Regression Models with Interval Censoring

Jian Huang and Jon A. Wellner [1]

University of Washington

October 6, 1993

Abstract In this paper we discuss estimation in semiparametric regression models with interval censoring, with emphasis on estimation of the regression parameter θ. The first section surveys some of the existing literature concerning these models and the various existing approaches to estimation, including a selected subset of the enormous literature on the binary choice model used in econometrics. In section 2 we calculate efficient score functions and information bounds for the regression parameter θ in the linear regression model with interval censoring, the binary choice model, and the Cox model with interval censoring. Profile likelihood approaches to maximum likelihood estimates are discussed and reviewed in section 3, and profile likelihod is used in section 4 to discuss maximum likelihood estimation for the linear regression model with interval censoring, and some of the remaining problems.

[1] Research supported in part by National Science Foundation grant DMS-9108409 and by NIAID grant 2R01 AI291968-04

AMS 1991 *subject classifications.* Primary: 60F17, 62E20; secondary 60B12.

Key words and phrases. Accelerated failure time, asymptotic distribution, binary choice model, consistency, Cox model, efficiency, greatest convex minorant, information bound, linear regression model, maximum likelihood, monotone, nonparametric maximum likelihood, profile likelihood

1. Regression Models with Interval Censoring. In this paper, we consider estimation in two important regression models, the linear regression model and the Cox's proportional hazard model, with interval censored data. These two models will be studied in a semiparametric context: the error distribution of the linear regression model, and the baseline hazard function of the proportional hazard model, will be assumed completely unknown (subject, perhaps, to regularity conditions).

Regression analysis concerns the relationship between two random vectors $T \in \mathbb{R}^k$ and $Z \in \mathbb{R}^d$, where T is a vector of response variables and Z is a vector of explanatory variables. We will only consider a one dimensional response variable, $T \in \mathbb{R}$. In many applications, for example, in biomedical studies or in reliability studies, the response variable T is usually not fully observable. Instead, there is an additional random variable $Y \in \mathbb{R}$, and only some kind of functional of T and Y are observable. This additional random variable Y will be called a *censoring variable*.

Suppose that the observable random vector is:

$$X = (\delta, Y, Z) \equiv (1_{\{T \le Y\}}, Y, Z), \tag{1.1}$$

where 1_S is the indicator function of the event S, that is, 1_S is equal to 1 if S is true, otherwise it equals to 0. So the only information about T is from the censoring variable Y, and whether T is greater or less than Y. This type of observation is called "case 1" interval censored data, or current status data.

Interval censored data arises naturally in some applications. For example, it arises in animal tumorigenicity experiments, see, e.g., Finkelstein and Wolfe (1985), and Finkelstein (1986). The goal of such studies is to analyze the effect of a suspected carcinogen on the time to tumor onset. However, the onset time cannot be observed. Rather, animals die or are sacrificed at certain times, and are examined for the presence or absence of a tumor. If the tumor is irreversible and nonlethal, the observed times of death or sacrifice yield interval censored observations.

Closely related models are of interest in studies of AIDS; see for example, Shiboski and Jewell (1992).

Notice the difference between interval censoring and the usual right censoring. In a right censorship model, the observed data is $(\min(T, Y), 1_{\{T \le Y\}}, Z)$. There is probability $P\{T \le Y\}$ of observing the survival time exactly. But with interval censoring, we are not able to observe the exact value of the survival time at all, just $1_{\{T \le Y\}}$. It is therefore expected that statistical inference with interval censored data is more difficult.

There has been a tremendous amount of research on the proportional hazards model under right censoring in the last twenty years since Cox's (1972) milestone work; see e.g. Andersen and Gill (1982), and the recent books by Fleming and Harrington (1991) and Andersen, Borgan, Keiding, and Gill (1993). However, we are not aware of any systematic treatment of the proportional hazards model under interval censoring.

To set up the semiparametric regression model under interval censoring, suppose (T, Y, Z) is distributed as $Q_{\theta,F}$, where $Q_{\theta,F}$ is a probability measure indexed by a finite dimensional parameter $\theta \in \Theta \subset \mathbb{R}^d$ and an infinite dimensional parameter F belonging to some class of functions $\mathcal{F}$. θ is usually called the regression parameter, which measures the influence of the explanatory variable Z on the response variable T. It is the parameter of primary interest. In many situations, F is a distribution function or a transformation of a distribution function, such as a cumulative hazard function. It is the secondary or "nuisance" parameter. Let $\mathcal{Q}$ be the collection of all the $Q_{\theta,F}$'s, that is,

$$\mathcal{Q} = \{Q_{\theta,F} : \theta \in \Theta, \quad F \in \mathcal{F}\}. \tag{1.2}$$

The collection $\mathcal{Q}$ is called a *model*.

Let $P_{\theta,F}$ be the induced law of $X \equiv (\delta, Y, Z)$ on $\{0,1\} \times \mathbb{R} \times \mathbb{R}^d$, i.e., for any measurable subsets $A \in \mathbb{R}$, and $B \in \mathbb{R}^d$,

$$P_{\theta,F}(\delta = 1, Y \in A, Z \in B) = \int_{y \in A, z \in B} 1_{\{t \le y\}} dQ_{\theta,F}(t, y, z). \tag{1.3}$$

We let

$$\mathcal{P} = \{P_{\theta,F} \; : \; \theta \in \Theta, \; F \in \mathcal{F}\} \tag{1.4}$$

denote the induced model for the observation (of a sample of size one).

1.1. The linear regression model under interval censoring In a linear regression model, T equals the sum of a linear combination of the covariate vector Z and a random error term,

$$T = Z'\theta + \varepsilon, \tag{1.5}$$

where $\theta \in \mathbb{R}^d$, and ε has an unspecified distribution function F with density f. Then the underlying model is

$$\mathcal{Q} = \{Q_{\theta,F} : \ \theta \in \mathbb{R}^d, F \text{ is a distribution function}\}$$

is determined by the (1.5). There is tremendous amount of literature on the estimation of this model when (T, Z) is fully observable. In survival analysis, T is usually taken to be a log failure time, and the model is called the "accelerated failure time" model. Estimation of θ under right censoring has been considered by Buckley and James (1979), Ritov (1990), Tsiatis (1990), and Ying (1993), among many other authors.

Under interval censoring, we only observe X defined in (1.1), i.e. , $X = (\delta, Y, Z)$. Consequently, the linear regression model under interval censoring is $\mathcal{P}$ as obtained from $\mathcal{Q}$ via (1.3) and (1.4).

Little has been done for the estimation of a linear regression model under interval censoring. One exception is Rabinowitz, Tsiatis, and Aragon (1993). Their approach is based on score statistics motivated as follows: if F and $E(Z|Y - Z'\theta)$ were known, then the estimated conditional covariance between δ and Z is

$$\frac{1}{n}\sum_{i=1}^{n}[\delta_i - F(Y_i - Z_i'\theta)][Z_i - E(Z|Y_i - Z_i'\theta)],$$

which has mean zero when $\theta = \theta_0$, where θ_0 is the "true" regression parameter. Hence it could be used as a score function for hypothesis testing and estimation. For the unknown F and $E(Z|Y_i - Z_i'\theta)$, they propose to estimate them via appropriately partitioning and binning of the data. They proved that the estimated score statistics are asymptotically normal at $\theta = \theta_0$. Furthermore, they proved that if an estimator defined by the estimated score is $\sqrt{n}$-consistent, then it is asymptotically normal.

However, in applying their approach, one has to subjectively choose the partitioning parameter (similar to the bandwidth in density estimation) in the first place. Moreover, it appears that these authors did not give justification for their estimators being $\sqrt{n}$-consistent.

1.2. The binary choice model Suppose that, in the linear regression model under interval censoring, the censoring variable Y is degenerate; i.e., $P\{Y = 0\} = 1$. Then the observed variable is $X = (\delta = 1_{\{T\le 0\}}, Z)$. In this case, the "case I" interval censoring regression model reduces to what is known as the *binary choice model* in the econometrics literature. The binary choice model has been studied by many authors. For example, see Manski (1975, 1985), Cosslett (1983), Han (1987), Horowitz (1992), Klein and Spady (1993), and Sherman (1993).

Although the structure of the linear regression model and the binary choice model is completely similar, there is a fundamental difference between these two models. In the binary choice model, the parameter θ is not identifiable (and hence not consistently estimable) unless there is some constraint on the parameter space, i.e., the length of θ is known or a component of θ is known. No such restrictions are needed when Y is not degenerate.

Manski (1975, 1985) proposed and studied the "maximum score estimator" defined by maximization of the object function

$$\sum_{i=1}^{n}[1_{\{T_i>0, Z_i'\theta>0\}} + 1_{\{T_i\le 0, Z_i'\theta\le 0\}}].$$

This is equivalent to the minimization of

$$\sum_{i=1}^{n} |\mathrm{sgn}(T_i) - \mathrm{sgn}(Z_i'\theta|| = \sum_{i=1}^{n} |(2\delta_i - 1) - \mathrm{sgn}(Z_i'\theta)| \,. \tag{1.6}$$

Kim and Pollard (1990) obtained the limiting distribution for the maximum score estimator under some conditions. Interestingly, and somewhat disappointingly, the convergence rate of the maximum score estimator is $n^{1/3}$ instead of the usual rate $\sqrt{n}$.

Cosslett (1983) proposed "distribution free maximum likelihood" estimators for the regression parameter θ and the error distribution F. Assuming that the error ε is independent of Z, it can be deduced that the log-likelihood function is, up to an additive constant,

$$l_n(\theta, F) = \sum_{i=1}^{n} \{\delta_i \log F(-Z_i'\theta) + (1-\delta_i)\log(1 - F(-Z_i'\theta)).$$

Then Cosslett's "distribution free maximum likelihood" estimator of θ is obtained as follows: for any fixed θ, let $F_n(\cdot;\theta)$ be the maximizer of $l_n(\theta, F)$ with respect to F, then the estimator of θ is any $\hat{\theta}_n$ that maximizes $l_n(\theta; F_n(\cdot;\theta))$. He proved that his estimators are consistent. But the asymptotic distribution of $\widehat{\theta}_n$ is still an open problem. In most of the statistics literature, Cosslett's "distribution free maximum likelihood" estimator would usually be called a *maximum profile likelihood* estimator; see e.g. Andersen, Borgan, Keiding, and Gill (1993), page 482, or Severini and Wong (1992). Maximum profile likelihood estimation is frequently a useful approach in dealing with semiparametric models. For example, for the proportional hazards model under right censoring it yields Cox's (1972) partial likelihood estimator. We will use this terminology in the following.

Cosslett (1987) calculated the information for θ. He showed that when the covariate Z and the random error ε are independent, then the information for θ (properly constrained) is positive. This suggests that it is possible to construct an estimator of θ with the $\sqrt{n}$-convergence rate. Actually, we conjecture that the maximum profile likelihood estimator for the binary choice model is asymptotically efficient. However, when Z and ε are correlated, then the information for θ is zero, see, e.g., Chamberlain (1986). The implication of this is that it is impossible to construct $\sqrt{n}-$consistent estimator in this situation.

Han (1987), in studying a more general regression model with the binary choice model as a special case, proposed the maximum rank correlation estimator defined as the θ that maximize the following U-processes of order two:

$$\psi_n^H(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \delta_i (1-\delta_j) 1_{\{Z_i'\theta < Z_j'\theta\}}. \tag{1.7}$$

Under appropriate conditions, he also proved that his estimator is consistent. With the independence assumption of Z and ε, Sherman (1993) proved that Han's maximum rank correlation estimator is $\sqrt{n}$-consistent and satisfies a central limit theorem, but not efficient.

Yet another estimator based on rank consideration was introduced by C. Cavanagh, see Sherman (1991). This estimator maximizes the following U-processes of order three:

$$\frac{1}{n \cdot (n-1) \cdot (n-2)} \sum_{(i,j,k)} \delta_i (1-\delta_j) 1_{\{Z_i'\theta < Z_k'\theta\}}.$$

Here (i,j,k) ranges over the odered triples of distinct integers from the set $\{1, 2, \ldots, n\}$. Sherman (1991) provided sufficient conditions for Cavanagh's estimator to be consistent and asymptotically normal.

To improve the convergence rate of Manski's maximum score estimator, Horowitz (1992) proposed the smoothed maximum score estimator, i.e., replace $\mathrm{sgn}(Z'\theta)$ in (1.6) by a smooth function $K(Z'\theta/\sigma_n)$, where

K is a twice differentiable function satisfies some other conditions, it is analogous to a cumulative distribution function; and where σ_n is an appropriately chosen bandwidth, depending on the sample size n. He showed that

$$n^{\alpha}(\widetilde{\theta} - \theta_0) \to_d N(b, V),$$

where $2/5 \leq \alpha < 1/2$ depending on smoothness of the distributions of ε and $Z'\theta_0$, and where b is the asymptotic bias which can be estimated consistently. The advantage of Horowitz's (1992) (also Manski's) estimator is that it does not assume the independence between ε and the covariate Z. Hence it is applicable when ε and Z are dependent.

Klein and Spady (1993) considered the estimator by maximizing

$$\sum_{i=1} n(\hat{\tau}_i/2\{\delta_i \log \widehat{F}(-Z_i'\theta)^2 + (1-\delta_i)\log(1-\widehat{F}(-Z_i'\theta))^2\},$$

where $\widehat{F}(Z'\theta)$ is a kernel estimator of the probability $P(\delta = 1|Z'\theta)$, and $\hat{\tau}_i, i = 1, \ldots, n$ are some trimming factors depending on the data and some preliminary consistent estimator for θ_0. They proved that, under the independence assumption of Z and ε and other regularity conditions, their estimator is asymptotically efficient.

However, in either Horowitz (1992) or Klein and Spady (1993), the estimation procedures involve choices of a kernel function K and a bandwidth. So there is a question of which kernel to choose and how to decide the bandwidth in the first place. This is a rather delicate matter in practice. It is well known that different kernels and bandwidths may give quite different estimators for the underlying density; see, e.g., Silverman (1986), chapter 2, pages 15-19; and chapter 3. Hence the sensitivity of the estimators with respect to the choices of kernel and the bandwidth with moderate to moderately large sample size need to be carefully studied.

Since the binary choice model and the linear regression model under interval censoring have very similar structure, the methods developed for estimation of the binary choice model can be applied to the regression model under interval censoring. For the case of extending Manski's estimator or Han's estimator to the regression model under interval censoring, see Huang (1993b).

1.3. The Cox model under interval censoring The proportional hazards model, or the Cox model (Cox, 1972), is probably the most widely used model in the analysis of failure time data. It has been extensively studied for the case of right censored data; see Andersen and Gill (1982), Fleming and Harrington (1991), and Andersen, Borgan, Gill, and Keiding (1993). However, there has been little rigorous theoretical study of the Cox model when the failure time is interval censored.

The Cox proportional hazard model assumes the hazard function $\lambda(t|z)$ of T given Z is related to the baseline hazard function $\lambda(t)$ in the following way.

$$\lambda(t|z) = e^{\theta z}\lambda(t).$$

Under this specification, we have the following relationships for the cumulative hazard functions and the survival functions.

$$\Lambda(t|z) = e^{\theta z}\Lambda(t),$$

$$\overline{F}(t \mid z) = \overline{F}(t)^{\exp(\theta z)}.$$

The semiparametric statistical model determined by one of the above relationships is

$$\mathcal{Q} = \left\{Q_{\theta,\Lambda} : \ \theta \in \mathbb{R}^d, \Lambda \text{ is a cumulative hazard function}\right\}.$$

With interval censoring of the Cox model, we do not observe $(T, Z) \sim Q_{\theta,\Lambda} \in \mathcal{Q}$, but instead we observe only $X = (\delta, Y, Z) = (1_{\{T \le Y\}}, Y, Z) \sim P_{\theta,\Lambda} \in \mathcal{P}$. Here we will suppose that (Y, Z) have a joint distribution (possibly unknown), but that the conditional distribution of T given (Y, Z) depends only on Z as specified above. The goal, as before, is to estimate the regression parameter θ and the cumulative hazard function Λ based on a sample of interval censored data $X_1, \ldots, X_n$.

Finkelstein and Wolfe (1985) proposed a semiparametric model which, instead of parametrizing the conditional probability density of the failure time T given the covariate Z in standard way, assumes a parametric form of the conditional probability of Z given T. They suggested using EM algorithm and the Newton-Raphson algorithm to obtain the estimates of the regression parameters and the baseline survival function. Finkelstein (1986) considered the Cox model with interval censoring. She proposed to use Newton-Raphson iteration to compute the maximum likelihood estimates of the regression parameter and the baseline cumulative hazard function. However, in both papers, the authors did not consider the properties of the computational procedure and the asymptotic properties of the estimators.

In the case of right censoring, there is a nice counting process martingale associated with the Cox model, and estimation of the regression parameter θ via partial likelihood is much easier than in the accelerated failure time model. However, with interval censored data, there are no obvious martingales associated with the Cox model, nor does partial likelihood work as simply. Moreover, the approaches developed in the binary choice model are not applicable here, because of the very different structure the model. Until very recently, there has been virtually no theory available for the Cox model with interval censoring. Huang (1993c) has succeeded in proving that the maximum likelihood estimator of the regression parameter θ is, in fact, consistent, asymptotically normal, and efficient. For further brief comments, see section 4.

1.4. Other types of interval censoring Besides the "case 1" interval censored data (or current status data) mentioned above, there are other types of interval censoring.

"Case 2" interval censoring. For "case 2" interval censored data, we only know that T has occurred either within some random time interval, or before the left end point of the time interval, or after the right end point of the time interval. More precisely, there are two censoring variables U and V, the observed is:

$$(\delta, \gamma, U, V, Z) = (1_{\{T \le U\}}, 1_{\{U < T \le V\}}, U, V, Z).$$

Finkelstein (1986) used Newton-Raphson algorithm to compute the maximum likelihood estimators of the regression parameter and the cumulative hazard function in the Cox model under this censoring scheme. But virtually no theoretical results are available for regression models under "case 2" interval censoring. However, in nonparametric setting, i.e. , when the observed i.i.d. samples are $(\delta_i, \gamma_i, U_i, V_i), i = 1, \ldots, n$, Groeneboom and Wellner (1992) considered nonparametric maximum likelihood estimator (NPMLE) of the distribution function F of T. They showed that the NPMLE of F is characterized by the greatest convex minorant of a self-induced cumulative sum diagram, and they proved that the NPMLE is consistent. But the distribution theory of the NPMLE has not been completely resolved.

A general interval censoring scheme. In some clinical trials, a patient may go through a sequence of examinations. Let

$$Y_i = (Y_{i,1}, Y_{i,2}, \cdots, Y_{i,n_i})$$

denote the ith patient's ordered examination times. $Y'_{i,j}$s may be monotonically transformed so that it may range from $-\infty$ to ∞. The failure time T_i (or transformed in the same manner as $Y'_{i,j}$s) of the ith patient is only known to be bracketed between a pair of examination times $(Y_{i,L}, Y_{i,U})$, where $Y_{i,L}$ is the last examination time preceding T_i, and $Y_{i,U}$ is the first examination time following T_i. Adopt the convention that $Y_{i,0} = -\infty$, and $Y_{i,n_i+1} = \infty$, we will have $L = j_i, U = j_i + 1$ for some $0 \le j_i \le n_i$. Let Z_i be the covariate vector of the ith patient. Then the effective observations are

$$(Y_{i,L}, Y_{i,U}, Z_i), \quad , i = 1, \ldots, n.$$

Rabinowitz, Tsiatis, and Aragon (1993b) considered analysis of a linear regression model under this interval censoring scheme. They proposed a class of score statistics that can be used in estimation and confidence procedure. Their approach is similar to Rabinowitz, Tsiatis, and Aragon (1993). They also provided some heuristics justifying their approach.

2. Efficient Scores and Information Bounds. In this section, we calculate efficient score functions and information bounds for the regression parameter θ in the three models discussed in section 1.

As illustrated in Bickel, Klaassen, Ritov, and Wellner (1993), it is of considerable importance to calculate the information bound in any semiparametric model. The information calculation serves four purposes: (a) to see if it is possible at all to obtain $\sqrt{n}-$consistent estimators; (b) to see how difficult it is to estimate the parameter; (c) to suggest estimation procedures that would lead to regular estimation in the sense that the estimators have $\sqrt{n}$-convergence rate. (d) to suggest estimation procedures that would lead to efficient estimation, i.e., estimators with asymptotic variance equal to the lower bound based on the information calculation. In the regression model under interval censoring, we expect that there is severe loss of information because we can not measure the response variable T exactly at all, and in fact, it is not clear in the first place whether regular estimators for the regression parameter can be obtained. Hence the information calculation seems to be particularly imperative.

Here are the calculations for the linear regression model (1.5) with interval censoring. First we calculate the density of the data (for a sample of size one): by the assumption that (Y, Z) is independent of $\epsilon = T - Z'\theta$, $P\{\delta = 1|Y, Z\} = F(Y - Z'\theta)$. Hence the density of (δ, Y, Z) with respect to the product of counting measure on $\{0, 1\}$ and Lebesgue measure on $\mathbb{R}^2$ is

$$p_{\theta,F}(\delta, y, z) = [F(y - z'\theta)]^{\delta}[1 - F(y - z'\theta)]^{1-\delta} g(z, y), \qquad \text{for} \qquad \delta \in \{0, 1\}, (y, z) \in \mathbb{R}^2.$$

Since $g(z, y)$ does not involve F or θ, it can be treated as known and will not be considered in the following. The log-likelihood function is, up to an additive constant,

$$l(\theta, F) = \log p_{\theta,F}(\delta, Y, Z) = \delta \log F(Y - Z'\theta) + (1 - \delta) \log(1 - F(Y - Z'\theta)).$$

Then the information bound for θ can be calculated from the efficient score of θ derived from the log-likelihood function. The following is the result concerning the efficient score and the information bound for θ in model (1.5).

THEOREM 2.1. *(Linear regression with interval censoring). Let* $h(s) = E(Z|Y - Z'\theta = s)$. *Suppose that* $\lim_{s\to\infty} f(s) = 0$ *and* $h(s)$ *is bounded. Then the efficient score for* θ *is*

$$l_\theta^*(x) = f(y - z'\theta)[z - E(Z|Y - Z'\theta = y - z'\theta)] \left[\frac{1-\delta}{1 - F(y - z'\theta)} - \frac{\delta}{F(y - z'\theta)}\right]. \tag{2.8}$$

Moreover, the information for θ *is* $I(\theta)$, *where*

$$I(\theta) = E\left\{\frac{f(Y - Z'\theta)^2}{F(Y - Z'\theta)(1 - F(Y - Z'\theta))}[Z - E(Z|Y - Z'\theta)]^{\otimes 2}\right\}. \tag{2.9}$$

To see how much information is lost under interval censoring, we calculated $I(\theta)$, and $I_{uc}(\theta)$, the information without censoring, for a simple normal model where ϵ and Z are independent one-dimensional normal random variables; $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $Z \sim N(0, \sigma_Z^2)$. Furthermore, assume that the censoring variable

Y is independent of Z and ϵ (and hence also independent of (T, Z)) and $Y \sim N(0, \sigma_Y^2)$. With this model, it is straightforward to calculate that

$$E(Z|Y - Z\theta) = -\frac{\theta\sigma_Z^2}{\sigma_Y^2 + \theta^2\sigma_Z^2}(Y - \theta Z)\,,$$

and

$$Z - E(Z|Y - Z\theta) = \frac{\sigma_Y^2}{\sigma_Y^2 + \theta^2\sigma_Z^2}Z - \frac{\theta\sigma_Z^2}{\sigma_Y^2 + \theta^2\sigma_Z^2}Y\,.$$

Furthermore, note that $Z - E(Z|Y - Z\theta)$ and $Y - \theta Z$ are independent. Hence the information for θ given in (2.9) splits into the product of two terms:

$$\begin{aligned} I(\theta) &= E\frac{f^2}{F\overline{F}}(Y - \theta Z) \cdot E(Z - E(Z|Y - \theta Z))^2 \\ &= E\frac{f^2}{F\overline{F}}(Y - \theta Z) \cdot \frac{\sigma_Y^2\sigma_Z^2}{\sigma_Y^2 + \theta^2\sigma_Z^2} \end{aligned}$$

where f is the $N(0, \sigma_\epsilon^2)$ density and $Y - \theta Z \sim N(0, \sigma_Y^2 + \theta^2\sigma_Z^2)$. Hence $I(\theta)$ can be computed numerically. Furthermore, it is straightforward to calculate that the information for θ without censoring is $I_{uc}(\theta) = \mathrm{Var}(Z)I_f$ where $I_f \equiv E_f(f'/f)^2$. In particular, for the simple normal model here, $I_{uc}(\theta) = \sigma_Z^2/\sigma_\epsilon^2$.

It is interesting to note that if the error distribution F is given by

$$F(x) = \frac{1}{2}(1 + \sin(x))1_{[-\pi/2,\pi/2]}(x) + 1_{(\pi/2,\infty)}(x)\,,$$

then $-f'(x)/f(x) = \tan(x)$ and $I_f = \infty$, but $f^2(x)/(F(x)\overline{F}(x)) = 1$ for all x, and hence, in the situation of Y and Z independent and normally distributed as above, $I(\theta) = \sigma_Y^2\sigma_Z^2/(\sigma_Y^2 + \theta^2\sigma_Z^2)$ is finite. Similarly, for the error distribution F given by

$$F(x) = \frac{1}{2}(1 - \mathrm{sgn}(x)\sin(-x^2/2))1_{[-\sqrt{\pi},\sqrt{\pi}]}(x) + 1_{(\sqrt{\pi},\infty)}(x)\,,$$

then $-f'(x)/f(x) = -\mathrm{sgn}(x)\{1/|x| + |x|\tan(-x^2/2)\}$ and $I_f = \infty$, but $f^2(x)/(F(x)\overline{F}(x)) = x^2$ for $|x| \le \sqrt{\pi}$, and hence, in the situation of Y and Z independent and normally distributed as above, $I(\theta) = \sigma_Y^2\sigma_Z^2/\sigma_\epsilon^2$ is finite and does not depend on θ.

The following table was computed under the assumption that $\sigma_\epsilon^2 = \sigma_Y^2 = \sigma_Z^2 = 1$.

Table 1: Comparison of information with and without censoring

Error density	θ	0.0	0.2	0.4	0.6	0.8	1.0	1.5	2.0
Normal	$I(\theta)$	0.4805	0.4581	0.4004	0.3283	0.2588	0.2002	0.1047	0.0577
	$I_{uc}(\theta)$	1	1	1	1	1	1	1	1
Logistic	$I(\theta)$	.2066	.1975	.1740	.1444	.1155	.0908	.0495	.0282
	$I_{uc}(\theta)$	1/3	1/3	1/3	1/3	1/3	1/3	1/3	1/3
$\cos(x)/2$	$I(\theta)$	1.0	.9615	.8621	.7353	.6098	.5000	.3077	.2000
	$I_{uc}(\theta)$	∞	∞	∞	∞	∞	∞	∞	∞
$\|x\|\cos(-x^2/2)/2$	$I(\theta)$	1.	1.	1.	1.	1.	1.	1.	1.
	$I_{uc}(\theta)$	∞	∞	∞	∞	∞	∞	∞	∞

Now we present information calculations for the regression parameter θ in the binary choice model discussed in section 1. This is really a special case of the linear regression model, except for the identifiability problem mentioned previously in section 1: since $Y = 0$ with probability 1, θ is unidentifiable without some restriction. A convenient restriction which makes the connection with the calculation for the linear regression model is to suppose that the first component of $\theta \in R^d$ is a fixed (known) value, e.g. $\theta_1 = -1$; then we relabel the remaining $d-1$ components of θ as simply θ (so that $\theta = (-1, \theta)$ where the θ appearing on the right side is in R^{d-1}. Then, if $Z = (Z_1, Z_2)$ where $Z_1 \in R$ and $Z_2 \in R^{d-1}$, $-\theta' Z = Z_1 - \theta' Z_2$, and the density of the observation (of a sample of size one) becomes

$$p_{\theta,F}(\delta, y, z) = [F(z_1 - z_2'\theta)]^{\delta}[1 - F(z_1 - z_2'\theta)]^{1-\delta} g(z), \qquad \text{for} \qquad \delta \in \{0,1\}, z \in \mathbb{R}^d.$$

Since $g(z)$ does not involve F or θ, it can be treated as known and will not be considered in the following. Here is the efficient score function and the information for θ in this model.

THEOREM 2.2. *(The binary choice model). Let $h(s) = E(Z_2 | Z_1 - Z_2'\theta = s)$. Suppose that $\lim_{s\to\infty} f(s) = 0$ and $h(s)$ is bounded. Then the efficient score for θ is*

$$l_\theta^*(x) = f(z_1 - z_2'\theta)[z_2 - E(Z_2 | Z_1 - Z_2'\theta = z_1 - z_2'\theta)] \left[\frac{1-\delta}{1 - F(z_1 - z_2'\theta)} - \frac{\delta}{F(z_1 - z_2'\theta)} \right]. \tag{2.10}$$

Moreover, the information for θ is $I(\theta)$, where

$$I(\theta) = E\left\{ \frac{f(Z_1 - Z_2'\theta)^2}{F(Z_1 - Z_2'\theta)(1 - F(Z_1 - Z_2'\theta))} [Z_2 - E(Z_2 | Z_1 - Z_2'\theta)]^{\otimes 2} \right\}. \tag{2.11}$$

Now we calculate the information for the regression parameter in Cox model under interval censoring.

The complete data is $X^0 = (T, Y, Z)$, where T is the failure time, Y is the censoring time, $Z \in \mathbb{R}^d$ is the covariate. Suppose that given Z, T and Y are independent. The Cox proportional hazard model assumes the hazard function $\lambda(t|z)$ of T given Z is related to the baseline hazard function $\lambda(t)$ in the following way.

$$\lambda(t|z) = e^{z'\theta}\lambda(t).$$

Under this model, we have the following relationships for the cumulative hazard functions and the survival functions.

$$\Lambda(t|z) = e^{z'\theta}\Lambda(t), \qquad \overline{F}(t \mid z) = \overline{F}(t)^{\exp(z'\theta)}.$$

With interval censoring, we only observe

$$X = (\delta = 1_{\{T<Y\}}, Y, Z) \in \{0,1\} \times \mathbb{R}^+ \times \mathbb{R}.$$

The problem is to estimate the parameter θ and the baseline hazard function. To obtain some insight into the difficulty of this estimation problem, we first compute the information contained for the regression parameter θ.

The probability density function of X is

$$\begin{aligned} p_{\theta,F}(x) &= p_{\theta,F}(y,\delta,z) \\ &= F(y|z)^{\delta}\overline{F}(y|z)^{1-\delta}g(y|z)h(z) \\ &= [1-\overline{F}(y)^{\exp(z'\theta)}]^{\delta}\overline{F}(y)^{(1-\delta)\exp(z'\theta)}g(y|z)h(z). \end{aligned}$$

Since g and h do not involve θ and F, and enter the density as factors of products, the information for θ (and for F) are the same whether g and h are known or unknown. So we assume that they are known in the following.

THEOREM 2.3. *(The Cox model with interval censoring). The efficient score function for θ is*

$$l_\theta^*(x) = \exp(z'\theta)Q(y,\delta,z)\Lambda(y)\left\{z - \frac{E[(Z\exp(2z'\theta)O(Y \mid Z) \mid Y=y]}{E[(\exp(2Z'\theta)O(Y \mid Z) \mid Y=y]}\right\}. \tag{2.12}$$

where

$$Q(y,\delta,z) = \delta\frac{\overline{F}(y \mid z)}{1-\overline{F}(y \mid z)} - (1-\delta)$$

and

$$O(y|z) = \frac{\overline{F}(y \mid z)}{1-\overline{F}(y \mid z)}.$$

Moreover, the information for θ is $I(\theta)$ where

$$\begin{aligned} I(\theta) &= E[l_\theta^*(X)l_\theta^*(X)] \\ &= E\left\{\Lambda^2(Y \mid Z)Q^2(Y,\delta,Z)\left[Z - \frac{E(Z\exp(2Z'\theta)O(Y \mid Z) \mid Y]}{E(\exp(2\theta Z)O(Y \mid Z) \mid Y]}\right]^{\otimes 2}\right\} \\ &= E\left\{R(Y,Z)\left[Z - \frac{E(ZR(Y,Z) \mid Y)}{E(R(Y,Z) \mid Y)}\right]^{\otimes 2}\right\}, \end{aligned}$$

where $R(Y,Z) \equiv \Lambda^2(Y \mid Z)O(Y \mid Z)$. Hence the information for θ is positive unless $ZE(R(Y,Z)|Y) = E(ZR(Y,Z)|Y)$ with probability one.

PROOF OF THEOREM 2.1. For simplicity, we will prove (2.8) and (2.9) for $\theta \in \mathbb{R}$. The generalization to $\theta \in \mathbb{R}^d$ is straightforward.

We first compute the score function for θ and F. The score function for θ is simply the derivative of the log-likelihood with respect to θ. That is,

$$\dot{l}_\theta(x) = -\delta\frac{f(y-z'\theta)z}{F(y-z'\theta)} + (1-\delta)\frac{f(y-z'\theta)z}{1-F(y-z'\theta)}.$$

Now suppose $\mathcal{F}_0 = \{F_\eta, |\eta| < 1\}$ is a regular parametric sub-family of
$\mathcal{F} = \{F : F \ll \mu, \mu = \text{Lebesgue measure}\}$. Set

$$\frac{\partial}{\partial\eta}\log f_\eta(t) \mid_{\eta=0} = a(t).$$

Then $a \in L_2^0(F)$, and

$$\frac{\partial}{\partial\eta}F_\eta(t)|_{\eta=0} = \int_{-\infty}^{t} adF, \quad \frac{\partial}{\partial\eta}\overline{F}_\eta(t)|_{\eta=0} = -\int_{-\infty}^{t} adF.$$

The score operator for f is:

$$\dot{l}_f(a)(x) = \delta\frac{\int_{-\infty}^{y-z'\theta} a(t)dF(t)}{F(y-z'\theta)} - (1-\delta)\frac{\int_{-\infty}^{y-z'\theta} a(t)dF(t)}{1-F(y-z'\theta)}.$$

To calculate the information for θ in this semiparametric model, we follow the general theory of Bickel, Klaassen, Ritov, and Wellner (1993). We first need to compute the efficient score function $\dot{l}_\theta^*$ for θ. Geometrically, $\dot{l}_\theta^*$ can be interpreted as the residual of $\dot{l}_\theta$ projected in the space spanned by $\dot{l}_f a$, where $a \in L_2^0(F) = \{a : \int adF = 0 \text{ and } \int a^2 dF < \infty\}$. Thus we need to find a function a_* with $\int a_* dF = 0$ so that $\dot{l}_\theta^* = \dot{l}_\theta - \dot{l}_f a_* \perp \dot{l}_f a$ for all $a \in L_2^0(F)$. That is

$$E(\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) = 0 \tag{2.13}$$

for all $a \in L_2^0(F)$. Now we proceed to find a_* such that (2.13) is true. We have

$$\begin{aligned}
&\dot{l}_\theta(x) - (\dot{l}_f a_*)(x) \\
&= \left[f(y-z'\theta)z + \int_{-\infty}^{y-z'\theta} a_*(t)dF(t)\right]\left[\frac{1-\delta}{1-F(y-z'\theta)} - \frac{\delta}{F(y-z'\theta)}\right].
\end{aligned}$$

Thus

$$\begin{aligned}
&-E((\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) \\
&= E\Bigg\{\left[\frac{1-\delta}{1-F(Y-Z'\theta)} - \frac{\delta}{F(Y-Z'\theta)}\right]^2 \\
&\qquad \times \left[f(Y-Z'\theta)Z + \int_{-\infty}^{Y-Z'\theta} a_*(t)dF(t)\right]\int_{-\infty}^{Y-Z'\theta} a(t)dF(t)\Bigg\} \\
&= E\Bigg\{\left[f(Y-Z'\theta)Z + \int_{-\infty}^{Y-Z'\theta} a_*(t)dF(t)\right] \\
&\qquad \times \left[\frac{1}{1-F(Y-Z'\theta)} + \frac{1}{F(Y-Z'\theta)}\right]\int_{-\infty}^{Y-Z'\theta} a(t)dF(t)\Bigg\} \\
&= E\Bigg\{\left[\frac{1}{1-F(s)} + \frac{1}{F(s)}\right]\int_{-\infty}^{s} a(t)dF(t) \\
&\qquad \times E\left[(f(s)Z + \int_{-\infty}^{s} a_*(t)dF(t))\,|Y-Z'\theta = s\right]\Bigg\}.
\end{aligned}$$

So we can set

$$E\left[(f(s)Z + \int_{-\infty}^{s} a_*(t)dF(t))\,|Y-Z'\theta = s\right] = 0$$

to ensure that (2.13) is true. We obtain a_* by solving the following equation:

$$f(s)E(Z|Y-Z'\theta = s) + \int_{-\infty}^{s} a_*(t)dF(t) = 0.$$

In other words, we can choose any a_* that satisfies the above equation. In particular, if $f(s)$ and $h(s) = E(Z|Y - Z'\theta = s)$ are differentiable, and $f(s) > 0$ for any $s \in \mathbb{R}$, then we have an explicit expression for a_*:

$$a_*(s) = -h'(s) - h(s)\frac{f'(s)}{f(s)}.$$

By the assumptions, we have

$$\int a_*(t)dF(t) = \lim_{s\to\infty}\int_{-\infty}^{s} a_*(t)dF(t) = \lim_{s\to\infty} f(s)E(Z|Y - Z'\theta = s) = 0.$$

It follows that the efficient score function for θ is

$$\begin{aligned} l_\theta^*(x) &= \dot{l}_\theta(x) - \dot{l}_f a_*(x) \\ &= f(y - z'\theta)[z - E(Z|Y - Z'\theta = y - z'\theta)]\left[\frac{1-\delta}{1 - F(y - z'\theta)} - \frac{\delta}{F(y - z'\theta)}\right]. \end{aligned}$$

The information for θ is

$$\begin{aligned} I(\theta) &= E[\dot{l}_\theta^*(X)]^2 \\ &= E\left\{f(Y - Z'\theta)^2\left[\frac{1}{1 - F(Y - Z'\theta)} + \frac{1}{F(Y - Z'\theta)}\right][Z - E(Z|Y - Z'\theta)]^2\right\} \\ &= E\left\{f(Y - Z'\theta)^2\left[\frac{1}{F(Y - Z'\theta)(1 - F(Y - Z'\theta))}\right][Z - E(Z|Y - Z'\theta)]^2\right\}. \end{aligned}$$

Hence the information for θ is positive unless $Z = E(Z|Y - Z'\theta)$ with probability one. $\square$

PROOF OF THEOREM 2.3. The log-likelihood function is

$$l(\theta, F) = \log p_{\theta,F}(X) = \delta\log(1 - \overline{F}(Y)^{e^{\theta Z}}) + (1-\delta)e^{\theta Z}\log\overline{F}(Y) + \text{constants}.$$

Now we first compute the score function for θ and F. The score function for θ is simply the derivative of the log-likelihood with respect to θ, that is,

$$\dot{l}_\theta(x) = \delta\frac{\overline{F}(y)^{\exp(\theta z)}e^{\theta z}\Lambda(y)z}{1 - \overline{F}(y)^{\exp(\theta z)}} + (1-\delta)ze^{\theta z}\log\overline{F}(y).$$

Using the formula $-\log\overline{F}(y) = \Lambda(y)$, it follows that

$$\dot{l}_\theta(x) = ze^{\theta z}\Lambda(y)\left[\delta\frac{\overline{F}(y|z)}{1 - \overline{F}(y|z)} - (1-\delta)\right].$$

Now suppose $\mathcal{F}_0 = \{F_\eta, |\eta| < 1\}$ is a regular parametric sub-family of $\mathcal{F} = \{F : F \ll \mu, \mu = \text{Lebesgue measure}\}$. Set

$$\frac{\partial}{\partial\eta}\log f_\eta(t)\mid_{\eta=0} = a(t),$$

then $a \in L_2^0(F)$ and

$$\frac{\partial}{\partial\eta}\overline{F}_\eta(t)|_{\eta=0} = \int_t^\infty a dF.$$

The score operator for f is

$$\begin{aligned}\dot{l}_f(a)(x) &= \delta\frac{-e^{\theta z}\overline{F}(y)^{\exp(\theta z)-1}\int_y^\infty adF}{1-\overline{F}(y\mid z)} + (1-\delta)e^{\theta z}\frac{1}{\overline{F}(y)}\int_y^\infty adF \\ &= e^{\theta z}\frac{\int_y^\infty adF}{\overline{F}(y)}\left[\delta\frac{-\overline{F}(y\mid z)}{1-\overline{F}(y\mid z)} + (1-\delta)\right].\end{aligned}$$

Let $Q(y,\delta,z) = \delta\frac{\overline{F}(y|z)}{1-\overline{F}(y|z)} - (1-\delta)$, then

$$\dot{l}_\theta(x) = ze^{\theta z}\Lambda(y)Q(y,\delta,z),$$

and

$$\dot{l}_f(a)(x) = -e^{\theta z}\frac{\int_y^\infty adF}{\overline{F}(y)}Q(y,\delta,z).$$

To calculate the efficient score $\dot{l}_\theta^*$ for θ, we need to find a function a_* with $\int a_* dF = 0$ so that

$$\dot{l}_\theta - \dot{l}_f a_* \perp \dot{l}_f a \text{ for all } a \in L_2^0(F),$$

that is, $E(\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) = 0$ for all $a \in L_2^0(F)$.

Letting $\exp(z) = e^z$, we have

$$\dot{l}_\theta(x) - (\dot{l}_f a_*)(x) = \exp(\theta z)Q(y,z,\delta)\left[z\Lambda(y) + \frac{\int_y^\infty a_* dF}{\overline{F}(y)}\right].$$

Thus it follows that

$$-E(\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) = E\left\{\exp(2\theta Z)Q^2(Y,Z,\delta)\left[Z\Lambda(Y) + \frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}\right]\frac{\int_Y^\infty adF}{\overline{F}(Y)}\right\},$$

where

$$Q^2(y,\delta,z) = \left[\delta\frac{\overline{F}(y\mid z)}{1-\overline{F}(y\mid z)} - (1-\delta)\right]^2 = \delta\left[\frac{\overline{F}(y\mid z)}{1-\overline{F}(y\mid z)}\right]^2 + (1-\delta).$$

The conditional expectation of $Q^2(Y,\delta,Z)$ given Y, Z is

$$E[Q^2(Y,\delta,Z)\mid Y=y, Z=z] = \frac{\overline{F}^2(y\mid z)}{1-\overline{F}(y\mid z)} + \overline{F}(y\mid z) = \frac{\overline{F}(y\mid z)}{1-\overline{F}(y\mid z)},$$

which is the conditional odds ratio of the conditional survival function of y and will be denoted as $O(y\mid z)$ below. So we have

$$\begin{aligned}&-E(\dot{l}_\theta - \dot{l}_f a_*)(\dot{l}_f a) \\ &= E_{Y,Z}E\left\{\exp(2\theta Z)Q^2(Y,\delta,Z)[Z\Lambda(Y) + \frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}]\frac{\int_Y^\infty adF}{\overline{F}(Y)}\mid Y,Z\right\} \\ &= E\left\{\exp(2\theta Z)O(Y\mid Z)[Z\Lambda(Y) + \frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}]\frac{\int_Y^\infty adF}{\overline{F}(Y)}\right\} \\ &= E_Y\left\{\frac{\int_Y^\infty adF}{\overline{F}(Y)}E[\exp(2\theta Z)O(Y\mid Z)[Z\Lambda(Y) + \frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}]\mid Y]\right\}.\end{aligned}$$

Suppose that

$$E\left\{\exp(2\theta Z)O(Y \mid Z)[Z\Lambda(Y) + \frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}] \mid Y\right\} = 0.$$

Then

$$\Lambda(Y)E[\exp(2\theta Z)O(Y \mid Z)Z \mid Y] = -\frac{\int_Y^\infty a_* dF}{\overline{F}(Y)}E[\exp(2\theta Z)O(Y \mid Z) \mid Y],$$

and hence

$$\int_y^\infty a_* dF = -\frac{\Lambda(y)\overline{F}(y)E[Z\exp(2\theta Z)O(Y \mid Z)|Y = y]}{E[\exp(2\theta Z)O(Y \mid Z) \mid Y = y]}.$$

Denoting the the term at the right side of the equation as $-L(y)$, we have

$$a_*(y) = \frac{\dot{L}(y)}{f(y)},$$

provided $f(y) > 0$ for $y \in R$.

Hence if we can prove that $a_* \in L_2^0(F)$, a_* meets our requirement. This is equivalent to prove that

$$\lim_{y \to 0} L(y) = 0.$$

This is true if the covariate Z has bounded support, i.e., there is constant $M > 0$ such that $|Z| \leq M$ with probability 1. Because in this case, noticing that $\exp(2\theta z)$ and $O(y \mid z)$ are both positive for all y and z, we have,

$$\left|\frac{E[Z\exp(2\theta Z)O(Y \mid Z)|Y = y]}{E[\exp(2\theta Z)O(Y \mid Z) \mid Y = y]}\right| \leq \frac{E[|Z|\exp(2\theta Z)O(Y \mid Z)|Y = y]}{E[\exp(2\theta Z)O(Y \mid Z) \mid Y = y]} \leq M.$$

This implies

$$\left|\int_y^\infty a_* dF\right| \leq M\Lambda(y)\overline{F}(y) \to 0,$$

as $y \to 0$. So the efficient score function for θ is

$$\begin{aligned}
\dot{l}_\theta^*(x) &= \dot{l}_\theta(x) - (\dot{l}_f a_*)(x) \\
&= \exp(\theta z)Q(y, \delta, z)\left\{z\Lambda(y) + \frac{\int_y^\infty a_* dF}{\overline{F}(y)}\right\} \\
&= \exp(\theta z)Q(y, \delta, z)\Lambda(y)\left\{z - \frac{E[(Z\exp(2\theta Z)O(Y \mid Z) \mid Y = y]}{E[(\exp(2\theta Z)O(Y \mid Z) \mid Y = y]}\right\}.
\end{aligned}$$

The information for θ is

$$\begin{aligned}
I(\theta) &= E[l_\theta^*(X)]^2 && (2.14)\\
&= E\left\{\Lambda^2(Y \mid Z)Q^2(Y, \delta, Z)\left[Z - \frac{E(Z\exp(2\theta Z)O(Y \mid Z) \mid Y]}{E(\exp(2\theta Z)O(Y \mid Z) \mid Y]}\right]^2\right\} && (2.15)\\
&= E\left\{R(Y, Z)\left[Z - \frac{E(ZR(Y, Z) \mid Y)}{E(R(Y, Z) \mid Y)}\right]^2\right\},
\end{aligned}$$

where $R(Y, Z) = \Lambda^2(Y \mid Z)O(Y \mid Z)$. □

3. Maximum Likelihood Estimation via Profile Likelihood. In this section, we discuss maximum likelihood estimates in semiparametric models obtained via *profile likelihoods*.

Consider a semiparametric model

$$\mathcal{P} = \{P_{\theta,F} : \theta \in \Theta, F \in \mathcal{F}\}, \tag{3.16}$$

where $\Theta \subset \mathbb{R}^d$, and $\mathcal{F}$ is an infinite dimensional function space. In many applications, the parameter of primary interest in the model $\mathcal{P}$ is the finite dimensional parameter θ. F is usually a parameter of secondary interest or a "nuisance" parameter. Our main goal here is to estimate θ. Let $p(\cdot;\theta,F)$ be the density function corresponding to $P_{\theta,F} \in \mathcal{P}$. Moreover, suppose $X_1,\ldots,X_n$ are i.i.d. with density $p(\cdot;\theta_0,F_0)$, where θ_0 and F_0 are the assumed true values of the parameters. Let $l_n(\theta,F) = \sum_{i=1}^n \log p(X_i;\theta,F)$ be the log-likelihood function. Computing the maximum likelihood estimates via the *profile likelihood* proceeds by the following steps:

(i) For any θ fixed, construct a function $F_n(\cdot;\theta)$ by maximizing $l_n(\theta;F)$ with respect to $F \in \mathcal{F}$: $F_n(\cdot;\theta) = \text{argmax}_F l_n(\theta,F)$; i.e. $\sup_{F\in\mathcal{F}} l_n(\theta,F) = l_n(\theta,F_n(\cdot;\theta))$. [If we knew $\theta = \theta_0$, then $F_n(\cdot;\theta_0)$ is the maximum likelihood estimator of F.]

(ii) Substitute $F_n(\cdot;\theta)$ back into the log-likelihood function to obtain the *profile log-likelihood function* $\widetilde{l}_n(\theta) \equiv l_n(\theta,F_n(\cdot;\theta))$.

(iii) Maximize $\widetilde{l}_n(\theta)$ to obtain an estimator $\widehat{\theta}_n$ of θ_0 and $\widehat{F}_n \equiv F_n(\cdot;\widehat{\theta}_n)$ of F_0: $\widehat{\theta}_n = \text{argmax}\, \widetilde{l}_n(\theta)$.

It is straightforward to show that in fact $(\widehat{\theta}_n,\widehat{F}_n)$ is the maximum likelihood estimator of (θ,F) (assuming existence and uniqueness). This method of computing maximum likelihood estimates seems to date back to Richards (1961); see also Kale (1963), Kalbfleisch and Sprott (1970), Patefield (1977), and the survey by Barndorff-Nielsen (1991). See Cox and Oakes (1984), pages 41-43 for a nice treatment of the Weibull model via profile likelihood. All of these authors are concerned with (regular) parametric models.

It is now well-known that in the case of the Cox model with right censoring, the Cox "partial likelihood" is also simply the profile likelihood; see e.g. Johansen (1978), Bailey (1984), and Andersen, Borgan, Gill, and Keiding (1993), page 482.

In the case of many semiparametric models $\mathcal{P}$ the maximization problem in step (i) is not sensible or well-defined. This often happens when the space $\mathcal{F}$ is "too large"; in these cases some sort of regularization is needed. However, in some important cases, such as the interval censoring problems considered here, the maximization problem in (i) *is* well defined. For example, in the linear regression model with interval censoring, $\mathcal{F}$ is the space of all distribution functions. The natural monotonicity constraints imposed on the maximization of $l_n(\theta,F)$ with respect to F make $F_n(\cdot;\theta)$ well defined for each $\theta \in \Theta$. These estimators were first proposed in the further special case of the binary choice model by Cosslett (1983). For the Cox model with interval censoring, $\mathcal{F}$ is the space of all cumulative hazard functions, again, the maximization problem is well defined.

In both of the interval censoring models studied here, the iterative convex minorant algorithms developed by Groeneboom (see Groeneboom and Wellner (1992), section 3.2, pages 69 - 73) converge sufficiently quickly to allow for rapid computation of the profile likelihood function $\widetilde{l}_n$ as a function of θ, and hence computation of $\widehat{\theta}_n$ numerically by a variety of methods. In fact, no iteration is needed in step (i) in the case of the linear regression model with interval censoring. When $\theta \in \Theta \subset R$, it is fruitful to plot $\widetilde{l}_n(\theta)$ and compute $\widehat{\theta}_n$ via a simple grid search.

An important question is the efficiency of the maximum profile likelihood estimator $\widehat{\theta}_n$. One would expect that if $\widehat{\theta}_n$ is obtained strictly according to steps (i) - (iii), then with some appropriate regularity conditions, it should be locally regular and (asymptotically) efficient. However, very little seems to be known about the properties of maximum likelihood estimators in the semiparametric model setting (when they exist). One exception is the work by Severini and Wong (1992). They hypothesized that $F_n(\cdot;\theta)$ converges to a function, say $F(\cdot;\theta)$, in an appropriate sense, and that $F(\cdot;\theta)$ satisfies $F(\cdot;\theta_0) = F_0(\cdot)$ and

$$E_{F_0}\left[\left|\frac{d}{d\theta}\frac{\partial l}{\partial F}(\theta, F(\cdot;\theta))\right|_{\theta=\theta_0}\right] = 0$$

where $l(\theta, F(\cdot;\theta)$ is the limit of $l_n(\theta, F_n(\cdot;\theta))$. They call such $F(\cdot;\theta)$ a *least favorable curve.* Under additional regularity and smoothness conditions on $F_n(\cdot;\theta)$ including existence of second derivatives with respect to θ, they showed that $\widehat{\theta}_n$ is asymptotically efficient,

Unfortunately however, the regularity and smoothness conditions of Severini and Wong's (1992) theorem are too severe to apply to maximum likelihood estimators in the interval censoring models of interest here: $F_n(\cdot;\theta)$, obtained by maximizing $l_n(\theta, F)$ over distribution functions F (or, in the case of the Cox model, over cumulative hazard functions Λ), is not a smooth function with respect to θ in either of the regression models we consider. However, we conjecture that the maximum likelihood estimators $\widehat{\theta}_n$ is asymptotically efficient in all of the models considered here.

One way to get around these difficulties is to take broader view toward step (i): to obtain an asymptotically efficient estimator of θ, we might relax the requirement that $F_n(\cdot;\theta)$ be the solution of $\max_{F\in\mathcal{F}} l_n(\theta, F)$. It could be constructed in other ways tailored to the specific problem. For example, for the linear regression model with interval censoring, $F_n(\cdot;\theta)$ could be a histogram estimator or a kernel estimator. Another way around these difficulties is to study estimators obtained from the "estimated efficient score equations" or suitable modifications thereof. Several different strategies are considered in the context of regression models with interval censoring in Huang (1993c).

4. Profile Likelihood for Regression Models with Interval Censoring. Now consider the profile likelihood function in the case of linear regression with interval censoring.

Suppose we observe

$$(\delta_1 = 1_{\{T_1\le Y_1\}}, Y_1, Z_1), \ldots, (\delta_n = 1_{\{T_n\le Y_n\}}, Y_n, Z_n),$$

where $(T_i, Y_i, Z_i), i = 1, \ldots, n$ are i.i.d. random variables with distribution Q_{θ_0,F_0}. Here θ_0 and F_0 are the true regression parameter and the true error distribution respectively. The main goal is to estimate the regression parameter θ in the presence of the infinite dimensional nuisance parameter F. Recall that the log-likelihood function is, up to an additive constant,

$$\begin{aligned}
l_n(\theta, F) &= \sum_{i=1}^n \log p_{\theta,F}(\delta_i, Y_i, Z_i) \\
&= \sum_{i=1}^n \delta_i \log F(Y_i - Z_i'\theta) + (1-\delta_i)\log(1 - F(Y_i - Z_i'\theta)) \\
&= n\mathbb{P}_n \{\delta \log F(y - z'\theta) + (1-\delta)\log(1 - F(y - z'\theta))\} \\
&= n\mathbb{Q}_n \left\{1_{\{t\le y\}} \log F(y - z'\theta) + 1_{\{t>y\}} \log(1 - F(y - z'\theta))\right\},
\end{aligned}$$

where $\mathbb{P}_n$ is the empirical measure of the samples $(\delta_i, Y_i, Z_i), i = 1, \ldots, n$; $\mathbb{Q}_n$ is the empirical measure of the unobservable random variables $(T_i, Y_i, Z_i), i = 1, \ldots, n$.

The $F_n(\cdot;\theta)$ that maximizes $l_n(\theta; F)$ (for each fixed θ) is a right continuous increasing step function. It can be expressed by the min-max formula of isotonic regression theory, see e.g., Robertson, Wright, and Dykstra (1988), chapter 1. Let $r_i(\theta) = Y_i - Z_i'\theta$, and let $r_{(i)}(\theta)$ be the ordered values of $r_i(\theta), i = 1, \ldots, n$, and let $\delta_{(i)}(\theta)$ be the corresponding δ, i.e., if $r_j(\theta) = r_{(i)}(\theta)$, then $\delta_{(i)}(\theta) = 1_{\{T_j \le Y_j\}}$. Then

$$F_n(r_{(i)}(\theta);\theta) = \max_{s\le i}\min_{t\ge i}\frac{\sum_{s\le k\le t}\delta_{(k)}(\theta)}{t-s+1}, \quad i = 1,\ldots,n.$$

The computation can be easily carried out using the "pool adjacent violators" algorithm. Geometrically, $F_n(\cdot;\theta)$ can be described as the slope of the greatest convex minorant (GCM) of the points $(i, \sum_{j=1}^{i}\delta_{(j)}(\theta))$ on $[0,n]$. See, e.g., Groeneboom and Wellner (1992), Proposition 1.2, page 41. Equivalently, this can be stated as in the following lemma. For $(t,\theta) \in \mathbb{R}\times\mathbb{R}^d$, define

$$\begin{aligned} V_n(t;\theta) &= \frac{1}{n}\sum_{i=1}^{n}\{T_i \le Y_i\}\{Y_i - Z_i'\theta \le t\} \\ &= \int\{u \le y\}\{y - z'\theta \le t\}d\mathbb{Q}_n(u,y,z), \end{aligned} \tag{4.17}$$

$$W_n(t;\theta) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - Z_i'\theta \le t\} = \int\{y - z'\theta \le t\}d\mathbb{Q}_n(u,y,z). \tag{4.18}$$

Lemma 4.1. *Let V_n and W_n be defined by (4.17) and (4.18), respectively. Then $F_n(\cdot;\theta)$ is the left derivative of the greatest convex minorant of the cumulative sum diagram $(\overline{W}_n(s;\theta), \overline{V}_n(s;\theta))$, defined by the points*

$$P_j = \left(W_n(r_{(j)}(\theta);\theta), V_n(r_{(j)}(\theta);\theta)\right), \quad j = 0,1,\ldots,n;$$

where $P_0 = (0,0)$ and $\overline{V}_n$ and $\overline{W}_n$ are obtained by linear interpolation of the points $P_0,\ldots,P_n$.

Figures 1 and 3 show the profile log likelihoods $\widetilde{l}_n(\theta) = l_n(\theta, F_n(\cdot,\theta))$ for several sample sizes of simulated data from the linear regression model with interval censoring for the following situations: in Figure 1, ϵ, Y, and Z are all independent with $N(0,1)$ distributions and the true $\theta_0 = 1$. In Figure 3, $\epsilon \sim N(0,1)$, $Y \sim U(-2,2)$, and $Z \sim U(-1,1)$ are all independent and again the true $\theta_0 = 1$. Figures 2 and 4 show the estimated and true error distributions for the four different sample sizes and two situations: figure 2 accompanies figure 1 and figure accompanies figure 3. Table 2 shows the computed estimates $\widehat{\theta}_n$ for the two situations and four sample sizes.

Table 2: Estimates $\widehat{\theta}_n$ of $\theta_0 = 1$

sampling situation	n	100	200	400	800
Normal Error, Normal Y, Z		0.7839	1.1256	0.9949	1.0351
Normal Error, Uniform Y, Z		0.8191	0.9146	0.9045	0.9748

INSERT FIGURES HERE

Although we are not yet able to show that the maximum likelihood estimator $\widehat{\theta}_n = \text{argmax}\widetilde{l}_n(\theta) \equiv \text{argmax}l_n(\theta, F_n(\cdot;\theta))$ is asymptotically efficient, we do have the following preliminary consistency result.

Let G denote the joint distribution of (Y, Z), let H be the distribution of $W \equiv Y - Z'\theta_0$, and let G_0 be the joint distribution of (W, Z). We also let $P \equiv P_{\theta_0, F_0}$ denote the joint law of each of the i.i.d. triples (δ_i, Y_i, Z_i), $i = 1, 2, \ldots$.

Theorem 4.1. *Suppose that:*
(i) The true value of the regression parameter $\theta_0 \in \Theta$, *where* Θ *is a bounded subset of* $\mathbb{R}^d$.
(ii) $H\{w : 0 < F_0(w) < 1\} = 1$.
(iii) The distribution of Z *is not concentrated on a hyperplane.*
Then

$$\widehat{\theta}_n \to_p \theta_0, \tag{4.19}$$

and

$$F_n(t; \widehat{\theta}_n) \to_p F_0(t) \tag{4.20}$$

for almost all (with respect to H*)* $t \in S \equiv \{t : 0 < F_0(t) < 1, t \text{ a continuity point of } F_0\}$.

Proof. Since Θ is bounded, for any subsequence of $\widehat{\theta}_n$, we can find a further subsequence converging to $\theta_* \in \overline{\Theta}$, the closure of Θ. On the other hand, by Helly's selection theorem, for any subsequence of $F_n(\cdot; \widehat{\theta}_n)$, we can find a further subsequence converging in distribution to some subdistribution function F_*; i.e. pointwise convergence at continuity points of F_*. Apparently, we can choose the convergent subsequence of $\widehat{\theta}_n$ and the convergent subsequence of $F_n(\cdot; \widehat{\theta}_n)$ so that they have the same indices. Without causing confusion, we assume that $\widehat{\theta}_n$ converges to θ_* and that $F_n(\cdot; \widehat{\theta}_n)$ converges to F_*. To prove the theorem, it suffices to prove that $\theta_* = \theta_0$ and $F_* = F_0$.

Let

$$a_n = \inf\{y - z'\widehat{\theta}_n : (y, z) \in supp(G)\}, \quad b_n = \sup\{y - z'\widehat{\theta}_n : (y, z) \in supp(G)\},$$
$$a_* = \inf\{y - z'\theta_* : (y, z) \in supp(G)\}, \quad b_* = \inf\{y - z'\theta_* : (y, z) \in supp(G)\},$$

where *supp*(G) is the support of G, the joint distribution of (Y, Z). We now prove that for any $a > a_*$ and $b < b_*$, there exist finite positive constants $0 < M_1 < M_2 < 1$ such that

$$M_1 \le F_n(a; \widehat{\theta}_n) \le F_n(b; \widehat{\theta}_n) \le M_2 \tag{4.21}$$

with probability one for all n sufficiently large. Since $(\widehat{\theta}_n, F_n)$ maximizes $l_n(\theta, F)$, we have

$$\begin{aligned} &\int [\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n) + (1 - \delta) \log(1 - F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n))] d\mathbb{P}_n(\delta, y, z) \\ &\ge \int [\delta \log F_0(y - z'\theta_0) + (1 - \delta) \log(1 - F_0(y - z'\theta_0))] d\mathbb{P}_n(\delta, y, z). \end{aligned} \tag{4.22}$$

Since $a_n \to_{a.s.} a_*$, $a_n < a$ for n sufficiently large. Noticing that $\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n)$ and $(1-\delta)\log(1 - F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n))$ are non-positive for all (y, z), it follows that

$$\begin{aligned} &\int 1_{[a_n,a]}(y - z'\widehat{\theta}_n)\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n) d\mathbb{P}_n(\delta, y, z) \\ &\geq \int [\delta \log F_0(y - z'\theta_0) + (1-\delta)\log(1 - F_0(y - z'\theta_0))] d\mathbb{P}_n(\delta, y, z). \end{aligned}$$

Since

$$E[\delta \log F_0(Y - Z'\theta_0) + (1-\delta)\log(1 - F_0(Y - Z'\theta_0))] \leq 0,$$

and because $x \log x \geq -e^{-1}$ for $0 < x < 1$,

$$\begin{aligned} &E[\delta \log F_0(Y - Z'\theta_0) + (1-\delta)\log(1 - F_0(Y - Z'\theta_0))] \\ &= E[F_0(Y - Z'\theta_0)\log F_0(Y - Z'\theta_0) + (1 - F_0(Y - Z'\theta_0))\log(1 - F_0(Y - Z'\theta_0))] \\ &\geq -2/e > -1. \end{aligned}$$

It follows from the strong law of large numbers that, with probability one,

$$\int [\delta \log F_0(y - z'\theta_0) + (1-\delta)\log(1 - F_0(y - z'\theta_0))] d\mathbb{P}_n(\delta, y, z) \geq -2$$

for n sufficiently large. So by the monotonicity of $F_n(\cdot; \widehat{\theta}_n)$, for n sufficiently large,

$$\begin{aligned} &\log F_n(a; \widehat{\theta}_n) \int \delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n) d\mathbb{P}_n(\delta, y, z) \\ &\geq \int \delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n) \log F_n(y - z'\widehat{\theta}_n) d\mathbb{P}_n(\delta, y, z) \geq -2. \end{aligned}$$

By the Glivenko-Cantelli theorem for half-spaces, a VC class of sets,

$$(\mathbb{P}_n - P)(\delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n)) \to_{a.s.} 0.$$

Furthermore, by the bounded convergence theorem,

$$\begin{aligned} P(\delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n)) &= \int \delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n) dP(\delta, y, z) \\ &\to_{a.s.} \int \delta 1_{[a_*,a]}(y - z'\theta_*) dP(\delta, y, z) = P(\delta 1_{[a_*,a]}(y - z'\theta_*)), \end{aligned}$$

and hence

$$\mathbb{P}_n(\delta 1_{[a_n,a]}(y - z'\widehat{\theta}_n)) \to_{a.s.} P(\delta 1_{[a_*,a]}(y - z'\theta_*)).$$

Moreover, it follows from condition (ii) that

$$P(\delta 1_{[a_*,a]}(y - z'\theta_*)) = E[F_0(Y - Z'\theta_0) 1_{[a_*,a]}(Y - Z'\theta_*)]$$

is positive. It follows that, with probability one, for n sufficiently large,

$$F_n(a; \widehat{\theta}_n) \geq M_1 \equiv \exp(-4/E[F_0(Y - Z'\theta_0) 1_{[a_*,a]}(Y - Z'\theta_*)]) > 0.$$

Thus the first inequality in (4.21) follows. The second inequality of (4.21) can be proved similarly.

By (4.21), and by the fact that the class of all bounded monotone functions is a VC-hull class, see Dudley (1987) or Van der Vaart and Wellner (1993), we have

$$\int 1_{[a,b]}(y - z'\widehat{\theta}_n) \left[\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n) + (1-\delta) \log(1 - F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n))\right] d(\mathbb{P}_n - P)(\delta, y, z) \to_{a.s.} 0.$$

By the bounded convergence theorem,

$$\begin{aligned} &\int 1_{[a,b]}(y - z'\widehat{\theta}_n) \left[\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n) + (1-\delta) \log(1 - F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n))\right] dP(\delta, y, z) \\ &\quad \to_{a.s.} \int 1_{[a,b]}(y - z'\theta_*) \left[\delta \log F_*(y - z'\theta_*)(1-\delta) \log(1 - F_*(y - z'\theta_*))\right] dP(\delta, y, z) \end{aligned}$$

Combining the above, we have

$$\begin{aligned} &\int 1_{[a,b]}(y - z'\widehat{\theta}_n) \left[\delta \log F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n) + \delta \log(1 - F_n(y - z'\widehat{\theta}_n; \widehat{\theta}_n))\right] d\mathbb{P}_n(\delta, y, z) \\ &\quad \to_{a.s.} \int 1_{[a,b]}(y - z'\theta_*) \left[\delta \log F_*(y - z'\theta_*) + (1-\delta) \log(1 - F_*(y - z'\theta_*))\right] dP(\delta, y, z). \end{aligned}$$

In view of (4.22), this implies, for any $a_* < a < b < b_*$,

$$\begin{aligned} &E\left\{1_{[a,b]}(Y - Z'\theta_*)(\delta \log F_*(Y - Z'\theta_*) + (1-\delta) \log(1 - F_*(Y - Z'\theta_*)))\right\} \\ &\quad \geq \; E\left\{\delta \log F_0(Y - Z'\theta_0) + (1-\delta) \log(1 - F_0(Y - Z'\theta_0))\right\}. \end{aligned}$$

Letting $a \downarrow a_*$, and $b \uparrow b_*$, it follows that

$$\begin{aligned} &E\left\{\delta \log F_*(Y - Z'\theta_*) + (1-\delta) \log(1 - F_*(Y - Z'\theta_*))\right\} \\ &\quad \geq \; E\left\{\delta \log F_0(Y - Z'\theta_0) + (1-\delta) \log(1 - F_0(Y - Z'\theta_0))\right\}, \end{aligned}$$

and by subtraction this yields

$$E_P\left\{\delta \log \frac{F_*(Y - Z'\theta_*)}{F_0(Y - Z'\theta_0)} + (1-\delta) \log \frac{1 - F_*(Y - Z'\theta_*)}{1 - F_0(Y - Z'\theta_0)}\right\} \geq 0$$

where $P = P_{\theta_0, F_0}$. But, with $P_* = P_{\theta_*, F_*}$, the expression on the left side is just $-K(P, P_*)$, the Kullback-Leibler "distance" from P to P_*. Since $K(P, P_*) \geq 0$ with equality if and only if $P_* = P$, see e.g. Shorack and Wellner (1986), proposition 24.3.1, page 790, the two inequalities together imply that $K(P, P_*) = 0$ and $P_* = P$. This implies that $F_*(y - z'\theta_*) = F_0(y - z'\theta_0)$ for almost all (y, z) (with respect to G, the joint distribution of (Y, Z)), or $F_*(w + z'(\theta_0 - \theta_*)) = F_0(w)$ for almost all (w, z) (with respect to G_0, the joint distribution of $(W, Z) \equiv (Y - Z'\theta_0, Z)$). Since the distribution of Z is not concentrated on any hyperplane by (iii), this implies that $\theta_* = \theta_0$ and hence that $F_*(w) = F_0(w)$ for almost all w (with respect to H, the distribution of $W \equiv Y - Z'\theta_0$). $\Box$

It would be very desirable to relax the assumption that Θ is bounded in theorem 4.1.

Since $F_n(s; \theta)$ is the slope of the convex minorant of the diagram $(\overline{W}_n(s; \theta), \overline{V}_n(s; \theta))$, so intuitively, it should converge to the slope $F(s; \theta)$ of the diagram $(W(s; \theta), V(s; \theta))$. In fact, it can be shown that $F_n(s, \theta)$ converges to

(4.23) $$F(s; \theta) = \frac{\partial V(s; \theta)}{\partial W(s; \theta)} = \frac{\int F_0(s + z'\theta - z'\theta_0) g(s + z'\theta, z) dz}{\int g(s + z'\theta, z) dz}.$$

Notice that $F(s; \theta_0) = F_0(s)$. Furthermore, this is the *least favorable curve* in the sense of Severini and Wong (1992) for this model.

We do not yet have a proof of asymptotic efficiency of the maximum likelihood estimator $\widehat{\theta}_n$ (or even that $\widehat{\theta}_n - \theta_0 = O_p(n^{-1/2})$). We conjecture that under minimal regularity conditions $\sqrt{n}(\widehat{\theta}_n - \theta_0) \to_d N(0, I(\theta_0)^{-1})$ where $I(\theta_0)$ is as given in Theorem 2.1. However, as discussed in section 1 it is possible to construct inefficient estimators of θ which do satisfy, under more regularity conditions, $\widetilde{\theta}_n - \theta_0 = O_p(n^{-1/2})$; several different $\sqrt{n}$−consistent estimators are studied in Huang (1993c). Then it is natural to estimate F_0 by $F_n(\cdot; \widetilde{\theta}_n)$. Here $F_n(\cdot;\theta) = \text{argmax}_F l_n(\theta, F)$ is the estimator described in lemma 4.1.

Note that if G has density g, then $H(s) = P(Y - Z'\theta_0 \le s) = \int\int g(u + z'\theta_0, z) 1_{\{u \le s\}} dz du$, and H has density $h(s) = H'(s) = \int g(s + z'\theta_0, z) dz$.

THEOREM 4.2. *Let s_0 be such that $0 < F_0(s_0), H(s_0) < 1$. Suppose that*

(i) *$F_0(s)$ and $H(s)$ are differentiable, with bounded derivatives $f_0(s)$ and $h(s)$, and have strictly positive derivatives $f_0(s_0)$ and $h(s_0)$ at $s = s_0$, respectively. Moreover, $f_0(s)$ and $k(s)$ are continuous in a neighborhood of θ_0.*

(ii) *The support of the covariate Z is bounded, i.e., there exists $z_0 > 0$ such that $|Z| \le z_0$ with probability one.*

(iii) $\widetilde{\theta}_n - \theta_0 = O_p(n^{-1/2})$.

Then:

$$n^{1/3}[F_n(s_0;\theta_n) - F_n(s_0;\theta_0)] \to_p 0 \ \textit{as } n \to \infty\,, \tag{4.24}$$

and

$$\left[\frac{2h(s_0)}{f_0(s_0)F_0(s_0)(1 - F_0(s_0))}\right]^{1/3} n^{1/3}[F_n(s_0;\theta_n) - F_0(s_0)] \Rightarrow 2\mathbb{Z}, \ \textit{as } n \to \infty, \tag{4.25}$$

where $\mathbb{Z}$ is the last time where standard two-sided Brownian motion $B(s)$ minus the parabola $x(s) = s^2$ reaches its maximum.

Theorem 4.2 is proved in Huang (1993c).

From this theorem, the maximum likelihood estimator of the distribution F_0 converges at rate $n^{1/3}$ even if the regression parameter θ_0 is known, which is slower than the usual $\sqrt{n}$-convergence rate. Groeneboom and Wellner (1992) showed that the information operator for F_0 is not invertible, hence there is no hope to construct $\sqrt{n}$-convergence rate estimator for F_0. Also from this theorem, when we have an estimator $\widetilde{\theta}_n$ of θ_0 such that $\widetilde{\theta}_n$ has $\sqrt{n}$-convergence rate, then asymptotically, $F_n(s;\widetilde{\theta}_n)$ is as good as $F_n(s;\theta_0)$. Actually, the conclusion remains valid if $\widetilde{\theta}_n - \theta_0 = O_p(n^{-\alpha})$ for some $\alpha > 1/3$.

Results similar to the above also hold for the Cox model with interval censoring. In fact, by comparing likelihoods, it is clear that the likelihood in the case of the Cox model with interval censoring is a smoother function of θ than is the likelihood in the case of the linear regression model with interval censoring. Exploiting this smoothness of the underlying model, Huang (1993c) has succeeded in proving that the maximum likelihood estimator $\widehat{\theta}_n$ is consistent, asymptotically normal, and efficient under mild regularity assumptions, even though $\Lambda_n(\cdot;\theta)$ is *not* a smooth function of θ, and the least favorable curve $\Lambda(\cdot;\theta)$ cannot be calculated explicitly in this case: it turns out to be the solution of

$$\int_y^\infty \left[1_{\{t \le y'\}} \frac{\exp(-e^{z'\theta}\Lambda(y';\theta))}{1 - \exp(-e^{z'\theta}\Lambda(y';\theta))} - 1_{\{t > y'\}}\right] e^{z'\theta} dQ(t, y', z) = 0 \qquad \text{for all } y\,. \tag{4.26}$$

Full details will appear in Huang (1993c).

ACKNOWLEDGEMENTS: The second author owes thanks to Piet Groeneboom for many stimulating discussions and conversations about interval censoring and related problems.

REFERENCES

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10**, 1100-1120.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Bailey, K. R. (1984). Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *Ann. Statist.* **12**, 730 - 736.

Barndorff-Nielsen, O. E. (1991) Likelihood theory, In *Statistical Theory and Modeling, In honour of Sir David Cox, FRS*, chapter 10, pages 232 - 264. Chapman and Hall, London. D. V. Hinkley, N. Reid, and E. J. Snell, editors.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Buckley, J., and James, I. R. (1979). Linear regression with censored data. *Biometrika* **66**, 429-436.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *J. Econometrics* **32**, 189-218.

Cosslett, S. R. (1983) Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51**, 765-782.

Cosslett, S. R. (1987). Efficiency bounds for distribution-free estimators of the binary choice and censored regression models. *Econometrica* **55**, 559-585.

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

Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Ann. Probability* **15**, 1306 - 1326.

Finkelstein, D. M. and Wolfe R. A. (1985) . A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-945.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis.* Wiley, New York.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar Band 19, Birkhäuser, Basel.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. Econometrics*, **35**, 303-316.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60**, 505-531.

Huang, J. (1993a). Central limit theorems for M-estimates. *Tech. Report*, No. **251**, Department of Statistics University of Washington.

Huang, J. (1993b). Maximum scored likelihood estimation of a linear regression model with interval censored data. *Tech. Report*, No. **253**, Department of Statistics, University of Washington.

Huang, J. (1993c). Estimation in Regression Models with Interval Censoring. *Ph. D. dissertation* in preparation, Dept. of Statistics, University of Washington

Huang, J. and Wellner, J. A. (1993). Estimation of a monotone density or a monotone hazard function under random censoring. *Tech. Report*, No. **252**, Department of Statistics, University of Washington.

Johansen, S. (1978). The product limit estimator as maximum likelihood. *Scand. J. Statist.* **5**, 195 - 199.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. Roy. Statist. Soc. B* **32**, 175 - 208.

Kale, B. K. (1963). Some remarks on a method of maximum-likelihood estimation proposed by Richards. *J. Roy. Statist. Soc. B* **25**, 209 - 212.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387-421.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *J. Econometrics* **3**, 205-228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *J. Econometrics* **27**, 313-333.

Pakes, D. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, **57**, 1027-1057.

Patefield, W. M. (1977). On the maximized likelihood function. *Sankhyā Ser. B* **39**, 92 - 96.

Pollard, D. (1989). Asymptotics via empirical processes (with discussion). *Statistical Science* **4**, 341-366.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Conference Board of the Mathematical Sciences, Regional Conference Series in Applied Mathematics, Washington, D.C. .

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1993). Regression with current status data. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Rabinowitz, D., Tsiatis, A., and Aragon, J. (1993). Regression with interval censored data. Technical Report, Department of Biostatistics, Harvard School of Public Health.

Richards, F. S. G. (1961). A method of maximum-likelihood estimation. *J. Roy. Statist. Soc.* **23**, 469 - 475.

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

Robertson, T., Wright, F. T., Dykstra, R. L. (1988). *Order Restricted Statistical Inference.* Wiley, New York.

Rockafellar, R. T. (1970). *Convex Analysis.* Princeton University Press.

Severini, T. A. and Wong, W. H. (1992). Profile likelihood and conditionally parametric models. *Ann. Statist.* **20**, 1768 - 1802.

Sherman, R. P. (1991). Maximal inequalities for degenerated U-processes with applications to optimization estimators. *Bellcore* Preprint.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123-137.

Shiboski, S. C. and Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *J. Amer. Statist. Assoc.* **87**, 360 - 372.

Shorack, G. R. and Wellner, J. A. (1986). Empirical Processes with Applications to Statistics. Wiley, New York.

Silverman, B. M.1986. *Density Estimation for Statistics and Data Analysis* Chapman and Hall, New York.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354-372.

Van der Geer, S. (1990). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14 - 44.

Van der Vaart, A. W. and Wellner, J. A. (1993). *Weak Convergence and Empirical Processes.* To appear, *IMS Lecture Notes-Monograph Series.*

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76 - 99.

Department of Statistics, GN-22
B313 Padelford Hall
University of Washington
Seattle, Washington 98195