Electronic Journal of Statistics
Vol. 17 (2023) 2317–2342
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2146

Maximum profile binomial likelihood estimation for the semiparametric Box–Cox power transformation model*

Pengfei Li

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada, N2L 3G1
e-mail: pengfei.li@uwaterloo.ca

Tao Yu

Department of Statistics and Data Science, National University of Singapore, Singapore, 117546
e-mail: stayt@nus.edu.sg

Baojiang Chen†

Department of Biostatistics and Data Science, University of Texas Health Science Center at Houston, School of Public Health, Austin, Texas, 78731, U.S.A.
e-mail: baojiang.chen@uth.tmc.edu

Jing Qin

National Institute of Allergy and Infectious Diseases, National Institutes of Health, MD 20892, U.S.A.
e-mail: jingqin@niaid.nih.gov

Abstract: The Box–Cox transformation model has been widely applied for many years. The parametric version of this model assumes that the random error follows a parametric distribution, say the normal distribution, and estimates the model parameters using the maximum likelihood method. The semiparametric version assumes that the distribution of the random error is completely unknown; existing methods either need strong assumptions, or are less effective when the distribution of the random error significantly deviates from the normal distribution. We adopt the semiparametric assumption and propose a maximum profile binomial likelihood method. We theoretically establish the joint distribution of the estimators of the model parameters. Through extensive numerical studies, we demonstrate that our method has an advantage over existing methods when the distribution of the random error deviates from the normal distribution. Furthermore, we compare the performance of our method and existing methods on an HIV data set.

MSC2020 subject classifications: Primary 62N99; secondary 62N02.

*Dr. Li's work is supported in part by the Natural Sciences and Engineering Research Council of Canada (grant number RGPIN-2020-04964). Dr. Yu's work is supported in part by Singapore Ministry of Education Academic Research Tier 1 Fund: A-8000413-00-00.

†Corresponding author.

Keywords and phrases: Binomial likelihood, Box–Cox transformation, empirical processes, M-estimation, semiparametric inference, U-processes.

Received March 2022.

Contents

1. Introduction

Since the seminal work of [8], the Box–Cox power transformation model has been extensively studied and applied in various disciplines. Let $(Y_i, X_i), i = 1, \ldots, n$ be independent and identically distributed (i.i.d.) observations with Y_i the response and $X_i = (X_{i1}, \ldots, X_{ip})^T$ the corresponding covariates. The Box–Cox model assumes that

$$Y_i^{(\lambda)} = \gamma + X_i^T \beta + \epsilon_i, \tag{1.1}$$

where $Y^{(\lambda)} = (Y^\lambda - 1)/\lambda$ if $\lambda \neq 0$ and $\log Y$ otherwise; λ, γ, and β are the parameters of interest; and $\epsilon_i, i = 1, \ldots, n$, are i.i.d. mean 0 random errors.

When the distribution of ϵ_i is assumed to be known only up to an unknown finite-dimensional parameter, we have the parametric Box–Cox power transformation model. This model has been studied extensively under the assumption that the ϵ_i's are i.i.d. equal-variance normal random variables; see, for example, [8, 6, 23, 11, 33, 34, 35, 31]. The maximum likelihood principle has proved a powerful tool, but the parametric assumption may be too strong. It could be severely violated in many practical applications, leading to biased inference results; see our numerical studies for details.

It is not uncommon for the distribution of the random error in the Box–Cox transformation model to deviate from normal. For example, in survival analysis, the well-known proportional hazard model [15, 16] is equivalent to the Box–Cox transformation model with the error following an extreme value distribution if the baseline hazard function is the Weibull distribution. See [25] and [18] for more discussion of the connection between the Box–Cox transformation model and the proportional hazard model. The proportional odds model [4, 5] is another example. It assumes that $\log[\{1-S_0(Y)\}/S_0(Y)] = X^T\beta + \epsilon$, where $S_0(\cdot)$ is the baseline survival function; the random error ϵ follows the logistic distribution. Therefore, when $\log[\{1-S_0(Y)\}/S_0(Y)]$ is assumed to be a power function of Y, this is the Box–Cox transformation model with the error following the logistic distribution. Additionally, in cases where the error distribution is skewed or heavy-tailed, quantile regression methods have been developed; see [9, 27, 17], and the references therein for details.

In this paper, we assume that the distribution of ϵ_i is completely unknown; parametric models where the error distribution deviates from normal are special cases of our approach. [3, 28, 30] have proposed quasi-likelihood estimating equation methods for this semiparametric Box–Cox power transformation model. However, [19] showed that the root of the expectation of the corresponding estimating equation is generally not unique, and therefore the resulting estimator is not consistent. They instead proposed a "minimum distance" estimator for λ and a least-square estimator for β, and they established the joint asymptotic distribution for these estimators.

[19] successfully established the asymptotic normality of their (λ, β) estimator under the assumption that the distribution of ϵ_i is completely unknown. However, their approach has two limitations. First, their estimator for β is based on the least-square method. This method performs well when the underlying distribution of ϵ_i is close to normal; but if it is not, the estimator may have less accurate numerical performance. This, in turn, affects the performance of the estimator for λ. Our simulation study demonstrates this; see Section 5 for details. Second, their method is based on the minimum distance method and does not have a likelihood interpretation. We study model (1.1) under the same assumptions used in [19]. In other words, we consider the case where the error distribution is completely unknown. Based on the distribution of $I(Y_i \leq t)$, we propose a profile binomial likelihood method. Our method has three main advantages. (1) It is a likelihood-based method, which is known to be more efficient than other methods in many scenarios. For example, for parametric models, it has been proven to be efficient under mild conditions; it can also achieve semiparametric efficiency for many semiparametric models. For further details, see [7, 24]. (2) Our binomial likelihood is a joint objective function for (λ, β), allowing us to estimate them simultaneously through the likelihood; in contrast, the method proposed by [19] requires a two-stage estimation procedure. (3) Our binomial likelihood incorporates all the $I(Y_i \leq Y_j)$, $i = 1, \ldots, n$; $j = 1, \ldots, n$, which encompass all the rank information of the responses. Hence, we anticipate that our method may have the benefits of rank-based methods. Theoretically, because our binomial likelihood function is a U-process with a plugged-in nonparametric component, existing U-process theory is not applicable. With the help of the advanced empirical processes theory, we derive the joint asymptotic normality of our estimators for λ and β. These developments may benefit research into M-estimators with objective functions being U-processes.

Our simulation studies demonstrate that when the distribution of ϵ_i deviates from normal, our method achieves more accurate parameter estimates than existing methods. However, when the distribution of ϵ_i is normal, the parametric methods perform the best, while the performance of our method and the method proposed by [19] is mixed.

The paper is organized as follows. Section 2 gives a brief review of the methods that will be compared with our approach in the numerical studies. Section 3 proposes the maximum profile binomial likelihood method for estimating the parameters under the Box–Cox power transformation model and presents an algorithm for obtaining our estimates numerically. Section 4 studies the joint asymptotic properties of our estimates. Section 5 discusses the simulation studies, Section 6 presents the HIV application, and Section 7 concludes the paper with a discussion. For convenience of presentation, the technical details are provided in two Appendices and the supplementary material.

2. Existing methods

With a parametric assumption on the distribution of ϵ, the Box–Cox model (1.1) can be analyzed by the classical maximum likelihood principle; see, for example, [8, 6, 23, 11, 33, 34, 35, 31]. The most popular parametric assumption is that $\epsilon_i, i = 1, \dots, n$ are i.i.d. $N(0, \sigma^2)$ random variables. Under this assumption, the classical maximum likelihood estimators of $(\lambda, \gamma, \beta, \sigma)$ maximize the log-likelihood function given by

$$-\frac{1}{2}\sum_{i=1}^{n}(Y_i^{(\lambda)} - \gamma - X_i^T\beta)^2/\sigma^2 - \frac{n}{2}\log(2\pi\sigma^2) + (\lambda - 1)\sum_{i=1}^{n}\log Y_i.$$

We can use existing R functions, such as the "powerTransform" function in the package *car*, to compute these estimates numerically. In the numerical studies, we will compare this parametric method with our method.

[19] proposed a semiparametric estimation approach that proceeds as follows. For a given λ, the model parameters $(\gamma, \beta^T)^T$ in Model (1.1) can be estimated by the classical least-square principle, namely,

$$\left(\widehat{\gamma}(\lambda), \widehat{\beta}^T(\lambda)\right)^T = \left(\sum_{i=1}^{n} X_i^* X_i^{*T}\right)^{-1} \sum_{i=1}^{n} X_i^* Y_i^{(\lambda)},$$

where $X_i^* = (1, X_i^T)^T$. Then, since $P(Y \le t) = F_\epsilon(t^{(\lambda)} - \gamma - X_i^T\beta)$ with $F_\epsilon(\cdot)$ being the cumulative distribution function (c.d.f.) of ϵ_i, λ can be estimated by a "minimum distance" estimator that minimizes $S_n(\lambda, \widehat{\gamma}(\lambda), \widehat{\beta}(\lambda))$, where

$$\begin{aligned} S_n(\lambda, \gamma, \beta) &= n^{-1}\sum_{i=1}^{n}\int_0^\infty \left\{I(Y_i \le t) - \widetilde{G}_{\lambda,\beta}(t^{(\lambda)} - \gamma - X_i^T\beta)\right\}^2 dW(t), \\ \widetilde{G}_{\lambda,\beta}(t) &= \frac{1}{n}\sum_{j=1}^{n} I\left\{Y_j^{(\lambda)} - \gamma - X_j^T\beta \le t\right\}, \end{aligned}$$

and $W(\cdot)$ is a positive, differentiable, strictly increasing, deterministic, and bounded weight function. In their numerical study, [19] set $W(\cdot)$ to a normal density with the mean and standard derivation being the sample mean and sample standard error of the Y_i's. Since $S_n(\lambda, \widehat{\gamma}(\lambda), \widehat{\beta}(\lambda))$ is a function of the one-dimensional parameter λ, a grid search can be used to find this λ estimate. In the numerical studies, we will also compare this semiparametric method with our approach.

3. Maximum profile binomial likelihood estimation

With the observed data $(Y_i, X_i), i = 1, \dots, n$, we consider the Box–Cox transformation model (1.1). We assume that the errors ϵ_i are i.i.d. and independent of X_i. Let $F(\cdot)$ be the c.d.f. of $\epsilon^* = \epsilon + \gamma$. For any $t > 0$, we have

$$P(Y_i \leq t | X_i) = P\Big(\epsilon_i^* \leq t^{(\lambda)} - X_i^T\beta \Big| X_i, Y_j\Big) = F(t^{(\lambda)} - X_i^T\beta).$$

Conditioning on X_i, $I(Y_i \leq t)$ follows a Bernoulli distribution with the probability of success for this Bernoulli distribution is $F(t^{(\lambda)} - X_i^T\beta)$; here $I(\cdot)$ is the indicator function. We observe that the similar idea has been applied in other statistical models, e.g., [29, 26, 36]. Therefore, conditioning on $X_i, i = 1, \dots, n$, the log-likelihood of $\{I(Y_i \leq t)\}_{i=1}^n$ is given by

$$\begin{aligned}\tilde{l}(\lambda, \beta, F; t) \;=\; \sum_{i=1}^n \Big[& I(Y_i \leq t) \log\Big\{F\Big(t^{(\lambda)} - X_i^T\beta\Big)\Big\} \\ & + I(Y_i > t) \log\Big\{1 - F\Big(t^{(\lambda)} - X_i^T\beta\Big)\Big\}\Big].\end{aligned}$$

We suggest choosing the values of t as the observed responses $\{Y_j\}_{j=1}^n$ and taking the summation of $\tilde{l}(\lambda, \beta, F; Y_j)$ over j; this leads to the binomial likelihood

$$\begin{aligned}\tilde{l}_B(\lambda, \beta, F) \;=\; \sum_{j=1}^n \sum_{i=1}^n \Big[& I_{i,j} \log\Big\{F\Big(Y_j^{(\lambda)} - X_i^T\beta\Big)\Big\} \\ & + (1 - I_{i,j}) \log\Big\{1 - F\Big(Y_j^{(\lambda)} - X_i^T\beta\Big)\Big\}\Big], \qquad (3.1)\end{aligned}$$

where $I_{i,j} = I(Y_i \leq Y_j)$.

Note that $F(\cdot)$ is an infinite-dimensional parameter. Estimating (F, λ, β) simultaneously by maximizing $\tilde{l}_B(\lambda, \beta, F)$ is possible but computationally demanding; this also leads to theoretical difficulties in the subsequent development of the asymptotic distributions of the estimates ([13]). Since $F(\cdot)$ is the distribution function of ϵ_i^*, we can instead use the following profile approach to estimate it by the empirical distribution function. For given λ and β, based on (1.1), we have $\epsilon_i^* = Y_i^{(\lambda)} - X_i^T\beta$; therefore, we consider

$$\widehat{G}_{\lambda,\beta}(t) \;=\; \frac{1}{n}\sum_{i=1}^n I\left\{Y_i^{(\lambda)} - X_i^T\beta \leq t\right\},$$

$$\widehat{F}_{\lambda,\beta}(t) \quad = \quad \left\{\widehat{G}_{\lambda,\beta}(t) \vee n^{-2}\right\} \wedge \left(1-n^{-2}\right), \tag{3.2}$$

where n^{-2} is added to ensure that $\widehat{F}_{\lambda,\beta}(\cdot)$ stays away from 0 and 1 to avoid complications in both the numerical analyses and the technical development. Substituting (3.2) into (3.1), we obtain the profile binomial likelihood:

$$\begin{aligned}\ell(\lambda,\beta) \quad = \quad & \sum_{j=1}^{n}\sum_{i=1}^{n}\left[I_{i,j}\log\left\{\widehat{F}_{\lambda,\beta}\left(Y_j^{(\lambda)}-X_i^T\beta\right)\right\}\right. \\ & \left.+(1-I_{i,j})\log\left\{1-\widehat{F}_{\lambda,\beta}\left(Y_j^{(\lambda)}-X_i^T\beta\right)\right\}\right]. \end{aligned} \tag{3.3}$$

Consequently, we define

$$\left(\widehat{\lambda},\widehat{\beta}^T\right)^T = \arg\max_{(\lambda,\beta^T)^T\in\Theta}\ell(\lambda,\beta), \tag{3.4}$$

where Θ is a compact subset of $\mathbb{R}^{p+1}$, and γ is then estimated by

$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^{(\widehat{\lambda})} - X_i^T\widehat{\beta}\right\}.$$

The estimator in (3.4) does not have an explicit form. We implemented the following algorithm in R to compute it numerically.

Step 1. For given λ, we define

$$\beta_\lambda = \arg\max_{\beta}\ell(\lambda,\beta), \tag{3.5}$$

which leads to the profile likelihood for λ, given by

$$p\ell(\lambda) = \ell(\lambda,\beta_\lambda).$$

In our numerical studies, we solve the optimization (3.5) using `optim()` with the default Nelder–Mead method. For the initial values of β, we treated λ as a constant in the model $Y^{(\lambda)} = X^T\beta + \epsilon$ and considered two possibilities: the least-square estimate implemented by `lm()` and the rank-based estimate from `rfit()` in the package `Rfit`.

Step 2. Since $p\ell(\lambda)$ is a function of a one-dimensional parameter λ, we compute $\widehat{\lambda}$ via a grid search maximization.

Step 3. With $\widehat{\lambda}$, we obtain $\widehat{\beta}$ from (3.5).

Remark 1. *As far as we are aware, the work in the literature that is most closely related to our work is [19]. We use the same model assumptions and have included the component $I(Y_i \le t)$ in the objective functions. We incorporate this component to establish the binomial likelihood, while [19] use it to construct the L_2-distance. We observe that they estimate (γ,β) by the least-square method for a given λ, and in the construction of their objective function $S_n(\lambda,\gamma,\beta)$*

for the estimation of λ, they suggest the normal distribution as the weights. These choices do not affect the convergence rates of their estimators and should increase the estimation accuracy of the model parameters when the responses and errors are approximately normally distributed. However, when normality is violated, the performance of their method may be affected. In contrast, our method estimates the model parameters by maximizing a profile binomial likelihood, which is unrelated to the normal distribution. We therefore expect that the method of [19] may have better performance when both Y and the random errors are close to the normal distribution, but our method may have the advantage when normality is violated. The observations in our numerical studies reinforce this conjecture; see Section 5 for details.

Remark 2. *Our method can be viewed as a rank-based method because the binomial likelihood (3.3) contains $I_{i,j} = I(Y_i \le Y_j)$ for all $i, j = 1, \dots, n$, which carry all the rank information of the responses. Various rank-based methods for data of the same structure as this article have been proposed in the literature; for example, the maximum rank correlation (MRC) estimator ([21, 32]), the monotone rank (MR) method ([12]), and the pairwise-difference rank (PDR) method ([1, 2]). But these methods are different from our method and may not be appropriate for the Box-Cox model for two reasons. (1) They were constructed not for the Box-Cox model, but for the transformation model:*

$$H(Y_i) = X_i^T \beta + \epsilon_i, \tag{3.6}$$

with $H(\cdot)$ being assumed to be a monotonic nonparametric function. If the data are truly from the Box-Cox transformation model, the analysis results from these methods may be less efficient. Furthermore, to be identifiable, Model (3.6) needs an assumption on the model parameter β, $\|\beta\|_2 = 1$ say; in the other words, the β estimates are directions, and are of different meaning from those based on the Box-Cox models. (2) Our method is constructed based on the conditional distribution of I_{ij}, but other methods are not.

4. Joint asymptotic distribution of estimators

In this section, we derive the joint asymptotic distribution of $\left(\widehat{\lambda}, \widehat{\beta}^T\right)^T$ defined by (3.4). We need the following notation. Let $\theta = (\lambda, \beta^T)^T$ and $\widehat{\theta} = \left(\widehat{\lambda}, \widehat{\beta}^T\right)^T$; and let $\theta_0 = (\lambda_0, \beta_0^T)^T$ be the true values of the corresponding parameters. Denote $V_\theta = Y^{(\lambda)} - X^T\beta$, $V_{\theta,i} = Y_i^{(\lambda)} - X_i^T\beta$, and $V_{\theta,i,j} = Y_i^{(\lambda)} - X_j^T\beta$. Define

$$F_\theta(t) = P(Y^{(\lambda)} - X^T\beta \le t) = P(V_\theta \le t).$$

When $\theta = \theta_0$, we write $F_0 = F_{\theta_0}$, $V_0 = V_{\theta_0}$, $V_{0,i} = V_{\theta_0,i}$, $V_{0,i,j} = V_{\theta_0,i,j}$. Let $\dot{F}_\theta(t) = \frac{\partial F_\theta(t)}{\partial \theta}$ and $F'_\theta(t) = \frac{\partial F_\theta(t)}{\partial t}$, if they exist; and denote $\dot{F}_0(t) = \dot{F}_{\theta_0}(t)$,

$F_0'(t) = F_{\theta_0}'(t)$. Let

$$\dot{V}_\theta = \frac{\partial V_\theta}{\partial \theta} = \begin{cases} \begin{pmatrix} \lambda^{-2}\left\{\lambda Y^\lambda \log Y - Y^\lambda + 1\right\} \\ -X \end{pmatrix} & \text{if} \quad \lambda \neq 0 \\ \begin{pmatrix} (\log Y)^2/2 \\ -X \end{pmatrix} & \text{if} \quad \lambda = 0 \end{cases}, \tag{4.1}$$

and define $\dot{V}_0$, $\dot{V}_{0,i}$, and $\dot{V}_{0,i,j}$ similarly.

Furthermore, we denote $Z = (Y, X)$ and $\boldsymbol{z} = (y, \boldsymbol{x})$. Define

$$\varphi(\boldsymbol{z}) = E\left[\frac{\dot{F}_0(V_{0,2,1}) + F_0'(V_{0,2,1})\dot{V}_{0,2,1}}{F_0(V_{0,2,1})\left\{1 - F_0(V_{0,2,1})\right\}}\left\{I(Y_1 \leq Y_2) - F_0(V_{0,2,1})\right\}\middle| Z_1 = \boldsymbol{z}\right], \tag{4.2}$$

$$\psi(\boldsymbol{z}) = -E\left[\frac{\dot{F}_0(V_{0,2,1}) + F_0'(V_{0,2,1})\dot{V}_{0,2,1}}{F_0(V_{0,2,1})\left\{1 - F_0(V_{0,2,1})\right\}} I\left(V_{0,3} \leq V_{0,2,1}\right)\middle| Z_3 = \boldsymbol{z}\right], \tag{4.3}$$

$$\Sigma_1 = E\left(\left[\frac{\left\{\dot{F}_0(V_{0,2,1}) + F_0'(V_{0,2,1})\dot{V}_{0,2,1}\right\}\left\{\dot{F}_0(V_{0,2,1}) + F_0'(V_{0,2,1})\dot{V}_{0,2,1}\right\}^T}{F_0(V_{0,2,1})\left\{1 - F_0(V_{0,2,1})\right\}}\right]\right), \tag{4.4}$$

$$\Sigma_2 = \mathrm{var}\left\{\varphi(Z) + \psi(Z)\right\}. \tag{4.5}$$

The following theorem establishes the joint asymptotic distribution of $\left(\widehat{\lambda}, \widehat{\beta}^T\right)^T$.

Theorem 4.1. *Assume Conditions 1–5 in Appendix A; then*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow N(0, \Sigma),$$

where $\Sigma = \frac{1}{4}\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}$ *with* Σ_1 *and* Σ_2 *defined by (4.4) and (4.5) respectively.*

Note that deriving the asymptotic properties for $\widehat{\theta}$ is a challenging task. The main difficulty is the complicated structure of the profile binomial likelihood $\ell(\cdot)$ defined by (3.3). Clearly, it is a U-process, with a plugged-in nonparametric component $\widehat{F}_{\lambda,\beta}(\cdot)$. Existing U-process theory is not applicable in our context. We use advanced empirical process theory ([37, 24]) to derive the asymptotic normality of $\widehat{\theta}$ presented in Theorem 4.1. For continuity of presentation, we sketch the lengthy proof of this theorem in Appendix B and relegate the full details to the supplementary document.

Remark 3. *Our method remains applicable when* λ_0 *is a known quantity. In such cases, we only estimate* β *from the model. Theorem 4.1 still holds, but with* $\widehat{\theta}$ *and* θ_0 *replaced by* $\widehat{\beta}$ *and* β_0*, respectively. Likewise, in equations* (4.2)–(4.5)*,* $\varphi(\boldsymbol{z})$*,* $\psi(\boldsymbol{z})$*,* Σ_1*, and* Σ_2 *are respectively replaced by:*

$$\varphi(\boldsymbol{z}) = -E\left[\frac{\{X_1 - E(X)\}F_0'(V_{0,2,1})}{F_0(V_{0,2,1})\left\{1 - F_0(V_{0,2,1})\right\}}\left\{I(Y_1 \leq Y_2) - F_0(V_{0,2,1})\right\}\middle| Z_1 = \boldsymbol{z}\right],$$

$$\psi(\boldsymbol{z}) = E\left[\frac{\{X_1 - E(X)\}F_0'(V_{0,2,1})}{F_0(V_{0,2,1})\left\{1 - F_0(V_{0,2,1})\right\}} I\left(V_{0,3} \leq V_{0,2,1}\right)\middle| Z_3 = \boldsymbol{z}\right],$$

$$\Sigma_1 = E\left[\frac{F_0'^2(V_{0,2,1})\{X_1 - E(X)\}\{X_1 - E(X)\}^T}{F_0(V_{0,2,1})\,\{1 - F_0(V_{0,2,1})\}}\right],$$
$$\Sigma_2 = var\{\varphi(Z) + \psi(Z)\}\,.$$

Numerical examples show that, compared to the case where λ_0 is unknown and estimated from the model, when λ_0 is known, the corresponding variances of the β estimates are significantly reduced; the details are omitted. This complies with the discussion of [6, 23].

5. Simulation study

5.1. Data simulation

We use the following simulation examples to examine the numerical performance of our method. We compare our method (labeled "Our") with the method of [19] ("Foster") and the classical parametric method ("Parametric").

We simulate the covariates X_1, X_2, X_3, X_4 as follows. Let $S_1 = (S_{11}, S_{12})^T$ and $S_2 = (S_{21}, S_{22})^T$ be independent random vectors from

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.6\\ 0.6 & 1\end{pmatrix}\right).$$

Set $X_1 = -\log\{1 - \Phi(S_{11})\}$, $X_2 = I(S_{21} > 0)$, $X_3 = -\log\{1 - \Phi(S_{12})\}$, and $X_4 = I(S_{22} > 0)$. Then X_1 and X_3 follow the Exponential(1) distribution, while X_2 and X_4 follow the Bernoulli(0.5) distribution. Based on these covariates, we consider six simulation models:

Model 1: $\log Y = X_1 + X_2 + \epsilon$;
Model 2: $\log Y = X_1 + X_2 + X_3 + X_4 + \epsilon$;
Model 3: $Y = 4 + 2.5X_1 + 2.5X_2 + \epsilon$;
Model 4: $Y = 4 + 1.2X_1 + 1.2X_2 + 1.2X_3 + 1.2X_4 + \epsilon$;
Model 5: $5/Y = 4 + 2.5X_1 + 2.5X_2 + \epsilon$;
Model 6: $5/Y = 4 + 1.2X_1 + 1.2X_2 + 1.2X_3 + 1.2X_4 + \epsilon$.

For Models 1 and 2, $\lambda = 0$; for Models 3 and 4, $\lambda = 1$; and for Models 5 and 6, $\lambda = -1$. For each model, we consider two distributions for ϵ, $N(0, 0.5^2)$ and $0.5(\chi_1^2 - 1)$, and two sample sizes, $n = 100$ and $n = 200$. For each scenario, we use 1000 repetitions.

5.2. Estimation results

We examine the performance of the three methods by evaluating their bias, mean squared error (MSE), coverage proportion (CP) and average length (AL) of the 95% bootstrap percentile confidence intervals (BPCIs) in the estimation of the model parameters λ, β_1, and β_2; here β_1 and β_2 are the coefficients of X_1 and X_2 in our simulation models. The results for β_3 and β_4, i.e., the coefficients for X_3 and X_4 in Models 2, 4, and 6, are similar to those for β_1 and β_2 and are omitted.

Table 1 presents the bias and MSE values, and Table 2 gives the CPs and ALs of the BPCIs when ϵ is simulated as $N(0, 0.5^2)$. From Table 1, we observe that all the methods have small biases. The parametric method results in the smallest MSEs in every scenario. This is not surprising since the assumption that the random error follows the normal distribution is satisfied; the other methods do not need this assumption. For our method and Foster: (1) when $\lambda = 0$ (Models 1 and 2), our method has slightly smaller MSEs; (2) when $\lambda = 1$ (Models 3 and 4), Foster performs slightly better; (3) when $\lambda = -1$, the MSE values are similar. This supports our remark in Section 3 that Foster may perform well when the distribution of the random error is close to normal. The results presented in Table 2 are consistent with those shown in Table 1. All methods have produced reliable coverage probabilities for all models and parameters. The parametric method has yielded the shortest ALs, while the ALs of our method and the Foster method are similar.

TABLE 1
Bias and MSE for the estimates of λ, β_1, and β_2: $\epsilon \sim N(0, 0.5^2)$. The reported MSEs for Models 1–6 are MSE$\times 100$.

		Parametric		Foster		Our		Parametric		Foster		Our	
n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		Model 1						Model 2					
100	λ	0.00	0.21	0.00	0.79	0.00	0.37	0.00	0.04	0.01	0.24	0.00	0.07
100	β_1	0.01	1.33	0.01	4.17	0.01	1.94	0.01	1.06	0.05	5.80	0.01	1.72
100	β_2	0.00	1.49	0.00	2.64	0.00	2.01	0.01	1.66	0.03	3.81	0.01	2.16
200	λ	0.01	0.09	0.00	0.37	0.00	0.17	0.00	0.01	0.00	0.10	0.00	0.03
200	β_1	0.01	0.62	0.02	2.30	0.01	0.96	0.01	0.42	0.02	2.37	0.00	0.72
200	β_2	0.01	0.72	0.02	1.31	0.01	0.97	0.01	0.74	0.01	1.61	0.00	0.97
		Model 3						Model 4					
100	λ	0.00	0.71	0.00	1.01	−0.01	1.29	0.01	1.36	0.01	2.06	0.00	2.36
100	β_1	0.05	23.01	0.04	32.75	0.05	40.13	0.07	11.61	0.08	20.47	0.08	20.56
100	β_2	0.04	19.77	0.03	27.20	0.03	33.89	0.07	10.46	0.06	16.64	0.07	18.40
200	λ	0.01	0.31	0.00	0.47	0.01	0.60	0.00	0.50	−0.01	0.85	−0.01	1.05
200	β_1	0.05	10.46	0.05	15.97	0.07	19.20	0.02	3.83	0.01	6.21	0.01	7.32
200	β_2	0.05	8.74	0.05	13.16	0.07	16.22	0.02	3.76	0.01	5.74	0.02	6.86
		Model 5						Model 6					
100	λ	0.00	0.71	0.00	1.23	0.01	1.32	−0.01	1.36	−0.01	2.51	−0.01	2.37
100	β_1	0.00	0.07	0.00	0.12	0.00	0.11	0.00	0.04	0.00	0.06	0.00	0.06
100	β_2	0.00	0.07	0.00	0.09	0.00	0.09	0.00	0.06	0.00	0.07	0.00	0.08
200	λ	−0.01	0.31	−0.01	0.55	−0.01	0.59	0.00	0.50	0.01	1.10	0.01	1.05
200	β_1	0.00	0.03	0.00	0.06	0.00	0.05	0.00	0.02	0.00	0.03	0.00	0.03
200	β_2	0.00	0.03	0.00	0.04	0.00	0.04	0.00	0.03	0.00	0.03	0.00	0.04

Tables 3 and 4 display the results obtained when simulating ϵ as $0.5(\chi_1^2 - 1)$, resulting in a non-normal distribution of the random error. The parametric method exhibits larger biases and MSEs than the other methods and consistently yields coverage probabilities below the nominal level of 95% in all scenarios. In contrast, our method and Foster's method demonstrate small and comparable biases, achieving coverage probabilities close to or greater than 95%. Our method exhibits over-coverage in some models, but it also has significantly

TABLE 2
CP(× 100) and AL of BPCI of λ, β_1, and β_2: $\epsilon \sim N(0, 0.5^2)$.

		Parametric		Foster		Our		Parametric		Foster		Our	
n		CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
		Model 1						Model 2					
100	λ	92.00	0.18	97.50	0.39	96.90	0.29	95.10	0.08	98.70	0.21	97.70	0.12
100	β_1	93.70	0.45	96.90	0.86	95.70	0.62	94.50	0.41	98.20	0.98	97.60	0.55
100	β_2	93.50	0.47	97.50	0.68	96.90	0.59	93.60	0.49	97.80	0.80	96.00	0.59
200	λ	92.60	0.11	98.40	0.25	96.40	0.18	91.70	0.05	98.80	0.14	94.20	0.07
200	β_1	94.80	0.30	98.10	0.60	95.70	0.40	93.20	0.26	98.80	0.65	95.60	0.35
200	β_2	93.70	0.32	97.40	0.46	94.80	0.39	95.00	0.33	97.80	0.53	95.70	0.39
		Model 3						Model 4					
100	λ	94.50	0.34	95.20	0.46	97.40	0.52	93.80	0.46	95.80	0.61	97.90	0.69
100	β_1	95.00	1.98	95.70	2.62	97.40	2.91	93.50	1.33	96.10	1.78	97.60	1.99
100	β_2	95.40	1.81	95.60	2.39	97.30	2.65	93.80	1.27	95.80	1.66	97.70	1.89
200	λ	94.30	0.22	95.40	0.29	97.40	0.33	93.40	0.29	94.50	0.39	96.60	0.43
200	β_1	94.20	1.27	95.20	1.65	97.30	1.80	93.60	0.81	94.30	1.05	96.50	1.14
200	β_2	94.40	1.15	95.30	1.49	96.70	1.65	93.70	0.79	94.20	0.99	97.00	1.09
		Model 5						Model 6					
100	λ	92.30	0.34	95.40	0.50	97.00	0.52	94.20	0.46	95.60	0.68	97.80	0.69
100	β_1	93.20	0.11	95.60	0.15	96.00	0.15	93.50	0.08	94.90	0.10	96.70	0.11
100	β_2	93.30	0.10	95.20	0.12	95.50	0.13	92.50	0.10	94.90	0.11	96.00	0.12
200	λ	94.80	0.22	95.20	0.32	96.80	0.33	94.70	0.29	95.50	0.43	97.60	0.43
200	β_1	95.30	0.07	94.80	0.10	96.60	0.10	93.30	0.05	95.20	0.07	95.20	0.07
200	β_2	94.20	0.07	95.20	0.08	95.70	0.08	94.00	0.07	94.50	0.07	95.10	0.08

smaller MSEs and comparable or smaller ALs than the other methods across all scenarios, supporting our Remark 1.

TABLE 3
Bias and MSE for the estimates of λ, β_1, and β_2: $\epsilon \sim 0.5(\chi_1^2 - 1)$. The reported MSEs for Models 1–4 are MSE$\times 100$; those for Models 5 and 6 are MSE$\times 1000$.

		Parametric		Foster		Our		Parametric		Foster		Our	
n		Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
		Model 1						Model 2					
100	λ	−0.18	4.15	−0.01	1.02	0.01	0.12	−0.04	0.29	−0.01	0.39	0.00	0.02
100	β_1	−0.31	11.13	−0.01	4.01	0.01	0.21	−0.15	3.86	−0.01	8.32	0.01	0.33
100	β_2	−0.22	6.14	−0.01	3.44	0.01	0.17	−0.10	3.36	−0.01	5.51	0.01	0.44
200	λ	−0.19	4.11	−0.01	0.44	0.00	0.03	−0.04	0.23	−0.01	0.14	0.00	0.01
200	β_1	−0.34	11.96	−0.01	2.09	0.00	0.05	−0.15	3.11	−0.03	2.75	0.01	0.07
200	β_2	−0.22	5.73	−0.01	1.85	0.00	0.04	−0.10	2.21	−0.02	2.24	0.01	0.09
		Model 3						Model 4					
100	λ	−0.13	3.94	0.00	0.59	0.01	0.33	−0.21	8.23	−0.01	1.39	0.02	0.64
100	β_1	−0.54	59.66	0.02	15.49	0.07	10.08	−0.38	23.88	0.01	8.55	0.06	4.64
100	β_2	−0.51	51.00	0.02	14.97	0.06	7.89	−0.35	21.37	0.02	10.17	0.06	4.27
200	λ	−0.14	2.99	0.00	0.24	0.00	0.07	−0.21	6.63	0.00	0.48	0.01	0.16
200	β_1	−0.60	53.05	−0.01	5.70	0.02	1.83	−0.41	21.59	0.00	2.90	0.03	1.04
200	β_2	−0.55	44.10	−0.01	6.21	0.02	1.53	−0.38	19.10	0.00	3.52	0.03	0.93
		Model 5						Model 6					
100	λ	0.13	39.40	0.00	6.26	−0.01	3.31	0.21	82.28	0.01	15.22	−0.02	6.46
100	β_1	0.04	2.65	0.00	0.42	0.00	0.17	0.03	1.30	0.00	0.45	0.00	0.11
100	β_2	0.02	1.54	0.00	0.93	0.00	0.10	0.02	1.20	0.00	1.08	0.00	0.16
200	λ	0.14	29.93	0.00	2.57	0.00	0.69	0.21	66.28	0.01	5.32	−0.01	1.68
200	β_1	0.04	2.27	0.00	0.16	0.00	0.04	0.03	1.00	0.00	0.17	0.00	0.02
200	β_2	0.02	1.08	0.00	0.51	0.00	0.03	0.02	0.73	0.00	0.49	0.00	0.03

Table 4
CP(× 100) and AL of BPCI of λ, β_1, and β_2: $\epsilon \sim 0.5(\chi_1^2 - 1)$.

		Parametric		Foster		Our		Parametric		Foster		Our	
n		CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL
		Model 1						Model 2					
100	λ	27.30	0.28	98.90	0.47	98.30	0.20	77.00	0.12	99.10	0.28	98.60	0.09
100	β_1	23.30	0.37	98.60	0.87	98.30	0.30	77.50	0.45	98.40	1.33	98.00	0.36
100	β_2	50.60	0.41	96.20	0.78	99.50	0.26	87.30	0.56	97.50	1.08	97.70	0.41
200	λ	3.60	0.21	97.00	0.29	97.20	0.10	55.60	0.09	98.80	0.17	97.70	0.04
200	β_1	3.10	0.28	97.80	0.60	97.80	0.14	61.40	0.33	97.90	0.79	98.70	0.17
200	β_2	25.60	0.31	94.90	0.54	98.20	0.12	80.10	0.40	95.50	0.65	97.90	0.19
		Model 3						Model 4					
100	λ	84.90	0.48	94.70	0.36	99.40	0.34	78.90	0.65	96.00	0.53	98.90	0.50
100	β_1	83.60	1.91	95.20	1.89	99.40	1.93	77.00	1.09	95.30	1.35	99.00	1.44
100	β_2	83.10	1.70	94.40	1.79	99.40	1.75	79.60	1.06	95.80	1.35	98.90	1.37
200	λ	69.20	0.35	94.10	0.21	99.70	0.16	55.30	0.48	95.10	0.31	98.90	0.24
200	β_1	67.80	1.34	94.40	1.07	99.80	0.86	54.60	0.76	94.90	0.78	98.70	0.63
200	β_2	64.40	1.21	94.10	1.06	99.70	0.79	57.10	0.74	93.90	0.81	98.70	0.60
		Model 5						Model 6					
100	λ	83.00	0.48	96.60	0.38	98.50	0.35	79.40	0.65	97.60	0.57	98.70	0.50
100	β_1	79.90	0.12	96.30	0.10	97.40	0.08	81.60	0.09	95.40	0.09	98.50	0.07
100	β_2	86.70	0.11	93.60	0.12	98.50	0.07	90.20	0.11	93.20	0.12	97.50	0.08
200	λ	64.80	0.35	93.80	0.22	96.80	0.15	58.80	0.48	95.40	0.33	96.80	0.23
200	β_1	62.90	0.09	95.70	0.06	97.40	0.04	64.90	0.06	95.20	0.06	97.80	0.03
200	β_2	79.20	0.08	94.00	0.08	97.30	0.03	83.20	0.08	93.70	0.09	97.10	0.04

To compare the computational speed of the three methods, we repeated the simulation/estimation 10 times using a single core of the same computer for all six models and two sample sizes ($n = 100$ and $n = 200$). The CPU times (in seconds) to compute the estimate of λ are presented in Table 5. As expected, the parametric method was the fastest, followed by the Foster method. Our method was slower due to the need to maximize a non-smooth objective function $\ell(\lambda, \beta)$, which involves a second-order U-statistic with the empirical c.d.f. in each term. However, we found that the computation time required by our method was reasonable. With the rapid advancement of computational technology, we do not anticipate computation speed to be an obstacle for practical application of our method.

In summary, we observe that the performance of the parametric method relies heavily on the distribution of the random error. Foster may be slightly better than our method when the distribution of the random error is close to normal. Otherwise, our method has much better performance.

6. HIV application

We now apply our method to analyze human immunodeficiency virus (HIV) data from the AIDS Clinical Trials Group Protocol 175 (ACTG175) ([20, 39]) in which $n = 2139$ HIV-infected patients were enrolled. The patients were randomly divided into four arms according to their treatment regimen: (I) zidovudine monotherapy, (II) zidovudine + didanosine, (III) zidovudine + zalcitabine, and (IV) didanosine monotherapy. The data record various measurements from

TABLE 5
CPU time (in seconds) to compute the estimate of λ based on 10 repititions.

	Parametric		Foster		Our		Parametric		Foster		Our	
n	100	200	100	200	100	200	100	200	100	200	100	200
Model	$\epsilon \sim N(0, 0.5^2)$						$\epsilon \sim 0.5(\chi_1^2 - 1)$					
1	< 0.1	< 0.1	1.3	6.3	5.2	14.4	< 0.1	< 0.1	1.3	6.3	5.1	14.0
2	< 0.1	< 0.1	1.3	6.4	11.2	34.5	< 0.1	< 0.1	1.3	6.2	11.7	36.6
3	< 0.1	< 0.1	1.4	6.4	5.6	16.3	< 0.1	< 0.1	1.4	6.4	5.4	15.7
4	< 0.1	< 0.1	1.4	6.4	10.2	33.5	< 0.1	< 0.1	1.4	6.4	10.9	33.0
5	< 0.1	< 0.1	1.4	6.5	5.7	15.2	< 0.1	< 0.1	1.4	6.4	5.6	15.5
6	< 0.1	< 0.1	1.4	6.6	11.1	34.1	< 0.1	< 0.1	1.3	6.3	10.5	31.9

each patient, including age (in years), weight (in kilograms), CD4 cell count at baseline (cd40), CD4 cell count at 20±5 weeks (cd420), CD4 cell count at 96±5 weeks (cd496), CD8 cell count at baseline (cd80), CD8 cell count at 20±5 weeks (cd820), and arm number (arms). The data are available in the `R` package `speff2trial`. The effectiveness of an HIV treatment can be assessed by monitoring the CD4 cell counts of HIV-positive patients: an increased count indicates an improvement in the patient's condition. It is of particular interest to estimate the average CD4 cell count in each arm after 96 weeks of treatment. We take this variable (cd496) plus 1 as the response variable in our analysis. We consider six covariates, age/10, weight/10, cd40/10, cd420/10, cd80/100, and cd820/100, and focus on the complete data for the patients in arm IV.

We apply the three methods from our simulation study to this data set. Table 6 summarizes the point estimate (Est), the corresponding bootstrap standard deviation (BSD), and the 95% BPCIs. Based on the estimates of λ and β from our method, Figure 1 shows the normal probability plot of the F estimate (3.2). We test the normality of the residuals using the Shapiro–Wilk test, which gives a p-value of 0.0015. Both Figure 1 and this test result suggest that the distribution of the random error might deviate from normal. It is therefore not surprising that in Table 6, the estimates of λ and β based on the parametric method are significantly different from those based on the other methods; the former estimates may not be reliable. Our method and Foster lead to λ estimates that are very close to 1 and similar β estimates, but our method has much smaller BSD values and shorter BPCIs for all the parameter estimates. Since the distribution of the random error might deviate from normal, we expect that our method has produced more accurate results than Foster in this real-data example.

TABLE 6
Analysis of ACTG data.

	Parametric			Foster			Our		
	Est	BSD	BPCI	Est	BSD	BPCI	Est	BSD	BPCI
λ	0.76	0.05	$(0.68, 0.89)$	1.00	0.13	$(0.81, 1.30)$	0.95	0.08	$(0.80, 1.10)$
β_1	-0.40	2.14	$(-5.74, 3.21)$	-2.18	15.23	$(-39.14, 21.89)$	-4.17	7.31	$(-22.24, 7.60)$
β_2	1.51	1.51	$(-0.92, 4.88)$	4.94	10.89	$(-6.26, 33.31)$	3.88	5.09	$(-3.59, 14.17)$
β_3	0.86	0.41	$(0.41, 2.05)$	3.36	5.10	$(0.85, 18.49)$	2.63	1.55	$(0.85, 6.62)$
β_4	1.83	0.65	$(1.09, 3.82)$	7.63	10.09	$(2.58, 38.05)$	5.27	2.84	$(2.20, 12.93)$
β_5	0.07	0.81	$(-1.62, 1.38)$	1.66	5.55	$(-5.16, 12.50)$	1.19	2.39	$(-3.52, 5.87)$
β_6	-0.55	0.74	$(-2.04, 0.67)$	-3.40	6.76	$(-27.65, 1.85)$	-2.65	2.80	$(-8.28, 1.05)$

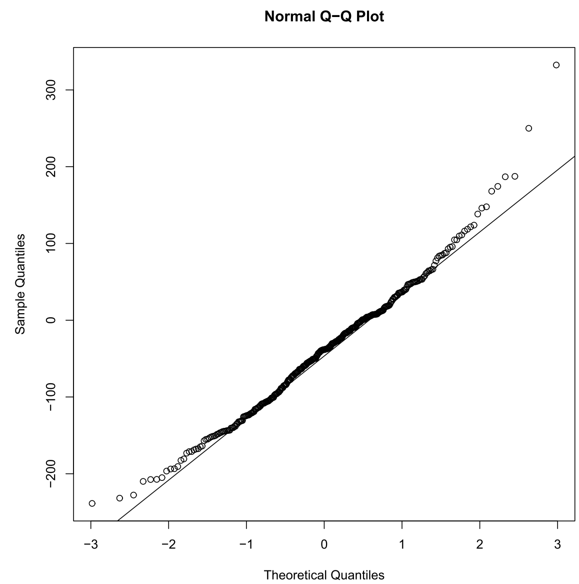

Fig 1. *Q-Q plot of residuals after Box–Cox transformation.*

7. Discussion

We have focused on the Box–Cox model, which has been extensively studied. Classical methods assume that the distribution of the random error is parametric, say normal, and apply the maximum likelihood method to estimate the model parameters. These methods may give misleading results when the parametric assumption is violated. Semiparametric methods assume that the distribution of the random error is unknown. They may be based on the estimating equation method [28, 30], the validity of which relies on a strong and possibly unrealistic assumption; see [19] for a detailed discussion. Alternatively, they may use least-square estimates [19], with lower efficiency when the distribution of the random error deviates from normal; this has been observed in our numerical studies.

We have adopted the semiparametric assumption and proposed a binomial likelihood method for this model. Via extensive numerical analyses, we have compared the performance of our method with the classical parametric method and the method of [19]. When the random error is normally distributed, the parametric method performs the best, and the method of [19] is sightly better than our method only when $\lambda = 1$. However, when the distribution of the random error deviates from normal, our method consistently outperforms the other approaches.

Our proposed pseudo-likelihood (3.3) is a U-process with a nonparametric plug-in component $\widehat{F}_{\lambda,\beta}(\cdot)$. The existing theory for U-processes is not applicable, so developing the theoretical properties of the estimators is a challenging task. We have used advanced empirical process techniques. We believe that these developments will benefit research into M-estimators where the objective function is a U-process. Such estimators are not uncommon; they include the objective function from the pairwise likelihood (e.g., [22]) and that from the binomial/multinomial likelihood [36].

There are several potential research topics for future exploration. Firstly, we have assumed that the effect of the covariates on $Y^{(\lambda)}$ is linear. We could explore this assumption by considering models with more complicated structures. Secondly, the Box-Cox model with the response Y right-censored can also be considered [10, 14]. Thirdly, smoothing techniques can be incorporated into the estimation of the nonparametric function $F(\cdot)$. Finally, our method may be integrated with the quantile regression methods [9, 27, 17] to enhance the stability when the random error distribution is skewed or heavy-tailed.

Appendix A: Regularity conditions

We impose the following regularity conditions to establish our asymptotic results. They are not necessarily the weakest possible.

Condition 1: $\theta = (\lambda, \beta) \in \Theta$, which is a compact subset of $\mathbb{R}^{p+1}$. $F_X(\boldsymbol{x})$ is supported on $\mathcal{X}$ and $F_Y(y)$ is supported on $\mathcal{Y}$. $\mathcal{Z} \equiv \mathcal{X} \times \mathcal{Y}$ is a compact subset of $\mathbb{R}^{p+1}$. Furthermore, $\inf_{y\in\mathcal{Y}} |y| > 0$. Here $F_X(\cdot)$ and $F_Y(\cdot)$ are the c.d.f.s of X and Y respectively.
As a consequence, $t = y^{(\lambda)} - \boldsymbol{x}^T\beta$ is supported on $\mathcal{T}$, which is a compact subset of $\mathbb{R}$.
Condition 2: There exists $\eta_0 > 0$ such that $F_\theta(t)$ is second-order continuously differentiable for $\|\theta - \theta_0\|_2 \leq \eta_0$ and $t \in \mathcal{T}$. Furthermore,

$$0 < \inf_{\boldsymbol{z}\in\mathcal{Z}, \|\theta-\theta_0\|_2\leq\eta_0} F_\theta(\boldsymbol{v}_\theta) \leq \sup_{\boldsymbol{z}\in\mathcal{Z}, \|\theta-\theta_0\|_2\leq\eta_0} F_\theta(\boldsymbol{v}_\theta) < 1$$

and

$$\inf_{\|\theta-\theta_0\|_2\leq\eta_0} \left| \frac{\partial F_\theta(\boldsymbol{v}_\theta)}{\partial\theta} \right| > 0,$$

where $\boldsymbol{v}_\theta = y^{(\lambda)} - \boldsymbol{x}^T\beta$.
Condition 3: For any $t_1, t_2 \in \mathbb{R}$,

$$\sup_{\beta\in\mathcal{B}} \left| F_{X^T\beta}(t_1) - F_{X^T\beta}(t_2) \right| \lesssim |t_1 - t_2|.$$

Condition 4: If $F_\theta(\boldsymbol{v}_\theta) = F_0(\boldsymbol{v}_0)$ almost surely in $F_Y(y)F_X(\boldsymbol{x})$, then $\theta = \theta_0$.
Condition 5: Both Σ_1 and Σ_2 defined by (4.4) and (4.5) are invertible.

Appendix B: Sketch of the Proof of Theorem 4.1

We give a blueprint of the proof of Theorem 4.1; the lengthy details are relegated to the supplementary document.

In addition to the notation of Section 4, we need the following. Throughout the development, "$\lesssim$" denotes smaller than, up to a universal constant; C denotes a large universal constant; and c denotes a small positive universal constant.

For any positive integer i, j, let $Z_{i,j} = (Y_i, X_j)$ and $\boldsymbol{z}_{i,j} = (y_i, \boldsymbol{x}_j)$. Therefore, $Z_{i,i} = Z_i = (Y_i, X_i)$ and likewise $\boldsymbol{z}_{i,i} = \boldsymbol{z}_i = (y_i, \boldsymbol{x}_i)$. Recall that $V_\theta = Y^{(\lambda)} - X^T\beta$, $V_{\theta,i,j} = Y_i^{(\lambda)} - X_j^T\beta$ and define accordingly $\boldsymbol{v}_\theta = y^{(\lambda)} - \boldsymbol{x}^T\beta$, $\boldsymbol{v}_{\theta,i,j} = y_i^{(\lambda)} - \boldsymbol{x}_j^T\beta$. Set $\boldsymbol{v}_0 = \boldsymbol{v}_{\theta_0}$, $\boldsymbol{v}_{0,i,j} = \boldsymbol{v}_{\theta_0,i,j}$.

Recalling the definition of $\dot{V}_\theta$ given by (4.1), we define accordingly

$$\dot{\boldsymbol{v}}_\theta = \frac{\partial \boldsymbol{v}_\theta}{\partial \theta} = \begin{cases} \begin{pmatrix} \lambda^{-2}\left\{\lambda y^\lambda \log y - y^\lambda + 1\right\} \\ -\boldsymbol{x} \end{pmatrix} & \text{if} \quad \lambda \neq 0 \\ \begin{pmatrix} (\log y)^2/2 \\ -\boldsymbol{x} \end{pmatrix} & \text{if} \quad \lambda = 0 \end{cases},$$

and we define $\dot{\boldsymbol{v}}_{\theta,i,j}$, $\dot{\boldsymbol{v}}_0$ similarly.

Let $\{Z_i\}_{i=1,\dots,n}$ be our observations; recall that we have the following definition in Section 3:

$$\begin{aligned} \widehat{G}_\theta(t) &= \frac{1}{n}\sum_{i=1}^n I(Y_i^{(\lambda)} - X_i^T\beta \le t) = \frac{1}{n}\sum_{i=1}^n I(V_{\theta,i} \le t) \\ \widehat{F}_\theta(t) &= \left\{\widehat{G}_\theta(t) \vee n^{-2}\right\} \wedge (1 - n^{-2}). \end{aligned}$$

Let $\widehat{F}_0(t) = \widehat{F}_{\theta_0}(t)$.

The proof has three main steps.

Step 1: Consistency of $\widehat{\boldsymbol{\theta}}$

In Step 1, we show that

$$\widehat{\theta} - \theta_0 = o_p(1). \tag{7.1}$$

To this end, we define

$$M(\theta) = \int \left\{F_0(y_2^{(\lambda_0)} - \boldsymbol{x}_1^T\beta_0) - F_\theta\left(y_2^{(\lambda)} - \boldsymbol{x}_1^T\beta\right)\right\}^2 dF_X(\boldsymbol{x}_1)dF_Y(y_2).$$

Then, based on the arguments in [38], to show (7.1), we need only to show that

(i) $M(\widehat{\theta}) = o_p(1)$;
(ii) $M(\theta) = 0$ implies that $\theta = \theta_0$;
(iii) $M(\theta)$ is continuous in $\theta \in \Theta$.

Note that (ii) holds because of Condition 4 and (iii) holds based on Condition 2. We need to show (i): it follows from Lemmas 1 and 2 given below, which are Lemmas 9 and 10 of the supplementary document. Therefore, the proof of Step 1 is complete.

We need the following notation:

$$\begin{aligned}\gamma_1(y,\boldsymbol{x};F,\lambda,\beta) &= 4\left\{\sqrt{\frac{F_\theta\Big(y^{(\lambda)}-\boldsymbol{x}^T\beta\Big)}{F_0\Big(y^{(\lambda_0)}-\boldsymbol{x}^T\beta_0\Big)}}-1\right\},\\ \gamma_2(y,\boldsymbol{x};F,\lambda,\beta) &= 4\left\{\sqrt{\frac{1-F_\theta\Big(y^{(\lambda)}-\boldsymbol{x}^T\beta\Big)}{1-F_0\Big(y^{(\lambda_0)}-\boldsymbol{x}^T\beta_0\Big)}}-1\right\}.\end{aligned}$$

Lemma 1. *Assume Conditions 1 and 2. We have*

$$\begin{aligned}&\int\Big\{F_0(y_2^{(\lambda_0)}-\boldsymbol{x}_1^T\beta_0)-F_{\widehat{\theta}}\Big(y_2^{(\widehat{\lambda})}-\boldsymbol{x}_1^T\widehat{\beta}\Big)\Big\}^2 dF_X(\boldsymbol{x}_1)dF_Y(y_2)\\ \le\ &\int\Big\{I(y_1\le y_2)\gamma_1(y_2,\boldsymbol{x}_1;\widehat{F},\widehat{\lambda},\widehat{\beta})+I(y_1>y_2)\gamma_2(y_2,\boldsymbol{x}_1;\widehat{F},\widehat{\lambda},\widehat{\beta})\Big\}\\ &\quad\times\Big\{d\mathbb{F}_{X,Y}(\boldsymbol{x}_1,y_1)d\mathbb{F}_{X,Y}(\boldsymbol{x}_2,y_2)-dF_{X,Y}(\boldsymbol{x}_1,y_1)dF_{X,Y}(\boldsymbol{x}_2,y_2)\Big\}+o_p(1).\end{aligned}$$

Lemma 2. *Assume Conditions 1 and 2. We have*

$$\begin{aligned}&\int\Big\{I(y_1\le y_2)\gamma_1(y_2,\boldsymbol{x}_1;\widehat{F},\widehat{\lambda},\widehat{\beta})+I(y_1>y_2)\gamma_2(y_2,\boldsymbol{x}_1;\widehat{F},\widehat{\lambda},\widehat{\beta})\Big\}\\ &\quad\times\Big\{d\mathbb{F}_{X,Y}(\boldsymbol{x}_1,y_1)d\mathbb{F}_{X,Y}(\boldsymbol{x}_2,y_2)-dF_{X,Y}(\boldsymbol{x}_1,y_1)dF_{X,Y}(\boldsymbol{x}_2,y_2)\Big\}=o_p(1).\end{aligned}$$

Step 2: Root n consistency of $\widehat{\boldsymbol{\theta}}$

In Step 2, we apply Lemma 3 below to show that

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)=O_p(1). \tag{7.2}$$

This lemma is adapted from Theorem 3.4.1 of [37].

Lemma 3. *For each n, let $\mathbb{M}_n$ and M_n be stochastic processes indexed by Θ. Let $0\le\delta_n<\eta$ be arbitrary. Suppose that for every n and $\delta_n<\delta\le\eta$*

$$\sup_{\delta/2<\|\theta-\theta_0\|_2\le\delta,\theta\in\Theta} M_n(\theta)-M_n(\theta_0)\lesssim-\delta^2; \tag{7.3}$$

$$E^*\left[\sup_{\delta/2<\|\theta-\theta_0\|_2\le\delta,\theta\in\Theta}\sqrt{n}\Big\{(\mathbb{M}_n-M_n)(\theta)-(\mathbb{M}_n-M_n)(\theta_0)\Big\}^+\right]\lesssim\phi_n(\delta), \tag{7.4}$$

for functions ϕ_n such that $\delta\to\phi_n(\delta)/\delta^\tau$ is decreasing on (δ_n,η), for some $\tau<2$. Let $r_n\lesssim\delta_n^{-1}$ satisfy

$$r_n^2\phi_n\left(\frac{1}{r_n}\right)\le\sqrt{n},\qquad \textit{for every } n. \tag{7.5}$$

If $\widehat{\theta}_n$ takes its values in Θ and satisfies $\mathbb{M}_n(\widehat{\theta})\ge\mathbb{M}_n(\theta_0)-O_p(r_n^{-2})$ and $\|\widehat{\theta}-\theta\|_2$ converges to zero in probability, then $r_n\|\widehat{\theta}-\theta\|_2=O_p^*(1)$.

Recalling that

$$\ell(\lambda,\beta)=\sum_{j=1}^{n}\sum_{i=1}^{n}\left[I_{i,j}\log\widehat{F}_{\theta}(V_{\theta,j,i})+(1-I_{i,j})\log\left\{1-\widehat{F}_{\theta}(V_{\theta,j,i})\right\}\right],$$

we define

$$\widetilde{\ell}(\lambda,\beta)=\sum_{j=1}^{n}\sum_{i=1}^{n}\left[I_{i,j}\log F_{\theta}(V_{\theta,j,i})+(1-I_{i,j})\log\left\{1-F_{\theta}(V_{\theta,j,i})\right\}\right].$$

Accordingly,

$$\begin{aligned}\ell(\lambda_0,\beta_0)&=\sum_{j=1}^{n}\sum_{i=1}^{n}\left[I_{i,j}\log\widehat{F}_{0}(V_{0,j,i})+(1-I_{i,j})\log\left\{1-\widehat{F}_{0}(V_{0,j,i})\right\}\right],\\ \widetilde{\ell}(\lambda_0,\beta_0)&=\sum_{j=1}^{n}\sum_{i=1}^{n}\left[I_{i,j}\log F_{0}(V_{0,j,i})+(1-I_{i,j})\log\left\{1-F_{0}(V_{0,j,i})\right\}\right].\end{aligned}$$

We will apply Lemma 3 to show (7.2). According to Lemma 3, $\mathbb{M}_n(\theta)$ and $M_n(\theta)$ are defined to be

$$\begin{aligned}\mathbb{M}_n(\theta)&=\frac{1}{n^2}\ell(\lambda,\beta)\\ M_n(\theta)&=\frac{1}{n^2}E\left\{\widetilde{\ell}(\theta)\right\}\\ &=E\left[I_{i,j}\log\left\{F_{\theta}(V_{\theta,j,i})\right\}+(1-I_{i,j})\log\left\{1-F_{\theta}(V_{\theta,j,i})\right\}\right].\end{aligned}$$

Then, based on the definition of $\widehat{\theta}$,

$$\mathbb{M}_n(\widehat{\theta})\geq\mathbb{M}_n(\theta_0),$$

and we have shown the consistency of $\widehat{\theta}$ in Step 1. To apply Lemma 3 to show the root n consistency of $\widehat{\beta}$, we need to specify "δ_n, η, τ", and verify (7.3) and (7.4). Furthermore, for $\phi_n(\delta)$ from (7.4), we need to verify that it satisfies (7.5) for $r_n=\sqrt{n}$ and that $\phi_n(\delta)/\delta^{\tau}$ is decreasing on (δ_n,η).

Note that (7.3) is verified by Lemma 4, which is Lemma 12 of the supplementary document. To verify (7.4), we decompose

$$\begin{aligned}&(\mathbb{M}_n-M_n)(\theta)-(\mathbb{M}_n-M_n)(\theta_0)\\ =&\ \frac{1}{n^2}\left(\widetilde{\ell}(\lambda,\beta)-E\left\{\widetilde{\ell}(\lambda,\beta)\right\}-\left[\widetilde{\ell}(\lambda_0,\beta_0)-E\left\{\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right]\right)\\ &+\frac{1}{n^2}\left[\ell(\lambda,\beta)-\widetilde{\ell}(\lambda,\beta)-\left\{\ell(\lambda_0,\beta_0)-\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right].\end{aligned}\tag{7.6}$$

In Lemma 5, which is Lemma 13 of the supplementary document, we verify that for any $\delta<\eta_0$,

$$E\left(\sup_{\theta\in\Theta,\|\theta-\theta_0\|_2\leq\delta}\left|\widetilde{\ell}(\lambda,\beta)-E\left\{\widetilde{\ell}(\lambda,\beta)\right\}-\left[\widetilde{\ell}(\lambda_0,\beta_0)-E\left\{\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right]\right|\right)$$

$$\lesssim n + n^{3/2}\delta. \quad (7.7)$$

Moreover, in Lemma 6, which is Lemma 14 of the supplementary document, we show that

$$\begin{aligned} & E\left(\sup_{\theta\in\Theta, \|\theta-\theta_0\|_2\le\delta}\left[\ell(\lambda,\beta)-\widetilde{\ell}(\lambda,\beta)-\left\{\ell(\lambda_0,\beta_0)-\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right]^+\right) \\ \lesssim\ & n\left(1+\sqrt{\log n}\delta^\alpha+\delta^\alpha\sqrt{-\log\delta}\right)+n^{3/2}\delta. \end{aligned} \tag{7.8}$$

Combining (7.6)–(7.8), we verify (7.4) with

$$\phi_n(\delta) = \frac{1+\sqrt{\log n}\delta^\alpha+\delta^\alpha\sqrt{-\log\delta}}{\sqrt{n}} + \delta,$$

for $\alpha \in (0, 0.25)$. We then have that $\delta \to \phi_n(\delta)/\delta^{1.5}$ is decreasing for $\delta \in (\delta_n, \eta_2)$ for some small $\eta_2 > 0$, where δ_n is defined in the proof of Lemma 14 in the supplementary document. In particular, $\delta_n = n^{-1/\{2(1-\alpha)\}}$ satisfies $\delta_n^{-1} > \sqrt{n}$. Now set $\eta = \min\{\eta_0, \eta_1, \eta_2\}$ so that it plays the role of "η" in Lemma 3, where η_0 is given by Condition 2 and η_1 is defined by (74) in the proof of Lemma 14 in the supplementary document. Clearly, $r_n = \sqrt{n}$ satisfies (7.5). We have finished checking the conditions for Lemma 3, and this completes the proof of Step 2.

Lemma 4. *Assume Condition 2. For any $\delta \in (0, \eta_0)$, we have*

$$\sup_{\delta/2<\|\theta-\theta_0\|_2\le\delta, \theta\in\Theta} M_n(\theta) - M_n(\theta_0) \lesssim -\delta^2.$$

Lemma 5. *Assume Conditions 1 and 2. For any $\delta \in (0, \eta_0)$, we have*

$$E\left(\sup_{\|\theta-\theta_0\|_2\le\delta}\left|\widetilde{\ell}(\lambda,\beta)-E\left\{\widetilde{\ell}(\lambda,\beta)\right\}-\left[\widetilde{\ell}(\lambda_0,\beta_0)-E\left\{\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right]\right|\right) \lesssim n+n^{3/2}\delta.$$

Lemma 6. *Assume Conditions 1–3. We have*

$$\begin{aligned} & E\left(\sup_{\theta\in\Theta, \|\theta-\theta_0\|_2\le\delta}\left[\ell(\lambda,\beta)-\widetilde{\ell}(\lambda,\beta)-\left\{\ell(\lambda_0,\beta_0)-\widetilde{\ell}(\lambda_0,\beta_0)\right\}\right]^+\right) \\ \lesssim\ & n\left(1+\sqrt{\log n}\delta^\alpha+\delta^\alpha\sqrt{-\log\delta}\right)+n^{3/2}\delta, \end{aligned}$$

for some $\alpha \in (0, 0.25)$ and $\delta_n < \delta < \min(\eta_0, \eta_1)$ with $\delta_n = n^{-1/\{2(1-\alpha)\}}$, η_0 given by Condition 2, and η_1 defined by (74) in the proof of this lemma (i.e., Lemma 14 in the supplementary document).

Step 3: Asymptotic normality of $\widehat{\theta}$

In Step 3, we establish the asymptotic normality of $\widehat{\theta}$. In particular, we aim to show that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \rightsquigarrow N(0, \Sigma), \tag{7.9}$$

where $\Sigma = \frac{1}{4}\Sigma_1^{-1}\Sigma_2\Sigma_1^{-1}$ with Σ_1 and Σ_2 defined by (4.4) and (4.5) respectively.

We need Lemma 7 below, which is adapted from Theorem 14.1 in [24]; see also Theorem 3.2.2 in [37].

Lemma 7. *Let $\mathbb{W}_n$, $\mathbb{W}$ be stochastic processes indexed by a metric space $\mathcal{H}$, such that $\mathbb{W}_n \rightsquigarrow \mathbb{W}$ in $L^\infty(H)$ for every compact $H \subset \mathcal{H}$. Suppose also that almost all sample paths $h \mapsto M(h)$ are upper semicontinuous and possess a unique maximum at a (random) point $\widehat{h}$, which as a random map in $\mathcal{H}$ is tight. If the sequence $\widehat{h}_n$ is uniformly tight and satisfies $\mathbb{W}_n(\widehat{h}_n) \geq \sup_{h\in H} \mathbb{W}_n(h) - o_p(1)$, then $\widehat{h}_n \rightsquigarrow \widehat{h}$ in $\mathcal{H}$.*

We apply the argmax theorem above to show (7.9). Denote $\widehat{h}_n = \sqrt{n}(\widehat{\theta} - \theta_0)$ and let $h = (h_1, h_2^T)^T$, $\theta_{n,h} = \theta_0 + h/\sqrt{n}$, $\lambda_{n,h} = \lambda_0 + h_1/\sqrt{n}$, $\beta_{n,h} = \beta_0 + h_2/\sqrt{n}$. Define

$$\mathbb{W}_n(h) = \frac{1}{n}\left\{\ell(\theta_{n,h}) - \ell(\theta_0)\right\}.$$

Clearly, $\widehat{h}_n$ is the maximizer of $\mathbb{W}_n(h)$, and therefore $\mathbb{W}_n(\widehat{h}_n) \geq \sup_{h\in\mathbb{R}^{p+1}} \mathbb{W}_n(h)$. In Step 2, we have shown that $\widehat{h}_n$ is uniformly tight.

For H an arbitrary compact subset of $\mathbb{R}^{p+1}$, consider the process

$$\mathbb{W}_n(h) = \frac{1}{n}\left\{\ell(\theta_{h,n}) - \ell(\theta_0)\right\} = \mathbb{W}_{n,1}(h) + \mathbb{W}_{n,2}(h), \tag{7.10}$$

with $h \in H$, where

$$\begin{aligned} \mathbb{W}_{n,1}(h) &= \frac{1}{n}\left[\ell(\theta_{n,h}) - \ell(\theta_0) - \left\{\widetilde{\ell}(\theta_{n,h}) - \widetilde{\ell}(\theta_0)\right\}\right], \\ \mathbb{W}_{n,2}(h) &= \frac{1}{n}\left\{\widetilde{\ell}(\theta_{n,h}) - \widetilde{\ell}(\theta_0)\right\}. \end{aligned}$$

We consider $\mathbb{W}_{n,1}(h)$ and $\mathbb{W}_{n,2}(h)$ separately. For $\mathbb{W}_{n,2}(h)$, we show in Lemma 9, which is Lemma 17 of the supplementary document, that

$$\left\|\mathbb{W}_{n,2}(h) - \left(h^T\mathbb{G}_n\varphi - h^T\Sigma_1 h\right)\right\|_{h\in H} = o_p(1), \tag{7.11}$$

where $\varphi(\cdot)$ is defined by (4.2) and Σ_1 by (4.4). For $\mathbb{W}_{n,1}(h)$, we have

$$\begin{aligned} \mathbb{W}_{n,1}(h) &= \frac{1}{n}\left[\ell(\theta_{h,n}) - \ell(\theta_0) - \left\{\widetilde{\ell}(\theta_{h,n}) - \widetilde{\ell}(\theta_0)\right\}\right] \\ &= \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n I_{i,j}\log\left\{\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})}\right\} \\ &\quad + \frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n(1 - I_{i,j})\log\left\{\frac{\left(1 - \widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})\right)(1 - F_0(V_{0,j,i}))}{\left(1 - \widehat{F}_0(V_{0,j,i})\right)(1 - F_{\theta_{n,h}}(V_{\theta_{n,h},j,i}))}\right\} \\ &= \mathcal{I}_5 + \mathcal{I}_6. \end{aligned} \tag{7.12}$$

Consider $\mathcal{I}_5$. By the Taylor expansion for $\log x$ at $x = 1$, we have

$$\begin{aligned}
\mathcal{I}_5 \quad &= \quad \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} I_{i,j}\left\{\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})} - 1\right\} \\
&\quad - \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} I_{i,j}\frac{1}{2\xi_{n,h,i,j}}\left\{\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})} - 1\right\}^2,
\end{aligned}$$

where $\xi_{n,h,i,j}$ is between $\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})}$ and 1. Based on Lemma 8, which is Lemma 8 of the supplementary document, and Condition 2, when n is sufficiently large, we have

$$\sup_{1\le i,j\le n; h\in H}|\xi_{n,h,i,j} - 1| \le \sup_{1\le i,j\le n; h\in H}\left|\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})} - 1\right| \to 0$$

in probability. This implies that

$$\sup_{1\le i,j\le n; h\in H}\frac{1}{\xi_{n,h,i,j}} = \frac{1}{1 - o_p^*(1)},$$

where $o_p^*(1)$ is uniform in $1 \le i, j \le n$ and $h \in H$. Therefore,

$$\begin{aligned}
&\left|\mathcal{I}_5 - \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} I_{i,j}\left\{\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})} - 1\right\}\right| \\
\lesssim \quad & \frac{n}{1 - o_p^*(1)}\sup_{\boldsymbol{z}\in\mathcal{Z}, h\in H}\left|\frac{\widehat{F}_{\theta_{n,h}}(\boldsymbol{v}_{\theta_{n,h}})F_0(\boldsymbol{v}_{\theta_0})}{\widehat{F}_0(\boldsymbol{v}_{\theta_0})F_{\theta_{n,h}}(\boldsymbol{v}_{\theta_{n,h}})} - 1\right|^2.
\end{aligned}$$

This together with Lemmas 10 and 11, which are Lemmas 18 and 19 in the supplementary document, leads to

$$\sup_{h\in H}\left|\mathcal{I}_5 - \sqrt{n}\mathbb{G}_n\{f_{1,n,h}(\cdot)\}\right| = o_p(1), \tag{7.13}$$

where $f_{1,n,h}(\cdot)$ comes from Lemma 11 and is given by

$$f_{1,n,h}(\boldsymbol{z}) = E\left\{\frac{F_0(V_{0,2,1})}{F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})} I\left(\boldsymbol{v}_{\theta_{n,h}} \le V_{\theta_{n,h},2,1}\right) - I\left(\boldsymbol{v}_0 \le V_{0,2,1}\right)\right\}. \tag{7.14}$$

Using exactly the same derivation, we can verify that

$$\sup_{h\in H}\left|\mathcal{I}_6 - \sqrt{n}\mathbb{G}_n\{f_{2,n,h}(\cdot)\}\right| = o_p(1), \tag{7.15}$$

with

$$f_{2,n,h}(\boldsymbol{z})$$

$$= \quad E\left[\frac{1-F_0(V_{0,2,1})}{1-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})}\left\{1-I\left(\boldsymbol{v}_{\theta_{n,h}}\le V_{\theta_{n,h},2,1}\right)\right\}-\left\{1-I\left(\boldsymbol{v}_0\le V_{0,2,1}\right)\right\}\right].$$

Combining (7.12), (7.13), and (7.15) we have

$$\sup_{h\in H}\left|\mathbb{W}_{n,1}(h)-\sqrt{n}\mathbb{G}_n\left\{f_{1,n,h}(\cdot)+f_{2,n,h}(\cdot)\right\}\right|=o_p(1). \tag{7.16}$$

Furthermore, noting that for any constant C, $\mathbb{G}_nC=0$, we have

$$\mathbb{G}_n\left\{f_{1,n,h}(\cdot)+f_{2,n,h}(\cdot)\right\}=\mathbb{G}_n\psi_{n,h}(\cdot), \tag{7.17}$$

where

$$\begin{aligned}\psi_{n,h}(\boldsymbol{z}) &= E\left[\left\{\frac{F_0(V_{0,2,1})}{F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})}-\frac{1-F_0(V_{0,2,1})}{1-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})}\right\}I\left(\boldsymbol{v}_{\theta_{n,h}}\le V_{\theta_{n,h},2,1}\right)\right]\\ &= E\left[\frac{F_0(V_{0,2,1})-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})}{F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})\left\{1-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})\right\}}I\left(\boldsymbol{v}_{\theta_{n,h}}\le V_{\theta_{n,h},2,1}\right)\right].\end{aligned}$$

Then, based on Lemma 12, which is Lemma 20 in the supplementary document, we have

$$E\left\|\sqrt{n}\mathbb{G}_n\psi_{n,h}(\boldsymbol{z})-h^T\mathbb{G}_n\psi(\boldsymbol{z})\right\|_{h\in H}=o(1), \tag{7.18}$$

where

$$\psi(\boldsymbol{z})=-E\left[\frac{\dot{F}_0(V_{0,2,1})+F_0'(V_{0,2,1})\dot{V}_{0,2,1}}{F_0(V_{0,2,1})\left\{1-F_0(V_{0,2,1})\right\}}I\left(\boldsymbol{v}_0\le V_{0,2,1}\right)\right],$$

as defined by (4.3). Combining (7.16), (7.17), and (7.18) we have

$$\sup_{h\in H}\left|\mathbb{W}_{n,1}(h)-h^T\mathbb{G}_n\psi(\boldsymbol{z})\right|=o_p(1). \tag{7.19}$$

This combined with (7.10) and (7.11) gives

$$\sup_{h\in H}\left|\mathbb{W}_n(h)-h^T\mathbb{G}_n(\varphi+\psi)+h^T\Sigma_1h\right|=o_p(1).$$

Furthermore, by the central limit theorem and the fact that Σ_2 is invertible (Condition 5), we have

$$\mathbb{G}_n(\varphi+\psi)\rightsquigarrow N(0,\Sigma_2), \tag{7.20}$$

where Σ_2 is given by (4.5). Now define $\mathbb{W}(h)=h^T\mathcal{N}-h^T\Sigma_1h$ where $\mathcal{N}$ is a random vector following the $N(0,\Sigma_2)$ distribution; then $\mathbb{W}(h)$ has a unique maximum at $\widehat{h}=0.5\Sigma_1^{-1}\mathcal{N}$ since Σ_1 is invertible (Condition 5). Combining (7.19) and (7.20), we have $\mathbb{W}_n(h)\rightsquigarrow\mathbb{W}(h)$, which indicates that $\mathbb{W}(h)$ plays the role of "$\mathbb{W}(h)$" in Lemma 7. This immediately leads to (7.9) by an application of Lemma 7. Our proof is complete.

Lemma 8. *Assume Conditions 1 and 2. For any $\delta\in(0,\eta_0)$, we have, for large n,*

$$\sqrt{n}E\left\{\sup_{\|\theta-\theta_0\|_2\le\delta;t\in\mathcal{T}}\left|\widehat{F}_\theta(t)-F_\theta(t)\right|\right\}\lesssim 1,$$

$$\sqrt{n}E\left\{\sup_{\|\theta-\theta_0\|_2\le\delta;t\in\mathcal{T}}|\widehat{F}_\theta(t)-F_\theta(t)|^2\right\}\lesssim 1/\sqrt{n}.$$

Lemma 9. *Assume Conditions 1 and 2. We have*

$$\left\|\frac{1}{n}\left\{\widetilde{\ell}(\theta_{n,h})-\widetilde{\ell}(\theta_0)\right\}-\left(h^T\mathbb{G}_n\varphi-h^T\Sigma_1 h\right)\right\|_{h\in H}=o_p(1),$$

where $\varphi(\cdot)$ is defined by (4.2) and Σ_1 is defined by (4.4).

Lemma 10. *Assume Conditions 1 and 2. We have*

$$\sup_{\boldsymbol{z}\in\mathcal{Z},h\in H}\left|\frac{\widehat{F}_{\theta_{n,h}}(\boldsymbol{v}_{\theta_{n,h}})F_0(\boldsymbol{v}_{\theta_0})}{\widehat{F}_0(\boldsymbol{v}_{\theta_0})F_{\theta_{n,h}}(\boldsymbol{v}_{\theta_{n,h}})}-1\right|=o_p(n^{-1/2}).$$

Lemma 11. *Assume Conditions 1 and 2. We have*

$$\begin{aligned}&\sup_{h\in H}\left|\frac{1}{n}\sum_{j=1}^n\sum_{i=1}^n I_{i,j}\left\{\frac{\widehat{F}_{\theta_{n,h}}(V_{\theta_{n,h},j,i})F_0(V_{0,j,i})}{\widehat{F}_0(V_{0,j,i})F_{\theta_{n,h}}(V_{\theta_{n,h},j,i})}-1\right\}-\sqrt{n}\mathbb{G}_n\left\{f_{1,n,h}(\cdot)\right\}\right|\\ =\;& o_p(1),\end{aligned}$$

where $f_{1,n,h}(\cdot)$ is defined by (7.14).

Lemma 12. *Assume Conditions 1–3. We have*

$$E\left\|\sqrt{n}\mathbb{G}_n\psi_{n,h}(\boldsymbol{z})-h^T\mathbb{G}_n\psi(\boldsymbol{z})\right\|_{h\in H}=o(1),$$

where

$$\begin{aligned}\psi_{n,h}(\boldsymbol{z})&=E\left[\frac{F_0(V_{0,2,1})-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})}{F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})\left\{1-F_{\theta_{n,h}}(V_{\theta_{n,h},2,1})\right\}}I\left(\boldsymbol{v}_{\theta_{n,h}}\le V_{\theta_{n,h},2,1}\right)\right];\\ \psi(\boldsymbol{z})&=-E\left[\frac{\dot{F}_0(V_{0,2,1})+F_0'(V_{0,2,1})\dot{V}_{0,2,1}}{F_0(V_{0,2,1})\left\{1-F_0(V_{0,2,1})\right\}}I\left(\boldsymbol{v}_0\le V_{0,2,1}\right)\right].\end{aligned}$$

Note that the definition of $\psi(\boldsymbol{z})$ complies with (4.3).

Acknowledgment

The authors thank the editor, the associate editor, and two referees for constructive comments and suggestions that lead to a significant improvement over the article. The first two authors contribute equally to this work.

Supplementary Material

Maximum profile binomial likelihood estimation for the semiparametric Box–Cox power transformation model: Supplementary Materials (doi: 10.1214/23-EJS2146SUPP; .pdf). The supplementary materials contain the full technical details of the proof of Theorem 4.1.

References

[1] Abrevaya, J. (1999a). Computation of the maximum rank correlation estimator. *Economics Letters*, 62, 279–285. MR1684870

[2] Abrevaya, J. (1999b). Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *Journal of Econometrics*, 93, 203–228. MR1721098

[3] Amemiya, T. (1985). Instrumental variable estimator for the nonlinear errors-in-variable models. *Journal of Econometrics*, 38, 273–289. MR0805460

[4] Bennett, S. (1983a). Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2, 273–277.

[5] Bennett, S. (1983b). Log-logistic regression models for survival data. *Applied Statistics*, 32, 165–171.

[6] Bickel, P. J. and Doksum, K. A. (1981). An analysis of transformations revisited. *Journal of the American Statistical Association*, 76, 296–311. MR0624332

[7] Bickel, P. J., Klaasen, C. A., Ritov, Y, and Wellner J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, MD: Johns Hopkins University Press. MR1245941

[8] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–252. MR0192611

[9] Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963–1987. *Journal of Econometrics*, 65, 109-154. MR1323055

[10] Cai, T., Tian, L., and Wei, L. J. (2005). Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika*, 92, 619–632. MR2202650

[11] Carroll, R. J. and Ruppert, D. (1985). Transformations in regression: A robust analysis. *Technometrics*, 27, 1–12. MR0772893

[12] Cavanagh, C. and Sherman, R. P. (1998). Rank estimators for monotonic index models. *Journal of Econometrics*, 84, 351–381. MR1630210

[13] Chen, B., Li, P., Qin, J., and Yu, T. (2016). Using a monotonic density ratio model to find the asymptotically optimal combination of multiple diagnostic tests. *Journal of the American Statistical Association*, 111, 861–874. MR3538711

[14] Chen, S. (2012). Distribution-free estimation of the Box–Cox regression model with censoring. *Econometric Theory*, 28, 680–695. MR2927924

[15] Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34, 187–220. MR0341758

[16] Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276. MR0400509

[17] Fitzenberger, B., Wilke, R. A., and Zhang, X. (2009). Implementing Box–Cox quantile regression. *Econometric Reviews*, 29, 158–181. MR2747497

[18] Flinn, C. and Heckman, J. (1982). New methods for analyzing struc-

tural models of labor force dynamics. *Journal of Econometrics*, 18, 115–168. MR0661666

[19] Foster, A. M., Tian, L., and Wei, L. J. (2001). Estimation for Box–Cox transformation model without assuming parametric error distribution. *Journal of the American Statistical Association*, 96, 1097–1101. MR1947257

[20] Hammer S. M., Katzenstein D. A., Hughes M. D., Gundacker H., Schooley R. T., Haubrich R. H., Henry W. K., Lederman M. M., Phair J. P., Niu M., Hirsch M. S., and Merigan T. C. for the AIDS Clinical Trials Group Study 175 Study Team (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335, 1081–1090.

[21] Han, A. K. (1987). Non-parametric analysis of a generalized regression model: The maximum rank correlation estimator. *Journal of Econometrics*, 35, 303–316. MR0903188

[22] Heller, G. and Qin, J. (2001). Pairwise rank-based likelihood for estimation and inference on the mixture proportion. *Biometrics*, 57, 813–817. MR1859816

[23] Hinkley, D. V. and Runger, G. (1984). The analysis of transformed data. *Journal of the American Statistical Association*, 79, 302–309. MR0755087

[24] Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* New York: Springer. MR2724368

[25] Lancaster, T. (1990). *The Econometric Analysis of Transition Data.* Cambridge: Cambridge University Press. MR1167199

[26] Lee, K., Bhattacharya, B. B., Qin, J., and Small, D. S. (2021). *A nonparametric likelihood approach for inference in instrumental variable models.* arXiv:1605.03868.

[27] Mu Y. M. and He X. M. (2007). Power transformation toward a linear regression quantile. *Journal of the American Statistical Association*, 102, 269–279. MR2293308

[28] Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58, 809–837. MR1064846

[29] Qin, J., Garcia, T. P., Ma, Y., Tang, M. X., Marder, K., and Wang, Y. (2014). Combining isotonic regression and EM algorithm to predict genetic risk under monotonicity constraint. *The Annals of Applied Statistics*, 8, 1182–1208. MR3262550

[30] Robinson, P. M. (1991). Best nonlinear three-stage least squares estimation of certain econometric models. *Econometrica*, 59, 755–786. MR1106511

[31] Sakia, R. M. (1992). The Box–Cox transformation technique: A review. *The Statistician*, 41, 169–178.

[32] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, 61, 123–137. MR1201705

[33] Taylor, J. M. G. (1985a). Measures of location of skew distributions obtained through Box–Cox transformations. *Journal of the American Statistical Association*, 80, 427–432. MR0792744

[34] Taylor, J. M. G. (1985b). Power transformations to symmetry. *Biometrika*, 72, 145–152. MR0790209

[35] Taylor, J. M. G. (1987). Using a generalized mean as a measure of location. *Biometrical Journal*, 29, 731–738. MR0919646

[36] Tian, Z., Liang, K., and Li, P. (2021). Maximum multinomial likelihood estimation in compound mixture model with application to malaria study. *Journal of Nonparametric Statistics*, 33, 31–38. MR4261896

[37] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer. MR1385671

[38] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20, 595–601. MR0032169

[39] Zhang, T. and Wang, L. (2020). Smoothed empirical likelihood inference and variable selection for quantile regression with nonignorable missing response. *Computational Statistics & Data Analysis*, 144, 106888. MR4038215