Published in final edited form as:
J Am Stat Assoc. 2020 ; 115(530): 810–821. doi:10.1080/01621459.2019.1585252.

Regression Analysis of Doubly Truncated Data

Zhiliang Ying*,
Department of Statistics, Columbia University, New York

Wen Yu,
Department of Statistics, School of Management, Fudan University, Shanghai, China

Ziqiang Zhao,
Novartis Pharmaceuticals

Ming Zheng
Department of Statistics, School of Management, Fudan University, Shanghai, China

Abstract

Doubly truncated data are found in astronomy, econometrics and survival analysis literature. They arise when each observation is confined to an interval, i.e., only those which fall within their respective intervals are observed along with the intervals. Unlike the one-sided truncation that can be handled by counting process-based approach, doubly truncated data are much more difficult to handle. In their analysis of an astronomical data set, Efron and Petrosian (1999) proposed some nonparametric methods for doubly truncated data. Motivated by their approach, as well as by the work of Bhattacharya et al. (1983) for right truncated data, we propose a general method for estimating the regression parameter when the dependent variable is subject to the double truncation. It extends the Mann-Whitney-type rank estimator and can be computed easily by existing software packages. Weighted rank estimation are also considered for improving estimation efficiency. We show that the resulting estimators are consistent and asymptotically normal. Resampling schemes are proposed with large sample justification for approximating the limiting distributions. The quasar data in Efron and Petrosian (1999) and an AIDS incubation data are analyzed by the new method. Simulation results show that the proposed method works well.

Keywords

Confidence interval; Empirical process; L_1 method; Linear programming; Rank estimation; Resampling; Wilcoxon-Mann-Whitney Statistic; U-process

1 Introduction

In their analysis of quasar data, Efron and Petrosian (1999) proposed nonparametric methods for doubly truncated data. Their methods deal with two common statistical issues: 1. testing independence between the explanatory variable and the dependent variable when the latter is subject to the double truncation; 2. estimating nonparametrically the marginal distribution of

*Contact: zying@stat.columbia.edu.

the response variable when the independence is true. For the first issue, they constructed an extension of Kendall's tau that corrects for possible bias due to the truncation. For the second issue, they applied the nonparametric EM algorithm to obtain a self-consistent estimator.

The existing literature contains many nonparametric methods for dealing with truncated data. Turnbull (1976) developed a general algorithm for finding the nonparametric maximum likelihood estimator of distribution for arbitrarily grouped, censored and truncated data. This estimator was obtained earlier by Lynden-Bell (1971) for singly truncated data. The large sample properties of Lynden-Bell's estimator were established by Woodroofe (1985). Wang, Jewell, and Tsai (1986), Keiding and Gill (1990) and Lai and Ying (1991a) applied the counting process-martingale techniques.

There is a substantial literature on regression analysis with the response variable subject to right or left truncation. Motivated from an application in astronomy, Bhattacharya, Chernoff, and Yang (1983) formulated the relationship between luminosity and red shift as a linear regression model in which the response variable is subject to right truncation. They extended the Mann-Whitney estimating function with a modification to correct for possible bias due to the truncation, and showed that their estimator is consistent and asymptotically normal. Tsui, Jewell, and Wu (1988) developed an iterative bias adjustment technique to estimate the regression parameter in the linear regression model. Tsai (1990) made use of Kendall's tau to construct tests for independence between the response and the explanatory variables. Lai and Ying (1991b) constructed a semiparametrically efficient estimator using rank based estimating functions. For modeling and analysis of truncated data in the econometrics literature, see Amemiya (1985) and Greene (2012), and references therein. For general biased sampling that contains truncation as special cases, we refer to recent works of Kim et al. (2013) and Liu et al. (2016).

Compared with singly truncated data, doubly truncated data are technically more challenging to deal with. Few results have been obtained for doubly truncated data due to lack of explicitly expressed estimating functions. Similar difficulties also arise for doubly censored data. Chang and Yang (1987) and Gu and Zhang (1993) discussed nonparametric estimators based on doubly censored data and established their asymptotic properties. Semiparametric regression M-estimators with doubly censored responses were studied by Ren and Gu (1997). For doubly truncated data, based on Efron and Petrosian (1999), Shen (2010) developed a nonparametric maximum likelihood estimator for the truncated variable distribution, and investigated the related asymptotic porperties. Moreira and Alvarez (2012) proposed a kernel-type density estimation approach and explored its asymptotic behavior. Bilker and Wang (1996) and Shen (2013) extended the two-sample Mann-Whitney test, with parametric modeling of the truncation variables. For regression with doubly truncated data, Shen (2013) considered semiparametric transformation models and used nonparametric EM algorithm as in Efron and Petrosian (1999) to obtain regression parameter estimation. Nonparametric regression analysis was considerd by Moreira and Alvarez (2016) by using a kernel-type approach.

This paper proposes a general approach to estimating the regression parameter in the linear regression model when the response variable is subject to the double truncation. An extended Mann-Whitney-type loss function is introduced that takes into consideration of the double truncation. A Mann-Whitney-type rank estimator is then defined as its minimizer. The minimization can be carried out easily and efficiently using existing software packages. A random perturbation approach is proposed for variance estimation and distributional approximation. By applying the large sample theory for U-processes, a quadratic approximation is developed for the loss function and, as a consequence, the usual asymptotic properties are established for the proposed estimator. Large sample justification for the random perturbation approach is also given. Extensive simulation results are reported to assess the finite sample performance of the proposed method. The method is applied to two real data sets. Additionally, extension to weighted Mann-Whitney-type pairwise comparisons that may improve efficiency is proposed.

The rest of the paper is organized as follows. The next section introduces some basic notation and defines the doubly truncated linear regression which is the focus of this paper. In Section 3, we introduce an extension of the Mann-Whitney-type objective function for regression parameter estimation that adjusts for double truncation, followed by a weighting scheme to improve efficiency. The usual large sample properties of the proposed method are established in Section 4. Sections 5 and 6 are devoted to simulation results and analyses of two real data sets, the quasar data and AIDS incubation data, respectively. Some concluding remarks are given in Section 7. The technical developments are summarized in the Appendix.

2 Notation and model specification

Consider the linear regression model

$$\tilde{Y} = \beta^{\top}\tilde{X} + \tilde{\varepsilon}, \tag{1}$$

where $\tilde{Y}$ is the response variable, $\tilde{X}$ the p-dimensional covariate vector with β the corresponding regression parameter vector and $\tilde{\varepsilon}$ the error term that is independent of covariates. This model becomes much more complicated when the response variable $\tilde{Y}$ is subject to double truncation. Specifically, let $\tilde{L}$ and $\tilde{R}$ denote the left and right truncation variables. The response $\tilde{Y}$, the truncation pair $(\tilde{L}, \tilde{R})$ and covariates $\tilde{X}$ are observed if and only if $\tilde{L} < \tilde{Y} < \tilde{R}$. Throughout this paper, we will make the usual (conditionally) independent truncation assumption: $\tilde{Y}$ and $(\tilde{L}, \tilde{R})$ are conditionally independent given $\tilde{X}$ or, equivalently, $\tilde{\varepsilon}$ is independent of $(\tilde{X}, \tilde{L}, \tilde{R})$. We will use f and F to denote respectively the density and distribution functions of $\tilde{\varepsilon}$.

Let $\tilde{Z} = \left(\tilde{Y}, \tilde{X}^{\top}, \tilde{L}, \tilde{R}\right)^{\top}$ and $\tilde{Z}_1, \ldots, \tilde{Z}_{\tilde{n}}$ be $\tilde{n}$ independent and identically distributed (i.i.d.) copies of $\tilde{Z}$. Because of truncation, for each i, $\tilde{Z}_i$ is observed if and only if $\tilde{L}_i < \tilde{Y}_i < \tilde{R}_i$. Let $n = \#\left\{i : \tilde{L}_i < \tilde{Y}_i < \tilde{R}_i\right\}$, the number of observations. Furthermore, let

$Z_i = \left(Y_i, X_i^\top, L_i, R_i\right)^\top$, $i = 1, \ldots, n$, be the observed $\widetilde{Z}_i$'s with ε_i the corresponding error terms.

There are two approaches to formulating the truncation data. The first one, as being used here, is from the missing data viewpoint with $\widetilde{Z}_i$, $i = 1, \ldots, \tilde{n}$ as the complete data. The second one is to directly model the observed data, i.e. to assume that Z_i, $i = 1, \ldots, n$ are i.i.d. observations with joint density

$$\frac{f\left(Y_i - \beta^\top X_i\right)}{F\left(R_i - \beta^\top X_i\right) - F\left(L_i - \beta^\top X_i\right)} h(L_i, R_i, X_i),\ L_i < Y_i < R_i, \tag{2}$$

where h is the joint density of $\left(L_i, R_i, X_i^\top\right)^\top$. We assume that the truncation variables and the covariates are ancillary, that is, h does not depend on β. It can be shown that these two approaches are equivalent. In the next section, we use the first formulation to motivate our estimator and the second one for rigorous justification.

The following notation is used throughout the rest. For each $i = 1, \ldots, n$, $L_i(\beta) = L_i - \beta^\top X_i$, $R_i(\beta) = R_i - \beta^\top X_i$ and $e_i(\beta) = Y_i - \beta^\top X_i$. Likewise, $\widetilde{L}_i(\beta) = \widetilde{L}_i - \beta^\top \widetilde{X}_i$, $\widetilde{R}_i(\beta) = \widetilde{R}_i - \beta^\top \widetilde{X}_i$ and $\tilde{e}_i(\beta) = \widetilde{Y}_i - \beta^\top \widetilde{X}_i$, $i = 1, \ldots, \tilde{n}$.

3 Methods

3.1 Main idea

We are concerned with inference about the regression parameter β. If $\widetilde{Z}_1, \ldots, \widetilde{Z}_{\tilde{n}}$ were observed, one could use the following Mann-Whitney-type estimating equation (Jin, Ying, and Wei, 2001)

$$\widetilde{U}_{\tilde{n}}(\beta) = \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left(\widetilde{X}_i - \widetilde{X}_j\right) \mathrm{sgn}\left\{\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right\} = 0, \tag{3}$$

where sgn $\{\cdot\}$ is the sign function. This estimating function is unbiased since, by symmetry, $\mathrm{E}\left(\mathrm{sgn}\left\{\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right\} \mid \widetilde{X}_i, \widetilde{X}_j\right) = 0$ when β takes the true value. Under the double truncation, only those $\tilde{e}_i(\beta)$ satisfying $\widetilde{L}_i(\beta) < \tilde{e}_i(\beta) < \widetilde{R}_i(\beta)$ are observed. $\widetilde{U}_{\tilde{n}}(\beta)$ would be biased if the summation on the right-hand-side of (3) only include those observed pairs. However, this bias can be corrected if we impose an artificial symmetrical truncation with further restriction $\widetilde{L}_j(\beta) < \tilde{e}_i(\beta) < \widetilde{R}_j(\beta)$. To this end, we define

$$U_n(\beta) = \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} I\left\{\widetilde{L}_i(\beta) \vee \widetilde{L}_j(\beta) < \tilde{e}_i(\beta) < \widetilde{R}_i(\beta) \wedge \widetilde{R}_j(\beta), \widetilde{L}_i(\beta) \vee \widetilde{L}_j(\beta) < \tilde{e}_j(\beta) < \widetilde{R}_i(\beta) \wedge \widetilde{R}_j(\beta)\right\}$$
$$\times \left(\widetilde{X}_i - \widetilde{X}_j\right) \mathrm{sgn}\left\{\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right\},$$

where $I\{\cdot\}$ is the indicator function and $\wedge$ $(\vee)$ is the minimum (maximum) operator. Again, by symmetry, $U_n(\beta)$ is an unbiased estimating function as its conditional expectation given the $\tilde{L}_i$, $\tilde{R}_i$, $\tilde{X}_i$ is zero. Furthermore, the non-zero terms in $U_n(\beta)$ are observed because of the constraints being imposed. In fact, we can write

$$U_n(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n} I\left\{L_j(\beta) < e_i(\beta) < R_j(\beta),\ L_i(\beta) < e_j(\beta) < R_i(\beta)\right\}\left(X_i - X_j\right)\operatorname{sgn}\left\{e_i(\beta) - e_j(\beta)\right\}.$$

Similar to Bhattacharya et al. (1983), we call $e_i(\beta)$ and $e_j(\beta)$ comparable only if $L_j(\beta) < e_i(\beta) < R_j(\beta)$ and $L_i(\beta) < e_j(\beta) < R_i(\beta)$. The proposed estimating function is the sum of the weights $\pm(X_i - X_j)$ of all the comparable pairs, with the sign being decided by whether $e_i(\beta) \leqslant e_j(\beta)$ or not.

Estimating function $U_n(\beta)$ is a step function, thus discontinuous. Finding root of a discontinuous function is typically not easy, especially for multidimensional cases. However, in the case of no truncation, finding root of $\tilde{U}_{\tilde{n}}(\beta)$ is equivalent to minimizing an L_1-type loss function $\tilde{G}_{\tilde{n}}(\beta) = \sum_{i=1}^{\tilde{n}}\sum_{j=1}^{\tilde{n}}\left|\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right| = \sum_{i=1}^{\tilde{n}}\sum_{j=1}^{\tilde{n}}\left|\tilde{Y}_i - \tilde{Y}_j - \beta^\top\left(\tilde{X}_i - \tilde{X}_j\right)\right|$, which is convex (Jin et al., 2001). In fact, this is a linear programming problem (Koenker and Bassett, 1978).

For doubly truncated data, we propose the following loss function

$$G_n(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\left[\left(e_i(\beta) - e_j(\beta)\right) \wedge \left(R_j - Y_j\right) \wedge (Y_i - L_i)\right] \vee \left(L_j - Y_j\right) \vee (Y_i - R_i)\right|. \tag{4}$$

Clearly, $G_n(\beta)$ becomes $\tilde{G}_{\tilde{n}}(\beta)$ when there is no truncation, i.e. $\tilde{L}_i \equiv -\infty$ and $\tilde{R}_i \equiv \infty$. Unlike $\tilde{G}_n(\beta)$, $G_n(\beta)$ is generally not a convex function. To see this, let $\underline{D}_{ij} = \left(L_j - Y_j\right) \vee (Y_i - R_i)$, $\overline{D}_{ij} = \left(R_j - Y_j\right) \wedge (Y_i - L_i)$, $Y_{ij} = Y_i - Y_j$ and $X_{ij} = X_i - X_j$. We have

$$G_n(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\left(Y_{ij} - \beta^\top X_{ij}\right) \wedge \overline{D}_{ij} \vee \underline{D}_{ij}\right|.$$

Since for any constants $a < b$, function $g(x) = |x \wedge a \vee b|$ is neither convex nor concave, $G_n(\beta)$ is generally not a convex function.

To see that minimizing the loss function $G_n(\beta)$ induces a consistent estimator, let

$$\overline{G}(\beta) = \mathrm{E}\left\{\left|\left[\left(e_i(\beta) - e_j(\beta)\right) \wedge \left(R_j - Y_j\right) \wedge (Y_i - L_i)\right] \vee \left(L_j - Y_j\right) \vee (Y_i - R_i)\right|\right\}. \tag{5}$$

It is shown in the Appendix A.1 that under mild conditions, $\overline{G}(\beta)$ is the limit of $[n(n-1)]^{-1}G_n(\beta)$ uniformly for β over a compact set. Differentiation of the right-hand side of (5) can be carried out by interchanging the differentiation and the expectation. Except on a set with zero probability, the derivative of the term inside the expectation sign is equal to

$$I\left\{\left(L_j - Y_j\right) \vee (Y_i - R_i) < e_i(\beta) - e_j(\beta) < \left(R_j - Y_j\right) \wedge (Y_i - L_i)\right\} \left(X_i - X_j\right)\operatorname{sgn}\left\{e_i(\beta) - e_j(\beta)\right\}. \quad (6)$$

From Lemma 1 in the Appendix, we can see that

$$\left(L_j - Y_j\right) \vee (Y_i - R_i) < e_i(\beta) - e_j(\beta) < \left(R_j - Y_j\right) \wedge (Y_i - L_i)$$

occurs if and only if $L_j(\beta) < e_i(\beta) < R_j(\beta)$ and $L_i(\beta) < e_j(\beta) < R_i(\beta)$. Thus, by symmetry, the expectation of (6) is equal to zero when β takes its true value, implying that $\bar{G}(\beta)$ has a minimizer at the true value of β.

Although $G_n(\beta)$ is generally not convex, in many cases it has a global minimizer, especially when the truncation is mild, making $G_n(\beta)$ close to $\tilde{G}_n(\beta)$. In our experience, we find that optimization functions in standard software packages can be used effectively to find the minimizer of $G_n(\beta)$ directly. For instance, '*fminsearch*' in the 'Optimization Toolbox' of MATLAB may be used to find the global minimizer.

Alternatively, the computation can be formulated as an iterative L_1-minimization problem. To be specific, consider the following modification of (4)

$$G_n^{(m)}(\beta, b) = \sum_{i=1}^{n}\sum_{j=1}^{n} I\left\{L_j(b) < e_i(b) < R_j(b), L_i(b) < e_j(b) < R_i(b)\right\}\left|e_i(\beta) - e_j(\beta)\right|.$$

Let $\hat{\beta}_{(0)}$ be an initial estimate, which may be taken as the naive estimate of β by ignoring the truncation. An iterative algorithm is given by $\hat{\beta}_{(k)} = \arg\min_\beta G_n^{(m)}\left(\beta, \hat{\beta}_{(k-1)}\right)$, $k \geqslant 1$. Note that in each iteration, $G^{(m)}\left(\beta, \hat{\beta}_{(k-1)}\right)$ is an L_1-type objective function, and $\hat{\beta}_{(k)}$ solves the equation

$$\sum_{i=1}^{n}\sum_{j=1}^{n} I\left\{L_j\left(\hat{\beta}_{(k-1)}\right) < e_i\left(\hat{\beta}_{(k-1)}\right) < R_j\left(\hat{\beta}_{(k-1)}\right), L_i\left(\hat{\beta}_{(k-1)}\right) < e_j\left(\hat{\beta}_{(k-1)}\right) < R_i\left(\hat{\beta}_{(k-1)}\right)\right\} \times \left(X_i - X_j\right)\operatorname{sgn}\left\{e_i(\beta) - e_j(\beta)\right\} = 0,$$

If $\hat{\beta}_{(k)}$ converges to a limit as the number of $k \to \infty$, then the limit must satisfy $U_n(\beta) = 0$.

Let $\hat{\beta}_n$ denote the minimizer of $G_n(\beta)$ over a suitable parameter space. We show in Section 4 that $\hat{\beta}_n$ is consistent and asymptotically normal under suitable regularity conditions. The limiting covariance matrix involves the error density f. Thus, direct variance estimation involves density estimation. In principle, one may apply the nonparametric method proposed by Efron and Petrosian (1999) to the residuals to first estimate F and then, via smoothing, f, under suitable conditions. Following Jin et al. (2001), we propose using resampling approach based on random weighting that bypasses density estimation. Specifically, we generate i.i.d. nonnegative random variables W_i, $i = 1, \ldots, n$, with mean μ and variance $4\mu^2$. Define the following perturbed version of $G_n(\beta)$

$$G_n^*(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(W_i + W_j\right)\left|\left[\left(e_i(\beta) - e_j(\beta)\right) \wedge \left(R_j - Y_j\right) \wedge \left(Y_i - L_i\right)\right] \vee \left(L_j - Y_j\right) \vee \left(Y_i - R_i\right)\right| \quad (7)$$

and let $\hat{\beta}^* = \mathrm{argmin}_\beta G_n^*(\beta)$. We show in Section 4 that the conditional distribution of $\sqrt{n}\left(\hat{\beta}^* - \hat{\beta}_n\right)$ given data converges to the same limiting distribution as that of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$, where β_0 is the true value of β. By repeatedly generating $\{W_i, i = 1, \ldots, n\}$, we can obtain a large number of replications of $\hat{\beta}^*$. Then the conditional distribution of $\sqrt{n}\left(\hat{\beta}^* - \hat{\beta}_n\right)$ given data can be approximated arbitrarily closely. In our numerical studies, we choose a gamma distribution to generate W_i. Other distributions, such as the beta distribution with proper parameters, can also be used (Jin et al., 2001).

3.2 Weighted estimation

It is well known that choosing proper weights can improve efficiency in rank estimation; see, for example, Hajek and Sidak (1967), Prentice (1978), Harrington and Fleming (1982) and Jin et al. (2003). For the full data, we may extend the estimating function $\widetilde{U}_{\tilde{n}}(\beta)$ in (3) by assigning weights to its summands. Specifically, we consider the following weighted estimating function

$$\widetilde{U}_{\tilde{n}, \widetilde{\mathrm{w}}}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\widetilde{\mathrm{w}}_{ij}(\beta)\left(\widetilde{X}_i - \widetilde{X}_j\right)\mathrm{sgn}\left\{\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right\}, \quad (8)$$

where the weights $\widetilde{\mathrm{w}}_{ij}(\beta)$, which may depend on β, are symmetric, i.e., $\widetilde{\mathrm{w}}_{ij}(\beta) = \widetilde{\mathrm{w}}_{ji}(\beta)$. By symmetry, we can easily see that the estimating function is unbiased, i.e., $\mathrm{E}\left[\widetilde{U}_{\tilde{n}, \widetilde{\mathrm{w}}}(\beta_0)\right] = 0$. The choice of $\widetilde{\mathrm{w}}_{ij}(\beta) \equiv 1$ corresponds to the Wilcoxon-Mann-Whitney statistic. It is asymptotically efficient when $\tilde{\varepsilon}$ in model (1) follows the standard logistic distribution. Under this weighting scheme, $\widetilde{U}_{\tilde{n}, \widetilde{\mathrm{w}}}(\beta)$ reduces to the unweighted estimating function $\widetilde{U}_{\tilde{n}}(\beta)$. Another commonly used weighting scheme in rank estimation is that of the log-rank, which is asymptotically efficient when $\tilde{\varepsilon}$ follows the extreme value distribution. Let $\widetilde{\mathrm{w}}_{ij}(\beta) = \widetilde{\psi}_{\tilde{n}}\left(\beta, \tilde{e}_i(\beta) \wedge \tilde{e}_j(\beta)\right)$, where $\widetilde{\psi}_{\tilde{n}}(b, t) = \left(\sum_{i=1}^{\tilde{n}} I\{\tilde{e}_i(b) \geqslant t\}\right)^{-1}$. We show in Lemma 2 in the Appendix that with such $\widetilde{\mathrm{w}}_{ij}(\beta)$, $\widetilde{U}_{\tilde{n}, \widetilde{\mathrm{w}}}(\beta)$ becomes the log-rank estimation function for β.

For the doubly truncated data, similar to (8), we can also introduce weights to the proposed estimating function $U_n(\beta)$, that is, to consider

$$U_{n, \mathrm{w}}(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{w}_{ij}(\beta)I\left\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\right\}\left(X_i - X_j\right)\mathrm{sgn}\left\{e_i(\beta) - e_j(\beta)\right\},$$

where the $w_{ij}(\beta)$ are again symmetric, i.e. $w_{ij}(\beta) = w_{ji}(\beta)$. For $w_{ij}(\beta) \equiv 1$, $U_{n,w}(\beta)$ reduces to $U_n(\beta)$. For the log-rank version, we let $w_{ij}(\beta) = \psi_n(\beta, e_i(\beta) \wedge e_j(\beta))$, where $\psi_n(b, t) = \left(\sum_{i=1}^{n} I\{e_i(b) \geqslant t\}\right)^{-1}$. Other weighting schemes can also be considered. Though the data are subject to double truncation, we still expect, as simulation results in Section 5 indicate, that proper choices of weights will generally improve the estimation efficiency.

Similar to $U_n(\beta)$, $U_{n,w}(\beta)$ is discontinuous and solving $U_{n,w}(\beta) = 0$ directly may not be easy. We consider loss function

$$G_{n,w}(\beta, b) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(b)\left|\left[\left(e_i(\beta) - e_j(\beta)\right) \wedge \left(R_j - Y_j\right) \wedge (Y_i - L_i)\right] \vee \left(L_j - Y_j\right) \vee (Y_i - R_i)\right|.$$

By differentiating with respect to β, it is easily seen that

$$\left.\frac{\partial G_{n,w}(\beta, b)}{\partial \beta}\right|_{b=\beta} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\beta) I\left\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\right\} \times \left(X_i - X_j\right)\mathrm{sgn}\left\{e_i(\beta) - e_j(\beta)\right\}, \tag{9}$$

which becomes the weighted estimating function $U_{n,w}(\beta)$. Therefore, we propose the following iterative algorithm. First set the initial b to be $\hat{\beta}_{(0)}^{w}$, and then find the estimator iteratively through $\hat{\beta}_{(k)}^{w} = \mathrm{argmin}_{\beta} G_{n,w}\left(\beta, \hat{\beta}_{(k-1)}^{w}\right)$, $k \geqslant 1$. From (9) we see that if $\hat{\beta}_{(k)}^{w}$ converges to a limit, say $\hat{\beta}_{n}^{w}$, as k goes to infinity, then the limit satisfies $U_{n,w}\left(\hat{\beta}_{n}^{w}\right) = 0$.

We show in Section 4 that under some regularity conditions, $\hat{\beta}_{n}^{w}$ is consistent and asymptotically normal. Moreover, as noted in Jin et al. (2003), when using the above algorithm, for each fixed k, $\hat{\beta}_{(k)}^{w}$ is itself a legitimate estimator, i.e. it is consistent and asymptotically normal. In view of this result, one may in practice consider the proposed iterative algorithm only for a relatively small number of the iterations to obtain a reasonable estimator. In our simulation study, we set the number of iterations to be 3 to get the log-rank estimate. We also iterated the algorithm until the difference between successive estimates attains a pre-specified accuracy as "convergence". We found that $\hat{\beta}_{(k)}^{w}$ converged in all the cases and the converged estimate was quite close to the $\hat{\beta}_{(k)}^{w}$ after 3 iterations.

For the variance estimation, we again apply the random weighting approach. We introduce the following perturbed version of $G_{n,w}(\beta, b)$:

$$G_{n,w}^{*}(\beta, b) = \sum_{i=1}^{n}\sum_{j=1}^{n} \left(W_i + W_j\right) w_{ij}(b)\left|\left[\left(e_i(\beta) - e_j(\beta)\right) \wedge \left(R_j - Y_j\right) \wedge (Y_i - L_i)\right] \vee \left(L_j - Y_j\right) \vee (Y_i - R_i)\right|,$$

where W_i, $i = 1, \ldots, n$, are i.i.d. nonnegative random variables with mean μ and variance $4\mu^2$. The perturbed estimate is solved by exactly following the above iterative algorithm. We first obtain $\hat{\beta}^*$ from minimizing $G^*_{n,\mathrm{w}}$ by setting $\mathrm{w}_{ij}(b) = 1$. Note that this $\hat{\beta}^*$ is just the minimizer of (7). Then let $\hat{\beta}^*_{(0)} = \hat{\beta}^*$, and iterate the value of the estimate by $\hat{\beta}^*_{(k)} = \mathrm{argmin}_\beta G^*_{n,\mathrm{w}}\left(\beta, \hat{\beta}^*_{(k-1)}\right)$. It is important to point out that here the number of iteration should stay the same as that for solving the point estimate. The asymptotic distribution of $\sqrt{n}\left(\hat{\beta}^{\mathrm{w}}_{(k)} - \beta_0\right)$ can be approximated by the conditional distribution of $\sqrt{n}\left(\hat{\beta}^*_{(k)} - \hat{\beta}^{\mathrm{w}}_{(k)}\right)$ given the observed data. By repeatedly generating the W_i sequences, we can obtain many realizations of $\hat{\beta}^*_{(k)}$ and make inference based on the empirical distribution of the realized $\hat{\beta}^*_{(k)}$'s.

4 Large sample theory

This section is devoted to the development of a large sample theory for the methods proposed in the preceding section. Assume that Z_i, $i = 1, \ldots, n$ are i.i.d. observations from (2). Let B be the parameter space. We shall assume that B is compact and β_0 is an interior point of B. We first discuss the asymptotic properties of $\hat{\beta}_n$. Let

$$\xi\left(Z_i, Z_j, \beta\right) = I\left\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\right\}\left(X_i - X_j\right)\mathrm{sgn}\left\{e_i(\beta) - e_j(\beta)\right\}$$

and $V = \mathrm{E}[\xi(Z_i, Z_j, \beta_0)\xi^{\mathsf{T}}(Z_i, Z_k, \beta_0)]$. Also, let $A = \partial^2 \bar{G} / \partial\beta\partial\beta^{\mathsf{T}}\big|_{\beta = \beta_0}$. The following regularity conditions will be used.

A1 The error density f is bounded and has a bounded and continuous derivative.

A2 The covariate vector has a bounded second moment, i.e., $\mathrm{E}(\|X\|^2) < \infty$.

A3 The true parameter value β_0 is the unique global minimizer of the limiting loss function $\bar{G}(\beta)$ over B.

A4 The second derivative of $\bar{G}(\beta)$ at β_0 is nonsingular, i.e., matrix A is strictly positive definite.

Conditions A1, A2 and A4 are mild conditions. In particular, when there is no truncation, A4 is always satisfied except for the degenerate case when covariates are co-linear with probability one. In general, the loss function will no longer be convex due to truncation, but we still expect that A4 should hold, at least when truncation is not heavy. Condition A3 is implied by condition A4 when there is no truncation. It is assumed to guarantee that the proposed estimator is consistent. The following theorem gives the asymptotic properties of $\hat{\beta}_n$.

Theorem 1.

Under conditions A.1–A.4, $\hat{\beta}_n$ is consistent and $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ converges in distribution to $N(0, A^{-1} V A^{-1})$.

The objective function $G_n(\cdot)$ is a typical U-process of order 2. Thus, we can apply results on quadratic approximations of U-processes to prove the above result. The details are provided in the Appendix.

As being proposed in Section 3.1, we approach the variance estimation through random weighting. The theoretical justification of this approach is given by the following theorem, which is again proved in the Appendix.

Theorem 2.

Let $\hat{\beta}^*$ be the minimizer of the perturbed loss function $G_n^*(\beta)$ as defined by (7). Then under conditions A.1–A.4, the conditional distribution of $\sqrt{n}\left(\hat{\beta}^* - \hat{\beta}_n\right)$ given $Z_1, \ldots, Z_\mathrm{n}$ converges in probability to $N(0, A^{-1}VA^{-1})$. In particular, the conditional covariance matrix of $\hat{\beta}^*$ given $Z_1, \ldots, Z_\mathrm{n}$ converges in probability to $A^{-1}VA^{-1}$.

Next we turn to the weighted estimators proposed in Section 3.2. For the weights $\mathrm{w}_{ij}(\beta)$ with form $\psi_n(\beta, e_i(\beta) \wedge e_j(\beta))$, where $\psi_n(b, t)$ may depend on the data, we assume the following condition.

A5 There exists a deterministic function $\psi(t)$ such that $\sup_t |\psi_n(\beta_0, t) - \psi(t)| = o_p(n^{-\eta})$ for some $\eta > 0$.

The asymptotic properties of the weighted estimators are given by the following theorem.

Theorem 3.

Under conditions A.1–A.5, (i) $\hat{\beta}_n^\mathrm{w}$ is consistent and $\sqrt{n}\left(\hat{\beta}_n^\mathrm{w} - \beta_0\right)$ converges in distribution to $N\left(0, A_\mathrm{w}^{-1}V_\mathrm{w}A_\mathrm{w}^{-1}\right)$; (ii) for each $k \geqslant 0$, $\sqrt{n}\left(\hat{\beta}_{(k)}^\mathrm{w} - \beta_0\right)$ converges in distribution to a normal distribution with zero mean and some variance-covariance matrix.

Matrices A_w and V_w are the asymptotic slope and covariance matrices for the weighted estimating function $U_{n,\mathrm{w}}$ that reduce to A and V when $\mathrm{w}_{ij}(\beta) = 1$. The proof is given in the Appendix.

5 Simulation study

We conducted simulation studies to assess the finite sample performance of the proposed method. For model (1), we considered a two-dimensional covariate vector, i.e., $\tilde{X} = \left(\tilde{X}_1, \tilde{X}_2\right)^\top$, where $\tilde{X}_1$ and $\tilde{X}_2$ were independently drawn from a binomial distribution with success probability 0.5 and uniform distribution on [0, 2], respectively. We set the two regression coefficients, denoted by β_1 and β_2, to be 0 and 1. For the error distribution F, three distributions, standard normal distribution, standard logistic distribution and extreme value (EV) distribution, were used. We considered two truncation schemes. The first one was covariate-dependent truncation with a random truncation interval, where the truncation variables $\tilde{L}$ and $\tilde{R}$ were independently generated from the uniform distribution on $\left[c_1, \tilde{X}_1 + \tilde{X}_2/2\right]$ and the uniform distribution on $\left[\tilde{X}_1 + \tilde{X}_2/2, c_2\right]$, respectively. The second one was covariate-dependent truncation with a fixed truncation interval, where $\tilde{L}$ was generated from the uniform distribution on $\left[c_1, \widetilde{X}_1 + \widetilde{X}_2/2\right]$ and $\widetilde{R} = \widetilde{L} + c_3$. The constants c_1, c_2 and c_3 were chosen to yield about 30% percentage of truncation under various error distributions (with both left and right truncation proportions being 15%). The observable sample size n was chosen to be 200, 300 and 400. Under each scenario, 1,000 replications were carried out. We first used loss function (4), which corresponds to the Wilcoxon weight. We then used the log-rank weighting scheme along with the proposed iterative algorithm with 3 iterations as suggested in Section 3.2. The minimization was implemented via the MATLAB function '*fminsearch*' in the 'Optimization Toolbox' of MATLAB, which uses a simplex search method to find the minimizer. For estimating standard errors, we used the proposed resampling approach with 500 sets of W_i, $i = 1, \ldots, n$ being generated from Gamma(0.25, 0.5).

Besides the proposed estimates, we also calculated "naive" estimates for the regression coefficients by ignoring the truncation. That is, we treated the observed data as data without double truncation, and solved the Mann-Whitney type estimating equation (3) for the estimates. The random weighting approach proposed by Jin et al. (2001) was applied to get the estimated standard errors. For all the estimates, we recorded the average bias, empirical standard error, the average of the standard errors estimated from the random weighting approach, and the empirical coverage probability of the 95% Wald-type confidence intervals. The results under random truncation interval scenario are summarized in Table 1, while the results under fixed truncation interval are in Table 2.

We found that under the both truncation schemes, both naive estimates for β_1 and β_2 were biased and the empirical coverage probabilities of the confidence intervals were far less than the nominal level. However, under all scenarios, the proposed estimates obtained from the loss function in (4) (i.e., Wilcoxon weight) and log-rank weight with $k = 3$ were both essentially unbiased. The average of the standard error estimates were quite close to the corresponding empirical standard errors. The empirical coverage probabilities of the Wald-type confidence intervals were close to the nominal level of 95%. For the normally distributed random error, the estimates with the two weighting schemes had comparable efficiency. For the logistic random error, the Wilcoxon weight gave slightly more efficient estimates than those with the log-rank weight, while for the extreme value random error, the estimate with log-rank weight was significantly more efficient. The results implied that for the doubly truncated data, one could still expect substantial efficiency improvement if a proper weighting scheme was chosen, as one would expect for the case with no truncation. In general, the simulation results showed that the proposed method worked well for practical sample sizes.

We also examined the difference between the log-rank estimates with 3 iterations versus those obtained after convergence. The algorithm was treated as convergence in the sense that the sum of absolute component differences between two consecutive estimates was less than 0.01. We took EV error distribution and random truncation interval case for illustration. The estimates with 3 iterations and at convergence were plotted for the two regression parameters under different sample sizes. In Figure 1, the top panel corresponds to the plots for β_1 and β_2 under $n = 200$, the middle panel corresponds to the plots for β_1 and β_2 under $n = 300$, and the bottom panel corresponds to the plots for β_1 and β_2 under $n = 400$. The two sets of estimates were quite similar, implying that a small number of iterations (such as 3) should be sufficient. The results were quite similar for the other error distributions and truncation mechanisms.

6 Applications

6.1 Quasar data

We applied the proposed methods to the quasar data analyzed by Efron and Petrosian (1999). The original dataset consists of quadruplets (z_i, m_i, a_i, b_i), $i = 1, \ldots, n$, where z_i is the redshift of the ith quasar, m_i is its apparent magnitude (with larger values of m corresponding to dimmer objects), and the two numbers a_i and b_i are lower and upper truncation bounds on apparent magnitude, respectively. Due to the experimental constraints, quasars with m_i above b_i were too dim to yield dependable redshift z_i, while the lower limit a_i was used to avoid confusion with nonquasar steller objects. Thus, any quasar with its apparent magnitude outside the lower and upper bounds was not visible (Bhattacharya et al., 1993), with no information being included in the dataset. Each observed m_i is subject to doubly truncation with truncation bounds a_i and b_i. In this study $a_i = 16.08$ remains the same for all i, and b_i varies between 18.494 and 18.93. The full dataset has $n = 1,052$ quasars.

Farther quasars tend to have bigger values of m_i, so they are dimmer. According to Hubble's law, one can transform apparent magnitudes into a luminosity measurement which should be independent of distance. The transformation depends on the cosmological model supposed. Following the Einstein-deSitter cosmological model (Weinberg, 1972), one can obtain the log luminosity values y_i from formula

$$y_i = t(z_i, m_i) = 19.894 - 2.303\frac{m_i}{2.5} + \log\left(Z_i - Z_i^{\frac{1}{2}}\right) - \frac{1}{2}\log(Z_i) \tag{10}$$

where $Z_i = 1 + z_i$. Larger values of y_i correspond to intrinsically brighter quasars. Since m_i is doubly truncated, so is y_i. The truncation limits for y_i, denoted by L_i and R_i, are obtained by applying (10) to b_i and a_i, respectively, i.e., $L_i = t(z_i, b_i)$ and $R_i = t(z_i, a_i)$.

The main purpose of the quasar investigation is to study luminosity evolution. Quasars may have been intrinsically brighter in the early universe and evolved toward a dimmer state as time went out. However, if there is no luminosity evolution, y_i should be independent of z_i except for truncation effects. Thus, testing the absence of luminosity evolution amounts to testing for independence. A convenient one-parameter model for luminosity evolution says that the expected log luminosity increases linearly as $\theta\log(1 + z)$, with $\theta = 0$ corresponding to no evolution. If θ is a hypothesized value of the evolution parameter, instead of directly testing for the independence of y_i and z_i, Efron and Petrosian (1999) tested the null hypothesis that H_θ: $y_i(\theta) = y_i - \theta\log(1 + z_i)$ is independent of z_i, using their proposed approach. Correspondingly, in their analysis, the truncation regions for $y_i(\theta)$ also changed with θ, that is, $L_i(\theta) = L_i - \theta\log(1 + z_i)$ and $R_i(\theta) = R_i - \theta\log(1 + z_i)$.

Since the one-parameter model for luminosity evolution assumes linear relationship between the expected log luminosity and $\log(1 + z)$, it is quite natural to consider the following linear model

$$y_i = \theta \log(1 + z_i) + \varepsilon_i, \quad (11)$$

where the response y_i is subject to double truncation with the truncation region $[L_i, R_i]$, ε_i is independent of z_i, and the evolution parameter θ becomes the unknown regression parameter. We can estimate θ by our proposed method. To make comparison, we used the same subset selected by Efron and Petrosian (1999) with $n = 210$ to do the analysis. Here we considered the loss function $G_n(\theta)$ defined in (4). The point estimate, denoted by $\hat{\theta}_n$, was obtained by minimizing $G_n(\theta)$. Figure 2 plots the curve $G_n(\theta)$ against θ within the range from 1 to 4.

The estimate $\hat{\theta}_n$, which is the minimizer of the displayed loss function, was 2.458. The proposed random weighting approach was used to estimate the standard error of $\hat{\theta}_n$. Five hundred draws of i.i.d. random variables following Gamma(0.25,0.5) were generated. The estimated standard error was 0.641. Consequently, an approximate 90% Wald-type confidence interval was [1.40, 3.51]. Under the linear model (11), the hypothesis of no evolution, i.e., H_0: y_i is independent of z_i, is equivalent to $H_0 : \theta = 0$. To test for $H_0 : \theta = 0$ against a positive evolution parameter $H_a : \theta > 0$, a Wald-type test statistic can be used. The test statistic equaled to the ratio of $\hat{\theta}_n$ and its estimated standard error, giving the value of 3.835. The corresponding one-sided p-value was about 6×10^{-5}, implying rejection of the null hypothesis of no evolution in favor of a positive value of θ at any commonly used significance level.

The tau test proposed by Efron and Petrosian (1999) for the no evolution hypothesis had an one-sided p-value 0.015. At 0.05 significance level, their test also rejected H_0 in favor of a positive value of θ, but failed to do so at 0.01 significance level. By inverting their test statistic, Efron and Petrosian (1999) obtained a point estimate for θ with the value of 2.38 and an approximate 90% central confidence interval [1.00, 3.20] which was slightly longer than the proposed Wald-type confidence interval.

The proposed approach is easy to handle multiple covariates. Here we further considered the following model with linear and quadratic term

$$y_i = \theta_1 \log(1 + z_i) + \theta_2 [\log(1 + z_i)]^2 + \varepsilon_i,$$

where ε_i is independent of z_i and θ_1 and θ_2 are unknown regression parameters. The regression parameters were estimated by minimizing (4), and the standard errors were estimated by the random weighting method with 500 i.i.d. Gamma(0.25,0.5) random variables being generated. The corresponding p-values of significance test for $H_0 : \theta_j = 0$ against $H_a : \theta_j \neq 0$, $j = 1, 2$, were calculated. The results are summarized in Table 3.

The significance tests showed that the effect of linear term, θ_1, was statistically significantly different from 0, while that of the quadratic term, θ_2, was apparently not. This provided some evidence to say the one-parameter model for luminosity evolution given by (11) is adequate for the current subset we analyzed.

6.2 AIDS incubation data

Another example we considered was an epidemiological data set on transfusion-related AIDS, collected by the Centers for Disease Control (CDC) in Atlanta, Georgia. For patients thought to be infected by HIV by blood or blood product transfusion, the original data recorded the gender and age, the date of reporting AIDS, the date of diagnosis, and the date of infection when it could be determined. The problems of interest include the process of infection, the distribution of the induction or "incubation" period (the time elapsed from HIV infection to the clinical manifestation of AIDS, confirmed by the AIDS diagnosis), and the dependence of the induction period on covariates such as age at the time of transfusion. The data include 494 cases reported to CDC prior to January 1, 1987, and diagnosed prior to July 1, 1986. Among them, 295 had consistent data on which we conduct our analysis. The information on infection time, the induction period, and the age at the time of transfusion for this subset were reported in Table 1 of Kalbfleisch and Lawless (1989). The times of infection and diagnosis were ascertained by reporting to the CDC.

Our primary interest centers on the relationship between the incubation period and the age at the time of transfusion. The period data can be viewed as being doubly truncated. Firstly, since HIV was unknown prior to 1982, any case of transfusion-related AIDS before this time would not have been properly classified and would have been unobserved, leading to left-truncation. Secondly, the data were retrospectively ascertained for all transfusion-associated AIDS cases in which the diagnosis of AIDS occurred prior to July 1, 1986, while cases diagnosed after this time were not included, thus leading to right-truncation. For the i-th observation, the left-truncation variable L_i was the duration (in month) from HIV infection to January 1, 1982; while the right-truncation variable R_i was defined as time from HIV infection to July 1, 1986 (the end of the diagnosis report). Thus the difference between R_i and L_i is always 54 months (4.5 years). The response variable Y_i is the incubation period duration and the covariate X_i is the age at the time of transfusion.

We applied the proposed method to estimate the regression parameter β in model (1). Minimizing the loss function $G_n(\beta)$ in (4) resulted in $\hat{\beta}_n = 0.73$. For standard error estimation, we again applied the proposed random weight approach with 500 independent samples from the gamma distribution. The estimated standard error was 0.29, giving an approximate 95% confidence interval [0.16, 1.30]. The results showed that the age at transfusion had a significant positive effect on the incubation period.

For comparison purpose, we also calculated the "naive" estimate of β from the Mann-Whitney type estimating equation (3) by ignoring the double truncation. The point estimate was 0.13. The standard error, also estimated by the random weighting approach, was 0.04, giving an approximate 95% confidence interval [0.05, 0.21]. The naive estimate seemed to underestimate the covariate effect to some extent. The similar phenomenon was also found in Moreira and Alvarez (2016) in their analysis of this data set.

7 Discussion

This paper is concerned with linear regression analysis when the response variable is subject to double truncation. Truncated data can be found in many applications, including those from biomedical researches, economics and astronomy. Most statistical methods for dealing with truncated data are for observations with left or right truncation. The left (right) truncation is relatively easy to handle due to the simple form of re-distribution-to-left (right) algorithm and applicability of counting process-martingale formulation. However, for the doubly truncated data, fewer technical tools and results are available.

We propose a novel method to estimate the regression parameter in the linear regression model with doubly truncated responses. To eliminate the bias introduced by double truncation, we extend the Mann-Whitney type loss function for estimating regression parameters by symmetrization. The proposed estimator is obtained by minimizing the extended Mann-Whitney type loss function. The minimization can be done by standard software packages directly, or by an iterative algorithm with an L_1-type minimization in each iteration. The proposed estimator is shown to be consistent and asymptotically normal under some regularity conditions. A simple random perturbation approach is used to get the variance estimator. We also provide a weighted estimation procedure for improving the estimation efficiency. Simulation studies show that the proposed approach works well for moderate sample sizes. The application to the quasar data gives new insights.

Under the conditionally independent truncation assumption and the condition that the covariates are not degenerate, the regression parameter β_0 is identifiable. However, in order to completely identify the error distribution F, some more conditions are needed. Let F_L and F_R be the (conditional) distribution of $\widetilde{L} - \beta_0^\top \widetilde{X}$ and $\widetilde{R} - \beta_0^\top \widetilde{X}$ given $\widetilde{X}$, respectively. If one assumes that for some values of $\widetilde{X}$, the left endpoint of the support of F is between the left endpoints of the support of F_L and F_R and the right endpoint of the support of F is between the right endpoints of the support of F_L and F_R, then all F, F_L, and F_R are identifiable (Woodroofe, 1985). Under such conditions, in principle one can also estimate the truncation probability given covariates based on the observed data and the proposed estimator $\hat{\beta}_n$.

In addition to handling multiple covariates, another major advantage of the proposed loss function-based approach to estimation over the test score-based approach of Efron and Petrosian (1999) is that it can easily incorporate a penalty function, such as LASSO, to do variable selection. Note that when LASSO penalty is used, our iterative algorithm is preferable since in each iteration the optimization can still be formulated into an L_1-minimization problem, facilitating the computation. It is also of interest to investigate the possibility of extending the approach to doubly censored responses, such as those considered in Ren and Gu (1997).

Acknowledgements

The authors thank the Editor, Associate editor, and two referees for their insightful comments and suggestions, which have led to many improvements. The authors also thank Professor Bradley Efron for providing the quasar data.

Funding

Zhiliang Ying's research was supported in part by NIH grant R37GM047845 and NSF grants DMS-1308566, SES-1323977. Wen Yu's research was supported in part by the Fudan Young Scholar Research Enhancement Program (20520133107) and the National Natural Science Foundation of China (11671097). Ming Zheng's research was supported in part by the National Natural Science Foundation of China (11771095).

A: Appendix

A.1 Two lemmas

The first lemma is crucial for the intuition about the validity of the proposed loss function $G_n(\beta)$ defined by (4).

Lemma 1.

Let $L_i(\beta)$, $R_i(\beta)$ and $e_i(\beta)$, $i = 1, \ldots, n$ be defined in Section 3. Then the event

$$\left(L_j - Y_j\right) \vee (Y_i - R_i) < e_i(\beta) - e_j(\beta) < \left(R_j - Y_j\right) \wedge (Y_i - L_i) \tag{12}$$

occurs if and only if $L_j(\beta) < e_i(\beta) < R_j(\beta)$ and $L_i(\beta) < e_j(\beta) < R_i(\beta)$.

Proof: We first show "if". From $e_i(\beta) < R_j(\beta)$, we have

$$e_i(\beta) - e_j(\beta) < R_j(\beta) - e_j(\beta) = R_j - Y_j. \tag{13}$$

From $L_i(\beta) < e_j(\beta)$, we have

$$e_i(\beta) - e_j(\beta) < e_i(\beta) - L_i(\beta) = Y_i - L_i. \tag{14}$$

Thus, the second inequality in (12) holds. The first inequality can be shown similarly.

Next we show "only if". This can be done by reversing the above argument. From (13), we obviously have $e_i(\beta) < R_j(\beta)$, while from (14), we get $L_i(\beta) < e_j(\beta)$. Additionally, from $(L_j - Y_j) \vee (Y_i - R_i) < e_i(\beta) - e_j(\beta)$, we get $e_i(\beta) > L_j$ and $e_j(\beta) < R_i(\beta)$.

The second lemma shows that the choice of $\widetilde{w}_{ij}(\beta) = \widetilde{\psi}_{\tilde{n}}\left(\beta, \tilde{e}_i(\beta) \wedge \tilde{e}_j(\beta)\right)$ makes the weighted estimation function becomes the log-rank estimation function.

Lemma 2.

When $\widetilde{w}_{ij}(\beta) = \widetilde{\psi}_{\tilde{n}}\left(\beta, \tilde{e}_i(\beta) \wedge \tilde{e}_j(\beta)\right)$, where, $\widetilde{\psi}_{\tilde{n}}(b, t) = \left(\sum_{i=1}^{\tilde{n}} I\{\tilde{e}_i(b) \geqslant t\}\right)^{-1}$, $\widetilde{U}_{\tilde{n}, \widetilde{w}}(\beta)$ becomes the log-rank estimating function for β.

Proof: When $\widetilde{w}_{ij}(\beta) = \widetilde{\psi}_{\tilde{n}}\left(\beta, \tilde{e}_i(\beta) \wedge \tilde{e}_j(\beta)\right)$, it can be seen that

$$\widetilde{U}_{\tilde{n}, \widetilde{\mathrm{w}}}(\beta) = \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left(\sum_{k=1}^{\tilde{n}} I\left\{\tilde{e}_k(\beta) \geqslant \tilde{e}_i(\beta) \wedge \tilde{e}_j(\beta)\right\} \right)^{-1} \left(\widetilde{X}_i - \widetilde{X}_j\right) \mathrm{sgn}\left\{\tilde{e}_i(\beta) - \tilde{e}_j(\beta)\right\}$$
$$= -2 \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left(\sum_{k=1}^{\tilde{n}} I\{\tilde{e}_k(\beta) \geqslant \tilde{e}_i(\beta)\} \right)^{-1} \left(\widetilde{X}_i - \widetilde{X}_j\right) I\left\{\tilde{e}_j(\beta) \geqslant \tilde{e}_i(\beta)\right\}$$
$$= -2 \sum_{i=1}^{\tilde{n}} \left(\frac{\sum_{j=1}^{\tilde{n}} \widetilde{X}_i I\left\{\tilde{e}_j(\beta) \geqslant \tilde{e}_i(\beta)\right\}}{\sum_{k=1}^{\tilde{n}} I\{\tilde{e}_k(\beta) \geqslant \tilde{e}_i(\beta)\}} - \frac{\sum_{j=1}^{\tilde{n}} \widetilde{X}_j I\left\{\tilde{e}_j(\beta) \geqslant \tilde{e}_i(\beta)\right\}}{\sum_{k=1}^{\tilde{n}} I\{\tilde{e}_k(\beta) \geqslant \tilde{e}_i(\beta)\}} \right)$$
$$= -2 \sum_{i=1}^{\tilde{n}} \left(\widetilde{X}_i - \frac{\sum_{j=1}^{\tilde{n}} \widetilde{X}_j I\left\{\tilde{e}_j(\beta) \geqslant \tilde{e}_i(\beta)\right\}}{\sum_{k=1}^{\tilde{n}} I\{\tilde{e}_k(\beta) \geqslant \tilde{e}_i(\beta)\}} \right).$$

This completes the proof.

A.2 Proof of Theorem 1

We first prove consistency. Let $\bar{G}_n = [n(n-1)]^{-1} G_n$. Under the assumption that $\widetilde{Z}_i$'s are i.i.d. and condition A2, by Corollary 7 in Sherman (1994), we have that $\bar{G}_n(\beta)$ converges uniformly to $\bar{G}(\beta)$ for β over B. Since, by condition A3, $\bar{G}(\beta)$ has a unique minimizer β_0, $\hat{\beta}_n$ must converge to β_0 as $\bar{G}(\beta)$ is obviously continuous.

The proof of asymptotic normality follows closely the technical developments given in Sherman (1993) for the maximum rank correlation estimator which is also defined as the optimizer of a U-type objective function. In fact, the situation there is more complicated as it deals with a discontinuous objective function. An essential ingredient of Sherman's approach is the quadratic approximation to the objective function.

Following Sherman (1993), define $\tau(z, \beta) = E\xi(Z_i, z; \beta)$. Let $\dot{\tau}(z, \beta)$ and $\ddot{\tau}(z, \beta)$ be its first and second derivatives with respect to β. Then it can be seen from conditions A1 and A2 that we have

$$E\left[\|\dot{\tau}(Z_i, \beta)\|^2 + \|\ddot{\tau}(Z_i, \beta)\|\right] < \infty$$

and there exists $K(z) \geq 0$ such that $EK(Z_i) < \infty$ and

$$\|\dot{\tau}(z, \beta) - \ddot{\tau}(z, \beta_0)\| \leq K(z)\|\beta - \beta_0\|.$$

From these and conditions A1–A4, we can verify the four assumptions in Sherman (1993, Theorem 4) from which the asymptotic normality of $\hat{\beta}_n$ follows.

A.3 Proof of Theorem 2

Because of scale invariance for $\hat{\beta}^*$ to change in W_i, we may assume, without loss of generality, that $E(W_i) = 1/2$. Similarly to the proof of consistency of $\hat{\beta}_n$, we can argue in the same way that $\hat{\beta}^*$ is consistent. Let

$$U_n^*(\beta) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(W_i + W_j\right)I\left\{L_j(\beta) < e_i(\beta) < R_j(\beta), L_i(\beta) < e_j(\beta) < R_i(\beta)\right\} \times \left(X_i - X_j\right)\text{sgn}\left\{e_i(\beta) - e_j(\beta)\right\}.$$

It is clear that $U_n^*(\beta)$ is the derivative of $G_n^*(\beta)$. Thus, by definition, $U_n^*\left(\hat{\beta}^*\right) = 0$. By the same argument as that of Jin et al. (2001), we can establish asymptotic linearity and therefore, up to an asymptotically negligible term,

$$0 = U_n^*\left(\hat{\beta}^*\right) \approx U_n^*\left(\hat{\beta}_n\right) + n^2 A\left(\hat{\beta}^* - \hat{\beta}_n\right),$$

or

$$\sqrt{n}\left(\hat{\beta}^* - \hat{\beta}_n\right) \approx -n^{-\frac{3}{2}}A^{-1}U_n^*\left(\hat{\beta}_n\right).$$

Since $U_n\left(\hat{\beta}_n\right) = 0$, we have

$$U_n^*\left(\hat{\beta}_n\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(W_i - \frac{1}{2} + W_j - \frac{1}{2}\right)I\left\{L_j\left(\hat{\beta}_n\right) < e_i\left(\hat{\beta}_n\right) < R_j\left(\hat{\beta}_n\right), L_i\left(\hat{\beta}_n\right) < e_j\left(\hat{\beta}_n\right) < R_i\left(\hat{\beta}_n\right)\right\} \times \left(X_i - X_j\right)\text{sgn}\left\{e_i\left(\hat{\beta}_n\right) - e_j\left(\hat{\beta}_n\right)\right\}. \tag{15}$$

Each summand on the right-hand side of (15) clearly has mean 0 conditional on data. Standard asymptotic normality for U-statistics can then be used to show that, conditional on the data, $n^{3/2}U_n^*\left(\hat{\beta}_n\right)$ to a limiting normal distribution. Simple calculation shows that the conditional covariance matrix of $n^{-3/2}U_n^*\left(\hat{\beta}_n\right)$ given data converges in probability to V. Hence Theorem 2 holds.

A.4 Proof of Theorem 3

To prove (i), we know that $\hat{\beta}_n^{\mathrm{W}}$ is the solution to the estimating equation $U_{n,\mathrm{w}}(\beta) = 0$. By the asymptotic linearity of $U_{n,\mathrm{w}}$, we have, ignoring an asymptotically negligible term,

$$0 = U_{n,\mathrm{w}}\left(\hat{\beta}_n^{\mathrm{W}}\right) \approx U_{n,\mathrm{w}}(\beta_0) + n^2 A_{\mathrm{w}}\left(\hat{\beta}_n^{\mathrm{W}} - \beta_0\right)$$

or $\sqrt{n}\left(\hat{\beta}_n^{\mathrm{W}} - \beta_0\right) \approx -n^{3/2}A_{\mathrm{w}}^{-1}U_{n,\mathrm{w}}(\beta_0)$. Since $n^{-3/2}U_{n,\mathrm{w}}(\beta_0)$ converges to $N(0, V_{\mathrm{w}})$ by the asymptotic normality of the U-statistics, we get the desired result.

To prove (ii), similarly to (A.5) of Jin et al. (2001), we can show that for each k, there exists a $p \times p$ matrix D_k such that

$$\sqrt{n}\left(\hat{\beta}_{(k)}^{\mathrm{w}} - \beta_0\right) = -n^{-\frac{3}{2}} D_k A^{-1} U_n(\beta_0) - n^{-\frac{3}{2}} (I - D_k) A_{\mathrm{w}}^{-1} U_{n,\mathrm{w}}(\beta_0) + o_p(1).$$

From this and the joint asymptotic normality of $n^{-3/2} U_n(\beta_0)$ and $n^{-3/2} U_{n,\mathrm{w}}(\beta_0)$), we conclude that $\sqrt{n}\left(\hat{\beta}_{(k)}^{\mathrm{w}} - \beta_0\right)$ is asymptotically normal.

References

Amemiya T (1985), Advanced Econometrics, Harvard University Press, Cambridge, MA.

Bhattacharya PK, Chernoff H, and Yang SS (1983),“Nonparametric Estimation of the Slope of a Truncated Regression,” The Annals of Statistics, 11, 505–514.

Bilker W, and Wang M-C (1996), “Generalized Wilcoxon Statistics in Semiparametric Truncation Models,” Biometrics, 52, 10–20. [PubMed: 8934583]

Chang MN, and Yang GL (1987), “Strong Consistency of a Nonparametric Estimator of the Survival Function with Doubly Censored Data,” The Annals of Statistics, 15, 1536–1547.

Efron B, and Petrosian V (1999), “Nonparametric Methods for Doubly Truncated Data,” Journal of the American Statistical Association, 94, 824–834.

Greene WH (2012), Econometric Analysis (7th Ed.), Prentice Hall, Upper Saddle River, NJ.

Gu MG, and Zhang C-H (1993), “Asymptotic Properties of Self-Consistent Estimators Based on Doubly Censored Data,” The Annals of Statistics, 21, 611–624.

Hajek J and Sidak Z (1967). Theory of Rank Tests, Academic Press, New York.

Harrington DP and Fleming TR (1982), “A Class of Rank Test Procedures for Censored Survival Data,” Biometrika, 69, 133–143.

Jin Z, Lin DY, Wei LJ, and Ying Z (2003), “Rank-based inference for the accelerated failure time model,” Biometrika, 90, 341–353.

Jin Z, Ying Z, and Wei LJ (2001), “A Simple Resampling Method by Perturbing the Minimand,” Biometrika, 88, 381–390.

Kalbfleisch JD, and Lawless JF (1989), “Inference Based on Retrospective Ascertainment: An Analysis of the Data on Transfusion-Related AIDS,” Journal of the American Statistical Association, 84, 360–372.

Keiding N, and Gill RD (1990), “Random Truncation Models and Markov Processes,” The Annals of Statistics, 18, 582–602.

Kim JP, Lu W, Sit T and Ying Z (2013), “A unified approach to semiparametric transformation models under generalized biased sampling schemes,” Journal of the American Statistical Association, 108, 217–227. [PubMed: 23667280]

Koenker R, and Bassett G (1978), “Regression Quantiles,” Econometrica, 46, 33–50.

Lai TL, and Ying Z (1991a), “Estimating a Distribution Function with Truncated and Censored Data,” The Annals of Statistics, 19, 417–442.

Lai TL, and Ying Z (1991b), “Rank Regression Methods for Left-truncated and Right-censored Data,” The Annals of Statistics, 19, 531–556.

Liu H, Ning J, Qin J and Shen Y (2016), “Semiparametric Maximum Likelihood Inference for Truncated or Biased-Sampling Data”, Statistica Sinica, 26, 1087–1115.

Lynden-Bell D (1971), “A Method of Allowing for Known Observational Selection in Small Samples Applied to 3CR Quasars,” Monthly Notices of the Royal Astronomical Society, 155, 95–118.

Moreira C and Una-Alvarez J (2012), “Kernel Density Estimation with Doubly Truncated Data,” Electronic Journal of Statistics, 6, 501–521.

Moreira C and Una-Alvarez J (2016), “Nonparametric Regression with Doubly Truncated Data,” Computational Statistics and Data Analysis, 93, 294–307.

Pollard D (1990), Empirical Processes: Theory and Applications Reginal Conference Series Probability and Statistics 2 Institute of Mathematical Statistics, Hayward, CA.

Prentice RL (1978), “Linear Rank Tests with Right Censored Data,” Biometrika, 65, 167–179.

Ren J-J, and Gu M (1997), “Regression M-Estimators with Doubly Censored Data,” The Annals of Statistics, 25, 2638–2664.

Shen P-S, (2010), “Nonparametric Analysis of Doubly Truncated Data,” Annals of the Institute of Statistical Mathematics, 62, 835–853.

Shen P-S (2013), “A Class of Rank-Based Tests for Doubly-Truncated Data,” TEST, 22, 83–102.

Shen P-S, (2013), “Regression Analysis of Interval Censored and Doubly Truncated Data with Linear Transformation Models,” Computational Statistics, 28, 581–596.

Sherman RP (1993), “The Limiting Distribution of the Maximum Rank Correlation Estimator,” Econometrica, 61, 123–138.

Sherman RP (1994), “Maximal inequalities for degenerate U-processes with applications to optimization estimators,” The Annals of Statistics, 22, 439–459.

Tsai W-Y (1990), “Testing the Independence of Truncation Time and Failure Time,” Biometrika, 77, 167–177.

Tsui K-L, Jewell NP, and Wu CFJ (1988), “A Nonparametric Approach to the Truncated Regression Problem,” Journal of the American Statistical Association, 83, 785–792.

Turnbull BW (1976), “The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data,” Journal of the Royal Statistical Society, Ser. B, 38, 290–295.

Wang M-C, Jewell NP, and Tsai W-Y (1986), “Asymptotic Properties Of The Product Limit Estimate Under Random Truncation,” The Annals of Statistics, 14, 1597–1605.

Weinberg S (1972), Gravitation and Cosmology, New York: Wiley.

Woodroofe M (1985), “Estimating a Distribution Function with Truncated Data,” The Annals of Statistics, 13, 163–177.

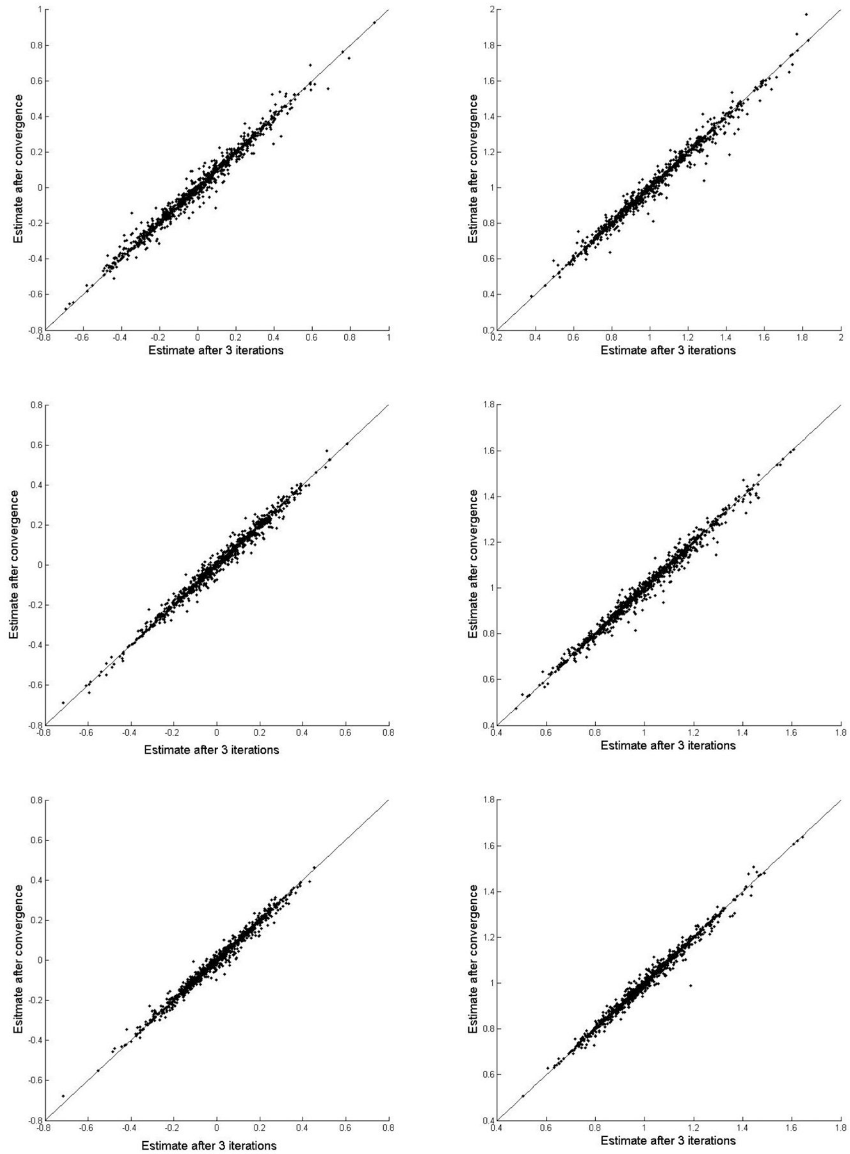

Figure 1:

Scatter plots of the estimates after 3 iterations against estimates after convergence. The error distribution was EV and the truncation interval was random. The top panel corresponds to $n = 200$, the middle panel corresponds to $n = 300$, and the bottom panel corresponds to under $n = 400$. The left ones are for β_1 and the right ones are for β_2.

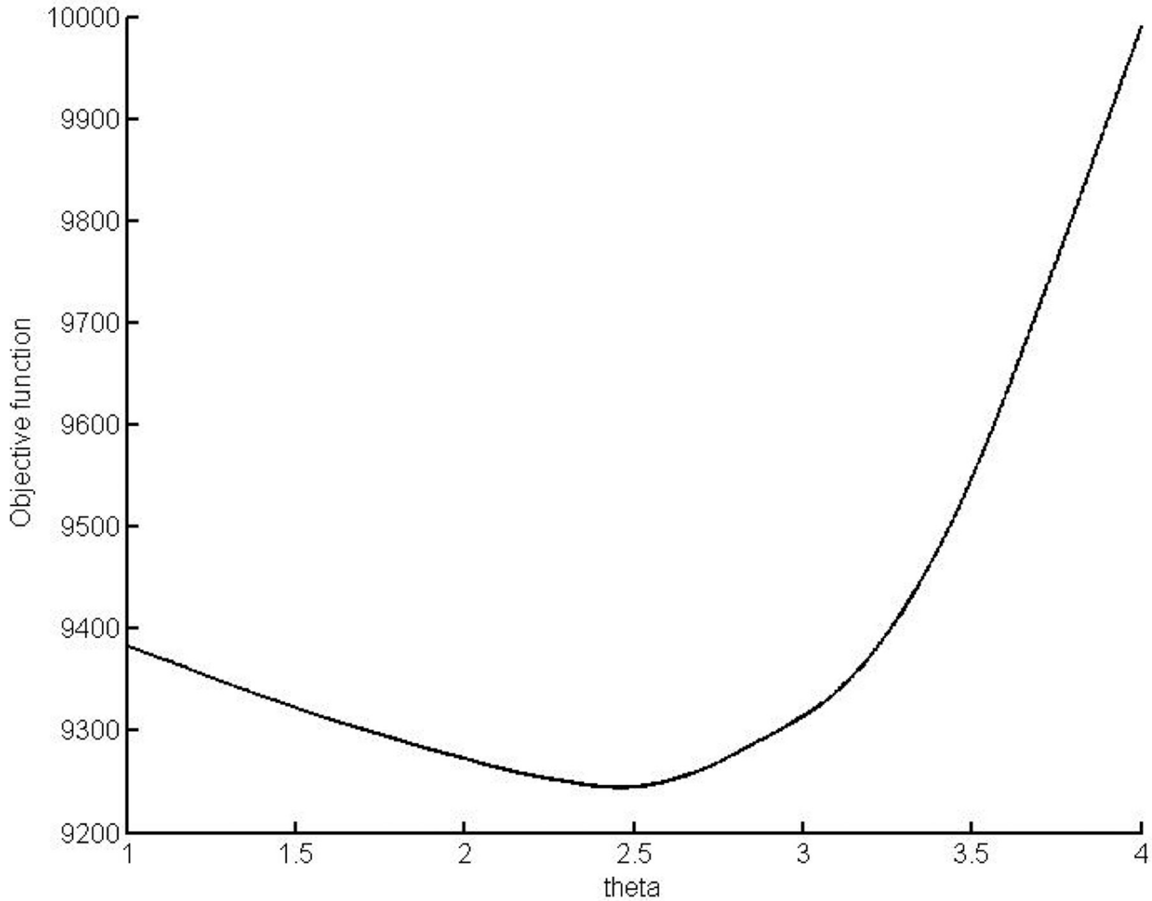

Figure 2:

Results from the quasar data analysis. The curve of the loss function $G_n(\theta)$ against θ within the range from 1 to 4.

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.

Summarized simulation results for the weighted estimates under random truncation interval.

			Naive				Wilcoxon				log-rank$_3$			
n	**Error Distribution**	**Weight Parameter**	**BIAS**	**SE**	**SEE**	**CP 95%**	**BIAS**	**SE**	**SEE**	**CP 95%**	**BIAS**	**SE**	**SEE**	**CP 95%**
200	Normal	β_1	0.2316	0.1139	0.1151	48.3%	0.0183	0.2024	0.2176	95.0%	0.0217	0.2035	0.2205	95.9%
		β_2	−0.2490	0.1022	0.0991	28.6%	−0.0042	0.1944	0.2053	95.1%	−0.0035	0.1993	0.2075	94.9%
	Logistic	β_1	0.2196	0.1834	0.1861	78.2%	−0.0070	0.3160	0.3381	95.5%	0.0016	0.3254	0.3508	96.6%
		β_2	−0.2358	0.1649	0.1611	69.1%	0.0129	0.2976	0.3287	93.9%	0.0168	0.3061	0.3360	93.9%
	EV	β_1	0.2292	0.1375	0.1308	56.7%	−0.0003	0.2694	0.3059	95.8%	0.0013	0.2305	0.2596	95.8%
		β_2	−0.2605	0.1203	0.1126	36.2%	0.0206	0.2711	0.3403	95.5%	0.0187	0.2357	0.2893	95.4%
300	Normal	β_1	0.2268	0.0957	0.0948	33.4%	0.0068	0.1689	0.1755	95.7%	0.0051	0.1733	0.1767	95.6%
		β_2	−0.2461	0.0817	0.0817	14.2%	0.0049	0.1633	0.1630	94.9%	0.0052	0.1671	0.1635	94.8%
	Logistic	β_1	0.2190	0.1503	0.1518	68.8%	−0.0039	0.2556	0.2661	95.3%	−0.0005	0.2650	0.2778	95.8%
		β_2	−0.2444	0.1346	0.1310	53.0%	−0.0064	0.2299	0.2353	94.6%	−0.0100	0.2412	0.2455	95.4%
	EV	β_1	0.2351	0.1051	0.1068	40.6%	0.0060	0.2083	0.2194	95.9%	0.0093	0.1827	0.1869	94.9%
		β_2	−0.2444	0.0936	0.0919	16.8%	0.0045	0.2019	0.2169	95.4%	0.0014	0.1811	0.1875	95.2%
400	Normal	β_1	0.2296	0.0810	0.0821	20.4%	0.0150	0.1453	0.1505	96.2%	0.0169	0.1470	0.1510	96.2%
		β_2	−0.2454	0.0712	0.0709	6.3%	0.0032	0.1342	0.1394	95.3%	0.0019	0.1325	0.1402	94.4%
	Logistic	β_1	0.2227	0.1362	0.1316	60.1%	−0.0023	0.2287	0.2294	95.0%	−0.0047	0.2358	0.2392	95.3%
		β_2	−0.2437	0.1123	0.1140	43.6%	−0.0059	0.1993	0.2033	95.6%	−0.0018	0.2094	0.2120	95.0%
	EV	β_1	0.2321	0.0942	0.0928	28.8%	0.0021	0.1824	0.1844	95.2%	0.0014	0.1566	0.1575	95.1%
		β_2	−0.2614	0.0811	0.0799	9.2%	0.0044	0.1678	0.1787	95.8%	0.0022	0.1509	0.1561	95.5%

Naive: naive estimate by ignoring double truncation; Wilcoxon: Wilcoxon weight estimate; log-rank$_3$: log-rank weight estimate with $k = 3$; BIAS: average bias of the estimates; SE: standard error of the estimates; SEE: average of the estimated standard errors; CP 95%: empirical coverage probabilities of Wald-type confidence intervals with 95% confidence level.

Table 2.

Summarized simulation results for the weighted estimates under fixed truncation interval.

			Naive				Wilcoxon				log-rank$_3$			
n	**Error Distribution**	**Weight Parameter**	**BIAS**	**SE**	**SEE**	**CP 95%**	**BIAS**	**SE**	**SEE**	**CP 95%**	**BIAS**	**SE**	**SEE**	**CP 95%**
200	Normal	β_1	0.1916	0.1165	0.1154	61.8%	0.0170	0.2235	0.2365	95.0%	0.0150	0.2266	0.2373	95.2%
		β_2	−0.2703	0.1050	0.1004	24.6%	−0.0088	0.2145	0.2135	94.8%	−0.0073	0.2182	0.2118	93.9%
	Logistic	β_1	0.1763	0.1828	0.1856	83.8%	−0.0070	0.3452	0.3637	95.0%	−0.0091	0.3607	0.3749	95.9%
		β_2	−0.2617	0.1681	0.1613	62.4%	0.0091	0.3174	0.3189	94.2%	0.0024	0.3235	0.3245	94.1%
	EV	β_1	0.1012	0.1457	0.1417	89.0%	−0.0009	0.2117	0.2132	95.4%	0.0014	0.1811	0.1761	93.8%
		β_2	−0.1580	0.1257	0.1232	74.8%	−0.0086	0.1838	0.1847	95.1%	−0.0078	0.1548	0.1512	93.7%
300	Normal	β_1	0.1895	0.0946	0.0948	48.6%	0.0152	0.1792	0.1884	95.5%	0.0180	0.1823	0.1885	94.9%
		β_2	−0.2700	0.0796	0.0829	9.0%	−0.0029	0.1638	0.1691	95.3%	−0.0042	0.1660	0.1671	94.5%
	Logistic	β_1	0.1873	0.1510	0.1510	75.0%	−0.0001	0.2798	0.2879	95.2%	0.0066	0.2916	0.2981	94.5%
		β_2	−0.2676	0.1344	0.1318	47.0%	−0.0016	0.2491	0.2524	94.8%	−0.0035	0.2609	0.2601	94.2%
	EV	β_1	0.1051	0.1170	0.1159	82.7%	−0.0014	0.1726	0.1718	94.8%	0.0015	0.1490	0.1431	93.9%
		β_2	−0.1604	0.0990	0.1007	64.2%	−0.0087	0.1525	0.1485	93.8%	−0.0105	0.1303	0.1229	92.5%
400	Normal	β_1	0.1901	0.0826	0.0823	37.1%	0.0167	0.1590	0.1619	95.9%	0.0189	0.1609	0.1618	95.5%
		β_2	−0.2664	0.0714	0.0720	4.0%	0.0026	0.1412	0.1452	95.2%	0.0033	0.1380	0.1435	95.8%
	Logistic	β_1	0.1809	0.1354	0.1313	72.1%	−0.0025	0.2473	0.2473	94.3%	−0.0004	0.2565	0.2562	93.9%
		β_2	−0.2686	0.1151	0.1143	33.4%	−0.0007	0.2147	0.2170	95.3%	−0.0042	0.2237	0.2237	94.2%
	EV	β_1	0.1034	0.1040	0.1002	81.6%	0.0043	0.1533	0.1474	93.6%	0.0032	0.1287	0.1234	93.0%
		β_2	−0.1536	0.0903	0.0879	58.8%	0.0044	0.1286	0.1286	95.6%	0.0035	0.1101	0.1069	93.6%

Naive: naive estimate by ignoring double truncation; Wilcoxon: Wilcoxon weight estimate; log-rank$_3$: log-rank weight estimate with $k = 3$; BIAS: average bias of the estimates; SE: standard error of the estimates; SEE: average of the estimated standard errors; CP 95%: empirical coverage probabilities of Wald-type confidence intervals with 95% confidence level.

Table 3.

Results from the quasar data: estimation for the model with linear and quadratic term.

Parameter	EST	SE	*p*-value
θ_1	7.6776	2.6396	0.0036
θ_2	−3.3173	2.2408	0.1388

EST: estimate of the parameter; SE: estimated standard error; p-value: asymptotic p-value of the significance test for $H_0 : \theta_j = 0$ against $H_a : \theta_j \neq 0$, $j = 1, 2$.