2, \dots, m-1;\\
a_1^{[0]} &= a_m^{[0]} = 0.
\end{aligned}$$

Step 2. Let

$$\Delta_i = (y_{i+1} - y_i)/(x_{i+1} - x_i),$$
$$\lambda_i = \mathbb{I}\{\Delta_i \neq 0\}\, a_i/\Delta_i,$$
$$\mu_i = \mathbb{I}\{\Delta_i \neq 0\}\, a_{i+1}/\Delta_i,$$
$$\tau_i = \sqrt{(\lambda_i^2 + \mu_i^2)/9}.$$

Then

$$a_i = \frac{a_i^{[0]}}{\max\{1, \tau_i\}}, \qquad a_{i+1} = \frac{a_{i+1}^{[0]}}{\max\{1, \tau_i\}}.$$

Suppose that $\widehat{y}_i$ is an estimator of y_i, $i = 1, \dots, m$, and $\widehat{y}_i \xrightarrow{\text{a.s.}} y_i$ as $n \longrightarrow \infty$, where n is the size of the sample used for obtaining $\widehat{y}_i$. All quantities with a hat (e.g., $\widehat{a}_i$) imply that y_i is substituted with $\widehat{y}_i$. Let

$$\widehat{g}(x) = \widehat{y}_i H_1(x) + \widehat{y}_{i+1} H_2(x) + \widehat{a}_i H_3(x) + \widehat{a}_{i+1} H_4(x).$$

Lemma 1 *If* $\widehat{y}_i \xrightarrow{\text{a.s.}} y_i$ *as* $n \longrightarrow \infty$, *then* $\sup_{x\in[x_1,x_m]} |\widehat{g}(x) - g(x)| \xrightarrow{\text{a.s.}} 0$ *as* $n \longrightarrow \infty$.

Proof Taking into account that each $\widehat{a}_i$ is a continuous function of $\widehat{y}_1, \dots, \widehat{y}_m$, it follows that $\widehat{a}_i \xrightarrow{\text{a.s.}} a_i$ as $n \longrightarrow \infty$.

Note that there is a constant c such that $\max_{1\le i\le m} |x_i - x_{i+1}| \le c$. Also, $\sup_{t\in[0,1]} |\varphi_1(t)| = 1$, $\sup_{t\in[0,1]} |\varphi_2(t)| = 4/27$. Then

$$\sup_{x\in[x_1,x_m]} |\widehat{g}(x) - g(x)| \le 2 \max_{1\le i\le m} |\widehat{y}_i - y_i| + \frac{8c}{27} \max_{1\le i\le m} |\widehat{a}_i - a_i| \xrightarrow{\text{a.s.}} 0.$$

□

A.2 Consistency of the proposed estimator

Lemma 2 *If Assumption 3 is satisfied, then*

$$\sup_{y\in\mathbb{R}} \left| \widetilde{G}_i(y \mid \mathbf{dat}_i) - G_i(y \mid \mathbf{dat}_i) \right| \xrightarrow{\text{a.s.}} 0 \quad \text{as} \quad n \longrightarrow \infty.$$

Proof The functions $\widetilde{G}_i$ and G_i are splines based on two different sets of data points. Assumption 3 guarantees that $\widetilde{p}_{j|h,k}$ is a strongly consistent estimator of $p_{j|h,k}$ (see Angelov and Ekström 2017). Therefore, the data points used for $\widetilde{G}_i$ converge almost surely to the data points used for G_i. Then, the claim follows from Lemma 1. □

Proof of Theorem 1 Using Lemma 2, we get

$$\sup_{\beta \in \Theta} \| n^{-1}\boldsymbol{\Psi}(\boldsymbol{\beta}) - n^{-1}\boldsymbol{\Psi}^{\bullet}(\boldsymbol{\beta}) \| \xrightarrow{\text{a.s.}} 0 \text{ as } n \longrightarrow \infty.$$

By definition, $\boldsymbol{\Psi}(\widehat{\boldsymbol{\beta}}) = 0$. Then $n^{-1}\boldsymbol{\Psi}^{\bullet}(\widehat{\boldsymbol{\beta}}) \xrightarrow{\text{a.s.}} 0$ as $n \longrightarrow \infty$. Also, we have $\boldsymbol{\Psi}^{\bullet}(\boldsymbol{\beta}^0) = 0$.

Applying Taylor's expansion (see Feng et al. 2013), we obtain

$$n^{-1}\boldsymbol{\Psi}^{\bullet}(\widehat{\boldsymbol{\beta}}) - n^{-1}\boldsymbol{\Psi}^{\bullet}(\boldsymbol{\beta}^0) = D(\boldsymbol{\beta}^0)(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) + o(\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|).$$

By Assumption 1, $D(\boldsymbol{\beta}^0)$ is negative definite for large n. Therefore $\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} \boldsymbol{\beta}^0$ as $n \longrightarrow \infty$. □

Acknowledgements We would like to thank Rupert Bustamante and the local government unit of Kalibo, Aklan for logistical assistance. We are also grateful to the reviewers for their valuable comments and suggestions.

Funding Open access funding provided by Swedish University of Agricultural Sciences. The work was supported by the Marianne and Marcus Wallenberg Foundation, Sweden (Project MMW 2017.0075).

Data availability Not available.

Code availability Available upon request.

Declarations

Conflict of interest No.

Consent to participate Informed consent was obtained from all individual participants included in the study.

Ethics approval Ethical review and approval was not required for this type of study according the local legislation and institutional requirements.

References

Angelov AG, Ekström M (2017) Nonparametric estimation for self-selected interval data collected through a two-stage approach. Metrika 80(4):377–399

Angelov AG, Ekström M (2019) Maximum likelihood estimation for survey data with informative interval censoring. AStA Adv Stat Anal 103(2):217–236

Belyaev Y, Kriström B (2010) Approach to analysis of self-selected interval data. Working Paper 2010:2, CERE, Umeå University and the Swedish University of Agricultural Sciences. https://doi.org/10.2139/ssrn.1582853

Belyaev Y, Kriström B (2012) Two-step approach to self-selected interval data in elicitation surveys. Working Paper 2012:10, CERE, Umeå University and the Swedish University of Agricultural Sciences. https://doi.org/10.2139/ssrn.2071077

Belyaev Y, Kriström B (2015) Analysis of survey data containing rounded censoring intervals. Inform Appl 9(3):2–16

Bogaerts K, Komarek A, Lesaffre E (2017) Survival analysis with interval-censored data: a practical approach with examples in R, SAS, and BUGS. CRC Press, Boca Raton

De Backer M, El Ghouch A, Van Keilegom I (2019) An adapted loss function for censored quantile regression. J Am Stat Assoc 114(527):1126–1137

Feng C, Wang H, Han Y, Xia Y, Tu XM (2013) The mean value theorem and Taylor's expansion in statistics. Am Stat 67(4):245–248

Fritsch FN, Carlson RE (1980) Monotone piecewise cubic interpolation. SIAM J Numer Anal 17(2):238–246

Frumento P (2022) A quantile regression estimator for interval-censored data. Int J Biostat. https://doi.org/10.1515/ijb-2021-0063

Håkansson C (2008) A new valuation question: analysis of and insights from interval open-ended data in contingent valuation. Environ Resour Econ 39(2):175–188

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd edn. Wiley, Hoboken

Kim Y-J, Cho H, Kim J, Jhun M (2010) Median regression model with interval censored data. Biom J 52(2):201–208

Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge

Li C, Li Y, Ding X, Dong X (2020) DGQR estimation for interval censored quantile regression with varying-coefficient models. PLoS ONE 15(11):e0240046

Mahieu P-A, Wolff F-C, Shogren J, Gastineau P (2017) Interval bidding in a distribution elicitation format. Appl Econ 49(51):5200–5211

McKeague IW, Subramanian S, Sun Y (2001) Median regression and the missing information principle. J Nonparametric Stat 13(5):709–727

Press SJ, Tanur JM (2004) An overview of the respondent-generated intervals (RGI) approach to sample surveys. J Mod Appl Stat Methods 3(2):288–304

Press SJ, Tanur JM (2004) Relating respondent-generated intervals questionnaire design to survey accuracy and response rate. J Off Stat 20(2):265–287

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Shen P-S (2013) Median regression model with left truncated and interval-censored data. J Korean Stat Soc 42(4):469–479

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Wang HJ, Wang L (2009) Locally weighted censored quantile regression. J Am Stat Assoc 104(487):1117–1128

Zhou X, Feng Y, Du X (2017) Quantile regression for interval censored data. Commun Stat Theor Methods 46(8):3848–3863