Journal of Biostatistics and Epidemiology

J Biostat Epidemiol. 2023;9(2):270-286

Original Article

Quantile Regression in Survival Analysis: Comparing Check-Based Modeling and the Minimum Distance Approach

Fereshteh Mokhtarpour[1], Mostafa Hosseini[1], Akram Yazdani[2*], Mehdi Yaseri[1*]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.
[2]Department of Biostatistics and Epidemiology, Faculty of Health, Kashan University of Medical Sciences, Kashan, Iran.

ARTICLE INFO

Received 26.01.2023
Revised 19.02.2023
Accepted 14.03.2023
Published 15.06.2023

Key words:
Quantile regression;
Minimum distance approach;
Survival;
Check-based modeling;
Inverse cumulative distribution function

ABSTRACT

Introduction: Quantile regression is a valuable alternative for survival data analysis, enabling flexible evaluations of covariate effects on survival outcomes with intuitive interpretations. It offers practical computation and reliability. However, challenges arise when applying quantile regression to censored data, particularly for upper quantiles. The minimum distance approach, utilizing dual-kernel estimation and the inverse cumulative distribution function, shows promise in addressing these challenges, especially with higher-dimensional covariates.

Methods: This study contrasts two methods within the realm of quantile linear regression for survival analysis: check-based modeling and the minimum distance approach. Effectiveness is assessed across various scenarios through comprehensive simulation.

Results: The simulation results showed that using the quantile regression model with the minimum distance approach reduces the percentage of root mean square error in parameter estimation compared to the quantile regression models based on the check loss function. Additionally, a larger sample size and reduced censoring percentage led to decreased root mean square error in parameter estimation.

Conclusion: The research highlights the benefits of using the minimum distance approach for quantile regression. It reduces errors, improves model predictions, captures patterns, and optimizes parameters even with complete data. However, this approach has limitations. The accuracy of estimated quantiles can be influenced by the choice of distance metric and weighting scheme. The assumption of independence between censoring mechanism and survival time may not hold in real-world scenarios. Additionally, dealing with large datasets can be computationally complex.

Introduction

Survival analysis, a key statistical tool, examines time-to-event data. Common applications include understanding the duration until a patient's demise or machinery breakdown.[1] Traditionally, the Cox proportional hazards model, which relates the hazard rate (the likelihood of an event at a given time) to predictor variables, has been a staple in this

*.Corresponding Author: myaseri@tums.ac & akram.yazdani@gmail.com

field. Yet, it falters when the hazard rate isn't proportionate over time or is influenced by variable interactions.[2]

The Accelerated failure time (AFT) models serve as an alternative. Assuming a fixed parametric distribution of survival times, they focus on parameter estimation of that distribution. However, they possess inherent limitations: the consistent shape assumption across covariate levels might not be always realistic, and they often struggle with the frequent challenge of censored data in survival analysis.[3, 4] Censored data in survival analysis refers to incomplete or partially observed information where the event of interest (e.g., death, failure, or recovery) has not yet occurred for all individuals in the study. Reasons for this incompleteness include limited study duration, loss to follow-up, or ongoing observation of participants. As a result, the exact timing of these events remains unknown, resulting in censored observations.[5]

Koenker and Bassett revolutionized regression analysis in 1978 with their introduction of quantile regression. Unlike traditional methods that estimate the conditional mean, quantile regression focuses on estimating the conditional quantiles of the response variable. Their work led to the development of a robust quantile regression estimator that provides a comprehensive understanding of the response variable's distribution, going beyond the mean.[6, 7]

Their inherent flexibility and resistance to outliers make them a potent tool in survival analysis, supplementing classical methods like Cox regression and AFT models.[8-10]

However, applying quantile regression, especially in censored contexts, hasn't been straightforward. Early works by Powell (1984, 1986) assumed observable censoring variables for all data, an unrealistic expectation given the prevalence of random censoring.[11]

In quantile regression, the check loss function is commonly used for model fitting and validation for cross-validation approaches.[12] In the check-based approach, it's an optimization problem to find a piecewise function (the check function) that best fits the desired quantile.[13] The check function divides data into segments and minimizes loss functions within each segment, measuring the deviation between observed and predicted quantiles. Iteratively adjusting the check function's parameters yields quantile-specific coefficients.[14] In quantile regression, three main approaches exist within the check-based framework:

1. Inverse-censoring-probability (ICP) Weighting:

Rooted in the check-based formulation of quantile regression, this approach leverages the expected loss of the observed response, specifically for uncensored observations. But it's limited by the need for smoothing the conditional distribution and often doesn't maximize the robustness of quantile regression in handling censored observations.[15-21]

2. Weighting Scheme for Quantile Regression:

Not treating all censored observations uniformly, this method proposed by Portnoy (2003) and refined by Wang and Wang (2009) is grounded on Efron's redistribution-of-mass idea.[8, 22]

3. Modification of Check-based Formulation:

Lindgren (1997) and De Backer et al. (2019) suggest leveraging all observations as if complete and modifying the target in the check-based formulation for more accurate results.[23, 24] Despite these advances, challenges remain. Censored quantile regressions, especially for upper

quantile levels, grapple with identifiability constraints.

De Backer (2020) critiqued check-function-based approaches, proposing the use of the inverse cumulative distribution function technique for linear regression. This method, although less common in linear regression, shows promise, particularly when combined with dimension reduction strategies for handling high-dimensional covariates.[25]

In essence, while survival analysis has evolved with diverse tools and techniques, finding the most effective and universally adaptable method remains a dynamic field of research.

The study aimed to compare check-based modeling and the minimum distance approach as two methods of quantile regression in survival analysis. This comparison was carried out through simulations across various conditions.

Models

Quantile regression, a statistical approach, enables inference about conditional quantile functions, estimating models for various quantile levels beyond the median. Our study contrasts two methods: Check-Based Modeling (Bang and Tsiatis, Wang and Wang) and the Minimum Distance approach (De Backer).

Bang and Tsiatis's method

Assuming that T_i represents the time of the *i*-th failure, or a monotonic transformation of it, and X_i is a (p-1)× 1 vector of covariates for T_i, the median regression establishes a relationship between the median of T_i and the covariates, given X_i:[17]

$$T_i = \beta_0' Z_i + \varepsilon_i$$

The vector Z_i is defined as $(1,X_i')'$, where $i = 1,...,n$, and β is a vector with p dimensions. It is assumed that ε_i, $i = 1,...,n$, has a conditional median of *0*. In the presence of censoring, the observations consist of bivariate vectors (T_i, δ_i), where $T_i = min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, C_i denote time to censoring and I(.) is the indicator function. The censoring variable C_i is assu med to be independent of T_i. Moreover, it is assumed that the survival function *G(.)* of C_i does not depend on Z_i, and $\{(T_i, C_i, X_i), i = 1,...,n\}$ is generated through random sampling. β is estimate by a root of

$$\tilde{U}_n(\beta) = \sum_{i=1}^{n} \frac{\delta_i}{\hat{G}(Y_i)} Z_i \left\{ I(Y_i - \beta' Z_i \geq 0) - \tau \right\} \approx 0$$

where $\hat{G}$ is the Kaplan-Meier estimator for *G*. In this model, the estimation of the expected loss for the unobservable response relies on using the expected loss of the observed response. This estimation is carried out by considering only uncensored observations and adjusting through the conditional distribution of *C* given *X*. However, a limitation of this approach is that it necessitates smoothing of the conditional distribution, which imposes a constraint on the number of covariates that can be considered. Additionally, this approach does not fully exploit the robustness benefits offered by quantile regression when dealing with censored observations.

Wang and Wang's method

Wang and Wang proposed a technique that uses local weighting to estimate a model for

quantile regression that is locally linear.[22]

Wang and Wang's proposal to modify the standard quantile loss function for random censoring by adapting the self-consistent Kaplan-Meier estimator is a significant breakthrough in the field. Their approach, which uses a local weighting scheme to redistribute the probability mass ($Pr(T_i>C_i \mid C_i, X_i)$) of censored cases to the right, is a powerful tool that can lead to more accurate estimators for $\beta(\theta)$. By adopting this technique and minimizing the objective $\beta(.)$, researchers can enhance the precision and reliability of their results. To obtain an estimator for $\beta(\theta)$, one can minimize the objective $\beta(.)$.

$$n^{-1}\sum_{i=1}^{n}\left[w_i\left(F_0\right)\rho_\theta\left(\tilde{Y}_i-X_i^Tb\right)+\left\{1-w_i\left(F_0\right)\right\}\rho_\theta\left(Y^*-X_i^Tb\right)\right]$$

$$F_0\left(t|x\right)\equiv Pr\left(T>t|X=x\right)$$

$$w_i\left(F_0\right)=\begin{cases}1 & F_0\left(C_i|X_i\right)>\theta \text{ or } \delta=1\\ \dfrac{\theta-F_0\left(C_i|X_i\right)}{1-F_0\left(C_i|X_i\right)} & F_0\left(C_i|X_i\right)<\theta \text{ or } \delta=0\end{cases}$$

In order to minimize the objective function, when $F_0(t|x)$ is unknown, Wang and Wang proposed replacing $F_0()$ with Beran's local Kaplan-Meier estimator and then minimizing the resulting function.

$$\hat{F}\left(t|x\right)=1-\prod_{j=1}^{n}\left\{1-\frac{B_{nj}\left(x\right)}{\sum_{k=1}^{n}I\left(\tilde{Y}_k>\tilde{Y}_j\right)B_{nk}\left(x\right)}\right\}^{N_j(t)}$$

Where $N\left(t\right)=I\left(\tilde{Y}\le t,\delta=1\right)$ and $B_{nk}(x)$ is a sequence of nonnegative weights that add up to 1.

$$B_{nk}\left(x\right)=\frac{K\left(\dfrac{x-x_k}{h_n}\right)}{\sum_{i=1}^{n}\dfrac{x-x_i}{h_n}}$$

$K()$ is a density kernel[(1)] function and h_n is a positive bandwidth[(2)] converging to 0 as $n\rightarrow\infty$. In order to handle random censoring in data with multivariate covariates, this method utilizes the concepts of redistribution of mass and effective dimension reduction. Asymptotically, this procedure achieves model selection consistency, meaning it can accurately identify the true model with a probability approaching one.[26]

This approach utilizes nonparametric estimation of $T's$ conditional distribution given X, with requisite weights, offering flexibility. Like Bang and Tsiatis's method, it implies covariate smoothing for modeling flexibility, regardless of initial parametric models. Two key assumptions underlie its validity: conditional independence of survival time and censoring given covariates, and linearity at the quantile of interest. In real data analysis, two limitations emerge: the curse of dimensionality hampers kernel smoothing with moderate covariates, and the method's design for continuous covariates challenges categorical variable cases, leading to an ill-defined situation.[27]

De Backer's method

Complete Data

Assume for every τ in the interval (0,1), $m_\tau(x)$

(1) In nonparametric statistics, kernels are weight functions used in estimation methods.
(2) Bandwidth is a parameter that controls the width of the kernel or smoothing function in techniques such as kernel density estimation or kernel regression.

represents the *τ-th* quantile of the distribution of a continuous dependent variable T, conditioned on $X = x$, where X is a vector of explanatory variables with at least one element and d dimensions.[25]

$$m_\tau = inf\left\{t : F_{T|X}\left(t|x\right) \geq \tau\right\}$$

The function $F_{T|X}$ represents the cumulative distribution function of T given X under certain conditions. With the utilization of Koenker and Bassett's alternative approach, which is achieved through targeted optimization:

$$m_\tau = \arg\min_a \mathbb{E}\left[\rho_\tau\left(T-a\right)\middle|X=x\right]$$

The "check" loss function(1), denoted as $\rho_\tau\left(x\right) = u\left(\tau - \mathbb{I}\left(u \leq 0\right)\right)$, is a powerful tool in statistical analysis. It optimizes the expected loss, improves accuracy, and can handle complex datasets efficiently. Incorporating it into your analysis can lead to enhanced results and more reliable conclusions.

Low-dimensional random variables.

This is a linear regression model that assumes:

$$m_\tau\left(X_i\right) = \beta_\tau^T X_i$$

for *i=1,...,n*, where X is a random vector of auxiliary variables with (d + 1) dimensions. The first element of X is set to 1, and β_τ is a vector of unknown coefficients with (d + 1) dimensions. For all *i=1,...,n*, it can be observed that $F_{T|X}\left(\beta_\tau^T X_i|X_i\right) = \tau$.

This is a natural extension of the technique of using the inverse cumulative distribution function to estimate β_τ:

$$\hat{\beta}_\tau = \arg\min_\beta \sum_{i=1}^{n}\left(\hat{F}\left(\beta_\tau^T X_i|X_i\right) - \tau\right)^2$$

In this scenario, $\hat{F}$ is a non-parametric estimator of $F_{T|X}$ that is appropriate and obtained through a 'double-kernel' method. It is estimated using $\hat{F}^s_{T|X}$, where $Y^u_{(i)}$ represents the i-th order statistic of the uncensored responses, and $n^u = \sum_{i=1}^{n}\delta_i$.

$$= \sum_{i=1}^{n^u}\left(\hat{F}_{T|X}\left(Y^u_{(i)}|x\right) - \hat{F}_{T|X}\left(Y^u_{(i-1)}|x\right)\right)H\left(\frac{t-Y^u_{(i)}}{h_T}\right)$$

In this context,

$$H\left(t\right) = \int_{-\infty}^{t}\tilde{k}\left(u\right)du$$

for some kernel density $\tilde{k}$,the positive bandwidth parameter is h_T, and $Y^u_{(0)} = 0$. Also, $\hat{F}_{T|X}$, is the local Kaplan-Meier Beran estimator (Beran (1981)). When defining the weighted sequence $B_{nj}\left(x\right)$, we add 1 as follows:

$$\hat{F}_{T|X}\left(t|x\right) =$$

$$1 - \prod_{i=1}^{n}\left\{1 - \frac{B_{nj}\left(x\right)}{\sum_{i=1}^{n}\mathbb{I}\left(Y_i \leq Y_J\right)B_{nj}\left(x\right)}\right\}^{\mathbb{I}\left(Y_i \leq t\ddot{a}_i=1\right)}$$

The Nadaraya-Watson type weights have been utilized in this instance:

$$B_{nj}\left(x\right) = \frac{K_d\left(\frac{x-X_j}{h_X}\right)}{\sum_{k=1}^{n}K_d\left(\frac{x-X_k}{h_X}\right)}, j = 1,\ldots,n$$

(1) The check loss function is utilized to define quantile regression, and it is also employed as a validation metric in cross-validation when the true distribution is unknown.

High-dimensional random variables.

When the dimension of the covariates becomes too, it becomes necessary to introduce an additional assumption in the model. In this regard, it is proposed to adopt a widely applicable global dimension reduction assumption, similar to the approach taken by Wang et al.[25, 26] The assumption of global dimension reduction posits that high-dimensional data can be effectively represented in a lower-dimensional space while preserving vital information. By capturing essential features and patterns, a lower-dimensional representation allows for simplification, visualization, and analysis of intricate datasets. The primary goal is to reduce dimensionality while minimizing the loss of crucial information.[28] According to this assumption, all the information regarding the dependence of T on X is effectively contained within q linear combinations of X. In other words, the relationship between T and X can be adequately captured by a set of q linear combinations. That is

$$T \perp X \left| \left(\gamma_{0,1}^T X, \ldots, \gamma_{0,q}^T X \right) \right.$$

In this context, "$\perp$ stands for independence, q < d + 1 for effective dimension reduction (EDR), and the γ_0 represent unknown (d+1)-dimensional linearly independent vectors."
In cases where auxiliary variables have a high dimension, we can make use of the following equation:

$$\hat{\beta}_\tau = \arg\min_{\beta} \sum_{i=1}^{n} \left(\hat{F}_{T|\hat{z}}^{s} \left(\beta^T X_i | \hat{Z}_i \right) - \tau \right)^2$$

Here, $\hat{Z}_i = \left(\hat{Z}_{i,1}, \ldots, \hat{Z}_{i,q} \right)^T$,

with $j=1,\ldots,q$ and $i=1,\ldots,n$, and $\hat{Z}_{i,j} = \hat{\gamma}_{0,j}^T X_i$. $\hat{F}_{T|\hat{Z}}^{s}$ is an estimator of the double kernel $F_{T|X}$, with the estimated images being replaced with X.

Simulation Scenarios

To compare the quantile regression approaches (check-based and minimum distance), we employ simulation. Figure 1 illustrates the simulation process method and various simulation scenarios. The simulations are based on a model that is frequently demonstrated in the literature by Wang and Wang, Leng and Tong, and De Backer et al.[18, 22, 24, 25]
In this simulation study, we explore different levels of censoring proportions for small-dimensional and multidimensional covariates. The study aims to simulate scenarios representative of domains with high levels of censoring, such as medical or survival analysis. The chosen censoring proportions are 40% and 30% for specific covariate types, while a 15% censoring proportion is used to examine the robustness of quantile regression estimators under moderate censoring levels. The focus is on two quantile levels, 0.3 and 0.5, with an emphasis on the median (0.5 quantile) for high-dimensional variables. These choices allow for a comprehensive understanding of the response variable's distribution, providing insights into the central tendency and aiding in risk assessment and decision-making processes. Analyzing these quantiles captures heterogeneity, uncovers valuable information, and identifies potential nonlinear or heterogeneous effects. In our study, we include two sample sizes: 100 and 200. These sample sizes are commonly used in various fields of study and offer a moderate representation of the population. They align with typical research practices and strike a balance between practicality and representativeness. The choice

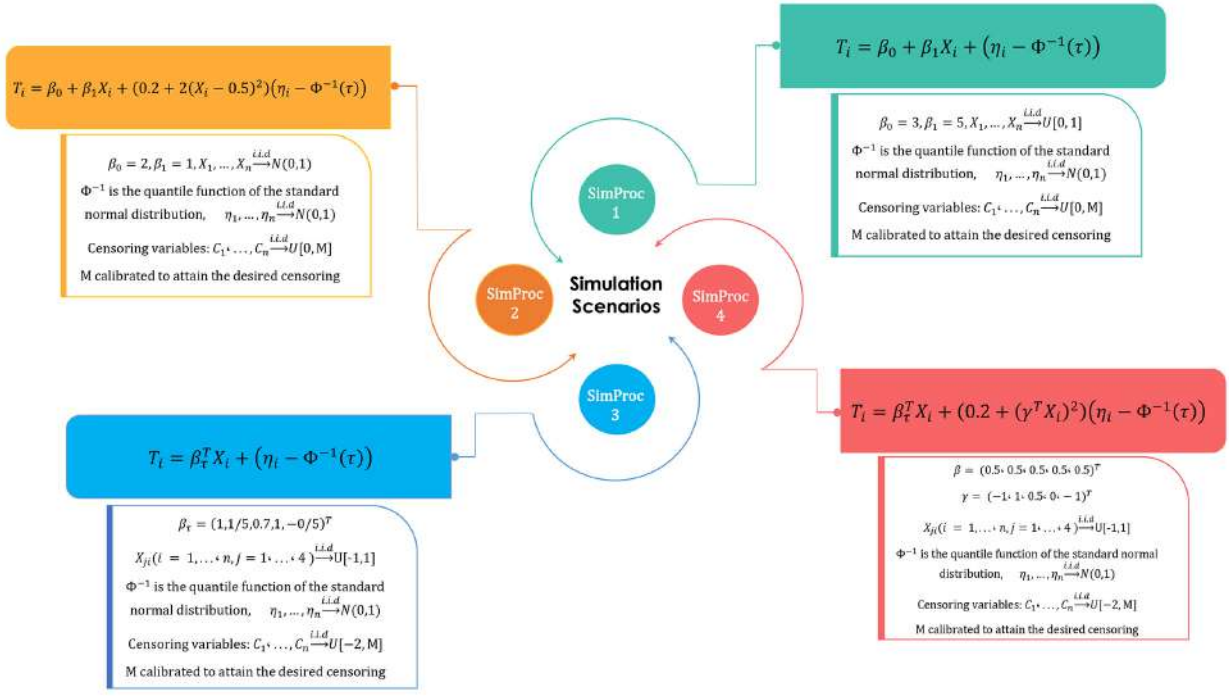

Figure 1. Scenario for simulation process

of these sample sizes also considers practical limitations in data collection, such as resource constraints. To compare the different estimators, we repeat each simulation process 500 times and assess their performance in terms of root mean squared errors (RMSE).

To ensure robust results about the optimization routine used in the estimation procedures, this study presents simulations by removing a few iterations for all estimators based on the settings that led to the worst mean absolute deviation between the estimated and true values (MAD) results. The MAD is defined for an estimator $n^{-1}\sum_{i=1}^{n}\left|\hat{\beta}^T X_i - \beta_\tau^T X_j\right|$.

By considering the worst MAD results, which indicate the highest deviations from the desired target, the study aims to identify and eliminate outliers or extreme values that may adversely impact the optimization process.

For complete data without censoring, the optimal check-based modeling of Koenker and Bassett ($\mathcal{O}_K$) and the minimum distance approach of De Backer ($\mathcal{O}_D$) are considered. The simulation $\mathcal{O}_K$ on uses candidate bandwidths for the De Backer and $\mathcal{O}_D$ estimators, with h_x ranging from 0.05 to 0.25 for SimPRoc1&2 and $h_x \epsilon\{0.2,0.25,\dots,0.7\}$ for SimPRoc3&4.

1. Model 1:
$$T_i = \beta_0 + \beta_1 X_i + \left(\eta_i - \Phi^{-1}(\tau)\right) \quad (\beta_0 = 3, \beta_1 = 5)$$

2. Model 2:
$$T_i = \beta_0 + \beta_1 X_i + (0.2 + 2(X_i - 0.5)^2)\left(\eta_i - \Phi^{-1}(\tau)\right) \quad (\beta_0 = 2, \beta_1 = 1)$$

Results

For the univariate analysis, two simulation process models were considered:

In both cases, terms were generated from a

uniform distribution *(0,1)* and a multivariate normal distribution, respectively. The censoring variable was generated from a uniform distribution *(0,M)*, where *M* was selected to achieve the desired censoring proportions.

For the multivariate analysis, two simulation process models were also considered:

1. Model 3:

$$T_i = \beta_\tau^T X_i + \left(\eta_i - \Phi^{-1}(\tau)\right);$$

$$\beta_\tau = (1,1.5,0.7,1,-0.5)^T$$

2. Model 4:

$$T_i = \beta_\tau^T X_i + (0.2 + (\gamma^T X_i)^2)\left(\eta_i - \Phi^{-1}(\tau)\right);$$

$$\beta = (0.5, 0.5, 0.5, 0.5, 0.5)^T,$$

$$\gamma = (-1, 1, 0.5, 0, -1)^T$$

which were conducted in both scenarios, X_{ji} *(i = 1,..., n, j=1, ..., 4*) are iid variables from *U[-1,1];* $\eta_1,\ldots,\eta_n$ are iid variables from *N(0,1)*, The censoring variables are also independent of auxiliary variables and are simulated from *U[-2,M]*.

The study explored various scenarios to analyze the effects of censoring and sample size, encompassing different total sample sizes, censoring rates, and quantile levels. Censoring levels in real-world datasets vary depending on the data and research field. Certain domains, like medical or survival analysis, often face high levels of censoring due to data collection processes or competing events.

Table 1 presents the RMSE of estimated β_0, and Table 2 displays RMSE values for estimated β_1, stemming from the analysis of *DGP1* and *DGP2*, which correspond to low-dimensional random variables.

Given the results in Table 2, regarding the coefficient β_1:In SimPRoc2 at quantile 0.3, with a sample size of 100 and 15% censoring proportions (P_C) the RMSE for complete data is 0.716 (Koenker ($\mathcal{O}_K$)) and 0.712 (De Backer ($\mathcal{O}_D$)). In censored data RMSE values for Bang-Tsiatis, Wang-Wang, and De Backer methods are 0.734, 0.720, and 0.718, respectively. With a sample size of 200, RMSE for complete data becomes 0.713 ($\mathcal{O}_K$)) and 0.709 ($\mathcal{O}_D$), and for censored data, RMSE values are 0.730, 0.718, and 0.715, respectively.

Increasing censorship to 40% with a sample size of 100 yields RMSE of 0.747 ($\mathcal{O}_K$) and 0.736 ($\mathcal{O}_D$) for complete data, and 0.778, 0.757, and 0.745 for censored data. Overall, RMSE rises with higher censorship.

When raising the quantile level to 0.5, (n=100, P_C=15%) in SimPRoc2, the RMSE for estimating the coefficient β_1 compared to the quantile level of 0.3 demonstrates reductions as follows: in complete data, $\mathcal{O}_D$ displays an approximately 9.4% decrease, $\mathcal{O}_K$ shows around 6.1%, and for censored data, the Bang-Tsiatis, Wang-Wang, and De Backer methods showcase reductions of about 5.6%, 6.3%, and 6.1%, respectively.

Figure 2 displays a boxplot of coefficient residuals for SimPRoc1, utilizing complete and censored data, with a sample size of 100, a quantile level of 0.5, and a censoring percentage of 15%. The findings notably indicate the consistent superiority of De Backer's method in terms of coefficient estimation for both complete and censored data situations.

The results are further compared for different scenarios in figure 3-6.

The findings indicate that the De Backer estimator performs slightly better than its

Table 1. The simulation results in Simulation Process 1 and 2, considering both censored and complete observations to compare the root mean square error of β_0 estimates with the initial value β_0=3.

SimProc	n	P_c	τ	Method				
				$\mathcal{O}_D$	$\mathcal{O}_K$	Bang & Tsiatis	De Backer	Wang & Wang
1	100	15%	0.3	0.763	0.768	0.794	0.772	0.783
			0.5	0.760	0.762	0.779	0.764	0.764
		40%	0.3	0.778	0.782	0.789	0.779	0.787
			0.5	0.768	0.770	0.793	0.778	0.785
	200	15%	0.3	0.762	0.767	0.778	0.770	0.771
			0.5	0.74	0.745	0.769	0.747	0.749
		40%	0.3	0.775	0.780	0.788	0.760	0.78
			0.5	0.745	0.752	0.787	0.750	0.754
2	100	15%	0.3	0.71	0.715	0.759	0.719	0.742
			0.5	0.701	0.713	0.722	0.709	0.715
		40%	0.3	0.718	0.721	0.757	0.723	0.748
			0.5	0.715	0.719	0.744	0.711	0.72
	200	15%	0.3	0.712	0.705	0.749	0.713	0.740
			0.5	0.611	0.622	0.651	0.620	0.622
		40%	0.3	0.716	0.715	0.746	0.722	0.742
			0.5	0.658	0.678	0.668	0.624	0.662

$\mathcal{O}_K$, Optimal check-based modeling of Koenker and Bassett for complete data without censoring ;
$\mathcal{O}_D$, The minimum distance approach of De Backer for complete data without censoring;
n , The sample size; P_c, Censoring proportions; τ , Quantile levels; SimProc, Simulation process;

Table 2. The simulation results in Simulation Process 1 and 2 based on both censored and complete observations to compare the root mean square error of β_1 estimates with the initial value β_1=5.

SimProc	n	P_c	τ	Method				
				$\mathcal{O}_D$	$\mathcal{O}_K$	Bang & Tsiatis	De Backer	Wang & Wang
1	100	15%	0.3	0.760	0.762	0.769	0.764	0.764
			0.5	0.744	0.753	0.755	0.732	0.744
		40%	0.3	0.791	0.814	0.823	0.774	0.771
			0.5	0.752	0.784	0.793	0.746	0.761
	200	15%	0.3	0.751	0.758	0.774	0.753	0.755
			0.5	0.737	0.750	0.761	0.752	0.752
		40%	0.3	0.781	0.811	0.781	0.769	0.766
			0.5	0.747	0.772	0.780	0.746	0.758
2	100	15%	0.3	0.712	0.716	0.734	0.718	0.720
			0.5	0.618	0.655	0.678	0.657	0.657
		40%	0.3	0.736	0.747	0.778	0.745	0.757
			0.5	0.702	0.705	0.748	0.712	0.718
	200	15%	0.3	0.709	0.713	0.730	0.715	0.718
			0.5	0.616	0.649	0.669	0.638	0.641
		40%	0.3	0.738	0.745	0.753	0.742	0.750
			0.5	0.700	0.703	0.721	0.701	0.714

$\mathcal{O}_K$, Optimal check-based modeling of Koenker and Bassett for complete data without censoring ;
$\mathcal{O}_D$, The minimum distance approach of De Backer for complete data without censoring;
n , The sample size; P_c, Censoring proportions; τ , Quantile levels; SimProc, Simulation process;

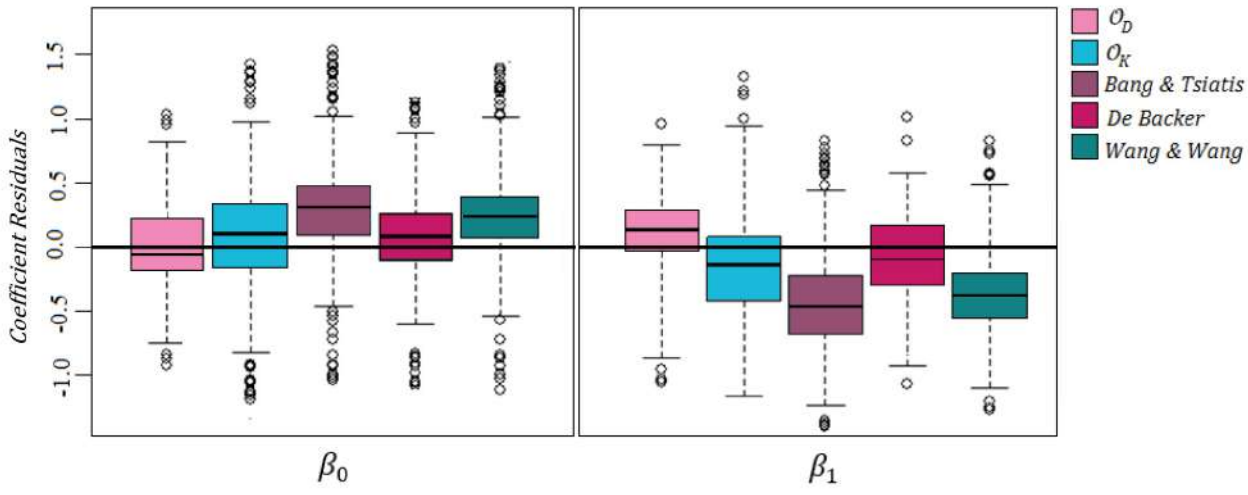

Figure 2: Comparison of residuals in estimated coefficients using complete and censored data approaches for simulation process 1 with $\tau = 0.5$, $n = 100$, and $P_c = 15\%$

$\mathcal{O}_K$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\mathcal{O}_D$ is the minimum distance approach of De Backer for complete data without censoring
n is the ample size; P_c is censoring proportions; τ is quantile levels;

competitors in terms of RMSE for both SimPRoc1 and SimPRoc2, especially for small censoring proportions.

This study focuses on two multivariate challenges: SimPRoc3 and SimPRoc4. For brevity, our analysis specifically targets the median. In SimPRoc3, the true model is linear across all quantile levels, while in SimPRoc4, linearity is confined to the τ-th quantile of interest.

Table 3 indicates that, for median response variable estimation with a sample size of 100 and a 15% censoring proportions using SimPRoc3, the $\mathcal{O}_D$ method reduces RMSE by around 1.45% compared to $\mathcal{O}_K$ for complete data, and by about 2.57% and 0.4% De Backer's method compared to Bang-Tsiatis and Wang-Wang for censored data. With a sample size of 200, RMSE decreases by around 0.01% on average, and with a 30% censoring proportions, it increases by about 0.03%.

Furthermore, Figure 7 provides a comparative evaluation of the RMSE concerning the estimated response variable, imparting valuable insights into the methods under review.

Our study highlights that the De Backer procedure exhibits comparable RMSE performance when compared to check-based estimators, particularly at the central quantile level of 0.5. Interestingly, we observe enhancements in performance when handling low levels of censoring. Notably, even in situations with complete data and no censoring, the De Backer minimum distance approach ($\mathcal{O}_D$) can outperform Koenker and Bassett's optimal check-based modeling ($\mathcal{O}_K$) at different quantile levels.

Discussion

This study compares two methods for censored data quantile linear regression: check-based

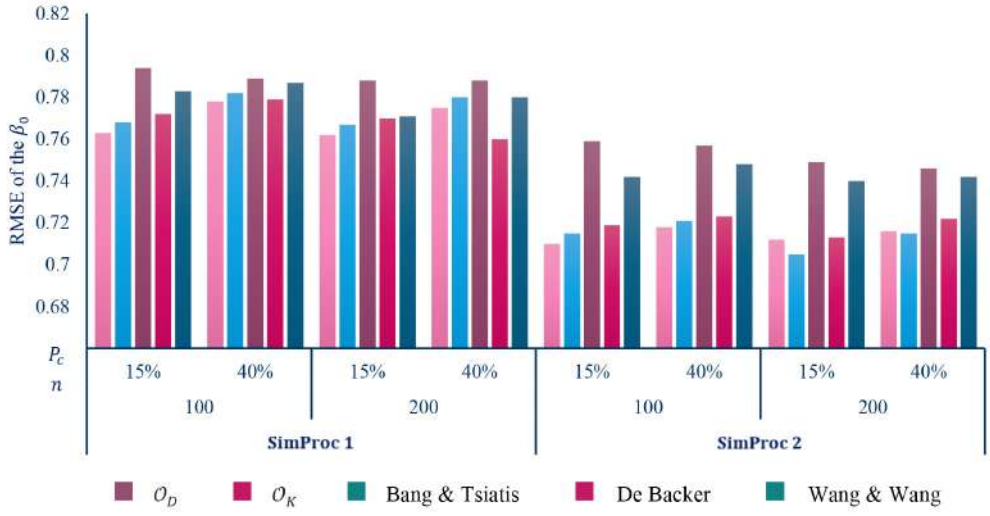

Figure 3: Comparing root mean squared errors of β_0 for simulation process 1 and 2 for both complete and censored data with $n = \{100,200\}, \tau = 0.3, P_c = \{15\%, 30\%\}$

$\boldsymbol{\mathcal{O}_K}$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\boldsymbol{\mathcal{O}_D}$ is the minimum distance approach of De Backer for complete data without censoring
$\boldsymbol{n}$ is the ample size; $\boldsymbol{P_c}$ is censoring proportions; $\boldsymbol{\tau}$ is quantile levels; $\boldsymbol{RMSE}$ is root mean squared errors; ; **SimProc** is simulation process

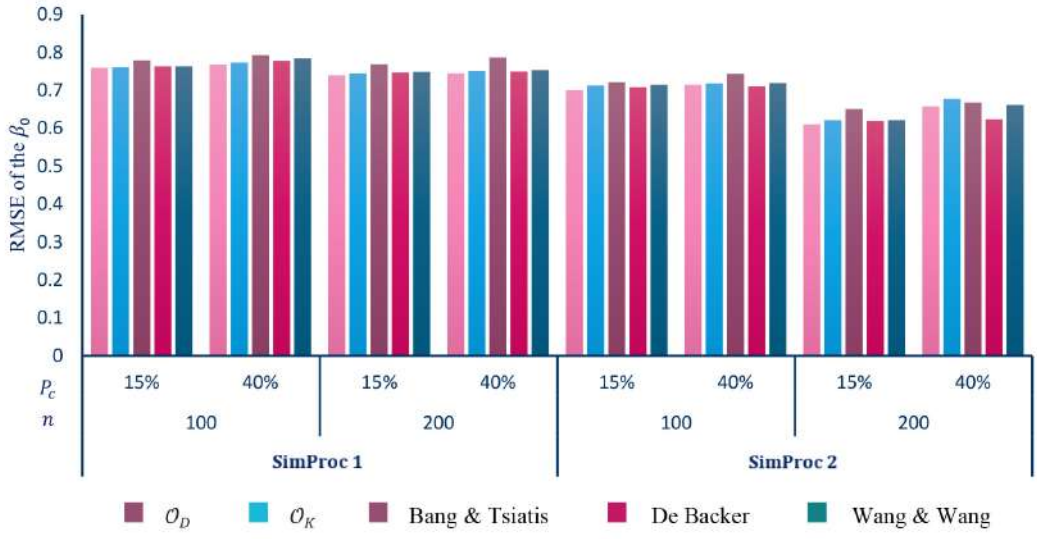

Figure 4: Comparing root mean squared errors of β_0 for simulation process 1 and 2 for both complete and censored data with $n = \{100,200\}, \tau = 0.5, P_c = \{15\%, 30\%\}$

$\boldsymbol{\mathcal{O}_K}$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\boldsymbol{\mathcal{O}_D}$ is the minimum distance approach of De Backer for complete data without censoring
$\boldsymbol{n}$ is the ample size; $\boldsymbol{P_c}$ is censoring proportions; $\boldsymbol{\tau}$ is quantile levels; $\boldsymbol{RMSE}$ is root mean squared errors; ; **SimProc** is simulation process

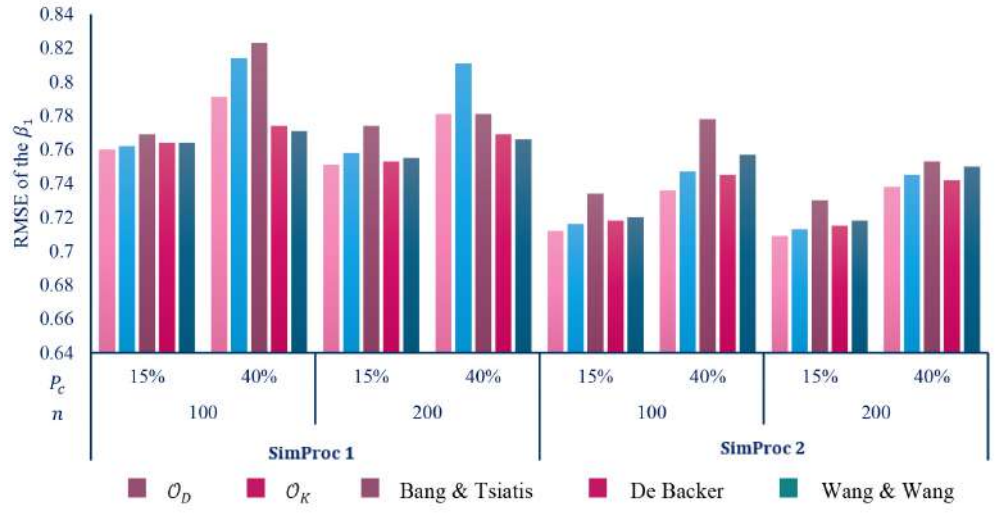

Figure 5: Comparing root mean squared errors of β_1 for simulation process 1 and 2 for both complete and censored data with $n = \{100,200\}, \tau = 0.3, P_c = \{15\%, 30\%\}$

$\boldsymbol{\mathcal{O}_K}$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\boldsymbol{\mathcal{O}_D}$ is the minimum distance approach of De Backer for complete data without censoring
$\boldsymbol{n}$ is the ample size; $\boldsymbol{P_c}$ is censoring proportions; $\boldsymbol{\tau}$ is quantile levels; $\boldsymbol{RMSE}$ is root mean squared errors; ; **SimProc** is simulation process

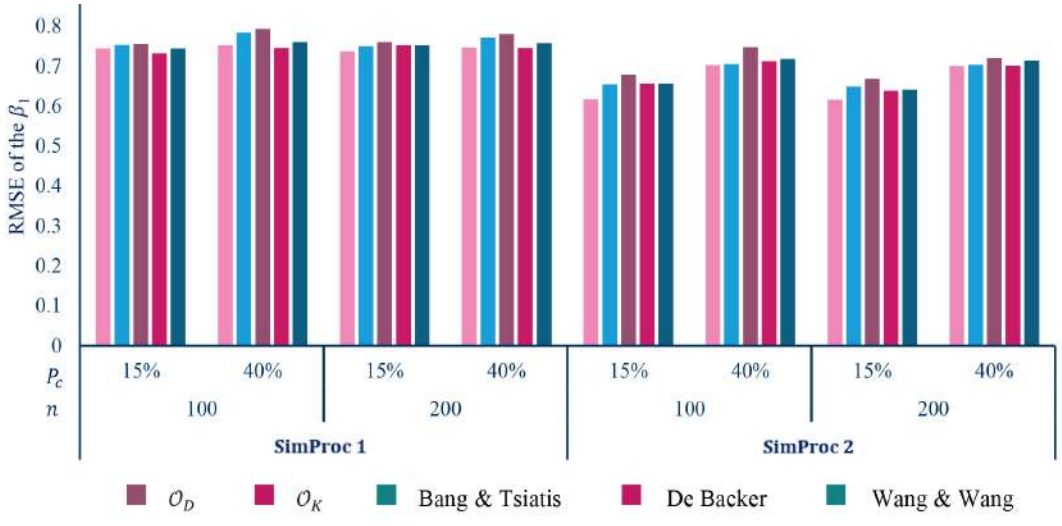

Figure 6: Comparing root mean squared errors of β_1 for simulation process 1 and 2 for both complete and censored data with $\mathrm{n} = \{100,200\}, \tau = 0.5, \mathrm{P_c} = \{15\%, 30\%\}$

$\boldsymbol{\mathcal{O}_K}$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\boldsymbol{\mathcal{O}_D}$ is the minimum distance approach of De Backer for complete data without censoring
$\boldsymbol{n}$ is the ample size; $\boldsymbol{P_c}$ is censoring proportions; $\boldsymbol{\tau}$ is quantile levels; $\boldsymbol{RMSE}$ is root mean squared errors; ; **SimProc** is simulation process

Table 3. The simulation results in Simulation Process 3 and 4 for median based on both censored and complete observations to compare the root mean square error of Response variable estimates.

SimProc	n	P_c	Method				
			$\mathcal{O}_D$	$\mathcal{O}_K$	Bang & Tsiatis	De Backer	Wang & Wang
3	100	15%	0.739	0.740	0.776	0.755	0.756
		30%	0.766	0.777	0.801	0.790	0.793
	200	15%	0.719	0.724	0.773	0.748	0.752
		30%	0.714	0.736	0.763	0.756	0.758
4	100	15%	0.716	0.725	0.744	0.720	0.724
		30%	0.725	0.749	0.779	0.751	0.759
	200	15%	0.644	0.658	0.674	0.660	0.661
		30%	0.678	0.679	0.697	0.673	0.674

$\mathcal{O}_K$, Optimal check-based modeling of Koenker and Bassett for complete data without censoring ;
$\mathcal{O}_D$, The minimum distance approach of De Backer for complete data without censoring;
n , The sample size; P_c, Censoring proportions; SimProc, Simulation process;

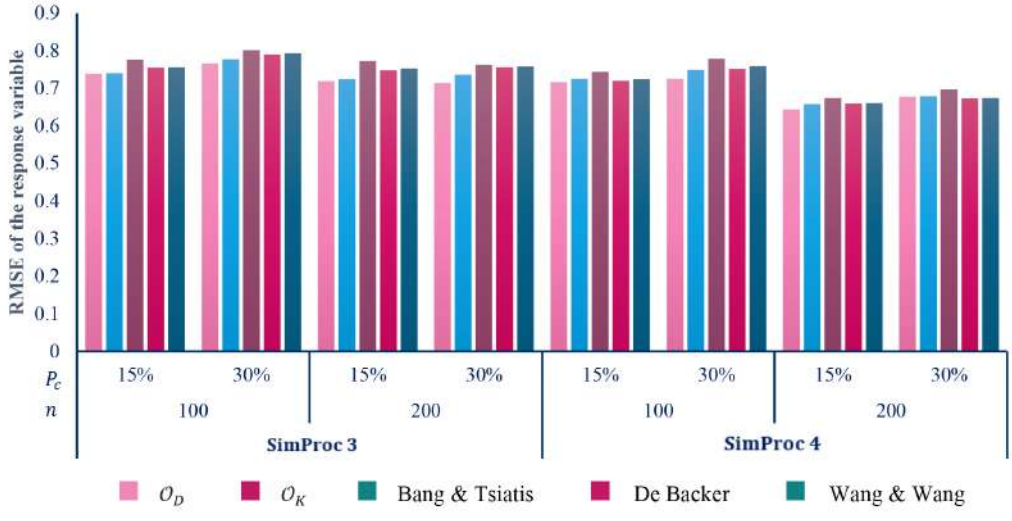

Figure 7: Comparing root mean squared errors of the response variable for simulation process 3 and 4 for both complete and censored data with $n = \{100,200\}, \tau = 0.5, P_c = \{15\%, 30\%\}$

$\boldsymbol{\mathcal{O}_K}$ is optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\boldsymbol{\mathcal{O}_D}$ is the minimum distance approach of De Backer for complete data without censoring
$\boldsymbol{n}$ is the ample size; $\boldsymbol{P_c}$ is censoring proportions; $\boldsymbol{\tau}$ is quantile levels; $\boldsymbol{RMSE}$ is root mean squared errors; ; **SimProc** is simulation process

modeling and the minimum distance approach. While both are parametric, check-based modeling uses nonparametric estimators for censored data, the minimum distance approach employs the inverse cumulative distribution function. The latter incorporates a double-kernel estimator and dimension reduction for multivariate covariates. Simulation results reveal similar model performance, yet the minimum distance approach consistently outperforms check-based methods with lower root mean square error across scenarios. Larger sample sizes decrease error, but higher censorship rates increase bias in coefficient estimates. Also, Increasing the sample size, even when accounting for a higher censorship percentage, results in a decrease in RMSE.

Lately, quantile regression has attracted substantial attention, leading to numerous studies conducted on this subject, some of which we highlight.

De backer et al. in their study indicated in an extensive simulation study that the resulting quantile regression estimator with respect to established check-based formulations has fewer variance results[25] Yazdani et al. compared five quantile regression methods for right-censored data, focusing on breast cancer patient survival. CQR consistently revealed prognostic factors, with coefficients similar below 0.1 quantiles and variations above.[29] In a recent study by Conde-Amboage, Was shown quantile regression's effectiveness in addressing complex biomedical questions.[30] Tedesco and Van Keilegom introduce a method to compare conditional quantile curves under right censoring, applicable to diverse data types. Validation with diabetic retinopathy data demonstrates its higher power across a range of quantile levels.[31] Rodrigues et al. innovatively combines quantile regression with an exponentiated odd log-logistic Weibull distribution, enhancing data analysis by addressing quantile estimation and distribution modeling.[32]

Beyhum introduces a semiparametric quantile regression model to handle endogeneity and random right censoring. It utilizes instrumental variables for this purpose, demonstrating effectiveness through analysis of the national Job Training Partnership Act study with robust performance and low estimator bias.[33]

Innovative quantile regression for discrete responses that introduce by Geraci, employs interpolation and a two-step estimator for conditional mid-quantiles and regression coefficients. The method proves strongly consistent and asymptotically normal, outperforming alternatives in simulations.[34] HE et al. proposes a smoothed martingale-based sequential estimating equations approach with scalable algorithms for enhanced performance in high-dimensional sparse settings. Simulations show improved results by relaxing the exponential sparsity term in existing CQR work.[35]

Conclusion

The simulation results show that all models perform similarly, but the minimum distance approach outperforms check-based models in all scenarios with a lower root mean square error. As the sample size increases, the root means square error decreases in all models. However, increasing the censorship rate leads to higher bias in the regression coefficient estimates.

Overall, the findings indicate that the minimum distance approach is a preferable option

for predicting results in various scenarios, particularly when precise parameter estimates are needed, despite its higher computational intensity compared to other methods. However, it is worth noting that check-based methods offer advantages in specific contexts, such as computational efficiency and ease of implementation. The selection between the minimum distance approach and check-based methods should consider the research question, data characteristics, and available resources. Further investigation into scenarios where check-based methods excel would contribute to a better understanding and aid researchers in making informed decisions.

In future studies, enhancing the method's versatility to handle left-censored data could broaden its applicability. Also, adapting the minimum distance approach to incorporate nonlinear quantile regression models introduces exciting avenues for analysis. This extension has the potential to offer novel insights and applications by capturing intricate variable relationships.

Conflict of Interests

The authors declare no conflict interests.

Abbreviations

AFT, Accelerated Failure Time
ICP , Inverse Censoring Probability
MAD, Mean Absolute Deviation
LCQR, Linear Censored Quantile Regression
RMSE , Root Mean Squared Errors
SimPRoc, Simulation Process
$\mathcal{O}_K$, Optimal check-based modeling of Koenker and Bassett for complete data without censoring
$\mathcal{O}_D$, Minimum distance approach of De Backer for complete data without censoring

References

1. Amuche I, George O, Edith U. Comparison of Parametric Models: Appication to Hypertensive Patients in a Teaching Hospital, Awka. Journal of Biostatistics and Epidemiology. 2020;5(2).

2. Lin D. Cox regression analysis of multivariate failure time data: the marginal approach. Statistics in medicine. 1994;13(21):2233-47.

3. Kleinbaum DG, Klein M. Survival analysis a self-learning text: Springer; 1996.

4. Alireza A, Bagher P, Farid Z, Taban B. Interpretation of exposure effect in competing risks setting under accelerated failure time models. Journal of Biostatistics and Epidemiology. 2018;4(2).

5. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data: Springer; 2003.

6. Koenker R. Quantile regression: Cambridge university press; 2005.

7. Koenker R, Bassett Jr G. Regression quantiles. Econometrica: journal of the Econometric Society. 1978:33-50.

8. Portnoy S. Censored regression quantiles. Journal of the American Statistical Association. 2003;98(464):1001-12.

9. Koenker R, Geling O. Reappraising medfly longevity: a quantile regression survival analysis. Journal of the American Statistical Association. 2001;96(454):458-68.

10. Koenker R, Bilias Y. Quantile regression for duration data: A reappraisal of the Pennsylvania reemployment bonus experiments. Economic applications of quantile regression: Springer; 2002. p. 199-220.

11. Powell JL. Censored regression quantiles. Journal of econometrics. 1986;32(1):143-55.

12. Ronchetti E, Field C, Blanchard W. Robust Linear Model Selection by Cross-Validation. Journal of the American Statistical Association. 1997;92(439):1017-23.

13. Koenker R, Chesher A, Society E, Jackson M. Quantile Regression: Cambridge University Press; 2005.

14. Lee Y, MacEachern SN, Jung Y. Regularization of Case-Specific Parameters for Robustness and Efficiency. Statistical Science. 2012;27(3):350-72, 23.

15. Koul H, Susarla V, Van Ryzin J. Regression analysis with randomly right-censored data. The Annals of statistics. 1981:1276-88.

16. Ying Z, Jung S-H, Wei L-J. Survival analysis with median regression models. Journal of the American Statistical Association. 1995;90(429):178-84.

17. Bang H, Tsiatis AA. Median regression with censored cost data. Biometrics. 2002;58(3):643-9.

18. Leng C, Tong X. A quantile regression estimator for censored data. Bernoulli. 2013;19(1):344-61.

19. Zhou L. A simple censored median regression estimator. Statistica Sinica. 2006:1043-58.

20. Shows JH, Lu W, Zhang HH. Sparse estimation and inference for censored median regression. Journal of statistical planning and inference. 2010;140(7):1903-17.

21. Gorfine M, Goldberg Y, Ritov Ya. A quantile regression model for failure-time data with time-dependent covariates. Biostatistics. 2017;18(1):132-46.

22. Wang HJ, Wang L. Locally weighted censored quantile regression. Journal of the American Statistical Association. 2009;104(487):1117-28.

23. Lindgren A. Quantile regression with censored data using generalized L1 minimization. Computational Statistics & Data Analysis. 1997;23(4):509-24.

24. De Backer M, Ghouch AE, Van Keilegom I. An adapted loss function for censored quantile regression. Journal of the American Statistical Association. 2019;114(527):1126-37.

25. De Backer M, El Ghouch A, Van Keilegom I. Linear censored quantile

regression: A novel minimum-distance approach. Scandinavian Journal of Statistics. 2020;47(4):1275-306.

26. Wang HJ, Zhou J, Li Y. Variable selection for censored quantile regresion. Statistica Sinica. 2013;23(1):145.

27. Wey A, Wang L, Rudser K. Censored quantile regression with recursive partitioning-based weights. Biostatistics. 2014;15(1):170-81.

28. Koenig-Archibugi M. Understanding the global dimensions of policy. Global Policy. 2010;1(1):16-28.

29. Yazdani A, Yaseri M, Haghighat S, Kaviani A, Zeraati H. The comparison of censored quantile regression methods in prognosis factors of breast cancer survival. Scientific Reports. 2021;11(1):18268.

30. Conde-Amboage M, Keilegom IV, González-Manteiga W. Application of Quantile Regression Models for Biomedical Data. Statistical Methods at the Forefront of Biomedical Advances: Springer; 2023. p. 83-113.

31. Tedesco L, Van Keilegom I. Comparison of quantile regression curves with censored data. Test. 2023:1-36.

32. Rodrigues GM, Ortega EM, Cordeiro GM, Vila R. Quantile Regression with a New Exponentiated Odd Log-Logistic Weibull Distribution. Mathematics. 2023;11(6):1518.

33. Beyhum J, Tedesco L, Van Keilegom I. Instrumental variable quantile regression under random right censoring. arXiv preprint arXiv:220901429. 2022.

34. Geraci M, Farcomeni A. Mid-quantile regression for discrete responses. Statistical Methods in Medical Research. 2022;31(5):821-38.

35. He X, Pan X, Tan KM, Zhou W-X. Scalable estimation and inference for censored quantile regression process. The Annals of Statistics. 2022;50(5):2899-924, 26.