Statistica Sinica **13**(2003), 311-325

ASYMPTOTIC PROPERTIES OF THE GENERALIZED SEMI-PARAMETRIC MLE IN LINEAR REGRESSION

Qiqing Yu and George Y. C. Wong

SUNY, Binghamton and Strang Cancer Prevention Center

Abstract: Consider the semi-parametric linear regression model, $Y = \beta'\mathbf{X} + \epsilon$, with sample size n, where ϵ has an unknown cdf F_o. The semi-parametric MLE (SMLE) $\tilde{\beta}_n$ of β under this set-up, called the generalized SMLE or GSMLE, has neither been studied in the literature nor an algorithm for it. We begin with an algorithm for the GSMLE. It is then shown that if F_o has a discontinuity point, $\mathrm{P}\{\tilde{\beta}_n = \beta$ if n is large$\} = 1$. Simulation suggests that under some discontinuous distributions, $\tilde{\beta}_n = \beta$ even for $n = 50$. In contrast the least squares estimator (LSE), $\hat{\beta}_n$, satisfies $\mathrm{P}\{\hat{\beta}_n \neq \beta$ i.o.$\} = 1$. We demonstrate via a real discontinuous data example that the GSMLE can be better than the LSE in applications. Properties of the GSMLE in the continuous case are also mentioned.

Key words and phrases: Algorithms, consistency, generalized likelihood, super efficiency.

1. Introduction

We study the semi-parametric maximum likelihood estimator (SMLE) in the linear regression model with complete data. Specifically, assume

A1. $Y = \beta'\mathbf{X} + \epsilon$, where only $(\mathbf{X}, Y)$ is observable, β is an unknown $p \times 1$ dimensional regression coefficient vector, β' is the transpose of β, and ϵ has an unknown cdf F_o.

Suppose $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are i.i.d. observations from $(\mathbf{X}, Y)$. Under this model, there are several estimators for β: the least squares estimator (LSE); the Theil-Sen estimator (Theil (1950) and Sen (1968)); various M-estimators (Huber (1964) and Ritov (1990)); adaptive estimators (Bickel (1982)); L-estimators and R-estimators (see, e.g., Montgomery and Peck (1992)).

The maximum likelihood method is an interesting estimation approach here. Kiefer and Wolfowitz (1956) define the generalized likelihood function of observations $T_1, \ldots, T_n$ as

$$Ł = \prod_{i=1}^{n} f(T_i), \quad f(t) = F(t) - F(t-) \quad \text{and} \quad F \in \mathcal{F}, \tag{1.1}$$

where $\mathcal{F} = \{F : F \text{ is an increasing function, } F(-\infty) = 0 \text{ and } F(\infty) = 1\}$. The MLE of the cdf, also called the generalized MLE (GMLE), is a value of F that maximizes Ł over $\mathcal{F}$. It is well known that the GMLE is the empirical distribution function. While f in Ł is discrete, the GMLE is consistent and (non-parametrically) efficient even if the cdf is continuous. Under A1, letting $T_i = Y_i - \mathbf{b}'\mathbf{X}_i$ in Ł, the generalized SMLE (GSMLE) of (F_o, β), denoted by $(\tilde{F}, \tilde{\beta}_n)$, is a value of $(F, \mathbf{b})$ that maximizes

$$Ł = \prod_{i=1}^{n} f(Y_i - \mathbf{b}'\mathbf{X}_i) \text{ over all } (F, \mathbf{b}) \in \mathcal{F} \times \mathcal{R}^p, \quad \text{where } f(t) = F(t) - F(t-). \tag{1.2}$$

In the literature, the GSMLE has not been studied and there is no algorithm for obtaining it. In fact, the GSMLE cannot be obtained by standard numerical methods.

For a fixed $\mathbf{b}$, the likelihood function Ł in (1.2) is maximized by the empirical density function $\hat{f}_{\mathbf{b}}$, where

$$\hat{f}_{\mathbf{b}}(Y_i - \mathbf{b}'\mathbf{X}_i) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{(Y_j - \mathbf{b}'\mathbf{X}_j = Y_i - \mathbf{b}'\mathbf{X}_i)} \tag{1.3}$$

and $\mathbf{1}_A$ is the indicator function of an event A. Thus the GSMLE of β actually maximizes

$$\mathcal{L}(\mathbf{b}) = \prod_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{(Y_j - \mathbf{b}'\mathbf{X}_j = Y_i - \mathbf{b}'\mathbf{X}_i)} \text{ over all } \mathbf{b}. \tag{1.4}$$

For simplicity, suppose $p = 1$ and that the n pairs (X_i, Y_i) are all distinct. If $Y_i - bX_i$ are distinct, $i = 1, \ldots, n$, $\mathcal{L}(b) = (\frac{1}{n})^n$, which is the minimum of $\mathcal{L}$. $\mathcal{L}(\cdot)$ increases if $Y_j - bX_j = Y_i - bX_i$ for some $i \neq j$. In the latter case, $b = b_{ij} = \frac{Y_i - Y_j}{X_i - X_j} \in \mathcal{B}_o$, where $\mathcal{B}_o = \{\frac{Y_i - Y_j}{X_i - X_j} : 1 \leq i < j \leq n, X_i \neq X_j\}$. There are at most $n(n-1)/2$ many distinct b_{ij}'s. That is, $\mathcal{L}$ equals its minimum almost everywhere (a.e.) in b, except at $b \in \mathcal{B}_o$, and thus $\frac{d \ln \mathcal{L}}{db} = 0$ a.e.. Newton-Raphson and Monte Carlo methods do not work for the GSMLE since $\mathcal{L}$ equals its minimum a.e..

The GSMLE is akin to a class of efficient M-estimators studied in Zhang and Li (1996). Their M-estimator is a zero-crossing point of a function $\Phi(\mathbf{b})$, which has the form

$$\Phi = \sum_{i=1}^{n} \hat{\phi}(Y_i - \overline{Y} - \mathbf{b}'(\mathbf{X}_i - \overline{\mathbf{X}}))(\mathbf{X}_i - \overline{\mathbf{X}}), \quad \phi(t) = \frac{\partial}{\partial t} \ln f(t), \tag{1.5}$$

$\hat{\phi}$ is an estimator of ϕ, f is a pdf, and $\overline{\mathbf{X}}$ is the sample mean of $\mathbf{X}_i$'s. In view of (1.5), this type of M-estimators is essentially a stationary point of a modification

of the log likelihood. Since $(\hat{f}_{\mathbf{b}}(t))' = 0$ a.e. (see (1.3)) and $\frac{\partial}{\partial \mathbf{b}} \ln \mathcal{L}(\mathbf{b}) = 0$ a.e., $\tilde{\beta}_n$ is an M-estimate with $\hat{\phi} = (\hat{f}_{\mathbf{b}}(t))'/\hat{f}_{\mathbf{b}}(t)$, as $\Phi = 0$ a.e.. Obviously, the M-estimation approach is non-informative to the GSMLE $\tilde{\beta}_n$.

In Section 2, we introduce an algorithm for obtaining the GSMLE. In Section 3, we show that if F_o has a discontinuity point then $\mathrm{P}\{\tilde{\beta}_n \neq \beta$ infinitely often (i.o.)$\} = 0$, i.e., $\tilde{\beta}_n$ is super-efficient. Some detailed proofs in Section 3 are relegated to the appendix. In Section 4, we discuss the properties of the GSMLE when F_o is continuous, compare the GSMLE to other estimators using real discontinuous data sets and present simulation results when $p = 1$ or 2.

The maximum likelihood method is an appealing one. For example, most textbooks on linear regression like to mention that the LSE, denoted by $\hat{\beta}_n$, is the (parametric) MLE if $\epsilon \sim N(\mu, \sigma^2)$. However, $\hat{\beta}_n$ is not the (semi-parametric) MLE under A1. For discontinuous data, the GSMLE $\tilde{\beta}_n$ is super-efficient, and simulation results indicate that $\tilde{\beta}_n = \beta$ even when $n = 50$ in some cases. Super-efficiency has not been reported for the existing estimators mentioned above under linear regression models. On the contrary, we note that the LSE satisfies $\mathrm{P}\{\hat{\beta}_n \neq \beta$ i.o.$\} = 1$. A real discontinuous data set is presented in Example 4.4.1, in which the GSMLE appears better than the LSE. Simulation demonstrates that our algorithm for the GSMLE is feasible for moderate n and small p. With fast-growing computing power, one may be able to find an efficient algorithm in the future, making the GSMLE feasible for large p.

2. The GSMLE

Denote $T(\mathbf{b}) = Y - \mathbf{b}'\mathbf{X}$ and $T_i = T_i(\mathbf{b}) = Y_i - \mathbf{b}'\mathbf{X}_i$. Since $\epsilon_i = T_i(\beta)$, $i = 1, \ldots, n$, are i.i.d. copies of ϵ, the generalized likelihood function is

$$Ł(F, \mathbf{b}) = \prod_{i=1}^{n} f(T_i(\mathbf{b})), \quad \mathbf{b} \in \mathcal{R}^p, \quad f(t) = F(t) - F(t-), \quad \text{and} \quad F \in \mathcal{F}. \tag{2.1}$$

It is well known that, given $\mathbf{b}$, the likelihood function is maximized by the empirical distribution function $\hat{F}_{\mathbf{b}}$ based on $T_i(\mathbf{b})$, $i = 1, \ldots, n$. Denote $\hat{f}_{\mathbf{b}}(t) = \hat{F}_{\mathbf{b}}(t) - \hat{F}_{\mathbf{b}}(t-)$. Then the GSMLE of β is a value of $\mathbf{b}$ that maximizes $\mathcal{L}(\mathbf{b})$ over all $\mathbf{b} \in \mathcal{R}^p$, where $\mathcal{L}(\mathbf{b}) = \prod_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{(Y_j - \mathbf{b}'X_j = Y_i - \mathbf{b}'X_i)}$ (see (1.4)) and $\mathcal{L}(\cdot)$ takes on only finitely many values. Thus the GSMLE exists.

If $p = 1$, b is an GSMLE only if $b \in \mathcal{B}_o$, where $\mathcal{B}_o = \{b : b = \frac{Y_i - Y_j}{X_i - X_j}$, $X_i \neq Y_j$ and $1 \leq i \leq j \leq n\}$. Thus it suffices to compare $\mathcal{L}(b)$ over $b \in \mathcal{B}_o$. Since $\mathcal{L}$ and $\mathcal{B}_o$ are both explicit, one has a non-iterative algorithm for the simple linear regression as follows. Find all $b \in \mathcal{B}_o$. Let $\mathcal{A}_o$ be the collection of all distinct

elements of $\mathcal{B}_o$, say $\mathcal{A}_o = \{a_1, \ldots, a_m\}$. Each value of b that maximizes $\mathcal{L}(b)$ over $b \in \mathcal{A}_o$ is an GSMLE of β. This is implemented as follows.

Algorithm 1.

Step 1. Let $\mathcal{M}_1 = \mathcal{L}(a_1)$, $\mathcal{M} = 1$ and $b_{\mathcal{M}} = a_1$.

Step k ($k=2, \ldots, m$). $\begin{cases} \text{If } \mathcal{L}(a_k) = \mathcal{M}_1 \text{ then increase } \mathcal{M} \text{ by 1 and set } b_{\mathcal{M}} = a_k; \\ \text{if } \mathcal{L}(a_k) > \mathcal{M}_1, \text{ then set } \mathcal{M}_1 = \mathcal{L}(a_k),\ \mathcal{M} = 1 \text{ and } b_{\mathcal{M}} = a_k. \end{cases}$

Step $m+1$ (outputs). $\mathcal{B} = \{b_1, \ldots, b_{\mathcal{M}}\}$ is the set of all GSMLE's of β.

The GSMLE of F_o is then $\hat{F}_{\mathbf{b}}$, where $\mathbf{b} \in \mathcal{B}$. Under the semi-parametric assumption, $\alpha = E(\epsilon)$ may not exist. If α exists, a naive GSMLE of α is $\tilde{\alpha} = \sum_t t \hat{f}_{\tilde{\beta}}(t)$, where $\tilde{\beta} \in \mathcal{B}$. According to our simulation experience, the algorithm is feasible if n is up to 1000 for $p = 1$. Algorithm 1 guarantees to yield all GSMLE's. However, it is very time-consuming if $n > 1000$. For practical purposes, we introduce another algorithm.

Algorithm 2. (Compute the mode of $b \in \mathcal{B}_o$).

Preliminary: Let $M(b)$ be the multiplicity of b in $\mathcal{B}_o$, i.e., $M(b) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \mathbf{1}_{(b_{ij}=b, b_{ij} \in \mathcal{B}_o)}$.

Step 1. Let $a_1, \ldots, a_m$ be all the distinct elements of the set $\mathcal{B}_o$. If $m = ||\mathcal{B}_o||$ (the cardinality of the set $\mathcal{B}_o$), then stop. Each a_i is an GSMLE of β. Otherwise, proceed.

Step 2. Reorder the a_i's so that $M(a_1) \geq \cdots \geq M(a_m)$.

Step 3. Find $\mathcal{N}$ such that $M(a_1) = \cdots = M(a_{\mathcal{N}}) > M(a_{\mathcal{N}+1})$.

Step 4. For $j \leq \mathcal{N}$, if a_j maximizes $\mathcal{L}(a_i)$ over $i \leq \mathcal{N}$ then take a_j as an GSMLE.

Algorithm 2 is motivated by super-efficiency, which suggests that if n is large enough, β is the mode of slopes of line segments connecting two points (X_i, Y_i) and (X_j, Y_j), $i < j$. This is not as rigorous as Algorithm 1 in the sense that it may sometimes not yield an GSMLE. However, for moderate sample sizes, it often generates an GSMLE and is much faster than Algorithm 1 (see Example 2.1 below). In applications, one may use Algorithm 1 if n is small, Algorithm 2 otherwise.

Example 2.1. Suppose there are $n = 5$ data: $(X_i, Y_i) = (0,1)$, $(1,2)$, $(2,3)$, $(4, -6)$ and $(3, -6)$. $\mathcal{B}_o = \{-9, -9/2, -4, -8/3, -7/3, -7/4, 0, 1, 1, 1\}$. Using Algorithm 2, we know immediately that the GSMLE is 1, as $\mathcal{N} = 1$, $M(1) = 3$ and $M(b) \leq 1$ if $b \neq 1$. However, Algorithm 1 needs to compute $\mathcal{L}(\cdot)$ eight times in order to find $\tilde{\beta}_n = 1$.

Remark 2.1. Algorithms 1 and 2 can be extended to the case $p > 1$ by replacing $\mathcal{B}_o$ by $\mathcal{B}_m$, the set of all possible $\mathbf{b}_{i_1,\ldots,i_p,j_1,\ldots,j_p}$ which solve the p linearly independent equations $Y_{i_k} - \mathbf{b}'\mathbf{X}_{i_k} = Y_{j_k} - \mathbf{b}'\mathbf{X}_{j_k}$, $k = 1, \ldots, p$, where $i_k, j_k \in \{1, \ldots, n\}$. The justification is similar to that for the case $p = 1$ and $\mathcal{B}_o$. Roughly speaking, (1) $T_i(\mathbf{b}) = Y_i - \mathbf{b}'\mathbf{X}_i$, (2) $\mathcal{L}(\mathbf{b}) = (1/n)^n$ when $T_i(\mathbf{b})$'s are all distinct, and (3) $\mathcal{L}(\mathbf{b})$ increases if $T_i(\mathbf{b}) = T_j(\mathbf{b})$ for some j (see the expression $\mathcal{L}$ in (1.4)). That is, $\tilde{\beta}_n$ should not belong to the subset for which the $T_i(\mathbf{b})$'s exhibit few ties. By construction, $\mathcal{B}_m$ is the subset for which there are at least $p + 1$ ties among the T_i's. We skip the details.

3. The Main Result

In this section, we investigate the properties of the GSMLE under the assumption that F_o is discontinuous. Let $\mathcal{D}$ be the collection of all discontinuity points of F_o. We make use of the following assumptions.

A2. ϵ and $\mathbf{X}$ are independent.

A3. $\mathrm{P}\left(\mathrm{rank}\begin{pmatrix} 1 & \cdots & 1 \\ \mathbf{X}_1 & \cdots & \mathbf{X}_{p+1} \end{pmatrix} = p + 1\right) > 0.$

A4. $\mathcal{D}$ is not empty and $|E(\xi \ln f_o(\epsilon))| < \infty$, where $f_o(t) = F_o(t) - F_o(t-)$ and $\xi = \mathbf{1}_{(\epsilon \in \mathcal{D})}$.

A2 and A3 are identifiability conditions. In Bickel (1982), A3 is replaced by the stronger condition $E(|(\mathbf{X}_1, \ldots, \mathbf{X}_n)(\mathbf{X}_1, \ldots, \mathbf{X}_n)'|) \neq 0$. It is easy to understand A3 in case $p = 1$: we need at least two distinct values $x_1 \neq x_2$ of $\mathbf{X}$ to identify β. Thus $\mathrm{rank}\left(\begin{pmatrix} 1 & 1 \\ x_1 & x_2 \end{pmatrix}\right) = 2$.

In proving consistency of an MLE, one often takes advantage of the Shannon-Kolmogorov inequality which requires $|E(\ln Ł)| < \infty$, rather than A4. However, under the current semi-parametric set-up, $E(\ln Ł(F_o, \beta)) = -\infty$ unless ϵ is discrete. This can be viewed as follows. Let $\xi_i = \mathbf{1}_{(\epsilon_i \in \mathcal{D})}$ and

$$Ł_d(F, \mathbf{b}) = \prod_{i=1}^{n} (f(T_i(\mathbf{b})))^{\xi_i} \quad \text{and} \quad Ł_c(F, \mathbf{b}) = \prod_{i=1}^{n} (f(T_i(\mathbf{b})))^{1-\xi_i}. \tag{3.1}$$

Then $Ł = Ł_d(F, \mathbf{b})Ł_c(F, \mathbf{b})$. Define $f_o(t) = F_o(t) - F_o(t-)$ (see (2.1)). If $\mathrm{P}\{\epsilon \in \mathcal{D}\} < 1$, then $f_o(t) = 0$ at continuity points of F_o. If we take $\ln 0 = -\infty$, $E(\ln Ł(F_o, \beta)) \leq E(\ln L_c(F_o, \beta)) = nE((1 - \xi) \ln f_o(\epsilon)) = -\infty$. Note that A4 is equivalent to $|E(\ln Ł_d(F_o, \beta))| = |nE(\xi \ln f_o(\epsilon))| < \infty$.

The following is the main result of the paper.

Theorem 3.1. *Suppose* A1-A4 *hold. If F_o has a discontinuous point, then $P\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} = 0$.*

Proof. We first prove the theorem in a simple case, then outline the proof in the general case in the Appendix.

First assume that $p = 1$, $\mathbf{X}$ is continuous and F_o has a unique discontinuity point, say at t_0. Then (1) $T(b) = \epsilon + (\beta - b)\mathbf{X}$, (2) $\mathrm{P}\{T(\beta) = t\} = 0$ for all $t \neq t_0$, (3) $c = \mathrm{P}\{T(\beta) = t_0\} > 0$, and (4) $\mathrm{P}\{T(b) = t\} = 0$ for all t if $b \neq \beta$. If $b \neq \beta$ then by (4), with probability one, except perhaps one pair of $T_i(b)$'s (if $b \in \mathcal{B}_o$), the rest $T_j(b)$'s are all distinct and $\mathcal{L}(b) \leq (\frac{1}{n})^{n-2}(\frac{2}{n})^2$. On the other hand, $\mathcal{L}(\beta) \approx (\frac{1}{n})^{n-nc}(\frac{nc}{n})^{nc} > (\frac{1}{n})^{n-2}(\frac{2}{n})^2$ by (3), if n is large. Thus $\tilde{\beta}_n = \beta$ in this simple case.

In general, we show that $\mathrm{P}\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} > 0$ leads to a contradiction. Under the given assumptions, with probability one (w.p.1),

$$\hat{f}_{\mathbf{b}}(t) \text{ converges uniformly in } (\mathbf{b}, t) \quad \text{(Lemma 5.1)}; \tag{3.2}$$

$$\varliminf_{n\to\infty} \frac{1}{n} \ln Ł_d(\hat{F}_\beta, \beta) \geq E(\xi \ln f_o(\epsilon)) \ (= E(\frac{1}{n} \ln Ł_d(F_o, \beta))) \quad \text{(Lemma 5.2)}; \tag{3.3}$$

$$\varliminf_{n\to\infty} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) \geq \lim_{n\to\infty} \frac{1}{n} \ln Ł_d(\hat{F}_\beta, \beta) \quad \text{(Lemma 5.3)}. \tag{3.4}$$

It follows from (3.3) and (3.4) that w.p.1,

$$\varliminf_{n\to\infty} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) \geq E(\xi \ln f_o(\epsilon)). \tag{3.5}$$

Let Ω_0 be the event that (3.2), (3.3), (3.4) and (3.5) hold. Then $\mathrm{P}\{\Omega_0\} = 1$. Now if $\mathrm{P}\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} > 0$, then $\Omega_0 \cap \{\tilde{\beta}_n \neq \beta \text{ i.o.}\}$ is not empty.

Suppose $\mathbf{X}$ is discrete. Then it is shown in Lemma 5.4 that

$$\lim_{k\to\infty} \sup_{n\geq k, \tilde{\beta}_n \neq \beta} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) < E(\xi \ln f_o(\epsilon)) \quad \text{if } \omega \in \Omega_0 \cap \{\tilde{\beta}_n \neq \beta \text{ i.o.}\}, \tag{3.6}$$

contradicting (3.5) since $\varliminf_{n\to\infty} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) \leq \lim_{k\to\infty} \sup_{n\geq k, \tilde{\beta}_n \neq \beta} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n)$. If $\mathbf{X}$ is not discrete, our Lemma 5.5 has

$$\lim_{k\to\infty} \sup_{n\geq k, \tilde{\beta}_n \neq \beta} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) < E(\xi \ln f_o(\epsilon)) \text{ with positive probability.} \tag{3.7}$$

Then (3.7) contradicts (3.5) again as $\mathrm{P}\{\Omega_0\} = 1$, concluding the proof of the theorem.

4. Discussion

We have established asymptotic properties of the GSMLE when F_o is discontinuous. In this section we discuss the continuous case. Moreover, we apply the

method to real data. Finally, we provide some simulation results on computing the GSMLE when $p = 1$ or 2 and discuss variance estimation of the GSMLE.

4.1. The continuous case

When F_o is continuous, Theorem 3.1 cannot apply. For example, if $p = 1$ and both X and ϵ are continuous then, w.p.1, there are exactly $n(n-1)/2$ distinct $b_{ij} = (Y_i - Y_j)/(X_i - X_j)$, $1 \le i < j \le n$. These are all the GSMLE's, as $\mathcal{L}(b)$ equals $(1/n)^n$ or $(1/n)^{n-2}(2/n)^2$.

Remark 4.1. If the GSMLE of β is not unique, one can choose an GSMLE closest to the LSE. Let this be $\tilde{\beta}$ and the LSE be $\hat{\beta}$. If both ϵ and $\mathbf{X}$ are continuous, $\tilde{\beta}$ and $\hat{\beta}$ can be expected to have the same asymptotic properties. This can be viewed as follows with $p = 1$. The LSE is between two consecutive GSMLE's of which there are $n(n-1)/2$. Thus one expects $|\tilde{\beta} - \hat{\beta}| = O(n^{-2})$ and $\mathrm{Var}\,(\tilde{\beta})/\mathrm{Var}\,(\hat{\beta}) \to 1$.

We present simulation results to support the last statement. Examples 4.1.1 and 4.1.2 are continuous cases, while Examples 4.1.3 and 4.1.4 are discontinuous.

Example 4.1.1. Suppose ϵ has a uniform distribution on the interval $(-1, 1)$ ($\epsilon \sim U(-1, 1)$), $X \sim U(0, 9)$ and $(\alpha, \beta) = (0, 2)$.

Example 4.1.2. Suppose $\epsilon \sim N(0, 0.09)$, $X \sim U(0, 9)$ and $(\alpha, \beta) = (0, 1)$.

Example 4.1.3. Suppose ϵ is a mixture of $U(0, 0.9)$ and the constant 0.45, with probabilities 0.9 and 0.1, respectively, $X \sim U(1, 2)$ and $(\alpha, \beta) = (0, 1)$.

Example 4.1.4. Suppose $\epsilon + \alpha$ is a mixture of $U(0, 0.6)$ and the constant 0.2, with probabilities 0.6 and 0.4, respectively, $X \sim U(1, 2)$ and $(\alpha, \beta) = (0.26, 1)$.

For each sample size and each situation, we carried out 1000 simulations and computed the sample mean and sample standard error (SE) of the 1000 estimates. It took less than a minute at a Pentium III PC in each case. The results of the above examples are summarized in Table 1.

It is seen in our continuous examples (see Examples 4.1.1 and 4.1.2), $\tilde{\beta}$ and $\hat{\beta}$ are equivalent in the sense that both of them are consistent and their asymptotic standard deviations are almost the same for moderate sample sizes ($n \ge 200$). These results support the statement in Remark 4.1. In particular, since $\hat{\beta}$ is efficient in Example 4.1.2, so is $\tilde{\beta}$. If F_o is discontinuous, the SE of $\tilde{\beta}$ is obviously smaller than the SE of $\hat{\beta}$. When n is large enough, the SE of $\tilde{\beta}$ is 0 while the SE of $\hat{\beta}$ is positive. In fact, it is so even when $n = 32$ in Example 4.1.4. This suggests that the GSMLE has some advantage over the LSE even when n is moderate, provides the underlying distribution is not continuous.

Table 1. Simulation results on estimating β when $p = 1$.

Example	sample size	β	GSMLE $\tilde{\beta}$(SE)	LSE (SE)
continuous F_o				
4.1.1	32	2	1.992 (0.047)	1.996 (0.040)
	100	2	1.997 (0.021)	1.996 (0.021)
4.1.2	32	1	0.998 (0.025)	1.000 (0.022)
	200	1	1.000 (0.009)	1.000 (0.009)
discontinuous F_o				
4.1.3	32	1	1.003 (0.121)	1.000 (0.156)
	200	1	1.000 (0.000)	0.997 (0.060)
4.1.4	32	1	1.000 (0.000)	1.002 (0.088)

4.2. Computation in multiple regression

We carried out simulations for $p > 1$ as well. For example, let X and ϵ have the same distribution as in Example 4.1.4 and let $Y = 2X + 4X^2 + \epsilon$. That is, $(\alpha, \beta_1, \beta_2) = (0.26, 2, 4)$ and $p = 2$. We took $n = 50$, 100 and 200, respectively, each with 1000 simulations. The results are in Table 2. It is seen that the sample average of the GSMLE of (β_1, β_2) is always $(2, 4)$, with SE $(0, 0)$ even when $n = 50$.

Table 2. Simulation results on estimating β when $p = 2$.

	$(\tilde{\beta}_{50,1}, \tilde{\beta}_{50,2})$	$(\tilde{\beta}_{100,1}, \tilde{\beta}_{100,2})$	$(\tilde{\beta}_{200,1}, \tilde{\beta}_{200,2})$	(β_1, β_2)
sample mean	(2.000,4.000)	(2.000,4.000)	(2.000,4.000)	(2,4)
SE	(0.000,0.000)	(0.000,0.000)	(0.000,0.000)	

It is obvious that the computing cost is expensive if n or p is large. However, for small p and $n \leq 1000$, the cost of computation is not bad since it only involves finding solutions of linear equations and empirical distribution functions. If $n > 1000$, say, then the LSE is pretty accurate and one does not need to compute the GSMLE.

4.3. Variance estimation

Theorem 3.1 suggests that if n is large and F_o is discontinuous, then $n\text{Var}\,(\tilde{\beta}_n) = 0$. Simulation results in Table 1 suggest that if there is severe discontinuity in F_o (i.e., $c = P\{X \in \mathcal{D}\}$ is not small, say $nc >> p$), then the sample variance of 1000 replications is 0 even for n as small as 30. In applications, there are two problems: (1) it is not clear in general how large the sample size should be so that $\tilde{\beta}_n = \beta$; (2) the model assumption $Y = \beta'\mathbf{X} + \epsilon$ is at best a linear

approximation of a non-linear model using the first-order Taylor expansion. Thus one needs some method to estimate the variance of the GSMLE. We propose two methods: (1) estimating $\mathrm{Var}\,(\tilde{\beta}_n)$ by the estimator of the variance of the LSE; (2) estimating $\mathrm{Var}\,(\tilde{\beta}_n)$ by the Bootstrap method.

In view of the super-efficiency of the GSMLE under discontinuous assumptions, by modifying the GSMLE it may be able to construct a semi-parametric efficient estimator under continuous assumptions. This is an interesting open problem.

4.4. Applications

The GSMLE is extremely good under the assumption that F_o is discontinuous, which is true if both Y and $\mathbf{X}$ are discontinuous, as $\epsilon = Y - \beta'\mathbf{X}$. The following are real discontinuous data examples. We compare the GSMLE to the LSE and/or the Theil-Sen estimator. We do not bring in M-estimators, R-estimators or L-estimators, as each of them refers to a wide class of estimators. In fact, the LSE is an M-estimator.

Example 4.4.1. In a study of revenue from advertising (see Chatterjee and Price (1991), p.257), data were collected from 41 magazines in 1986. They are plotted in Figure 1. Let X denote the number of pages of advertising (in hundreds) and Y the advertising revenue (in millions of dollars). Chatterjee and Price fitted them to a simple linear regression model.

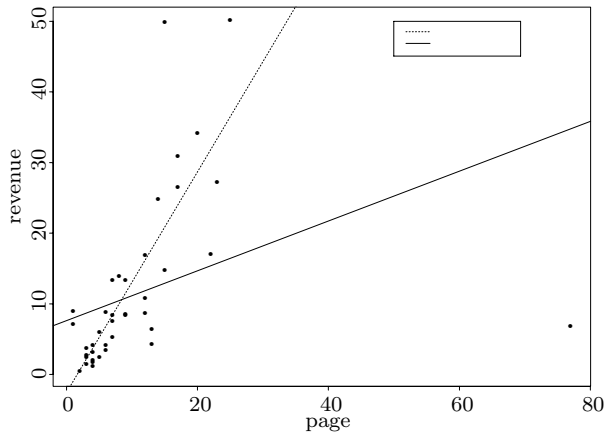

Figure 1. Plot of advertising pages and advertising revenue.

This is a discontinuous data set, as there are ties among the (X_i, Y_i)'s (e.g., two $(3, 1.3)$'s). Furthermore, X and Y are discrete by nature. The GSMLE of β is 1.2 and the LSE is 0.35 with an SE 0.14. It is seen from Figure 1 that the

GSMLE is more reasonable than the LSE due to an outlier. If it is deleted, the GSMLE remains the same and the LSE is 1.238 with an SE 0.14.

Example 4.4.2. The weight and the systolic blood pressure of 26 randomly selected males in the age group 25-30 were collected. The data can be found in Montgomery and Peck ((1992), p.63)). Letting X denote the weight and Y the systolic blood pressure, Montgomery and Peck fit the data to a simple linear regression model. While weight and blood pressure are continuous random variables in theory, they are discrete in practice. Moreover, there are ties among the X_i's or Y_i's in this data set (e.g., $(X_i, Y_i) = (180, 156)$, $(180, 150)$, or $(170, 150)$).

The LSE of $\hat{\beta}$ is 0.4194 with an SE 0.0674; the Theil-Sen estimate is 0.4857; the (unique) GSMLE of β is 1 with an SE 0.0674 (using the SE of the LSE). There are differences, but it is not clear which estimator is better.

5. Appendix

In this appendix, we assume A1-A4 and give proofs of some technical details in the proof of Theorem 3.1.

Lemma 5.1. *Statement* (3.2) *holds.*

Proof. Denote $F_{\mathbf{b}}(t) = P\{T(\mathbf{b}) \le t\}$ and $f_{\mathbf{b}}(t) = F_{\mathbf{b}}(t) - F_{\mathbf{b}}(t-)$. An estimator of $F_{\mathbf{b}}$ is $\hat{F}_{\mathbf{b}}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(T_i(\mathbf{b})\le t)}$. Since $\hat{f}_{\mathbf{b}}(t) = \hat{F}_{\mathbf{b}}(t) - \hat{F}_{\mathbf{b}}(t-)$ and $f_{\mathbf{b}}(t) = F_{\mathbf{b}}(t) - F_{\mathbf{b}}(t-)$ by definition, it suffices to show that w.p.1, $\hat{F}_{\mathbf{b}}(t)$ converges to $F_{\mathbf{b}}(t)$ uniformly in t and $\mathbf{b}$.

Note that $T(\mathbf{b}) = \epsilon + (\beta - \mathbf{b})'\mathbf{X}$, $F_{\mathbf{b}}(t) = \int\int \mathbf{1}_{e+(\beta-\mathbf{b})'\mathbf{x}\le t} dF_{\epsilon,\mathbf{X}}(e,\mathbf{x})$, and $\hat{F}_{\mathbf{b}}(t) = \int\int \mathbf{1}_{(e+(\beta-\mathbf{b})'\mathbf{x}\le t)} d\hat{F}_{\epsilon,\mathbf{X}}(e,\mathbf{x})$, where $\hat{F}_{\epsilon,\mathbf{X}}(e,\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(\epsilon_i\le e, \mathbf{X}_i\le \mathbf{x})}$. Thus,

$$\hat{F}_{\mathbf{b}}(t) - F_{\mathbf{b}}(t) = \int\int \mathbf{1}_{(e+(\beta-\mathbf{b})'\mathbf{x}\le t)} (d\hat{F}_{\epsilon,\mathbf{X}}(e,\mathbf{x}) - dF_{\epsilon,\mathbf{X}}(e,\mathbf{x})).$$

Since $\hat{F}_{\epsilon,\mathbf{X}}(e,\mathbf{x})$ converges to $F_{\epsilon,\mathbf{X}}(e,\mathbf{x})$ uniformly in $(e,\mathbf{x})$ w.p.1, $\hat{F}_{\mathbf{b}}(t)$ converges to $F_{\mathbf{b}}(t)$ uniformly in t and $\mathbf{b}$ w.p.1. This completes the proof of the lemma.

Lemma 5.2. *Statement* (3.3) *holds.*

Proof. Suppose that given n observations, there are n_d ϵ_i's that belong to $\mathcal{D}$, the non-empty set of all discontinuous point of F_o. Let $d_1, d_2, \ldots$, be the elements of $\mathcal{D}$, and N_i the replications at d_i, that is, $N_i = \sum_{k=1}^{n} \mathbf{1}_{(\epsilon_k = d_i)}$. Then

$$\sum_i \frac{N_i}{n_d} \ln \frac{N_i}{n_d} \ge \sum_i \frac{N_i}{n_d} \ln f(d_i), \quad \text{where } f \ge 0 \text{ and } \textstyle\sum_i f(d_i) \le 1$$

(the Shannon-Kolmogorov inequality). Consequently,

$$\begin{aligned}
&\frac{1}{n}\ln Ł_d(\hat{F}_\beta,\beta)\\
&=(\sum_i \frac{N_i}{n}\ln\frac{N_i}{n_d})+\frac{n_d}{n}\ln\frac{n_d}{n}\\
&\geq \sum_i \frac{N_i}{n}\ln(\frac{f_o(d_i)}{p})+\frac{n_d}{n}\ln\frac{n_d}{n} \quad \text{(by (5.2.1) with } f=\tfrac{f_o}{p} \text{ and } p=P(\epsilon\in\mathcal{D}))\\
&=\sum_{j=1}^{n}\frac{1}{n}\xi_j\ln f_o(\epsilon_j)+\frac{n_d}{n}\ln\frac{n_d}{np} \quad (\text{noting } \frac{n_d}{np}\to 1)\\
&\to E(\xi\ln f_o(\epsilon))\ w.p.1 \quad \text{(by the Strong Law of Large Numbers).}
\end{aligned}$$

This yields (3.3).

Lemma 5.3. $\underline{\lim}_{n\to\infty}\frac{1}{n}\ln Ł_d(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)\geq\lim_{n\to\infty}\frac{1}{n}\ln Ł_d(\hat{F}_\beta,\beta)$ *w.p.*1.

Proof. Since $\tilde{\beta}_n$ is an GSMLE, $\ln Ł(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)\geq\ln Ł(\hat{F}_\beta,\beta)$. Thus $\ln Ł_d(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)-\ln Ł_d(\hat{F}_\beta,\beta)\geq\ln Ł_c(\hat{F}_\beta,\beta)-\ln Ł_c(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)$, as $Ł=Ł_dŁ_c$ (see (3.1)). It suffices then to show that $\underline{\lim}_{n\to\infty}\frac{1}{n}[\ln Ł_c(\hat{F}_\beta,\beta)-\ln Ł_c(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)]=0$ *w.p.*1.

By our notations, $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. copies of ϵ, $T_i(\mathbf{b})=Y_i-\mathbf{b}'\mathbf{X}_i$ and $Y_i=\beta'\mathbf{X}_i+\epsilon_i$. For ease in understanding, we first consider the case that F_o is discontinuous and $p=1$, and then consider the general case.

The solution to $T_i(b)=T_j(b)$, say b_{ij}, satisfies $b_{ij}=\beta+(\epsilon_i-\epsilon_j)/(\mathbf{X}_i-\mathbf{X}_j)$. By the independent assumption on ϵ_i's and $\mathbf{X}_i$'s, it can be verified that $\mathrm{P}\{b_{12}=t|\xi_1=0\ \&\ \xi_2=1\}=0$ for all t. Hence $\mathrm{P}\{b_{12}=b_{ij}|\xi_1=\xi_i=0\ \&\ \xi_2=1\}=0$ if $i\notin\{1,2\}$. This implies that, w.p.1 among all $T_i(b)$'s with $\xi_i=0$, either there is only one element such that $T_i(b)=T_j(b)$ with $\xi_j=1$ (and $\xi_i=0$), or there is at most one pair $(T_i(b),T_j(b))$ with $\xi_i=\xi_j=0$ such that $T_i(b)=T_j(b)$. Consequently, w.p.1,

$$Ł_c(\hat{F}_b,b)=\begin{cases}(\frac{1}{n})^{n-n\bar{\xi}} & \text{if there is no tie among } T_i(b)\text{'s with } \xi_i=0,\\ (\frac{2}{n})^2(\frac{1}{n})^{n-n\bar{\xi}-2} & \text{if there is a tie between } T_i(b)\text{'s with } \xi_i=0,\\ \frac{k}{n}(\frac{1}{n})^{n-n\bar{\xi}-1} & \text{if } T_i(b)=T_j(b) \text{ for some } i \text{ and } j \text{ such that } \xi_i\neq\xi_j,\end{cases}$$

where $k\geq 2$. It follows that, w.p.1,

$$\begin{aligned}
&\underline{\lim_{n\to\infty}}\ \frac{1}{n}|\{\ln Ł_c(\hat{F}_\beta,\beta)-\ln Ł_c(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)\}|\\
&\leq \overline{\lim_{n\to\infty}}\ \frac{1}{n}|\ln\{(\frac{n}{n})^2(\frac{1}{n})^{n-n\bar{\xi}-2}\}-\ln(\frac{1}{n})^{n-n\bar{\xi}}|\\
&=\lim_{n\to\infty}\frac{1}{n}\ln n^2=0,
\end{aligned}$$

the required conclusion.

If F_o is discontinuous and $p > 1$, the same conclusion holds since

$$\begin{aligned}&\underline{\lim}_{n\to\infty}\frac{1}{n}|\{\ln \text{Ł}_c(\hat{F}_\beta,\beta)-\ln \text{Ł}_c(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)\}|\\ &\le \overline{\lim}_{n\to\infty}\frac{1}{n}|\ln\{(\frac{n}{n})^{p+1}(\frac{1}{n})^{n-n\bar{\xi}-p-1}\}-\ln(\frac{1}{n})^{n-n\bar{\xi}}|.\end{aligned}$$

The proof is similar and is skipped.

Lemma 5.4. *If* $\mathbf{X}$ *is discrete,* $\omega\in\Omega_0\cap\{\tilde{\beta}_n\ne\beta$ *i.o.*$\}$*, and* (3.6) *holds.*

Proof. Fix an $\omega\in\Omega_0$. By assumption the range of $\mathbf{X}$, $\mathcal{R}_{\mathbf{X}}$, is discrete. Write $\mathcal{R}_{\mathbf{X}}=\{\mathbf{x}_i : i\ge 1\}$. Note $\mathcal{D}=\{d_j : i\ge 1\}$. If $\tilde{\beta}_n\ne\beta$ i.o., by taking a subsequence, we can assume $\tilde{\beta}_n\ne\beta$ for all n. By taking a further subsequence, we can assume that $\tilde{\beta}_n\to\mathbf{b}_*$, where $\mathbf{b}_*$ may not be finite. By Helly's lemma, we can further assume that $\{\hat{f}_{\tilde{\beta}_n}(d_j+(\beta-\tilde{\beta}_n)'\mathbf{x}_i)\}_{n\ge1}$ converges to a function f_* for all $(d_j,\mathbf{x}_i)$, as $\hat{f}_{\tilde{\beta}_n}$ is bounded and there at most countably many d_j's and $\mathbf{x}_i$'s. Then

$$\begin{aligned}&\overline{\lim}_{n\to\infty}\frac{1}{n}\ln \text{Ł}_n(\hat{F}_{\tilde{\beta}_n},\tilde{\beta}_n)\\ &=\overline{\lim}_{n\to\infty}\sum_{i,j}\frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{(\epsilon_k=d_j,\mathbf{X}_k=\mathbf{x}_i)}\ln\hat{f}_{\tilde{\beta}_n}(d_j+(\beta-\tilde{\beta}_n)'\mathbf{x}_i)\\ &\le E(\xi\ln f_*(\epsilon+(\beta-\mathbf{b}_*)'\mathbf{X}))\quad\text{(by Fatou's lemma).}\end{aligned}\tag{5.4.1}$$

Note that (3.6) follows from (5.4.1) and

$$E(\xi\ln f_*(\epsilon+(\beta-\mathbf{b}_*)'\mathbf{X}))<E(\xi\ln f_o(\epsilon)).\tag{5.4.2}$$

We now prove (5.4.2). Hereafter, let $\mathcal{C}$ be the collection of all real-valued functions f such that $f\ge0$ and $\sum_j f(d_j)\le\sum_j f_o(d_j)$. Note that $E(\xi\ln f(\epsilon))=\sum_j f_o(d_j)\ln f(d_j)$ for each $f\in\mathcal{C}$. By the Shannon-Kolmogorov inequality, f_o uniquely maximizes $\sum_j f_o(d_j)\ln f(d_j)+p\ln p$ over all $f\in\mathcal{C}$, where $p=P\{\epsilon\notin\mathcal{D}\}$. Thus if $f\in\mathcal{C}$, $E(\xi\ln f(\epsilon))<E(\xi\ln f_o(\epsilon))$ unless $f(d_j)=f_o(d_j)$ for all j. Letting $f(d_j)=f_*(d_j+(\beta-\mathbf{b}_*)'\mathbf{x})$,

$$E(\xi\ln f_*(\epsilon+(\beta-\mathbf{b}_*)'\mathbf{x})|\mathbf{x})<E(\xi\ln f_o(\epsilon))\ \text{ if } \sup_j|f_*(d_j+(\beta-\mathbf{b}_*)'\mathbf{x})-f_o(d_j)|>0.\tag{5.4.3}$$

We prove in Lemma 5.6 that

$$\underline{\lim}_{n\to\infty}\inf_{\mathbf{b}\ne\beta,\mathbf{x}_i\ne0}\sup_j|\hat{f}_{\mathbf{b}}(d_j+(\beta-\mathbf{b})'\mathbf{x}_i)-f_o(d_j)|>0.\tag{5.4.4}$$

Since f_* is the limit of $\hat{f}_{\tilde{\beta}_n}$, where $\tilde{\beta}_n \neq \beta$, (5.4.4) implies that if $\mathbf{x} \neq 0$, then

$$\begin{aligned}&\sup_j |f_*(d_j + (\beta - \mathbf{b}_*)'\mathbf{x}) - f_o(d_j)| \\ &\geq \varliminf_{n\to\infty} \inf_{\mathbf{b}\neq\beta, \mathbf{x}_i \neq 0} \sup_j |\hat{F}_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) - f_o(d_j)| > 0.\end{aligned}$$

Thus, (5.4.3) yields $E(\xi \ln f_*(\epsilon + (\beta - \mathbf{b}_*)'\mathbf{X})|\mathbf{X}) < E(\xi \ln f_o(\epsilon))$. Taking expectation on both sides of the above inequality yields (5.4.2). This concludes the proof of the lemma.

Lemma 5.5. *If* $\mathbf{X}$ *is not discrete and* $P\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} > 0$, *then* (3.7) *holds.*

Proof. Write $\mathbf{X} = (X_1, \ldots, X_p)'$, $\beta = (\beta_1, \ldots, \beta_p)'$ and $\tilde{\beta}_n = (\tilde{\beta}_{n,1}, \ldots, \tilde{\beta}_{n,p})'$. By assumption, $\mathbf{X}$ is not discrete so there is a non-discrete coordinate X_k, $k \in \{1, \ldots, p\}$. There are two possibilities: (1) $\tilde{\beta}_{n,k} \neq \beta_k$ i.o.; (2) $\tilde{\beta}_{n,k} = \beta_k$ for all n.

If (2) holds for each non-discrete X_k, then it follows that $\tilde{\beta}_{n,h} \neq \beta_h$ i.o. implies X_h is discrete. By rearranging indices, we can assume that (a) X_h is discrete if $h = 1, \ldots, q$; (b) X_h is non-discrete and $\tilde{\beta}_{n,h} = \beta_h \ \forall\ n$ if $h > q$. It is conceivable that the proof of the lemma in such case is similar to the proof in the case that $\mathbf{X}$ is discrete (i.e., Lemma 5.4). We skip the details for the sake of simplicity.

Now assume that (1) holds. Thus X_k is not discrete and $\tilde{\beta}_{n,k} \neq \beta_k$ i.o.. Let $\mathcal{Q}_{\mathbf{b}}$ the be event $\{T(\mathbf{b}) \notin \mathcal{D}_{\mathbf{b}} : T(\mathbf{b}) = \epsilon + (\beta - \mathbf{b})'\mathbf{X}, \epsilon \in \mathcal{D}\}$, where $\mathcal{D}_{\mathbf{b}}$ is the collection of all discontinuity points of $F_{\mathbf{b}}$ Since $\mathbf{X}$ is not discrete, $\mathrm{P}\{\mathcal{Q}_{\mathbf{b}}\} \geq p_1$, where $p_1 = P(X_k \notin D_{X_k})$ and D_{X_k} is the set of discontinuity points of the cdf F_{X_k}. Note that p_1 does not depend on $\mathbf{b}$. Hereafter, we assume that $\mathbf{b}$ satisfies $b_k \neq \beta_k$, where $\mathbf{b} = (b_1, \ldots, b_p)'$. Since $\mathcal{Q}_{\mathbf{b}}$ is the set of continuous points of $F_{\mathbf{b}}$, $\mathrm{P}(T(\mathbf{b}) = t) = 0$ for each $t \in \mathcal{Q}_{\mathbf{b}}$. Thus, except on an event of zero probability, if $t \in \mathcal{Q}_{\mathbf{b}}$ there is no tie at t among $T_1(\mathbf{b}), \ldots, T_n(\mathbf{b})$. Without loss of generality, take $\sum_{i=1}^n \mathbf{1}_{(T_i(\mathbf{b})=t)} \leq 1$ if $t \in \mathcal{Q}_{\mathbf{b}}$. Since $\hat{f}_{\mathbf{b}}$ is the empirical pdf of $F_{\mathbf{b}}$, $\hat{f}_{\mathbf{b}}(T_i(\mathbf{b})) = \frac{1}{n}\sum_{j=1}^n \mathbf{1}_{(T_j(\mathbf{b})=T_i(\mathbf{b}))} = \frac{1}{n}$ if $T_i(\mathbf{b}) \in \mathcal{Q}_{\mathbf{b}}$. Thus, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(T_i(\mathbf{b})\in\mathcal{Q}_{\mathbf{b}})} \ln \hat{f}_{\mathbf{b}}(T_i(\mathbf{b})) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(T_i(\mathbf{b})\in\mathcal{Q}_{\mathbf{b}})} \ln \frac{1}{n}$. By assumption, $\mathrm{P}\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} > 0$. Along a subsequence, take $\tilde{\beta}_{n,k} \neq \beta_k$ for all n. It follows that

$$\begin{aligned}&\varliminf_{j\to\infty} \inf_{n\geq j, \tilde{\beta}_{n,k}\neq\beta_k} \frac{1}{n} \ln Ł_d(\hat{F}_{\tilde{\beta}_n}, \tilde{\beta}_n) \\ &= \varliminf_{j\to\infty} \inf_{n\geq j, \tilde{\beta}_{n,k}\neq\beta_k} \frac{1}{n}\Big\{ \sum_{i=1}^n \mathbf{1}_{(T_i(\tilde{\beta}_n)\in\mathcal{Q}_{\tilde{\beta}_n})} \ln \frac{1}{n} + \sum_{i=1}^n \xi_i \mathbf{1}_{(T_i(\tilde{\beta}_n)\notin\mathcal{Q}_{\tilde{\beta}_n})} \ln \hat{f}_{\tilde{\beta}_n}(T_i(\tilde{\beta}_n))\Big\} \\ &\leq \varlimsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{(T_i(\tilde{\beta}_n)\in\mathcal{Q}_{\tilde{\beta}_n})} \ln \frac{1}{n} \quad (\text{as } \ln \hat{f}_{\tilde{\beta}_n} \leq 0)\end{aligned}$$

$$= p_1 \lim_{n\to\infty} \ln \frac{1}{n} \quad (\text{as } \tilde{\beta}_{n,k} \neq \beta_k \text{ for all } n \text{ and } \mathrm{P}\{Q_{\mathbf{b}}\} \geq p_1)$$
$$= -\infty \quad \text{with probability } \mathrm{P}\{\tilde{\beta}_n \neq \beta \text{ i.o.}\} > 0.$$

Consequently, (3.7) is trivially true, as $E(\xi \ln f_o(\epsilon))$ is finite by A4. This completes the proof of the lemma.

Lemma 5.6. *Suppose* $\mathbf{X}$ *is discrete. Then* (5.4.4) *holds.*

Proof. Fix an $\omega \in \Omega_0$. In order to prove (5.4.4), it suffices to show that there exists an n_1 such that

$$\inf_{\mathbf{b}\neq\beta,\mathbf{x}_i\neq 0} \sup_j |\hat{F}_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) - f_o(d_j)| > 0 \quad \text{if } n \geq n_1. \tag{5.6.1}$$

To this end, we prove later that

$$t_0 > 0, \quad \text{where } t_0 = \inf_{\mathbf{b}\neq\beta,\mathbf{x}_i\neq 0} \sup_j |f_o(d_j) - f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i)|. \tag{5.6.2}$$

Since $\omega \in \Omega_0$, $\hat{f}_{\mathbf{b}}$ converges uniformly in $\mathbf{b}$ and t. Consequently, there exists n_1 such that $\sup_{j,\mathbf{b}\neq\beta,\mathbf{x}_i\neq 0} |\hat{f}_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) - f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i)| \leq t_0/2$, if $n \geq n_1$. Thus, it follows from (5.6.2) that for each $\mathbf{b}_0 \neq \beta$ and $\mathbf{x} \neq 0$,

$$\begin{aligned}
&\sup_j |\hat{f}_{\mathbf{b}_0}(d_j + (\beta - \mathbf{b}_0)'\mathbf{x}) - f_o(d_j)| \\
&\geq \inf_{\mathbf{b}\neq\beta,\mathbf{x}_i\neq 0} \sup_j |f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) - f_o(d_j)| \\
&\quad - \sup_{j,\mathbf{b}\neq\beta,\mathbf{x}_i\neq 0} |\hat{f}_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) - f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i)| \\
&\geq t_0/2, \quad \text{if } n \geq n_1.
\end{aligned}$$

Then (5.6.1) holds. The proof of the lemma will be completed after we prove (5.6.2).

By A3, there exists $\mathbf{x}_i \neq 0$ such that $f_{\mathbf{X}}(\mathbf{x}_i) > 0$. Suppose that there are exactly j_1 points in $\mathcal{D}$ at which f_o achieves the maximum. By rearranging indices, we can assume that $f_o(d_1) = \cdots = f_o(d_{j_1}) > f_o(d_{j_1+1}) \geq f_o(d_i)$ for each $i > j_1$, where $d_1 > \cdots > d_{j_1}$. Note that j_1 is finite as $\sum_i f_o(d_i) \leq 1$ and $f_o \geq 0$. Without loss of generality take $j_1 = 1$. Suppose $\mathbf{b} \neq \beta$. Since $f_{\mathbf{b}}(t) = \sum_{(\mathbf{x}_k,d_h):d_h+(\beta-\mathbf{b})'\mathbf{x}_k=t} f_o(d_h) f_{\mathbf{X}}(\mathbf{x}_k)$ if $t \in \mathcal{D}_{\mathbf{b}}$,

$$\begin{aligned}
f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i) &= \sum_{(\mathbf{x}_k,d_h):d_h+(\beta-\mathbf{b})'\mathbf{x}_h=d_j+(\beta-\mathbf{b})'\mathbf{x}_i} f_o(d_h) f_{\mathbf{X}}(\mathbf{x}_k) \\
&= \sum_{(\mathbf{x}_k,d_h):d_h-d_j=(\beta-\mathbf{b})'(\mathbf{x}_h-\mathbf{x}_i)} f_o(d_h) f_{\mathbf{X}}(\mathbf{x}_k).
\end{aligned}$$

In view of the last expression of $f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i)$,

$$\sup_j |f_o(d_j) - f_{\mathbf{b}}(d_j + (\beta - \mathbf{b})'\mathbf{x}_i)| \geq f_o(d_1) - \sum_{(\mathbf{x}_k, d_h): d_h - d_1 = (\beta - \mathbf{b})'(\mathbf{x}_h - \mathbf{x}_i)} f_o(d_h) f_{\mathbf{X}}(\mathbf{x}_k)$$
$$\geq f_o(d_1) - f_o(d_2) > 0, \quad \text{as } f(d_1) > f(d_2) \geq \cdots.$$

Thus (5.6.2) holds.

Acknowledgement

The authors thank the editors, an associate editor and two referees for their invaluable comments and suggestions. This research was partially supported by Army Grant DAMD17-99-1-9390 and DAMD17-00-1-0448.

References

Bickel, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10**, 647-671.

Chatterjee S. and Price B. (1991). *Regression Analysis by Example*. 2nd edition. John Wiley, New York.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. John Wiley, New York.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73-101.

Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887-906.

Montegomery, D. C. and Peck, E. A. (1992). *Introduction to Linear Regression Analysis*. John Wiley, New York.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18**, 303-328.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *J. Amer. Statist. Assoc.* **63**, 1379-1389.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis, *I. Proc. Kon. Ned. Akad. v. Wetensch. A* **53**, 386-392.

Zhang, C. H. and Li, X. (1996). Linear regression with doubly censored data. *Ann. Statist.* **24**, 2720-2743.

Department of Mathematical Sciences, SUNY, P.O. Box 6000, Binghamton, NY 13902-6000, U.S.A.

E-mail: qyu@math.binghamton.edu

Strang Cancer Prevention Center, 428 E 72nd Street, NY 10021, U.S.A.

E-mail: gwong@strang.org

(Received August 2001; accepted August 2002)