Statistica Sinica Preprint No: SS-2017-0451	
Title	Estimation of Single-index Models with Fixed Censored Responses
Manuscript ID	SS-2017-0451
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0451
Complete List of Authors	Hailin Huang Yuanzhang Li Hua Liang and Yanlin Tang
Corresponding Author	Hua Liang
E-mail	hliang@gwu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Statistica Sinica

ESTIMATION OF SINGLE-INDEX MODELS WITH FIXED CENSORED RESPONSES

Hailin Huang[1*], Yuanzhang Li[1], Hua Liang[1], Yanlin Tang[2*],

[1] *George Washington University and* [2] *East China Normal University*

Abstract: We propose a new procedure to estimate the index parameter and link function of single-index models, where the response variable is subject to fixed censoring. Under some regularity conditions, we show that the estimated index parameter is root-n consistent and asymptotically normal, and the estimated nonparametric link function achieves the optimal convergence rate and is asymptotically normal. In addition, we propose a linearity testing method for the nonparametric link function. A simulation study shows that the proposed procedures perform well in finite-sample experiments. An application to an HIV data set is presented for illustrative purposes.

Key words and phrases: Nonparametric censored regression, single-index model, semi-parametric least-squares.

1. Introduction

Because of the non-negativity or a detection limit, data with fixed cen-

*The two authors are co-first authors.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

sored responses are common in econometrics and biometrics studies (Maddala, 1986; Adesina and Zinnah, 1993; Nizar Al-Malkawi, 2007; Haab, Dunham, and Brown, 2001; Van der Pouw Kraan et al., 1995). For instance, in our motivating HIV data set, the viral load in the blood serum can only be observed if it is above 50 units (Kobie et al., 2012).

To explore the relationship between the fixed censored response variable and the covariates, several models and associated estimation methods are proposed. Earlier works focused on parametric regression models, including the Tobit model (Tobin, 1958) and its variants (Amemiya, 1984, 1979; Blundell and Meghir, 1987), which assume a linear relationship with normal errors. However, both linearity and normality assumptions can be violated in practice (Maddala and Nelson, 1975; Gawande, 1995; Chen, Dahl, and Khan, 2005). To make the model more flexible, several researchers have studied nonparametric regression models with fixed censored data. For example, Lewbel and Linton (2002) proposed a two-stage moment-based method to estimate the nonparametric conditional mean function; Chen, Dahl, and Khan (2005) studied the identification and estimation problems of the conditional median function in nonparametric location-scale models. These nonparametric methods achieve greater flexibility and, in general, do not require distributional assumptions. However, they suffer from "curse of

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

dimensionality", and their performance can be poor, even when the dimension of the covariates is moderate.

To amend the limitations of the existing methods, we consider single-index models with the fixed censored responses. Single-index models have been widely studied in the literature (Powell, Stock, and Stoker, 1989; Duan and Li, 1991; Härdle, Hall, and Ichimura, 1993; Ichimura, 1993; Horowitz and Härdle, 1996; Carroll et al., 1997; Xia and Härdle, 2006; Liang et al., 2010). The majority of these studies focuses on cases in which the response Y is fully observed, although some researchers have studied estimation when Y is randomly censored (Lopez, 2009; Bücher, El Ghouch and Van Keilegom, 2014; Chiang, Wang, and Huang, 2017; Kong and Xia, 2017). Note that the methods for single-index models with randomly censored responses implicitly assume that we can always observe uncensored observations below any given value of the censoring point (Lopez, 2009; Bücher, El Ghouch and Van Keilegom, 2014; Kong and Xia, 2017; Huang, 2017). However, this is not true for the fixed censoring case because the probability of observing uncensored observations below the given fixed censored point is zero. Thus, the associated methods cannot be applied. To the best of our knowledge, no estimation methods for single-index models with fixed censored responses are available in the literature.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

By establishing a relationship between fixed censored single-index models and uncensored single-index models, we propose a new procedure to estimate the index parameter. Under certain regularity conditions, the proposed estimator is root-n consistent and asymptotically normal. After substituting in the index parameter, the single-index model is simplified to a univariate fixed-censored nonparametric model, and we apply the method of Lewbel and Linton (2002) to estimate the nonparametric link function. The estimated nonparametric link function achieves the optimal convergence rate and is asymptotically normal. Finally, a hypothesis testing procedure is proposed to check the linearity of the nonparametric link function.

The rest of the paper is organized as follows. Section 2 presents the model, and gives the estimation and testing procedures. Section 3 presents the asymptotic properties. Section 4 explores the finite-sample performance by means of a simulation study, and an HIV data set is analyzed in Section 5 for illustrative purposes. All technical proofs are provided in the online Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

2. Model and Methods

2.1 Model

Consider the following single-index model for the latent responses:

$$Y_i^* = m(X_i^\top \beta) - \epsilon_i,\ i = 1, \ldots, n, \tag{2.1}$$

where $X_i = (X_{i,1}, \ldots, X_{i,d})^\top$ is a d-dimensional covariate vector, $\beta = (\beta_1, \ldots, \beta_d)^\top$ is an unknown index parameter vector, $m(u) = E(Y_i^* | X_i^\top \beta = u)$ is an unknown smooth function, and ϵ_i is the random error. Owing to fixed censoring, Y_i^* cannot be fully observed. Instead we can only observe (Y_i, δ_i), where $Y_i = \max(Y_i^*, c)$, $\delta_i = I(Y_i^* > c)$, c is the known lower detection limit, and $I(\cdot)$ is an indicator function. Without loss of generality, we assume $c = 0$. Instead of making parametric distribution assumptions, such as normality, we assume only that ϵ_i is independently and identically distributed (i.i.d.), from an unknown distribution symmetric around zero, and with finite variance. Furthermore, we assume that no intercept is included in the index function $X_i^\top \beta$, for $||\beta|| = 1$, and that the first element of β is positive, which ensures identification, where $\|\cdot\|$ denotes the L_2-norm. In addition, we assume that $\beta \in \Theta \subset \mathbb{R}^d$ for some compact set Θ, and $X \in D_X \subset \mathbb{R}^d$ for some compact set D_X.

Remark 1. To facilitate theoretical derivations, we consider an error term

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

of " $-\epsilon_i$" instead of "ϵ_i"; a similar model setting can be found in Lewbel and Linton (2002). Note that with the symmetry assumption on ϵ_i around 0, ϵ_i and $-\epsilon_i$ have the same distribution.

2.2 Profile least-squares estimator of β

Under model (2.1), the proposed estimation procedure for β is inspired by considering a connection between fixed censored single-index models and uncensored single-index models. Under mild assumptions, this connection changes the estimation of a single-index Tobit model to a standard single-index model; as a result, well-developed estimation procedures can be applied.

Assumption A.1 (i) The latent response Y^* has first $\nu(\geq 3)$ absolute moments. (ii) The common density function of ϵ_i, denoted as $f(\cdot)$, is symmetric around zero and its derivative is continuous.

Proposition 1. *Let $F(\cdot)$ be the distribution function of ϵ. Under Assumption A.1, if $\lim_{\epsilon\to-\infty}\epsilon F(\epsilon)=0$, then $E(Y_i|X_i^\top\beta) = \int_{-\infty}^{m(X_i^\top\beta)} F(\epsilon)d\epsilon$.*

Assumption A.1 and the assumption $\lim_{\epsilon\to-\infty}\epsilon F(\epsilon)=0$ are mild, and are most commonly used with symmetric distributions, such as the normal distribution, Student's t_υ distribution ($\upsilon \geq 4$), and the uniform distribution on a symmetric interval (Lewbel and Linton, 2002). Proposition 1 implies

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

that $E(Y_i|X_i^\top\beta)$ can be represented as a new uncensored single-index model with the same index parameter β, but with a new link function. More specifically, $E(Y_i|X_i^\top\beta = u) = r(u) = w \circ m(u)$, where $w(t) = \int_{-\infty}^{t} F(\epsilon)d\epsilon$, and "$\circ$" means the composition of two functions; a similar derivation can be found in Lewbel and Linton (2002).

According to Proposition 1, we can assign a new single-index model for the observed responses as

$$Y_i = r(X_i^\top\beta) - \epsilon_i^{'}, i = 1, 2, \ldots, n, \tag{2.2}$$

where $\epsilon_i^{'} = \epsilon_i + (Y_i^* - Y_i) + r(X_i^\top\beta) - m(X_i^\top\beta)$. By Proposition 1, $E(Y_i|X_i^\top\beta) = r(X_i^\top\beta)$. Thus we have $E(\epsilon_i^{'}|X_i^\top\beta) = 0$. Therefore, existing estimation methods for single-index models can be applied to estimate β. Here, we adopt the profile least-squares method of Liang et al. (2010), as follows. Given β, we employ the local linear regression technique to estimate $r(\cdot)$, that is, we minimize

$$\sum_{i=1}^{n}\{a + b(X_i^\top\beta - u) - Y_i\}^2 K_h(X_i^\top\beta - u) \tag{2.3}$$

with respect to a and b, where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot) \geq 0$ is a kernel function, and $h > 0$ is the bandwidth. Let $(\widehat{a}, \widehat{b})$ be the minimizer of (2.3); then, $\widehat{r}(u) = \widehat{a}$. As discussed in Jennrich (1969), there exists a profile

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

least-squares estimator $\widehat{\beta}$ that minimizes

$$Q(\beta) = \sum_{i=1}^{n} \{Y_i - \widehat{r}(X_i^\top \beta)\}^2$$

with respect to β, where the minimization problem can be solved using standard optimization algorithms, such as the Newton–Raphson algorithm, and convergence is guaranteed.

Remark 2. The estimation procedure above treats all covariates as important. In practice, especially when the dimension of X is high, it is quite possible that irrelevant covariates are included. This may motivate us to consider variable selection. Given expression (2.2), any variable selection method for single-index models can be used for variable selection, including the penalized profile least-squares method of Liang et al. (2010). A detailed discussion can be found in Huang (2017).

2.3 Nonparametric estimation of $m(\cdot)$

Given $\widehat{\beta}$, we can estimate the unknown link function $m(\cdot)$. For notational convenience, we rewrite model (2.1) as

$$Y_i^* = m(U_i) - \epsilon_i, \ i = 1, \ldots, n, \tag{2.4}$$

where $U_i = X_i^\top \beta$. Recall the definition of $r(\cdot)$ in (2.2), namely, $r(u) = E(Y|U = u) = E(Y|X^\top \beta = u)$, and define $s = r(u)$, $q(s) = q(r(u)) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

$P\{Y > 0|r(U) = r(u)\} = P(Y > 0|U = u)$; and $\widehat{U}_i = X_i^\top \widehat{\beta}$. We propose estimating $m(\cdot)$ in a similar manner to that of Lewbel and Linton (2002).

Step 1. Smooth the observed response Y_i over $\widehat{U}_i$ to estimate $r(\widehat{U}_i)$ using a local linear smoother (Fan and Gijbels, 1996); that is,

$$(\widehat{a}_{i,0}, \widehat{a}_{i,1}) = \underset{(a_0,a_1)\in\mathbb{R}^2}{\arg\min} \sum_{j=1}^{n} \{Y_j - a_0 - a_1(\widehat{U}_j - \widehat{U}_i)\}^2 K_{h_1}(\widehat{U}_j - \widehat{U}_i). \quad (2.5)$$

Then, $r(\widehat{U}_i)$ is estimated as $\widehat{r}(\widehat{U}_i) = \widehat{a}_{i,0}$, where $h_1 > 0$ is a bandwidth.

Step 2. Smooth $I(Y_i > 0)$ over $\widehat{r}(\widehat{U}_i)$ to estimate $q(\cdot)$ using a local linear smoother; that is,

$$(\widehat{b}_0, \widehat{b}_1) = \underset{(b_0,b_1)\in\mathbb{R}^2}{\arg\min} \sum_{i=1}^{n} [I(Y_i > 0) - b_0 - b_1\{\widehat{r}(\widehat{U}_i) - \widehat{r}(u)\}]^2 K_{h_2}(\widehat{r}(\widehat{U}_i) - \widehat{r}(u)).$$

Then, $q(\widehat{r}(u))$ is estimated as $\widehat{q}(\widehat{r}(u)) = \widehat{b}_0$, where $\widehat{r}(u)$ is estimated by replacing $\widehat{U}_i$ with u in (2.5), and $h_2 > 0$ is a bandwidth.

Step 3. Estimate $m(u)$ by $\widehat{m}(u) = \widehat{\lambda}_r - \int_{\widehat{r}(u)}^{\widehat{\lambda}_r} 1/\widehat{q}(s)ds$, where $\widehat{\lambda}_r = \max_{i=1,\ldots,n} \widehat{r}(X_i^\top \widehat{\beta})$. For the integration part, any one-dimensional numerical integration approach, such as Trapezoid rule, can be employed.

2.4 Testing the linearity of the link function

In practice, we may wish to determine whether $m(\cdot)$ is a linear function, because if it is, we can simplify the single-index model to a linear model.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

In this section, we study the hypothesis

$$H_0 : m(u) = \zeta_0 + \zeta_1 u \quad \text{versus} \quad H_1 : H_0 \text{ is not true.}$$

To test the linearity of $m(\cdot)$, we further assume $\epsilon_i \sim N(0, \sigma^2)$, where σ is an unknown scale parameter.

Recalling Proposition 1, we have

$$w'(u) = \partial r(u)/\partial m(u) = F(m(u)) = \Phi(m(u)/\sigma) > 0,$$

which indicates that $r(\cdot)$ is a strictly increasing function of $m(\cdot)$. As a result, testing H_0 against H_1 is equivalent to

$$K_0 : r_0(u) = \int_{-\infty}^{\zeta_0+\zeta_1 u} \Phi(\epsilon/\sigma)d\epsilon \quad \text{versus} \quad K_1 : K_0 \text{ is not true.}$$

We adopt the idea of Koul, Song, and Liu (2014) to test K_0 versus K_1. Given a root-n consistent estimator of β_0, say $\widehat{\beta}$, we define

$$\widehat{\epsilon}_i' = Y_i - \int_{-\infty}^{\zeta_0+\zeta_1 \cdot X_i^\top \widehat{\beta}} \Phi(\epsilon/\sigma)d\epsilon,$$

where ζ_0, ζ_1, and σ are estimated by the maximum likelihood method in the Tobit model (Tobin, 1958; Amemiya, 1984). Define

$$\begin{aligned} V_n &= \frac{1}{n(n-1)h} \sum_{i \neq j} K\Big(\frac{X_i^\top \widehat{\beta} - X_j^\top \widehat{\beta}}{h}\Big)\widehat{\epsilon}_i'\widehat{\epsilon}_j', \\ \widehat{\gamma}^2 &= \frac{2}{n(n-1)h} \sum_{i \neq j} K^2\Big(\frac{X_i^\top \widehat{\beta} - X_j^\top \widehat{\beta}}{h}\Big)\widehat{\epsilon}_i'^2\widehat{\epsilon}_j'^2, \end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

where $h > 0$ is the same bandwidth as that of the profile least-squares estimator of the index parameter, specified in equation (2.3). The test statistic is then defined as

$$T_n = nh^{1/2}V_n/\widehat{\gamma}.$$

Under certain regularity conditions, we can prove that T_n is asymptotically normal under the null hypothesis. Thus, a large value of T_n indicates a deviation from the Tobit model.

3. Asymptotic Properties

In this section, we present the asymptotic properties of the proposed estimators for the index parameter and the link function, as well as the properties of the test statistic. The true index parameter and unknown link function are denoted as $\beta_0 = (\beta_{1,0}, \ldots, \beta_{d,0})^\top$ and $m(\cdot)$, respectively. In addition to Assumption A.1, the following assumptions are needed for the asymptotic results.

Assumption A.2. (i) $r(\cdot)$ and $m(\cdot)$ are not constant on the support $\Omega = \{u|u = x^\top\beta, x \in D_X, \beta \in \Theta\}$; then their third derivatives are uniformly Lipschitz continuous for all $u \in \Omega$. (ii) Let $f_X(x)$ be the density function of X; then, the third derivative of $f_X(x)$ is continuous. (iii) The second derivative of the function $q(\cdot)$ is continuous, and $\inf_{u\in\Omega} q(r(u)) > 0$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Furthermore, $q(\lambda_r) = 1$, where $\lambda_r = \sup_{u\in\Omega} r(u)$, and the supremum is taken over $u = x^\top\beta_0$ for $x \in D_X$. (iv) $\tau^2(u) = E[\{Y - r(X^\top\beta_0)\}^2|X^\top\beta_0 = u]$ and $v^4(u) = E[\{Y - r(X^\top\beta_0)\}^4|X^\top\beta_0 = u]$ are bounded functions with continuous derivatives.

Assumption A.3. (i) $nh^8 \to 0$ and $nh^{3+3/(\nu-1)}/\log n \to \infty$ as $n \to \infty$, where $\nu \geq 3$ is specified in A.1. (ii) $nh_1^2/\log^2(n) \to \infty$, $nh_2^2/\log^2(n) \to \infty$ and $h_1/h_2 \leq C_1$, $nh_1^5 \leq C_2$, and $nh_2^5 \leq C_3$ for some positive constants C_1, C_2, and C_3.

Assumption A.4. The support of the kernel function $K(\cdot)$ is $[-1, 1]$, and its second derivative is Lipschitz continuous. Moreover, $\int_{-1}^{1} K(s)ds = 1$, $\int_{-1}^{1} sK(s)ds = 0$, and $\int_{-1}^{1} s^2K(s)ds > 0$.

Assumptions A.2 (i)–(ii) are similar to the regularity conditions in Carroll et al. (1997) and Liang et al. (2010) for uncensored data. A.2 (iii) is adopted from Assumption 2 in Lewbel and Linton (2002), which is necessary to ensure that the estimated nonparametric link function achieves the optimal convergence rate. Assumption A.2 (iv) is adopted from Assumption (C2) in Koul, Song, and Liu (2014), which is a necessary condition for the asymptotic normality of the test statistic. Assumption A.3 provides us with a guideline for selecting appropriate bandwidths. Furthermore, as pointed out by Liang et al. (2010), Assumption A.3 (i) implies that the estimation

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

performance remains stable in a reasonable range of bandwidth, especially when the sample size is large. In practice, the bandwidth can be chosen using cross-validation. Assumption A.4 is standard for nonparametric regressions.

Theorems 1–2 present the asymptotic properties of the estimated index parameter and the nonparametric link function.

Theorem 1. *Under Assumptions A.1–A.4, we have*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{D}} N(0, W_0^+), \tag{3.1}$$

where $W_0 = E\Big[r'^2(X^\top\beta_0)\{X - E(X|X^\top\beta_0)\}\{X - E(X|X^\top\beta_0)\}^\top\tau^2(X^\top\beta_0)\Big]$, *and* W_0^+ *denotes its Moore–Penrose inverse.*

Theorem 2. *Under Assumptions A.1–A.4, for an interior point* $u = x^\top\beta$, *where* $x \in D_X$ *and* $\beta \in \Theta_{c_0} = \{\beta : ||\beta - \beta_0|| \le c_0 n^{-1/2}\}$, *for some positive constant* c_0, *we have*

$$\sqrt{nh_1}\left\{\widehat{m}(u) - m(u) - k_0 - b_m(u)h_1^2\right\} \xrightarrow{\mathcal{D}} N\Big\{0, \frac{1}{s_0^2(u)}\sigma_u^2\Big\}. \tag{3.2}$$

Here, $\sigma_u^2 = \tau^2(u)f_U^{-1}(u)\int_{-1}^{1} K^2(t)dt$, *with* $f_U(\cdot)$ *being the density function of* $U = X^\top\beta_0$; $k_0 = \lambda_r - F_1^{-1}(\lambda_r)$, *with* $F_1(\lambda_r) = \int_{-\infty}^{\lambda_r} F(\epsilon)d\epsilon$; $s_0(u) = q(r(u))$; $b_m(\cdot)$ *is a bounded continuous function that is determined by the terms* T_2 *and* T_6 *in the online Supplementary Material. If we further assume that*

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

$\sup_{\epsilon\in\Omega_\epsilon}\epsilon \leq \lambda_r$, *where* Ω_ϵ *is the domain of* ϵ, *then the term* k_0 *disappears, and we have*

$$\sqrt{nh_1}\Big\{\widehat{m}(u) - m(u) - b_m(u)h_1^2\Big\} \xrightarrow{\mathcal{D}} N\Big\{0, \frac{1}{s_0^2(u)}\sigma_u^2\Big\}. \tag{3.3}$$

Theorem 1 shows that the estimator $\widehat{\beta}$ is root$-n$ consistent and asymptotically normal. Theorem 2 indicates that, up to a location constant, the proposed nonparametric estimator achieves the optimal convergence rate. Furthermore, note that although k_0 is theoretically nonzero, it is numerically negligible in many situations, based on our experience. In addition, Theorem 2 theoretically justifies that the location shift k_0 disappears with slightly stronger assumptions.

Remark 3. Constructing confidence intervals for β_0 and $m(\cdot)$ may require that we estimate the asymptotic variances involved in Theorems 1–2. The weighting function $\tau(\cdot)$ and asymptotic covariance matrix W_0^+ of $\widehat{\beta}$ can be estimated using typical variance estimation methods for heterogeneous single-index models (Ichimura, 1993; Härdle, Hall, and Ichimura, 1993; Chiou and Müller, 1998, 1999). The asymptotic variance $\sigma_u^2/s_0^2(u)$ of the link function estimator can be obtained by replacing $f_U^{-1}(\cdot)$ and $s_0(\cdot)$ with their consistent estimators (Lewbel and Linton, 2002). Considering the potential complexity in the estimation of the variances, the bootstrap method is a good alternative for constructing confidence intervals for β_0 and $m(\cdot)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Lastly, we state the asymptotic properties of the proposed test. We need two additional assumptions.

Assumption A.5. The random noise $\epsilon_i \sim N(0, \sigma^2)$, where $\sigma \in \Omega_\sigma$ is an unknown parameter.

Assumption A.6. For any given $\beta \in \Theta$, and any root-n consistent estimator $\widehat{\sigma}$ of σ, $\sup_{(x,\sigma)\in D_X\times\Omega_\sigma} |r(x^\top\beta, \widehat{\sigma}) - r(x^\top\beta, \sigma) - (\widehat{\sigma} - \sigma)r'(x^\top\beta, \sigma)| = O_p(1/n)$, where $r(x^\top\beta, \sigma) = \int_{-\infty}^{m(x^\top\beta)} \Phi(\epsilon/\sigma)d\epsilon$.

Assumption A.6 is adapted from Assumption (C.4) of Koul, Song, and Liu (2014). We have the following result for T_n.

Theorem 3. *Assume Assumptions A.1–A.6 hold. Then, under* H_0,

$$T_n = nh^{1/2}V_n/\widehat{\gamma} \xrightarrow{\mathcal{D}} N(0, 1).$$

4. Simulation Studies

In this section, we investigate the finite-sample performance of the proposed estimation and testing methods using Monte Carlo simulations. Examples 4.1 and 4.2 focus on the estimation of β_0 and $m(\cdot)$, respectively, and Example 4.3 studies the performance of T_n.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Example 4.1. In this example, we focus on the estimation of β_0. We generate 100 replicates from the following two models:

$$Y_i^* = e^{(X_{i1}+X_{i2})/\sqrt{2}} - \epsilon_i,\ i = 1, \ldots, n, \tag{4.1}$$

and

$$Y_i^* = \sin\Big(\pi\{(X_{i1} + X_{i2})/\sqrt{2}\}/(b-a)\Big) - \epsilon_i,\ i = 1, \ldots, n, \tag{4.2}$$

where X_{i1} and X_{i2} are i.i.d. from $Uniform(0, 1)$, ϵ_i follows either a $N(0, 0.1^2)$ or a Laplace distribution $\mathcal{L}(0, 0.1^2)$, and $a = \sqrt{2}/2$ and $b = \sqrt{3}/2 + 1.645/\sqrt{12}$. In both (4.1) and (4.2), the true index parameter is $\beta_0 = (\beta_{01}, \beta_{02})^\top = (0.701, 0.701)^\top$. The observed responses Y_i are set as $Y_i = \max(Y_i^*, c)$, where c is properly chosen to yield two censoring proportions (Cen), Cen=20% and Cen=40%. We consider two sample sizes, $n = 200$ and 400.

Because no estimation methods are available for such models, we compare our estimator with the widely used profile least-squares estimator, based on the latent data Y_i^* (corresponding to Cen=0). The performance is evaluated using the L_2 difference $||\beta_0 - \widehat{\beta}||_2$ across replicates. We select the bandwidth h using a grid search to minimize the simulation-based estimates of the L_2 differences, following Liang et al. (2010). The average CPU time for each replicate is 28 seconds for $n = 200$ and 101 second-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

s for $n = 400$, running on an Intel(R) Core(TM) i7-6700HQ CPU with 2.60GHz. Table 1 summarizes the averaged estimates (AVE) of β_0 and the corresponding MSE. From Table 1, we find that the biases based on Y_i are comparable with those from Y_i^*, whereas the MSE based on Y_i is larger, but still within a reasonable range.

Example 4.2. In this example, we focus on the estimation of $m(\cdot)$. We generate 200 replicates, where each replicate consists of $n = 400$ observations from models (4.1) and (4.2). We estimate $m(\cdot)$ at 400 grid points, uniformly spaced within the range of $X^\top \beta_0$. The censoring point c is set to yield Cen=20%, which mimics our real HIV data in Section 5. To alleviate the computational burden, the bandwidths for estimating the link function are chosen using a the rule of thumb (Silverman, 1986), that is, $h_1 = 1.06 s(X^\top \widehat{\beta}) n^{-1/5}$ and $h_2 = 1.06 s(\widehat{r}(X^\top \widehat{\beta})) n^{-1/5}$, where $s(\cdot)$ denotes the sample standard deviation.

Figures 1 and 2 present the point-wise median curve (solid line) of the estimated function $\widehat{m}(u)$ on the selected grid, point-wise 5% and 95% quantiles (dotted line) of $\widehat{m}(u)$, and the true $m(u)$ (dashed lines). The difference between the median curve and the true curve provides a measure of the bias, whereas the 5% and 95% lines provide measures of spread, which can be interpreted as simulation-based point-wise confidence bands.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Table 1: Example 4.1, average estimates (AVE) and MSE$\times 10^4$ of the index parameter.

n	Cen	Model (4.1)				Model (4.2)			
		AVE		MSE		AVE		MSE	
		β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
		$\epsilon_i \sim N(0, 0.1^2)$							
200	0%	0.7067	0.7080	4.84	5.07	0.7083	0.7057	5.37	5.64
	20%	0.7065	0.7076	5.16	4.51	0.7083	0.7054	6.92	7.35
	40%	0.7080	0.7061	7.91	7.43	0.7079	0.7052	10.80	12.12
400	0%	0.7078	0.7064	1.16	0.98	0.7069	0.7072	1.32	1.46
	20%	0.7067	0.7060	1.63	1.49	0.7074	0.7074	2.05	2.28
	40%	0.7056	0.7070	1.64	1.67	0.7093	0.7065	3.52	4.24
		$\epsilon_i \sim \mathcal{L}(0, 0.1^2)$							
200	0%	0.7067	0.7074	4.12	3.67	0.7079	0.7073	5.76	5.97
	20%	0.7075	0.7068	5.19	4.47	0.7076	0.7067	7.41	7.78
	40%	0.7071	0.7061	6.27	6.34	0.7071	0.7052	9.52	9.89
400	0%	0.7070	0.7072	1.27	1.30	0.7074	0.7071	3.16	3.25
	20%	0.7069	0.7730	1.28	1.33	0.7074	0.7068	4.01	4.16
	40%	0.7071	0.7070	1.76	1.72	0.7093	0.7065	5.09	5.69

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

In general, regardless of normal errors or Laplace errors, the fitted curves are close to the true curve, and the confidence bands cover the true curve, except for a small region. Finally, as pointed out by Lewbel and Linton (2002), if the assumption that $\sup_{\epsilon\in\Omega_\epsilon}\epsilon \leq \sup_u r(u) = \lambda_r$ in Theorem 2 is not satisfied, a location shift may be expected. However, for these scenarios, $\int_{-\infty}^{\widehat{\lambda}_r}\epsilon f(\epsilon)d\epsilon$ is almost zero and $F(\widehat{\lambda}_r) = 1$, numerically, which implies that $\int_{-\infty}^{\widehat{\lambda}_r}\epsilon f(\epsilon)d\epsilon = \widehat{\lambda}_r - \int_{-\infty}^{\widehat{\lambda}_r}F(\epsilon)d\epsilon = 0$ (i.e., $\widehat{\lambda}_r = F_1^{-1}(\widehat{\lambda}_r)$). Therefore, the location bias can be ignored.

Example 4.3. In this example, we focus on the linearity test. We generate 200 replicates from the model

$$Y_i^* = m\left((X_{i1} + X_{i2})/\sqrt{2}\right) - \sigma\epsilon_i,\ Y_i = \max(Y_i^*, c),\ i = 1, \ldots, n,$$

where X_{i1}, X_{i2} are i.i.d. from $N(0,1)$, $\epsilon_i \sim N(0,1)$, σ is equal to either 0.1 or 0.25, n is equal to either 200 or 400, and c is chosen to yield Cen=20%. The true $m(\cdot)$ function is

$$m(u) = u + c_2\exp(u),$$

where c_2 ranges from 0 to 0.16, with increment 0.04, and $c_2 = 0$, corresponding to the null hypothesis.

Table 2 summarizes the rejection rates for all cases, given the nominal level 0.05. We find the following: (i) under the null hypothesis, the em-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

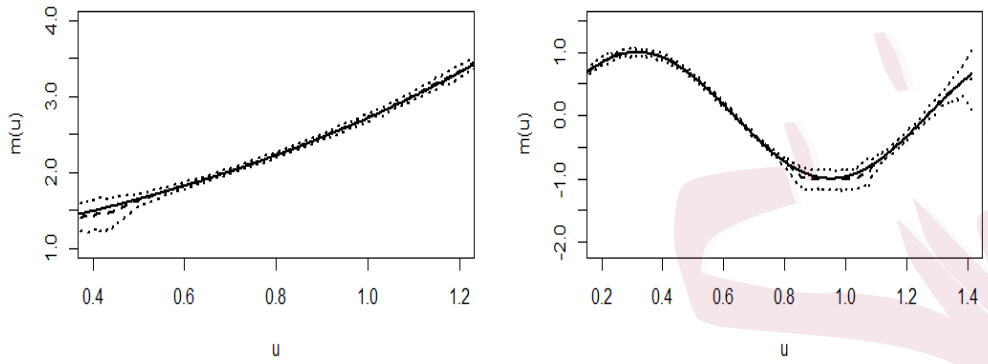

Figure 1: Simulation results for models (4.1) and (4.2) with the normal error: fitted curves (dashed lines) and true curves (solid lines) with 90% confidence bands (dotted lines)

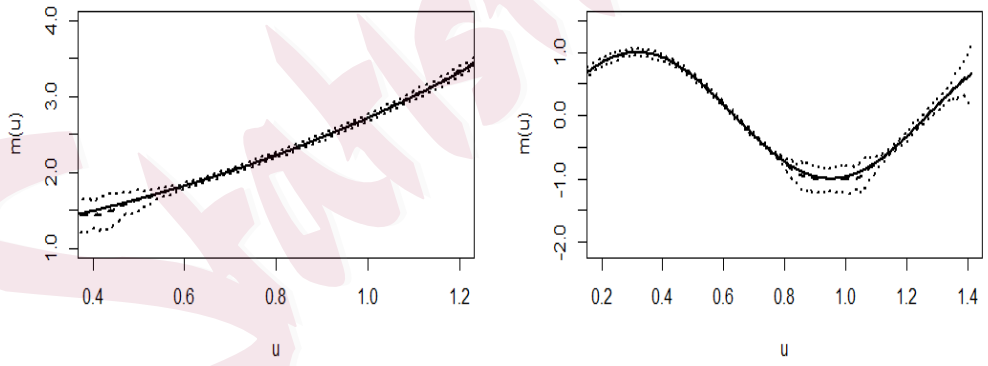

Figure 2: Simulation results for models (4.1) and (4.2) with the Laplace error: fitted curves (dashed lines) and true curves (solid lines) with 90% confidence bands (dotted lines)

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Table 2: Rejection rates for the linearity test of the link function when $n = 200$ or 400, and $\sigma = 0.1$ or $\sigma = 0.25$.

	n=200		n=400	
c_2	$\sigma = 0.1$	$\sigma = 0.25$	$\sigma = 0.1$	$\sigma = 0.25$
0	0.02	0.01	0.01	0.02
0.04	0.58	0.04	0.96	0.13
0.08	0.98	0.35	1	0.75
0.12	1	0.78	1	0.98
0.16	1	0.93	1	1

pirical sizes are less than the nominal level; hence, the proposed tests are conservative, which is common for nonparametric smoothing based tests (Zheng, 1996; Koul, Song, and Liu, 2014); (ii) when the alternative is true, the power approaches to one quickly.

5. Analysis of an HIV Study

A primary goal of vaccine strategies aimed at trying to prevent HIV infection is the induction of a protective humoral response. Some HIV-infected patients develop potent serum antibodies that are able to neutralize a broad range of HIV isolates. By studying the characteristics of the T-cells in such

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

patients, mechanisms for the induction of potent neutralizing antibodies may be revealed.

In this section, we apply the proposed methods to analyze a data set from a study that measures T-cell-related parameters in HIV patients with varying degrees of HIV viral load. The data set consists of observations of of four variables for 414 patients: CD4, CD8, the difference of CD4 (diffcd4), and difference of CD8 (diffcd8). Owing to detection limit, 20% of viral load values are left censored at 50 units. All covariates are standardized to $[0, 1]$, and a log-transformation is applied to the response variable.

We first apply the linearity test for the link function. The resulting p-value is 0.002, which provides strong evidence that the link function is nonlinear. As a result, the proposed model is more appropriate for this data set. We then estimate the index parameter and the link function. The bandwidth for estimating β is selected using 10-fold cross-validation, yielding $h_{real} = 0.14$, and the bandwidths for estimating the unknown link function are selected using a rule of thumb, as in the simulation study.

The estimated coefficients are 0.3970 (CD4), 0.0002 (CD8), 0.5919 (diffcd4), and -0.7015 (diffcd8). Figure 3 presents the estimated curve of the link function and the 90% point-wise confidence band at 50 grid points, uniformly spaced between $[0, 0.3]$. The figure indicates that the viral load

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

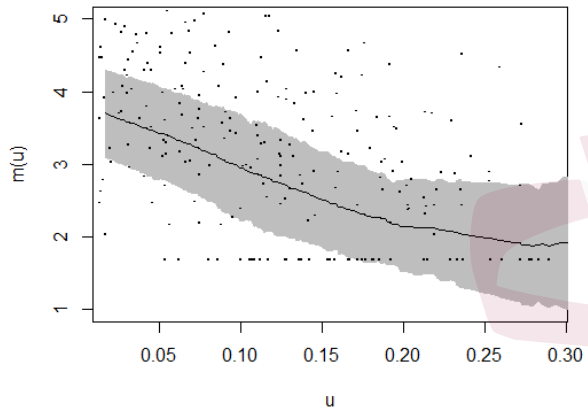

Figure 3: Fitted link function (solid line) and 90% confidence band (shaded area)

shows a logarithmically descending trend with the composite single-index. Combining the index parameter signs and the descending trend of the link function, we find that CD4, CD8, and diffcd4 have negative effects, whereas diffcd8 has a positive effect on the viral load, although the effect of CD8 is very small. These results are largely consistent with the conclusions in the scientific literature. For example, Jiao et al. (2006) discovered that there is a negative relation between CD4 and viral load.

Supplementary Material

The online Supplementary Material provides the proofs of Proposition

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

1 and Theorems 1–3.

Acknowledgments

The authors would like to thank the two reviewers, the associate editor, and the editor for their constructive comments and helpful suggestions. Liang's research was partially supported by National Science Foundation grant DMS-1418042 and DMS-1620898, National Natural Science Foundation of China grant, Award Number 11529101. Tang's research was partially supported by the OSR-2015-CRG4-2582 grant from KAUST, Shanghai Pujiang Program 18PJ1409800, and Key Laboratory for Applied Statistics of MOE, Northeast Normal University 130028849.

References

Adesina, A. A. and Zinnah, M. M. (1993). Technology characteristics, farmers' perceptions and adoption decisions: A Tobit model application in sierra leone. *Agricultural Econom.* **9**, 297–311.

Adesina, A. A. and Zinnah, M. M. (1993). Technology characteristics, farmers' perceptions and adoption decisions: A Tobit model application in sierra leone. *Agricultural Econom.* **9**, 297–311.

Amemiya, T. (1979). The estimation of a simultaneous-equation Tobit model. *Internat. Econom. Rev.* **20**,169–181.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

Amemiya, T. (1984). Tobit models: a survey. *J. Econometrics* **24**,3–61.

Blundell, R. and Meghir, C. (1987). Bivariate alternatives to the Tobit model. *J. Econometrics* **34**,179–200.

Bücher, A., El Ghouch, A. and Van Keilegom, I. (2014). Single-index quantile regression models for censored data. *Technical report*, UCL.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92**,477–489.

Chen, S., Dahl, G. B. and Khan, S. (2005). Nonparametric identification and estimation of a censored location-scale regression model. *J. Amer. Statist. Assoc.* **100**, 212–221.

Chiang, C.-T., Wang, S.-H. and Huang, M.-Y. (2017). Versatile estimation in censored single-index hazards regression. *Ann. Inst. Statist. Math.* **70**, 1–29.

Chiou, J.-M. and Müller, H.-G. (1998). Quasi-likelihood regression with unknown link and variance functions. *J. Amer. Statist. Assoc.* **93**, 1376–1387.

Chiou, J.-M. and Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Ann. Statist.* **27**, 36–64.

Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free regression method. *Ann. Statist.* **19**,505–530.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Vol. 66, Chapman & Hall, London.

Gawande, K. (1995). Are US nontariff barriers retaliatory? an application of extreme bounds

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

analysis in the Tobit model. *Rev. Econom. Statist.* **77**, 677–688.

Haab, B. B., Dunham, M. J. and Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biology* **2**,research0004–1.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**,157–178.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Amer. Statist. Assoc.* **91**,1632–1640.

Huang, H. (2017). *Semi-parametric and Structured Nonparametric Modeling with Censored Responses*. PhD thesis, George Washington University.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics* **58**,71–120.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Math. Statist.* **40**,633–643.

Jiao, Y., Xie, J., Li, T., Han, Y., Qiu, Z., Zuo, L. and Wang, A. (2006). Correlation between gag-specific cd8 t-cell responses, viral load, and cd4 count in hiv-1 infection is dependent on disease status. *J. Acquired Immune Def. Syn.* **42**,263–268.

Kobie, J. J., Alcena, D. C., Zheng, B., Bryk, P., Mattiacio, J. L., Brewer, M., LaBranche, C., Young, F. M., Dewhurst, S., Montefiori, D. C. et al. (2012). 9G4 autoreactivity is increased

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

in HIV-infected patients and correlates with HIV broadly neutralizing serum activity. *PLoS One* **7**,e35356.

Kong, E. and Xia, Y. (2017). Uniform bahadur representation for nonparametric censored quantile regression: A redistribution-of-mass approach. *Econometric Theory* **33**,242–261.

Koul, H. L., Song, W. and Liu, S. (2014). Model checking in Tobit regression via nonparametric smoothing. *J. Multivariate Anal.* **125**,36–49.

Lewbel, A. and Linton, O. (2002). Nonparametric censored and truncated regression. *Econometrica* **70**,765–779.

Liang, H., Liu, X., Li, R. and Tsai, C. L. (2010). Estimation and testing for partially linear single-index models. *Ann. Statist.* **38**, 3811–3836.

Lopez, O. (2009). Single-index regression models with right-censored responses. *J. Statist. Plann. Inference* **139**, 1082–1097.

Maddala, G. and Nelson, F. D. (1975). Specification Errors in Limited Dependent Variable Models, *NBER Working Papers 0096*, National Bureau of Economic Research.

Maddala, G. S. (1986). *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge university press.

Nizar Al-Malkawi, H.-A. (2007). Determinants of corporate dividend policy in Jordan: an application of the Tobit model. *J. Econom. Admin. Sci.* **23**,44–70.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coeffi-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

cients. *Econometrica* **57**,1403–1430.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Vol. 26, CRC press.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**,24–36.

Van der Pouw Kraan, T., Boeije, L., Smeenk, R., Wijdenes, J. and Aarden, L. A. (1995). Prostaglandin-e2 is a potent inhibitor of human interleukin 12 production. *J. Experiment. Med.* **181**,775–779.

Xia, Y. C. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97**, 1162–1184.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* **75**, 263–289.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* **75**, 263–289.

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: hhl1988@email.gwu.edu

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

E-mail: hliang@gwu.edu

Department of Statistics, George Washington University, Washington, D.C. 20052, USA

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0451

E-mail: yuanzhang.li@yahoo.com

Key Laboratory of Advanced Theory and Application in Statistics and Data Science - MOE, School of Statistics, East China Normal University, Shanghai, 200062, China

E-mail: yltang@fem.ecnu.edu.cn