The Annals of Statistics
1992, Vol. 20, No. 2, 1002–1021

ASYMPTOTIC NORMALITY OF THE 'SYNTHETIC DATA' REGRESSION ESTIMATOR FOR CENSORED SURVIVAL DATA

BY MAI ZHOU

University of Kentucky

This article studies the large sample behavior of the censored data least squares estimator derived from the synthetic data method proposed by Leurgans and Zheng. The asymptotic distributions are derived by representing the estimator as a martingale plus a higher-order remainder term. Recently developed counting process techniques are used. The results are then compared to the censored regression estimator of Koul, Susarla and Van Ryzin.

1. Introduction. The linear regression model has been successfully used as a statistical model in many areas. A well-developed theory and many computer software packages are available today. However, difficulties arise when regression models are used to analyze lifetime data in practice. Lifetime data are often censored, making ordinary least squares procedures inapplicable.

Suppose the lifetimes are $Y_i = \mathbf{X}_i\beta + \varepsilon_i$, where the $\mathbf{X}_i$'s are observable, β is the parameter we want to estimate and the error terms ε_i are i.i.d. with $E\varepsilon_i = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2 < \infty$. However, we only observe $(T_i, \delta_i, \mathbf{X}_i)$, where

$$T_i = \min(Y_i; C_i) = Y_i \wedge C_i \quad \text{and} \quad \delta_i = I_{[Y_i \le C_i]}, \qquad i = 1, 2, \ldots, n.$$

Here the C_i's are censoring times, i.i.d. random variables independent of Y_i. This model will be called the censored linear model in this paper.

Recently, several methods that can accommodate censored data in regression analysis have been proposed. The Buckley–James (1979) method relies on the i.i.d. assumption of the ε_i's but requires few assumptions on the censoring times C_i. On the other hand, methods due to Koul, Susarla and Van Ryzin (1981) and Leurgans (1987) rely on the i.i.d. assumption of the C_i's [although this can be relaxed somewhat—see remark of Leurgans (1987)], but impose less stringent assumptions upon the ε_i's. Also the latter two methods are computationally much easier.

Leurgans transforms the censored observations (T_i, δ_i) into synthetic data,

$$Y_i^* = \int_{-\infty}^{\infty} \left(\frac{I_{[T_i \ge s]}}{\hat{G}(s)} - I_{[s<0]} \right) ds, \tag{1.1}$$

and applies the usual least squares procedure to $(Y_i^*; \mathbf{X}_i)$. Here $\hat{G}(t)$ is the

Received November 1988; revised September 1991.
AMS 1980 *subject classifications*. Primary 62G10; secondary 62P10, 62N05.
Key words and phrases. Censored data, linear regression, asymptotic distribution.

Kaplan–Meier estimator of the survival function $G(t)$ of the i.i.d. censoring times C_i's, defined in (2.6). From a somewhat different point of view, Zheng (1984) suggested a class of transformations which includes (1.1) and (1.1) is the one he recommends. Zheng showed strong consistency of the estimator. Leurgans' (1987) paper contains a statement of the asymptotic distribution of the estimator in the two-sample case. However, asymptotic theory for the general case is lacking. In particular, the asymptotic variance is still unknown.

In this paper, counting process and martingale techniques are used to show that Leurgans' synthetic data method yields estimates that are asymptotically normally distributed and the asymptotic variances of the estimates are derived. In particular, we represent the synthetic data estimator as a martingale plus a high-order term (see 3.6–3.10). We then compare the variance with that of the Koul, Susarla and Van Ryzin (1981) result.

2. Notation and counting process preliminaries. Counting process techniques have been widely used in the analysis of lifetime data since Aalen (1978). Reviews of this approach include the books of Gill (1980) and Jacobsen (1982) and the article of Andersen and Borgan (1985).

First we introduce some more notation and establish some simple facts. For $i = 1, 2, \ldots, n$, let

$$F_i(t) = P(Y_i \geq t),\ G(t) = P(C_i \geq t) \text{ and } H_i(t) = P(T_i \geq t) = F_i(t)G(t).$$

For simplicity, we suppose that the survival functions F_i and G are continuous, although we do not need this assumption throughout. Let

$$\text{(2.1)} \quad \hat{H}_i(t) = I_{[T_i \geq t]}, \quad R^+(t) = \sum_{i=1}^{n} I_{[T_i \geq t]} = \sum_{i=1}^{n} \hat{H}_i \quad \text{and} \quad T^n = \max_i \{T_i\}.$$

Also, let

$$\Lambda_i^+(t) = -\int_{[0,t]} \frac{dH_i(s)}{H_i(s-)}, \qquad \Lambda_i^D(t) = -\int_{[0,t]} \frac{dF_i(s)}{F_i(s-)},$$

$$\Lambda^C(t) = -\int_{[0,t]} \frac{dG(s)}{G(s-)}.$$

It is well known that the three processes

$$\text{(2.2)} \qquad M_i^+(t) = I_{[T_i \leq t]} - \int_0^t I_{[T_i \geq s]}\, d\Lambda_i^+(s),$$

$$\text{(2.3)} \qquad M_i^D(t) = I_{[T_i \leq t;\, \delta_i = 1]} - \int_0^t I_{[T_i \geq s]}\, d\Lambda_i^D(s),$$

$$\text{(2.4)} \qquad M_i^C(t) = I_{[T_i \leq t;\, \delta_i = 0]} - \int_0^t I_{[T_i \geq s]}\, d\Lambda^C(s),$$

are square-integrable martingales on $[0, \infty]$ with respect to the filtration

$$\mathcal{F}_s = \sigma\{T_k I_{[T_k \leq s]};\, \delta_k I_{[T_k \leq s]};\, k = 1, 2, \ldots, n\}$$
$$= \sigma\{\text{everything observed up to time } s\}$$

and

$$(2.5)\quad \langle M_i^D\rangle(t) = \int_0^t I_{[T_i \geq s]}\, d\Lambda_i^D(s), \qquad \langle M_i^C\rangle(t) = \int_0^t I_{[T_i \geq s]}\, d\Lambda_i^C(s),$$

where $\langle M\rangle$ denotes the predictable variation process of the square-integrable martingale M. .

Clearly $M_i^+ = M_i^D + M_i^C$, since $\Lambda_i^+ = \Lambda_i^D + \Lambda^C$. Further, define

$$M_C^+ = \sum_{i=1}^n M_i^C.$$

The Kaplan–Meier estimator of the survival function $G(t)$ of the C_i's is given by

$$(2.6)\qquad \hat{G}(t) = \prod_{s \leq t \wedge T^n} \left\{1 - \frac{\Delta N^C(s)}{R^+(s)}\right\}.$$

Here $N^C(s) = \sum_{i=1}^n I_{[T_i \leq s,\, \delta_i = 0]}$ and $\Delta N^C(s) = N^C(s) - N^C(s-)$. Thus the processes

$$(2.7)\quad \frac{G(t) - \hat{G}(t)}{G(t)} = \int_0^{t \wedge T^n} \frac{\hat{G} -}{G}\, \frac{1}{R^+(s)}\, dM_C^+(s), \qquad t \in [0, T^n],$$

$$(2.8)\quad \frac{H_i(t) - \hat{H}_i(t)}{H_i(t)} = \int_0^t \frac{1}{H_i}\, dM_i^+(s), \quad \text{for } t \text{ such that } H_i(t) > 0,$$

are local martingales. For simplicity, we shall assume that $H_i(t) > 0$ for any $t < \infty$ in this paper, so that (2.8) holds on $[0, \infty)$.

LEMMA 2.1. *M_i^D and M_i^C are orthogonal martingales, that is, their predictable covariation process vanishes.*

PROOF. Since we have assumed continuity of the underlying distributions, this follows from Theorem 2.3.1 of Gill (1980). □

3. Main theorem. We first consider in detail the case of simple linear regression $E(Y_i|X_i) = \alpha + \beta X_i$. Recall that the Leurgans estimator is based on the synthetic data defined in (1.1). For simplicity we assume $T^n \geq 0$ for large n [otherwise make some changes in both (3.1) and (3.2) so that (3.3) below always holds]. It is not hard to see that in this case

$$(3.1)\qquad Y_i^* = \int_{-\infty}^{T^n} \left(\frac{I_{[T_i \geq s]}}{\hat{G}(s)} - I_{[s<0]} \right) ds.$$

The least squares estimators $\hat{\alpha}$ and $\hat{\beta}$ based on the synthetic data are [cf.

Leurgans (1987) or Zheng (1984)]

$$\hat{\beta} = \frac{\Sigma(X_i - \bar{X})Y_i^*}{\Sigma(X_i - \bar{X})^2}, \qquad \hat{\alpha} = \bar{Y}^* - \hat{\beta}\bar{X},$$

where $\bar{Y}^* = (1/n)\Sigma Y_i^*$, $\bar{X} = (1/n)\Sigma X_i$. Here and in the sequel, we suppress the index in the summation sign whenever there is only one subscript in the summand.

For easy comparison with Koul, Susarla and Van Ryzin's (1981) result, we adopt notation similar to theirs:

$$a_i = a_{ni} = \left(\frac{1}{n} - \frac{\bar{X}(X_i - \bar{X})}{\Sigma(X_j - \bar{X})^2}\right) \quad \text{and} \quad b_i = b_{ni} = \frac{X_i - \bar{X}}{\Sigma(X_j - \bar{X})^2}.$$

With this notation the estimators above are simply

$$\hat{\alpha} = \sum a_i Y_i^*, \qquad \hat{\beta} = \sum b_i Y_i^*.$$

For $T^n \geq 0$, the correct centering quantity turns out to be α^* and β^*, defined as follows

$$\alpha^* = \sum a_i \int_{-\infty}^{T^n} (F_i - I_{[t<0]})\, dt; \qquad \beta^* = \sum b_i \int_{-\infty}^{T^n} (F_i - I_{[t<0]})\, dt. \tag{3.2}$$

Notice that if we replace the upper limit of the integrals T^n by ∞ in (3.2) we will obtain exactly the expressions of α and β. In general it is impossible to center $\hat{\alpha}$, $\hat{\beta}$ by α, β without additional assumptions—see Remark 4.4 of Koul, Susarla and Van Ryzin (1981). Intuitively, though, it is clear that we can only hope to estimate the mean up to where the data are available, T^n.

REMARK 3.1. If we need to center the estimates by α and β, as we do in practice, we have to guarantee that the bias $\sqrt{n}(\alpha - \alpha^*) = o_p(1)$, $\sqrt{n}(\beta - \beta^*) = o_p(1)$. A set of sufficient conditions that will guarantee this is: (i) F_i has bounded mean residual life function; (ii) $G(t) > KF_i(t)^\beta$, $\forall i$ for some constants $K > 0$, $\beta < 1$; and (iii) either X_i are bounded or $\max_i(X_i) \to \infty$ with $G(\infty) = 0$. See also Remark 3.3 of Gill (1983) in the i.i.d. case. Notice that (i) is also implied by assumption (ii) of Theorem 3.1 [cf. (3.20)].

Thus the centered estimators are

$$\begin{aligned} \hat{\beta} - \beta^* &= \sum b_i \int_{-\infty}^{T^n} \left(\frac{I_{[T_i \geq t]}}{\hat{G}(t)} - F_i(t)\right) dt, \\ \hat{\alpha} - \alpha^* &= \sum a_i \int_{-\infty}^{T^n} \left(\frac{I_{[T_i \geq t]}}{\hat{G}(t)} - F_i(t)\right) dt. \end{aligned} \tag{3.3}$$

For easy formulation of the theorem, we assume $Y_i \geq 0$ from now on; thus the integration in (3.3) starts at 0 rather than $-\infty$. The general case can be treated similarly at a cost of some more conditions on the left tail of the distributions.

For technical simplicity, we restrict ourselves to the case of a random design; that is, we assume that the observable $\mathbf{X}_i$'s in the censored linear model are actually random, independent and identically distributed according to some distribution with a finite, nonzero variance. We also assume that $\mathbf{X}_i$'s are independent of everything else. Then all the definitions and proofs hold conditionally on the observed $\mathbf{X}_i$'s. The case of a fixed design can be treated using the same method, but the conditions are tedious to formulate and will be presented elsewhere. Also, the i.i.d. assumption of the ε_i's is not essential for the validity of the following theorem; we make this assumption for ease of presentation.

THEOREM 3.1. *For the simple linear model* $E(Y_i|X_i) = \alpha + \beta X_i$, *the synthetic data least squares estimates are asymptotically normally distributed, that is,*

$$\sqrt{n}\begin{pmatrix}\hat{\alpha} - \alpha^* \\ \hat{\beta} - \beta^*\end{pmatrix} \to_D N(0, \Sigma(\infty))$$

provided the following conditions are satisfied: (i) *the entries of the covariance matrix* $\Sigma(\tau) = (\sigma_{\alpha\beta}(\tau))$ [*see* (3.4) *below*] *are well defined for* $\tau \in [K, \infty]$ *for some* $K < \infty$ *and* $\sigma_{ij}(\tau) \to_{\tau\to\infty} \sigma_{ij}(\infty) < \infty$; (ii) $\sup_t E[\varepsilon_i - t|\varepsilon_i > t] < \infty$; (iii) *the tail conditions* (5.10) *and* (5.14) *hold*.

The components of the covariance matrix $\Sigma(\tau) = (\sigma_{\alpha\beta}(\tau))$ *are*

$$\begin{aligned}\sigma_{\alpha\alpha}(\tau) &= \sigma_{11}(\tau) \\ &= \lim n \sum a_i^2 \int_0^\tau \left[\int_t^\tau F_i\, ds\right]^2 \frac{d\Lambda_i^D}{H_i} \\ &\quad + \lim n \sum_{i=1}^n \int_0^\tau \left[\frac{\sum a_j \int_t^\tau F_j\, ds}{G\sum F_j} - \frac{a_k\left[\int_t^\tau F_i\, ds\right]}{H_i(t)}\right]^2 H_i\, d\Lambda^C,\end{aligned}$$

$$\begin{aligned}\sigma_{\alpha\beta}(\tau) &= \sigma_{12}(\tau) = \sigma_{21}(\tau) \\ &= \lim n \sum a_i b_i \int_0^\tau \left[\int_t^\tau F_i\, dx\right]^2 \frac{d\Lambda_i^D(t)}{H_i} \\ &\quad + \lim n \sum_{i=1}^n \int_0^\tau \prod_{c_i = a_i, b_i} \left[\frac{\sum c_j \int_t^\tau F_j\, ds}{G\sum F_j} - \frac{c_i \int_t^\tau F_i\, ds}{H_i}\right] H_i\, d\Lambda^C,\end{aligned} \tag{3.4}$$

$$\sigma_{\beta\beta}(\tau) = \sigma_{22}(\tau) = \textit{same as } \sigma_{11} \textit{ with } a_i, a_j \textit{ replaced by } b_i, b_j.$$

REMARK 3.2. By the random design assumption it is easy to see that $\bar{X}$, $(1/n)\sum g(X_i)$, $(1/n)\sum X_i g(X_i)$ and $(1/n)\sum X_i^2 g(X_i)$, where $g(\cdot)$ is a bounded real function, have limits as $n \to \infty$. This, in turn, guarantees that the above limiting covariance matrix exists for $\tau < \infty$, as can be seen by looking at the

alternative variance expression (4.1) and recalling the definition of b_i.

PROOF OF THEOREM 3.1. By (3.3), we have

$$\hat{\beta} - \beta^* = \sum b_i \int_0^{T^n} \left(\frac{I_{[T_i \geq t]}}{\hat{G}} - F_i \right) dt = \sum b_i \int_0^{T^n} \left[\frac{\hat{H}_i}{H_i} \frac{G}{\hat{G}} - 1 \right] F_i \, dt \tag{3.5}$$

[recalling that $H_i = F_i G$ and $\hat{H}_i(t) = I_{[T_i \geq t]}$]. Some simple algebra gives

$$\begin{aligned} \left[\frac{\hat{H}_i}{H_i} \frac{G}{\hat{G}} - 1 \right] &= \left[\left(1 + \frac{\hat{H}_i - H_i}{H_i} \right) \left(1 + \frac{G - \hat{G}}{\hat{G}} \right) - 1 \right] \\ &= \frac{\hat{H}_i(t) - H_i(t)}{H_i(t)} + \frac{G(t) - \hat{G}(t)}{\hat{G}(t)} + \frac{\hat{H}_i - H_i}{H_i} \frac{G - \hat{G}}{\hat{G}} \\ &= \frac{\hat{H}_i - H_i}{H_i} + \frac{G - \hat{G}}{G} + \frac{\hat{H}_i - H_i}{H_i} \frac{G - \hat{G}}{\hat{G}} + \frac{G - \hat{G}}{G} \frac{G - \hat{G}}{\hat{G}}. \end{aligned}$$

Plugging this back into (3.5), we obtain

$$\begin{aligned} \hat{\beta} - \beta^* &= \sum b_i \int_0^{T^n} \left(\frac{\hat{H}_i - H_i}{H_i} + \frac{G - \hat{G}}{G} \right) F_i \, dt + \text{a remainder term} \\ &= S_\beta(T^n) + SS_\beta(T^n) \quad \text{(say)}. \end{aligned} \tag{3.6}$$

In Section 5 we show that $\sqrt{n}\, SS_\beta(T^n)$ is negligible. It remains to establish the asymptotic normality of $\sqrt{n}\, S_\beta(T^n)$. Clearly,

$$S_\beta(T^n) = \int_0^{T^n} \sum b_i \frac{\hat{H}_i - H_i}{H_i} F_i \, dt + \int_0^{T^n} \frac{G - \hat{G}}{G} \sum b_i F_i \, dt. \tag{3.7}$$

We first show that $\sqrt{n}\, S_\beta(\tau)$ (replacing the upper limit T^n by an arbitrary but fixed constant $\tau < \infty$) is asymptotically normally distributed. Noticing that

$$F_i(t) \, dt = d\left[-\int_t^\tau F_i \, ds \right] \quad \text{for } t < \tau,$$

integration by parts yields [in view of (2.7) and (2.8)]

$$S_\beta(\tau) = \sum b_i \int_0^\tau \left[-\int_t^\tau F_i \, ds \right] \frac{dM_i^+(t)}{H_i} + \int_0^\tau \sum b_i \int_t^\tau F_i \, ds \frac{\hat{G}_-}{G} \frac{1}{R^+} \, dM_C^+(t).$$

Using the fact that $M_i^+(s) = M_i^D(s) + M_i^C(s)$ and $M_C^+(s) = \sum M_i^C(s)$ (cf. Section 2), we can show that

$$\begin{aligned} S_\beta(\tau) = \sum \int_0^\tau &- \left(\int_t^\tau b_i F_i \, ds \right) \frac{dM_i^D(t)}{H_i(t)} \\ &+ \sum_{i=1}^n \int_0^\tau \left[\left(\int_t^\tau \sum b_j F_j \, ds \right) \frac{\hat{G}_-}{G} \frac{1}{R^+} - \frac{\int_t^\tau b_i F_i \, ds}{H_i(t)} \right] dM_i^C(t). \end{aligned} \tag{3.8}$$

In very much the same way (with a_i instead of b_i) we can show that

$$\hat{\alpha} - \alpha^* = S_\alpha(T^n) + SS_\alpha(T^n) \quad \text{(say)}. \tag{3.9}$$

Again, using integrating by parts, $M_i^+ = M_i^C + M_i^D$ and $M_C^+ = \sum M_i^C$, we have, for $\tau < \infty$,

$$\begin{aligned} S_\alpha(\tau) = {} & \sum \int_0^\tau \left[-\int_t^\tau a_i F_i \, ds \right] \frac{dM_i^D(t)}{H_i(t)} \\ & + \sum_{i=1}^n \int_0^\tau \left[\left(\int_t^\tau \sum a_j F_j \, ds \right) \frac{\hat{G}_-}{G} \frac{1}{R^+} - \frac{\int_t^\tau a_i F_i \, ds}{H_i(t)} \right] dM_i^C(t). \end{aligned} \tag{3.10}$$

If we change the upper limit τ of the outermost integrals to $v \in [0, \tau]$ (but keep the τ in the inner integrals) in the two terms of (3.8) [or (3.10)], then each of them becomes a martingale in v for $v \in [0, \tau]$. Let us denote the martingales by $S_\beta(v)$ and $S_\alpha(v)$.

To establish the joint asymptotic normality of $S_\alpha(\tau)$ and $S_\beta(\tau)$, it suffices to show that the two martingales, $S_\alpha(v)$ and $S_\beta(v)$, converge jointly in $D[0, \tau]^2$. To prove this latter (and stronger) result, according to the martingale CLT, we need to show that

$$\langle \sqrt{n} S_\beta, \sqrt{n} S_\beta \rangle(v) \to_p G_\beta(v), \tag{3.11}$$

$$\langle \sqrt{n} S_\alpha, \sqrt{n} S_\alpha \rangle(v) \to_p G_\alpha(v), \qquad v \in [0, \tau], \tag{3.12}$$

$$\langle \sqrt{n} S_\alpha, \sqrt{n} S_\beta \rangle(v) \to_p G_{\alpha\beta}(v), \tag{3.13}$$

where G_α, G_β and $G_{\alpha\beta}$ are nonrandom continuous functions, and the following Lindeberg conditions hold:

$$\begin{aligned} \forall \varepsilon > 0, \qquad & \sum \int_0^\tau (*)^2 I_{[|*| > \varepsilon]} d\langle M_i^D \rangle(t) \\ & + \sum \int_0^\tau (\Delta)^2 I_{[|\Delta| > \varepsilon]} d\langle M_i^C \rangle(t) \to_p 0, \end{aligned} \tag{3.14}$$

where $*$ and Δ denote the two integrands in (3.8) and (3.10).

Let us now prove (3.11). First notice that by the independence assumption and Lemma 2.1, $M_1^D(t), \ldots, M_n^D(t)$, $M_1^C(t), \ldots, M_n^C(t)$ are mutually orthogonal martingales. Thus the predictable variation process $\langle \sqrt{n} S_\beta, \sqrt{n} S_\beta \rangle(v)$ is

$$\begin{aligned} & -n \sum \int_0^v \left(\int_t^\tau b_i F_i \, ds \right)^2 \frac{I_{[T_i \geq t]}}{[H_i(t)]^2} \frac{dF_i}{F_i} \\ & - n \sum_{i=1}^n \int_0^v \left[\left(\int_t^\tau \sum b_j F_j \, ds \right) \frac{\hat{G}_-}{G} \frac{1}{R^+} - \frac{\int_t^\tau b_i F_i \, ds}{H_i(t)} \right]^2 I_{[T_i \geq t]} \frac{dG}{G}. \end{aligned} \tag{3.15}$$

The first term above is easily seen to converge in probability to

$$-\lim_n n\sum b_i^2\int_0^v\left[\frac{\int_t^\tau F_i\,ds}{F_i}\right]^2\frac{dF_i}{G}$$

by assumption (i) and by the fact that the variance of this term tends to zero:

$$n^2\sum b_i^4\,\mathrm{Var}\left(\int_0^v\left(\int_t^\tau F_i\,ds\right)^2\frac{I_{[T_i\geq t]}}{[H_i(t)]^2}\frac{dF_i}{F_i}\right)$$

$$\leq n^2\sum b_i^4 E\left[\int_0^v\left(\int_t^\tau F_i\,ds\right)^2\frac{I_{[T_i\geq t]}}{[H_i(t)]^2}\frac{dF_i}{F_i}\right]^2$$

$$\leq n^2\sum b_i^4\sup_{0\leq t\leq\tau}\left[\frac{\int_t^\tau F_i\,ds}{H_i(t)}\right]^4 E\left(\int_0^v I_{[T_i\geq t]}\frac{dF_i}{F_i}\right)^2$$

$$\leq n^2\sum b_i^4\frac{K^4}{[G(\tau)]^4}E\left(\int_0^\tau I_{[T_i\geq t]}\frac{dF_i(t)}{F_i(t)}\right)^2 \tag{3.16}$$

$$\leq n\left(\max_i b_i^2\right)n\sum b_i^2\frac{2K^4}{[G(\tau)]^4}. \tag{3.17}$$

The reason for the inequality (3.17) is that

$$-\int_0^\tau I_{[T_i\geq t]}\frac{dF_i(t)}{F_i(t)}$$

is stochastically less than a unit exponential random variable and therefore the second moment is less than 2. The reason for (3.16) is essentially assumption (ii) and is spelled out in (3.20). Finally, (3.17) goes to zero since $n\max_i b_i^2$ does and the rest is bounded.

As for the second term of (3.15), by squaring out the big bracket and summing inside the integral we can show (by invoking Lemma 5.1) that it converges in probability to

$$-\lim_n n\sum\int_0^v\left[\frac{\sum b_j\int_t^\tau F_j\,ds}{\sum F_j(t)}-\frac{b_i\int_t^\tau F_i\,ds}{F_i(t)}\right]^2\frac{F_i}{G}\frac{dG}{G}.$$

Thus we verified (3.11) by showing that $\langle\sqrt{n}\,S_\beta,\sqrt{n}\,S_\beta\rangle(v)$ converges in probability to

$$G_\beta(v)=-\lim_n n\sum b_i^2\int_0^v\left[\frac{\int_t^\tau F_i\,ds}{F_i}\right]^2\frac{dF_i}{G}$$

$$-\lim_n n\sum\int_0^v\left[\frac{\sum b_j\int_t^\tau F_j\,ds}{\sum F_j(t)}-\frac{b_i\int_t^\tau F_i\,ds}{F_i(t)}\right]^2\frac{F_i}{G}\frac{dG}{G}.$$

Similarly, we can show (3.12) and (3.13) with

$$G_\alpha(v) = \lim n \sum a_i^2 \int_0^v \left[\int_t^\tau F_i\, ds\right]^2 \frac{d\Lambda_i^D}{H_i}$$

$$+ \lim n \sum_{i=1}^n \int_0^v \left[\frac{\sum a_j \int_t^\tau F_j\, ds}{G \sum F_j} - \frac{a_i \int_t^\tau F_i\, ds}{H_i(t)}\right]^2 H_i\, d\Lambda^C$$

and

$$G_{\alpha\beta}(v) = \lim n \sum_{i=1}^n a_i b_i \int_0^v \left[\int_t^\tau F_i\, ds\right]^2 \frac{d\Lambda_i^D(t)}{H_i}$$

$$+ \lim n \sum_{i=1}^n \int_0^v \prod_{c_i = a_i,\, b_i} \left[\sum c_j \int_t^\tau F_j\, ds \frac{1}{G \sum F_j} - \frac{c_i \int_t^\tau F_i\, ds}{H_i}\right] H_i\, d\Lambda^C.$$

Now, let us check the Lindeberg condition (3.14) for $S_\beta(v)$. We have to show that

$$\sum \int_0^\tau \left[\sqrt{n}\, b_i \frac{\int_t^\tau F_i\, ds}{H_i}\right]^2 I_{[|*|>\varepsilon]} I_{[T_i \geq t]}\, d\Lambda_i^D \to_p 0 \tag{3.18}$$

and

$$\sum_{i=1}^n \int_0^\tau \left[\left(\int_t^\tau \sum b_j F_j\, ds\right) \frac{\hat{G}_-}{G} \frac{\sqrt{n}}{R^+} - \frac{\sqrt{n} \int_t^\tau b_i F_i\, ds}{H_i(t)}\right]^2 I_{[|\Delta|>\varepsilon]} I_{[T_i \geq t]}\, d\Lambda^C \to_p 0. \tag{3.19}$$

By assumption (ii), $\sup_t E[\varepsilon_i - t | \varepsilon_i > t] < \infty$, we see that

$$E[Y_i - t | Y_i > t] = E[\varepsilon_i - (t - \alpha - \beta X_i) | \varepsilon_i > (t - \alpha - \beta X_i)]$$

is bounded, and thus

$$\frac{\int_t^\infty F_i\, ds}{F_i(t)} \leq K \qquad \forall t, \qquad \forall i, \tag{3.20}$$

which implies that

$$\sup_{i,\, t \in [0,\tau]} n b_i^2 \left[\frac{\int_t^\tau F_i\, ds}{H_i(t)}\right]^2 \leq \max_i n b_i^2 K^2 \frac{1}{[G(\tau)]^2} \to 0.$$

Hence, every integrand in (3.18) vanishes for sufficiently large n and (3.18) holds. For (3.19), first apply the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ to the big square bracket to get the upper bound

$$\sum_{i=1}^n \int_0^\tau \left[\frac{2 n b_i^2 \left[\int_t^\tau F_i\, ds\right]^2}{[H_i(t)]^2} + 2\left(\sum b_j \int_t^\tau F_j\, ds\right)^2 \left[\frac{\hat{G}_-}{G}\right]^2 \frac{n}{R^{+2}}\right] I_{[|\Delta|>\varepsilon]} I_{[T_i \geq t]}\, d\Lambda^C.$$

As we have just seen, the first term in the big square bracket above is tending

to zero. As for the second term, by the Schwarz inequality,

$$\left(\sum b_j \int_t^\tau F_j \, ds\right)^2 \left[\frac{\hat{G}_-}{G}\right]^2 \frac{n}{R^{+2}}$$

is bounded by

$$\begin{aligned}
&\sum b_j^2 \sum \left[\int_t^\tau F_j \, ds\right]^2 \left[\frac{\hat{G}_-}{G}\right]^2 \frac{1}{n(R^+/n)^2} \\
&\quad = \sum b_i^2 \frac{1}{n} \sum \left[\frac{\int_t^\tau F_j \, ds}{F_j(t)}\right]^2 [F_j(t)]^2 \left[\frac{\hat{G}_-}{G}\right]^2 \frac{1}{(R^+/n)^2} \\
&\quad \le K^2 \sum b_j^2 \frac{1}{n} \sum [F_j(t)]^2 \left[\frac{\hat{G}_-}{G}\right]^2 \frac{1}{(R^+/n)^2},
\end{aligned}$$

where the last inequality follows by (3.20). For $t \in [0, \tau]$, the above is bounded by

$$K^2 \sum b_j^2 \times 1 \times \left[\frac{1}{G(\tau)}\right]^2 \frac{n^2}{R^+(\tau)^2} = \frac{K^2 n \sum b_j^2}{[G(\tau)]^2} \frac{n}{[R^+(\tau)]^2}.$$

The first part of the above is clearly bounded and the second part goes to zero, so (3.19) holds.

For the Lindeberg condition on S_α, notice that the same proof, with b_i replaced by a_i, carries over. The conditions on b_i used in the proof, $\max_i n b_i^2 \to 0$ and $n \sum b_i^2$ bounded, are also valid with a_i. Thus, we have shown that $\sqrt{n}\, S_\alpha(\tau)$ and $\sqrt{n}\, S_\beta(\tau)$ are asymptotically jointly normal.

Now, similar to an argument in Section 5 [cf. (5.12) and (5.13)], $\forall \varepsilon > 0$, $P(\sqrt{n}\, |S_\alpha(T^n) - S_\alpha(\tau)| > \varepsilon)$ and $P(\sqrt{n}\, |S_\beta(T^n) - S_\beta(\tau)| > \varepsilon)$ can be made arbitrarily small when $n \to \infty$ by first choosing τ sufficiently large. Thus the convergence of $(\sqrt{n}\, S_\alpha(\tau), \sqrt{n}\, S_\beta(\tau))$ is still valid with τ replaced by T^n, by the fact that the distribution of a normal r.v. is continuous in its variance, $T^n \to_{\text{a.s.}} \infty$ and assumption (i). Thus the limiting normal distribution has covariance matrix (3.4). □

Clearly, estimation of the covariance matrix $\Sigma(\infty)$ is possible. Based on our alternative asymptotic variance formula (4.1) and Lemma 4.1, we suggest using

$$\begin{aligned}
&n \sum b_i^2 \left(\int_0^{T_i} \frac{dt}{\hat{G}(t)} - \hat{\alpha} - \hat{\beta} X_i\right)^2 \\
&\quad - n \int_0^\infty \frac{\left[\sum b_j \int_t^\infty \left(I_{[T_j \ge s]} / \hat{G}(s)\right) ds\right]^2}{R^+(t) - 1} \frac{dN^C(t)}{R^+(t)}
\end{aligned} \tag{3.21}$$

as an estimator of $\sigma_{22}(\infty)$. Estimators of $\sigma_{11}(\infty)$ and $\sigma_{12}(\infty)$ can be constructed similarly.

REMARK 3.3. To see that (3.21) is consistent, notice that

$$\int_0^{T_i} \frac{dt}{\hat{G}(t)} = \int_0^{T_i} \frac{dt}{G(t)} + o_p(1) \quad \text{and} \quad \hat{\alpha} + \hat{\beta} X_i = \alpha + \beta X_i + o_p(1).$$

Therefore, the first term of (3.21) equals

$$n \sum b_i^2 \left(\int_0^{T_i} \frac{dt}{G(t)} - \alpha - \beta X_i \right)^2 + o_p(1).$$

By the law of large numbers, this converges to the first term of (4.1). As for the consistency of the second term of (3.21), Lemma 5.1 gives that

$$\frac{n}{R^+(t) - 1} \quad \text{and} \quad \sum b_j I_{[\tilde{T}_j \geq s]}$$

are convergent (uniformly on compact intervals). The desired conclusion then follows from the consistency of the weighted Nelson–Aalen estimator:

$$\int W(t) \frac{dN^C(t)}{R^+(t)} \to \int W(t)\, d\Lambda^C(t).$$

where $W(t)$ is a continuous weight function.

The extension of the above theorem to multiple regression involves nothing new. The least squares estimate is given by

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}^*,$$

where $\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*)^T$, X is the design matrix and $(X^T X)^{-1}$ is assumed to exist.

It is clear that the estimate $\hat{\beta}_j$ is a weighted average of Y_i^*. In fact, $\hat{\beta}_j = \sum_i u_{ji} Y_i^*$, where u_{ji} are the jth row, ith column element of the matrix $(X^T X)^{-1} X^T$. Hence the technique used in proving Theorem 3.1 also works here.

THEOREM 3.2. *In the censored linear model with random design, suppose that*: (i) *the covariance matrix* $\Sigma(\tau)$ [*see* (3.22) *below*] *is well defined and finite for* $\tau \in [k, \infty]$ *and* $\Sigma(\tau) \to \Sigma(\infty)$ *as* $\tau \to \infty$; (ii) $\sup_t E[\varepsilon_i - t | \varepsilon_i > t] < \infty$; *and* (iii) (5.10) *and* (5.14) *hold with* $c_i = u_{ji}$, $j = 1, 2, \ldots, p$.

Then

$$\sqrt{n}(\hat{\beta} - \beta^*) \to_D N(0, \Sigma(\infty)) \quad \text{as } n \to \infty,$$

where $\beta^* = \{\beta_1^*, \ldots, \beta_p^*\}$ *with* $\beta_j^* = \sum_i u_{ji} \int_0^{T_n} F_i\, dt$ *and* $\Sigma(\tau) = (\sigma_{kl}(\tau))$ *with*

$$\begin{aligned} \sigma_{kl}(\tau) &= \lim n \sum_{i=1}^n u_{ki} u_{li} \int_0^\tau \left[\int_t^\tau F_i\, ds \right]^2 \frac{d\Lambda_i^D(t)}{H_i} \\ &\quad + \lim n \sum_{i=1}^n \int_0^\tau \prod_{c_i = u_{ki}, u_{li}} \left[\frac{\sum c_j \int_t^\tau F_j\, ds}{G \sum F_j} - \frac{c_k \int_t^\tau F_i\, ds}{H_i} \right] H_i\, d\Lambda^C. \end{aligned} \tag{3.22}$$

4. Discussion and comparison. In this section we take a closer look at the asymptotic variance of the synthetic data estimator derived in the previous section and compare it with that of the Koul, Susarla and Van Ryzin (1981) estimator.

We begin by rewriting our variance formula (3.4). This can be done in a way that makes the comparison of the two estimators easier, and offers another way of looking at the sources of variance. Specifically we shall focus on the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^*)$, namely σ_{22}. First expand the contents of the big square bracket in the second term of σ_{22} to get

$$-\lim n \sum \int_0^\infty \left[\left(\frac{\sum b_j \int_t^\infty F_j \, ds}{\sum F_j(t)} \right)^2 + \left(\frac{b_i \int_t^\infty F_i \, ds}{F_i(t)} \right)^2 - 2 \frac{\sum b_j \int_t^\infty F_j \, ds}{\sum F_j(t)} \frac{b_i \int_t^\infty F_i \, ds}{F_i(t)} \right] \frac{F_i \, dG}{G^2}.$$

By summing the above three terms, the first and third terms combine to give

$$-\lim_n n \sum \int_0^\infty \frac{b_i^2 \left[\int_t^\infty F_i \, ds \right]^2}{\left[H_i(t) \right]^2} F_i \, dG(t) + \lim_n n \int_0^\infty \frac{\left[\sum b_j \int_t^\infty F_j \, ds \right]^2}{\sum F_j} \frac{dG}{G^2}.$$

This, together with the first part of σ_{22}, yields

$$\begin{aligned} \sigma_{22} = & -\lim_n n \sum \int_0^\infty b_i^2 \left[\frac{\int_t^\infty F_i \, ds}{H_i} \right]^2 \{ G \, dF_i + F_i \, dG \} \\ & + \lim_n n \int_0^\infty \frac{\left[\sum b_j \int_t^\infty F_j \, ds \right]^2}{\sum F_j} \frac{dG}{G^2}. \end{aligned} \tag{4.1}$$

As the following lemma shows, the first term in (4.1) is just the limit of $\operatorname{Var}(\sqrt{n} \sum b_i Y_i^*) = n \sum b_i^2 \operatorname{Var}(Y_i^*)$, when the true G is used in (1.1). The second term, which is *negative*, represents the effect of replacing the true G by the Kaplan–Meier estimator $\hat{G}$.

LEMMA 4.1.

$$\operatorname{Var}\left(\int_0^{T_i} \frac{dt}{G(t)} \right) = -\int_0^\infty \left[\frac{\int_t^\infty F_i \, ds}{H_i} \right]^2 \{ G \, dF_i + F_i \, dG \}.$$

PROOF. The mean of the random variable $\int_0^{T_i} (dt/G(t))$ is, by Fubini,

$$E \int_0^{T_i} \frac{dt}{G(t)} = E \int_0^\infty \frac{I_{[T_i \geq t]} \, dt}{G(t)} = \int_0^\infty \frac{E I_{[T_i \geq t]}}{G(t)} \, dt = \int_0^\infty F_i \, dt.$$

Centering the random variable by its mean, we have

$$\begin{aligned} \int_0^{T_i} \frac{dt}{G} - \int_0^\infty F_i \, dt &= \int_0^\infty \frac{I_{[T_i \geq t]} \, dt}{G} - \int_0^\infty F_i \, dt \\ &= \int_0^\infty \left[\frac{I_{[T_i \geq t]}}{G F_i} - 1 \right] F_i \, dt \\ &= \int_0^\infty \left[\frac{\hat{H}_i}{H_i} - 1 \right] F_i \, dt = \int_0^\infty \left[\frac{\hat{H}_i - H_i}{H_i} \right] F_i \, dt. \end{aligned}$$

It is well known that the ratio in the square bracket is a local martingale. Integrating by parts, we get

$$\int_0^\infty \left(\int_t^\infty F_i\,ds\right) d\frac{\hat{H}_i - H_i}{H_i} = \int_0^\infty \left(\int_t^\infty F_i\,ds\right)\frac{1}{H_i}\,dM_i^+(t),$$

which has predictable variation process

$$\int_0^\infty \left(\int_t^\infty F_i\,ds\right)^2 \frac{I_{[T_i \geq t]}}{(H_i)^2}\,d\Lambda_i^+(t).$$

The expected value of the predictable variation process gives the desired variance. □

The Koul, Susarla and Van Ryzin (1981) estimator of β has an asymptotic variance formula [their (3.6)], which in our notation (when T_i are nonnegative) is

$$\sigma_{\mathrm{KSV}}^2 = \lim n\left[\sum b_i^2\,\mathrm{Var}\left(\frac{\delta_i T_i I_{[T_i \leq M_n]}}{G(T_i)}\right) + \int_0^{M_n} \frac{\left[\int_t^{M_n} s\,d\sum b_i F_i\right]^2}{\sum F_j}\,\frac{dG}{G^2}\right]. \tag{4.2}$$

It again consists of two terms, the second being negative. Notice that the negative term should have an extra n factor in (3.6) of their (1981) paper, [see Zhou (1989)]. Therefore their Remark 4.5 needs to be revised and in this connection see our Remark 4.1 below. The next lemma furnishes a comparison of the first part of the two variances, with M_n replaced by ∞ in (4.2).

LEMMA 4.2.

$$\mathrm{Var}\left(\int_0^{T_i} \frac{dt}{G(t)}\right) \leq \mathrm{Var}\left(\frac{\delta_i T_i}{G(T_i)}\right).$$

PROOF. Because the means of the two random variables are the same, as shown in the proof of our Lemma 4.1 and Koul, Susarla and Van Ryzin [(1981), (2.2)], it suffices to show that their second moments satisfy the same inequality.

To this end, we integrate by parts to get

$$E\left[\int_0^{T_i} \frac{dt}{G(t)}\right]^2 = \int_0^\infty \left[\int_0^t \frac{ds}{G(s)}\right]^2 dH_i = \int_0^\infty F_i 2\left(\int_0^t \frac{ds}{G(s)}\right) dt. \tag{4.3}$$

On the other hand,

$$E\left[\frac{\delta_i T_i}{G(T_i)}\right]^2 = \int_0^\infty \frac{t^2}{G(t)}\,dF_i \geq -\int_0^\infty t\int_0^t \frac{ds}{G(s)}\,dF_i,$$

since

$$\frac{t}{G(t)} = \int_0^t \frac{ds}{G(t)} \geq \int_0^t \frac{ds}{G(s)}.$$

Integrating by parts and using the latter inequality again, the right-hand side above becomes

$$\int_0^\infty F_i \left\{ \int_0^t \frac{ds}{G(s)}\, dt + \frac{t}{G(t)}\, dt \right\} \geq \int_0^\infty F_i 2 \left(\int_0^t \frac{ds}{G(s)} \right) dt$$

which is (4.3). □

We now compare the negative parts of the two variances. Again replace the M_n's in (4.2) by ∞. It is easy to see that the only difference is in the squared numerator in the integrand, namely

$$\text{(4.4)} \qquad \left[\sum b_i \int_t^\infty F_i\, ds \right]^2 \quad \text{versus} \quad \left[\sum b_i \int_t^\infty s\, dF_i \right]^2.$$

The two terms in (4.4) are both continuous functions of t and are equal at $t = 0$ $(= \beta^2)$ and $t = \infty$ $(= 0)$. Comparison of the two expressions in (4.4) is provided by the following lemma.

LEMMA 4.3. *If $t > 0$, we have*

$$\text{(4.5)} \qquad \left[\sum b_i \int_t^\infty F_i\, ds \right]^2 \leq \left[\sum b_i \int_t^\infty s\, dF_i \right]^2.$$

If $t < 0$, then

$$\text{(4.6)} \qquad \left[\sum b_i \int_t^\infty F_i\, ds \right]^2 \geq \left[\sum b_i \int_t^\infty s\, dF_i \right]^2.$$

PROOF. First, we integrate by parts to get

$$\text{(4.7)} \qquad -\int_t^\infty s\, dF_i(s) = tF_i(t) + \int_t^\infty F_i\, ds,$$

then proceed to show that both $\sum b_i F_i(t)$ and $\sum b_i \int_t^\infty F_i\, ds$ have the same sign as β.

Without loss of generality, assume $\beta > 0$. Notice that for any constant C, $\sum b_i C = 0$ which implies

$$\text{(4.8)} \qquad \sum b_i F_{\bar{X}}(t) = 0,$$

where $F_{\bar{X}}(t)$ is the survival function of $\alpha + \beta \bar{X} + \varepsilon$. Now let us compare $\sum b_i F_i(t)$ with (4.8). Since $X_i > \bar{X}$ when the b_i's are positive,

$$\alpha + \beta X_i > \alpha + \beta \bar{X} \quad \text{and} \quad F_i(t) \geq F_{\bar{X}}(t)$$

for positive b_i. For negative b_i's,

$$\alpha + \beta X_i < \alpha + \beta \bar{X} \quad \text{and} \quad F_i(t) \leq F_{\bar{X}}(t).$$

Thus the sum $\sum b_i F_i(t)$ inflates the positive terms and reduces the negative terms, as compared to the sum (4.8), so

$$\sum b_i F_i(t) \geq (4.8) = 0 \quad (\text{same sign as } \beta).$$

A similar argument yields

$$\sum b_i \int_t^\infty F_i \, ds \geq 0.$$

We have shown that the two sums always have the same sign as $\beta(> 0)$.

Now by summing (4.7),

$$-\sum b_i \int_t^\infty s \, dF_i(s) = t \sum b_i F_i(t) + \sum b_i \int_t^\infty F_i \, ds. \tag{4.9}$$

If $t > 0$, then

$$\left| \sum b_i \int_t^\infty s \, dF_i(s) \right| \geq \left| \sum b_i \int_t^\infty F_i \, ds \right|,$$

since the two terms on the right-hand side of (4.9) have the same sign. Thus, (4.5) holds. If $t < 0$, the two terms have opposite signs and (4.6) follows. □

From the above lemmas, it can be seen that the magnitude of the two negative terms will depend upon how $G(t)$ is related to the two terms in (4.4). For instance, if censoring only happens when $t > 0$, then clearly the absolute value of the negative term in the variance of the Koul, Susarla and Van Ryzin estimator (4.2) is larger than its counterpart in the Leurgans estimator (4.1). If, on the other hand, censoring only happens when $t < 0$, then the absolute value of the negative term in (4.2) is less than the corresponding term in (4.1) and (4.1) $\leq$ (4.2) in view of Lemma 4.2.

REMARK 4.1. In general, the negative terms [due to estimating $G(t)$] in the variances (4.1) and (4.2) are nonzero, since the respective integrands are nonnegative, continuous and equal to β^2 when $t = 0$. Notice that we have already taken into account the factor n in the front of (4.1) and (4.2); it cancels with $\sum F_j(0) = n$. This somewhat counterintuitive fact can be exploited to further reduce the variance in estimation of β. The idea is deliberately not to estimate $G(t)$ as well as we can, but only to base the estimator $\hat{G}$ on a subgroup of the data. For details see Fygenson and Zhou (1988).

5. Higher-order terms. In this section we show that the remainder terms in (3.6) and (3.8) are negligible. First, we prove a useful lemma on the uniform convergence of weighted empirical distribution functions in the case of nonidentically distributed random variables.

LEMMA 5.1. *Let X_i, $i = 1, \ldots, n$, be independent random variables with $P(X_i < t) = U_i(t)$ and f_{ni} be arbitrary constants then $\forall \varepsilon > 0$, for those n such that $\sum_{i=1}^n f_{ni}^2 \le (\varepsilon^2/2)$ we have*

$$P\left(\sup_t \left| \sum_{i=1}^n f_{ni}\left[I_{[X_i < t]} - U_i(t) \right] \right| > \varepsilon \right) \le 8(n+1)\exp\left[- \frac{\varepsilon^2}{32 \sum_{i=1}^n f_{ni}^2} \right]. \tag{5.1}$$

For the case where f_{ni} are functions of bounded variation in t with $V_{-\infty}^{\infty} f_{ni}(t) \le K < \infty$, K independent of n and i, then $\forall \varepsilon > 0$ and for all n such that $\sup_t \sum_{i=1}^n f_{ni}^2(t) \le (\varepsilon^2/2)$,

$$\begin{aligned} &P\left(\sup_t \left| \sum_{i=1}^n f_{ni}(t)\left[I_{[X_i < t]} - U_i(t) \right] \right| > \varepsilon \right) \\ &\qquad \le 8C_\varepsilon(n)\exp\left(- \frac{\varepsilon^2}{128 \sup_t \sum_{i=1}^n f_{ni}^2(t)} \right), \end{aligned} \tag{5.2}$$

where $C_\varepsilon(n) = (16K/\varepsilon)n^2 + n + 1$.

REMARK 5.1. Inequalities (5.1) and (5.2) are most useful when $\sum_{i=1}^n f_{ni}^2$ or $\sup_t \sum_{i=1}^n f_{ni}^2(t) = O(\log n^{-(1+\delta)})$ which make the right-hand side tend to zero.

PROOF. For the case where f_{ni} are constants, follow the argument of Pollard [(1984), pages 14–16].

If f_{ni} are functions of t, the above proof does not work but the symmetrization argument still carries though and leads to

$$\begin{aligned} &P\left(\sup_t \left| \sum_{i=1}^n f_{ni}(t)\left[I_{[X_i < t]} - U_i(t) \right] \right| > \varepsilon \right) \\ &\qquad \le 4P\left(\sup_t \left| \sum_{i=1}^n f_{ni}(t)\sigma_i I_{[X_i < t]} \right| > \frac{\varepsilon}{4} \right), \end{aligned} \tag{5.3}$$

where $\sigma_i = \pm 1$ with probability 1/2. Conditional on $X_1, \ldots, X_n$, $\sigma_i I_{[X_i < t]}$ does not vary with t whenever t is within two consecutive X_i's. The only variation of $\sum_{i=1}^n f_{ni}(t)\sigma_i I_{[X_i < t]}$ comes from $f_{ni}(t)$.

Since $f_{n1}(t)$ is of bounded variation, for any $\varepsilon > 0$, we can choose no more than $8(2K/\varepsilon)n$ points on the line such that $f_{n1}(t)$ varies by no more than $(\varepsilon/8n)$ in any of the intervals between consecutive points. Do the same thing with the other $f_{ni}(t)$ to get a total of no more than $n \times (16K/\varepsilon)n = (16K/\varepsilon)n^2$ points on the line. They form no more than $(16K/\varepsilon)n^2 + 1$ intervals.

Now within any of the refined intervals any one of the $f_{ni}(t)$ does not vary by more than $(\varepsilon/8n)$ and thus $\sum_{i=1}^n f_{ni}(t)$ does not vary by more than $\sum_{i=1}^n (\varepsilon/8n) = (\varepsilon/8)$ in any interval. We further refine the intervals by adding n points $X_1, \ldots, X_n$, making $I_{[X_i < t]}$ constant within each interval. Thus

$$\begin{aligned} P\left(\sup_t \left| \sum_{i=1}^n f_{ni}(t)\sigma_i I_{[X_i < t]} \right| > \frac{\varepsilon}{4} \right) &\le P\left(\max_{t_j} \left| \sum_{i=1}^n f_{ni}(t_j)\sigma_i I_{[X_i < t_j]} \right| > \frac{\varepsilon}{8} \right) \\ &\le \sum_{t_j} P\left(\left| \sum_{i=1}^n f_{ni}(t_j)\sigma_i I_{[X_i < t_j]} \right| > \frac{\varepsilon}{8} \right), \end{aligned} \tag{5.4}$$

where the points t_j are arbitrary except that there must be one in each interval. There are at most $(16K/\varepsilon)n^2 + n + 1)$ points t_j. Hoeffding's inequality [cf. Pollard (1984), pages 15–16] can be applied to finish the proof. □

There is a remainder term to be considered in connection with $\hat{\beta} - \beta^*$, as with $\hat{\alpha} - \alpha^*$ (cf. Section 3). We only treat the term connected with $\hat{\beta} - \beta^*$, the other being similar.

Recall from (3.6) that

$$
\begin{aligned}
\sqrt{n}\, SS_\beta(T^n) &= \sqrt{n} \sum b_i \int_0^{T_n} \frac{\hat{H}_i - H_i}{H_i} \frac{G - \hat{G}}{\hat{G}} F_i \, dt \\
&\quad + \sqrt{n} \sum b_i \int_0^{T_n} \frac{G - \hat{G}}{G} \frac{G - \hat{G}}{\hat{G}} F_i \, dt \\
&= \int_0^{T_n} \sum b_i \frac{\hat{H}_i - H_i}{G} \sqrt{n} \frac{G - \hat{G}}{G} \frac{G}{\hat{G}} \, dt \\
&\quad + \int_0^{T_n} \sqrt{n} \frac{G - \hat{G}}{G} \frac{G - \hat{G}}{\hat{G}} \sum b_i F_i \, dt.
\end{aligned} \tag{5.5}
$$

For any fixed $\tau < \infty$ it is easy to see that the integrals $\int_0^\tau (*)\, dt$, where $(*)$ is one of the integrands in (5.5), tend to zero as $n \to \infty$, since

$$
\sup_t \left| \sum b_i \left(H_i(t) - \hat{H}_i(t) \right) \right| \to_p 0 \quad \text{(Lemma 5.1)},
$$

$$
\sup_{t \le \tau} \left| \hat{G}(t) - G(t) \right| \to_p 0, \tag{5.6}
$$

$$
\sqrt{n} \frac{G(t) - \hat{G}(t)}{G(t)} \to_D B(C(t)) \quad \text{in space } D[0, \tau], \tag{5.7}
$$

and the functions $[G(t)]^{-1}$, $[\hat{G}(t)]^{-1}$ and $\sum b_i F_i(t)$ are bounded (in probability) for $t \le \tau$. The proofs of (5.6) and (5.7) are simple applications of Lenglart's inequality and the martingale central limit theorem. The fact that here the Y_i's are not identically distributed is of no consequence, as long as $(n/R^+(t))$ has a p-limit for $t \in [0, \tau]$, with $p - \lim(n/R^+(\tau)) < \infty$. The latter is in turn implied by Lemma 5.1 and i.i.d.-ness of the ε_i's and X_i's. The function $C(t)$ in (5.7) is then given by

$$
- \int_0^t \frac{1}{\lim(1/n) \sum F_i} \frac{dG}{G^2} = C(t). \tag{5.8}
$$

Thus, it only remains to show that the tails $\int_\tau^{T_n} (*)\, dt$ are asymptotically negligible (since $T^n \to \infty$ a.s., the event $\{T^n < \tau\}$ is asymptotically negligible).

To this end, first notice that

$$\sup_{t<T^n} \frac{G(t)}{\hat{G}(t)} \quad \text{and} \quad \sup_{t<T^n} \frac{G(t)-\hat{G}(t)}{\hat{G}(t)} \tag{5.9}$$

are bounded in probability [see Zhou (1991)].

Second, notice that the functions $\Sigma_1^n b_i F_i$ clearly have a limit as $n \to \infty$ (in fact, uniform in t). Denote this limit, with a little abuse of notation, by $\Sigma_1^\infty b_i F_i$. The following lemma can be proved.

LEMMA 5.2. *If*

$$\sup_n \int_0^\infty \sum_1^n c_i^2 \sum_1^n F_i^2 \, dC(t) < \infty, \qquad c_i = a_i \text{ or } b_i, \tag{5.10}$$

then

$$\sqrt{n}\,\frac{G(t)-\hat{G}(t)}{G(t)} \sum_{i=1}^n c_i F_i(t) \to_D B(C(t)) \sum_1^\infty c_i F_i(t), \quad \text{in } D[0,\infty], \tag{5.11}$$

where the function $C(t)$ is defined in (5.8).

PROOF. This is essentially Theorem 2.1 of Gill (1983) with some generalization of the weight function $h(t)$ [it takes the form $\Sigma_{i=1}^n c_i F_i(t)$ here] and the i.i.d. requirement. The fact that the Y_i's are not identically distributed here again poses no difficulties (we use van Zuijlen's inequality). The generalization of the weight function can be accomplished by writing it as the difference of two nonnegative, nonincreasing functions. Some of the details of this argument can be found in the proof of Theorem 2.4 of Zhou (1986). □

REMARK 5.2. Condition (5.10) basically requires that the tail of F_i be small compared to the tail of G. For instance, if $KF_i^\beta(t) \le G(t)$, $\forall i$ for some constants $K > 0$, $\beta < 1$, then (5.10) holds. See also Remark 3.1.

Let us deal with the tail part of the second term of (5.5) first. Because of (5.11),

$$\begin{aligned} &P\left(\int_\tau^{T^n} \left|\sqrt{n}\,\frac{G-\hat{G}}{G} \sum_1^n b_i F_i(t)\right| dt > \varepsilon\right) \\ &\quad \to P\left(\int_\tau^\infty \left|B(C(t)) \sum_1^\infty b_i F_i(t)\right| dt > \varepsilon\right) \\ &\quad \le \frac{1}{\varepsilon} E \int_\tau^\infty \left|B(C(t)) \sum b_i F_i(t)\right| dt. \end{aligned} \tag{5.12}$$

The latter can be made arbitrarily small, as a result of the assumption (5.14) below, by choosing a large τ. This together with (5.9) shows the tail part of the second term of (5.5) is negligible.

For the tail part of the first term of (5.5), notice that

$$
\begin{aligned}
&P\left(\int_{\tau}^{T^n}\left|\sqrt{n}\,\frac{\sum b_i\left[\hat{H}_i(t)-H_i(t)\right]}{G(t)}\right|dt>\varepsilon\right)\\
&\quad\le\frac{1}{\varepsilon}E\int_{\tau}^{\infty}\left|\sqrt{n}\,\frac{\sum b_i\left[\hat{H}_i(t)-H_i(t)\right]}{G(t)}\right|dt\\
&\quad\le\frac{1}{\varepsilon}\int_{\tau}^{\infty}\left\{E\left(\sqrt{n}\,\frac{\sum b_i\left[\hat{H}_i(t)-H_i(t)\right]}{G(t)}\right)^2\right\}^{1/2}dt\\
&\quad\le\frac{1}{\varepsilon}\int_{\tau}^{\infty}\left\{\frac{n\sum b_i^2F_i}{G(t)}\right\}^{1/2}dt.
\end{aligned}
\tag{5.13}
$$

Assume

$$
\begin{aligned}
&\int_0^{\infty}\left\{\frac{n\sum c_i^2F_i}{G(s)}\right\}^{1/2}ds\le M<\infty\quad\text{and}\\
&\int_0^{\infty}C^{1/2}(t)\left|\sum c_iF_i\right|dt<\infty\quad\text{for } c_i=a_i \text{ or } b_i.
\end{aligned}
\tag{5.14}
$$

Then (5.13) can be made arbitrarily small for large τ. In view of (5.9), this shows the tail of the first term of (5.5) is negligible.

Acknowledgments. I would like to thank Professor P. K. Sen for helpful discussions which lead to a simplification of the proof of Theorem 3.1 and comments by a referee which eliminated an unnecessary condition in Theorem 3.1. I also want to thank William Rayens and a referee for helping improve the presentation.

REFERENCES

AALEN, O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6** 701–726.

ANDERSEN, P. K. and BORGAN, O. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12** 97–158.

BILLINGSLEY, P. (1968). *Weak Convergence of Probability Measures*. Wiley, New York.

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

FYGENSON, M. and ZHOU, M. (1988). Analysis of linear regression models with censoring. Unpublished manuscript.

GILL, R. (1980). *Censoring and Stochastic Integrals. Math. Centre Tracts* **124**. Math. Centrum, Amsterdam.

GILL, R. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* **11** 49–58.

JACOBSEN, M. (1982). *Statistical Analysis of Counting Processes. Lecture Notes in Statist.* **12**. Springer, New York.

KALBFLEISCH, J. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

KAPLAN, E. and MEIER, P. (1958). Non-parametric estimator from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481.

KOUL, H., SUSARLA, V. and VAN RYZIN, J. (1981). Regression analysis with randomly right-censored data. *Ann. Statist.* **9** 1276–1288.

LEURGANS, S. (1987). Linear models, random censoring and synthetic data. *Biometrika* **74** 301–309.

MILLER, R. G. and HALPERN, J. (1982). Regression with censored data. *Biometrika* **69** 521–531.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

SHORACK, G. R. and WELLNER, J. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

SUSARLA, V. and VAN RYZIN, J. (1980). Large sample theory for an estimator of the mean survival time from censored samples. *Ann. Statist.* **8** 1002–1016.

VAN ZUIJLEN, M. C. A. (1978). Properties of the empirical distribution function for independent non-identically distributed random variables. *Ann. Probab.* **6** 250–266.

ZHENG, Z. (1984). Regression with randomly censored data. Ph.D. dissertation, Columbia Univ.

ZHOU, M. (1986). Some nonparametric two-sample tests with censored data. Ph.D. dissertation, Columbia Univ.

ZHOU, M. (1989). A new proof of CLT for the Koul–Susarla–Van Ryzin estimator. Mimeo Series 1770, Dept. Statistics, Univ. North Carolina.

ZHOU, M. (1991). Some properties of the Kaplan–Meier estimator for independent nonidentically distributed random variables. *Ann. Statist.* **19** 2266–2274.

DEPARTMENT OF STATISTICS
859 PATTERSON OFFICE TOWER
UNIVERSITY OF KENTUCKY
LEXINGTON, KENTUCKY 40506-0027