*The Annals of Statistics*
1997, Vol. 25, No. 6, 2665–2681

# ESTIMATIONS IN HOMOSCEDASTIC LINEAR REGRESSION MODELS WITH CENSORED DATA: AN EMPIRICAL PROCESS APPROACH

BY FUSHING HSIEH

*National Taiwan University*

Pertaining to the estimating equations proposed by Tsiatis, based on the linear rank test, we show the existence of local confounding between the baseline hazard function and the covariates. Due to the local confounding, an estimating equation in Tsiatis' family with a larger time point of truncation could contain less information about the regression parameter than the estimating equation with a smaller time point of truncation. This phenomenon further indicates significant loss of efficiency of Tsiatis' estimating equations as well as the power loss of log-rank type tests when the baseline hazard function goes up and down along the time scale. To take care of this local confounding without using nonparametric estimates of the derivative of the baseline hazard function, we propose the empirical process approach (EPA) based on an empirical process constructed from Tsiatis' log-rank estimating equation by varying its truncating time point. The EPA will provide very tractable estimations of the regression parameters as well as Pearson's chi-squared statistics for testing the model's assumptions. Specifically, the performance of the EPA estimator is shown to be very close to the best estimator in Tsiatis' family.

**1. Introduction.** The problem of estimating the regression coefficient in a linear model with censored data has received much attention in the statistical literature. Some earlier works of Miller (1976). Buckley and James (1979) and Koul, Susarla and Van Rayzin (1981) suggested estimators based on different modifications of the ordinary least squares method. Recently, Ritov (1990) generalized the method of Buckley and James (1979) and introduced a family of $M$-estimators, then Tsiatis (1990), building on some earlier works on the two-sample problem of Louis (1981) and Wei and Gail (1983), suggested a family of estimating equations based on linear rank tests. The asymptotic equivalence of the two types of estimators was shown in Ritov (1990). Lai and Ying (1992) and Ying (1993) improved upon the results of Tsiatis (1990) by eliminating the truncation previously imposed on estimating functions and by proving stronger results on uniform convergences in the development of asymptotic properties of the resulting estimators. Fygenson and Ritov (1994) studied a class of monotone rank-test-based estimating

Received October 1995; revised November 1996.
[1]Supported in part by the National Science Council of Taiwan.
*AMS* 1991 *subject classifications*. 62G05, 62G20.
*Key words and phrases.* Accelerated failure time model, log-rank test, local confounding, martingale central limit theorem, time-dependent covariates.

equations, which were also shown to be a subfamily of Tsiatis' estimating equations.

In both Ritov's and Tsiatis' families, the optimal estimating equations involve the unknown score function, the derivative of the logarithm of the baseline hazard function. Therefore, in the semiparametric setting, solving the estimating equations with an estimated score function for efficient estimates of the regression coefficient becomes a common practice. However, it is well known that this procedure suffers from severe computational difficulties, such as nonmonotonicity, sensitivity to different choices of smoothing parameters and many others.

To avoid these difficulties, we introduce a new inference approach in this paper called an empirical process approach (EPA) for estimating the regression parameter in a homoscedastic linear model with right-censored data. With the EPA we do not need nonparametric smoothed estimates either in point estimation or in interval estimation. Furthermore, as a by-product of the EPA, Pearson chi-square statistics can be constructed easily for a goodness-of-fit test of the semiparametric model. For an introductory account of the EPA, Hsieh and Turnbull (1996) discuss its applicability in many semiparametric models with very tractable inferences on both estimating and testing.

To motivate our EPA, in Section 2, we first show that the bias term pertaining to Tsiatis' log-rank estimating equation resulting from a small perturbation of the regression parameter reveals the existence of local confounding between the baseline hazard function and the covariates. The most important effect of this local confounding on members of Tsiatis' family is that an estimating equation with a larger truncated time point could contain less information about the regression parameter than the estimating equation with a smaller truncated time point. Heuristically, the log-rank test could lose some of its power right after the baseline hazard function achieves its local maximum or minimum. Hence, when the baseline hazard function goes up and down along its time scale $t$, the original log-rank statistic (without a truncation imposed) used in Ying (1993) could become very inefficient. Furthermore, members of Ritov's family as well as Fygenson and Ritov's (1994) estimator also suffer from this local confounding.

Under the setting of a homoscedastic regression model with censored data, in Section 3, we consider an empirical process defined as the martingale process constructed from Tsiatis' long-rank estimating function by varying its truncated time point $t$ from $-\infty$ to a prespecified constant $T^*$. The EPA approach is motivated as a way to take care of the local confounding by chopping this empirical process into several small pieces and then combining them into an approximated likelihood function constructed from its limiting Gaussian martingale process. By maximizing this likelihood function, our EPA estimator of the regression parameter is derived. The performance of the estimator is shown to be very close to the optimal estimator in Tsiatis' class. Therefore, we neither need to work with weighted log-rank estimating func-

tions nor to find the optimal weight function by using the unstable nonparametric score estimates.

In Section 4, we construct a Pearson chi-square statistic for testing the goodness of fit of the semiparametric linear model. Extension of the EPA approach to a multiple regression model is discussed in Section 5. In Section 6, we first discuss issues related to our EPA, such as what numerical methods and how many cutoff points to use. Then we suggest a practical algorithm and report results from a small simulation in which the EPA estimators are compared with estimators proposed in Fygenson and Ritov (1994) and in Tsiatis (1990). The effect of local confounding on the log-rank estimating equations is also evaluated. At the end further applications of our EPA are also mentioned.

**2. The local confounding.** In this section, we point out that local confounding between the baseline hazard function and the covariates exists in members of Tsiatis' family of estimating equations. Due to the local confounding, we conclude that local structures (especially the first-order derivative) of the baseline hazard function should be taken into account in order for members of Tsiatis' family to be able to accumulate information during the study or data collecting period. For expositional simplicity, we will follow the notation used in Tsiatis (1990).

Let $T_1, \ldots, T_N$ be a sequence of random variables, usually corresponding to responses of interest of $N$ subjects in a study. Ultimately, we wish to make inferences about the relationship between the response $T_i$ and another concomitant variable $Z_i$, say, through a system of homoscedastic linear equations

$$T_i = \beta^T Z_i + e_i, \qquad i = 1, \ldots, N, \tag{1}$$

where $\beta$ is a $p \times 1$ parameter vector and, conditional on $Z_i$, the $e_i$ are independent and identically distributed (i.i.d.) residual variables with a common distribution $F(x)$ and corresponding hazard function $\lambda(x) = -d \log S(x)/dx$, where $S(x)$ denotes the survival function $1 - F(x)$. Throughout this paper the distribution $F(x)$ is assumed to satisfy conditions (A) and (C) given in Section 3 and otherwise is unknown.

If $T_i$ in the model (1) represents the logarithms of survival times, then this model is often referred to as an accelerated failure time model. See Cox and Oakes (1984) and Kalbfleisch and Prentice (1980).

Since all of our developments are conditional on the variables $Z_i$, we shall assume that the $Z_i$ are nonrandom, and furthermore, satisfy conditions (D), (E) and (F) given in Section 3. For ease of exposition, we assume from now through Section 4 that $\beta$ is a scalar parameter.

The regression model (1) for survival data is often complicated by right-censoring; that is, the data that are observed will only consist of $N$ i.i.d. random vectors

$$(X_i, \Delta_i, Z_i), \qquad i = 1, \ldots, N, \tag{2}$$

where $X_i = \min(T_i, C_i)$ and

$$\Delta_i = \text{failure indicator} = \begin{cases} 1, & \text{if } T_i \le C_i, \\ 0, & \text{if } T_i > C_i. \end{cases}$$

In the accelerated failure time model, the random variables $C_i$ represent the logarithms of the censoring times. Here the $C_i$ are independent random variables whose distribution may depend on the covariates. To avoid any nonidentifiability problems, we shall further assume that $(T_i, C_i)$ are statistically independent. In this sense, the censoring scheme is noninformative.

For the regression model (1) with censored data (2), Tsiatis (1990) studied a family of estimating equations based on linear rank tests. Let $T^*$ be a prespecified constant satisfying condition (A) in the next section. Let $N_i(u)$ be the counting process for the $i$th individual, defined by

$$N_i(u) = \mathbf{I}(X_i \le u, \Delta_i = 1),$$

where $\mathbf{I}(A)$ denotes the indicator function for event $A$, and let

$$\bar{Z}(u, \beta) = \frac{\sum_{i=1}^N Z_i Y(u + \beta Z_i)}{\sum_{i=1}^N Y_i(u + \beta Z_i)},$$

where $Y_i(u) = \mathbf{I}\{X_i \ge u\}$. The estimating equation is

$$S_N(W_N, \beta) = \sum_{i=1}^N \int_{-\infty}^{T^*} W_N(u + \beta Z_i)\, dN_i(u + \beta Z_i)\{Z_i - \bar{Z}(u, \beta)\}, \tag{3}$$

where $W_N(u)$ denotes a prespecified weight function and an estimate of $\beta$ is defined as a solution of the equation

$$S_N(W_N, \beta) = 0.$$

It is known that, in general, solutions of the preceding equation are not unique.

Specifically, the member of Tsiatis' family with $W_N(u) \equiv 1$, that is, $S_N(1, \beta)$, is the well-known log-rank statistic truncated at $T^*$.

Consider the asymptotically linear approximation of $S_N(1, \beta)$ with $\beta$ near $\beta_0 = 0$ given in (3.4) and an approximation of (3.3) of Tsiatis (1990):

$$\begin{aligned} S_N(1, \beta) &\approx S_N(1, 0) \\ &\quad + \beta \int_{-\infty}^{T^*} \Big\{ \sum \big(Z_i - \bar{A}(u, \beta_0)\big)^2 Y_i(u + \beta_0 Z_i) \Big\} \lambda'(u)\, du \qquad (4) \\ &= S_N(1, 0) + \beta B(T^*, \beta_0), \quad \text{say.} \qquad (5) \end{aligned}$$

It is noted that the term $B(T^*, \beta_0)$ is not necessarily monotone with respect to $T^*$, since the first-order derivative of $\lambda(u)$ is involved. For example, if $T^*$ is a local extreme point of $\lambda$, that is, $\lambda'(T^*) = 0$, then the following inequality holds, for some positive $d$,

$$|B(T^*, \beta_0)| > |B(T^* + d, \beta_0)|, \tag{6}$$

where $|\cdot|$ denotes the absolute value.

Furthermore, we know that $n^{-1/2} S_N(1, 0)$ converges weakly to a Gaussian variable with its variance as an increasing function of $T^*$. Therefore, inequal-

ity (6) implies that the log-rank estimating equation with a larger truncation point may contain less information about $\beta$ than the estimating equation with a smaller truncating time point.

This phenomenon can also be understood from the following heuristics. Given $\beta_0 = 0$, then the rank of $T_i$ is independent of the rank of its covariate $Z_i$. However, if the baseline hazard function $\lambda(u)$ changes its pattern from increasing failure to decreasing failure at $T^*$, then the log-rank statistic will locally mistake the effect of decreasing failure as the effect of the covariate $Z$ on a small interval, such as $(T^*, T^* + d)$. This "false" local effect could be positive or negative depending on the particular permutation of the $Z_i$ corresponding to those subjects whose estimated residuals $Y_i - \beta Z_i$ are greater than $T^*$. Hence we call this phenomenon "local confounding" between the baseline hazard function $\lambda(u)$ and the covariates $Z$.

The significance of this local confounding is that it reveals the necessity of taking local structures of the baseline hazard function into a statistical inference approach in order to accumulate information about the regression parameter during the study or data collecting period. In view of this, the original log-rank estimating equation (without a truncation imposed) used in Ying (1993) could be very inefficient when the baseline hazard function goes up and down during the study of data collecting period.

One seemingly natural way to take care of the local confounding is to consider Tsiatis' estimating function $S_N(W_N, \beta)$ with the weight function $W_N$ being proportional to $\lambda'(u)$. In such a case, a nonparametric estimate of $\lambda'(u)$ is needed. However, it is known that this type of estimation has a very low rate of convergence and, in general, is rather sensitive to smoothing parameters. Furthermore, these undesirable properties will cause complications in computing as well as severe instability for the estimating equation with an estimated score function.

In the next section, we propose the empirical process approach (EPA) to take care of the local confounding without employing unstable nonparametric estimates of $\lambda'(u)$.

**3. The EPA approach for the homoscedastic linear model.** In view of the local confounding introduced previously, it is not only natural, but also necessary, to consider Tsiatis' log-rank estimating function truncated at all time points $t$, $t \in (-\infty, T^*]$, as an empirical process on the interval. Furthermore, in order to take care of the local confounding, it seems practically sufficient to chop this empirical process on the interval $(-\infty, T^*]$ into several small pieces. Hopefully, we can confine the effect of local confounding to just a few of those small pieces of the empirical process and extract almost all the information about $\beta$ available in those pieces on which the local confounding does not take place. To achieve this goal, we stitch all these small pieces together by combining them into an approximated likelihood function constructed from a limiting Gaussian martingale process. By maximizing this likelihood function, our EPA estimator of the regression parameter is derived. This is the idea behind our empirical process approach.

For expositional clarity, the truncation $T^*$ is used throughout this paper, as in Tsiatis (1990) and Andersen and Gill (1982). With this truncation we further avoid the technical difficulties due to possible tail instability treated in Ying (1993).

Now we list conditions (A)–(F) used in Tsiatis (1990).

(A) The density of the error term in model (1), $f(x) = dF(x)/dx$, exists and is bounded by $K_1$ for all $x \le T^* + \xi$, for some $\xi > 0$, and $\int_{-\infty}^{T^*+\xi} (f'(x)/f(x))^2 f(x)\,dx < \infty$ and $P(X_i - \beta Z_i \ge T^* + \xi) \ge \psi > 0$ for all $i$.

(B) The density of the censoring random variable $C_i$ is also uniformly bounded.

(C) There exists a function $\theta(u)$ such that

$$|\lambda(u + \varepsilon) - \lambda(u) - \varepsilon\lambda'(u)| \le \varepsilon^2\theta(u)$$

for $u \le T^*$ and $|\varepsilon| \le \xi$;

$$\int_{-\infty}^{T^*} |\theta(u)|\,du < \infty,$$

where $\lambda'(u) = d\lambda(u)/du$.

(D) The covariates are uniformly bounded, and without loss of generality we assume that $|Z_i| \le 1$ for all $i$.

(E) There exists a continuous function $\mu(u, \beta)$ for values $\beta$ in a neighborhood $B$ of $\beta_0$, such that

$$\sup_{\beta \in B,\, u \le T^* + \xi} \{|\bar{Z}(u, \beta) - \mu(u, \beta)|\} \to_p 0.$$

(F) There exists a continuous function $A(u, \beta)$ such that

$$\sup_{\beta \in B,\, u \le T^* + \xi} \left[ \left| N^{-1} \sum_{i=1}^{N} \{Z_i - \bar{Z}(u, \beta)\}^2 Y_i(u + \beta Z_i) - A(u; \beta) \right| \right] \to_p 0.$$

Now we define an empirical process on $(-\infty, T^*]$ as the normalizing log-rank statistic truncated at $t \in (-\infty, T^*]$, that is,

$$M(t; \beta) = N^{-1/2} \sum_{i=1}^{N} \int_{-\infty}^{t} dN_i(u + \beta Z_i)\{Z_i - \bar{Z}(u; \beta)\}, \qquad t \le T^*. \tag{7}$$

By applying Rebolledo's central limit theorem for a local square integrable martingale [see Appendix I in Andersen and Gill (1982)], the normalizing process $M(t; \beta_0)$ converges weakly to a continuous one-dimensional Gaussian martingale $W(t)$, say, with $W(-\infty) = 0$ and a covariance function $\operatorname{cov}(W(t), W(s)) = H(s \wedge t)$ for all $t, s \le T^*$ and

$$H(t) = \int_{-\infty}^{t} A(u; \beta_0)\lambda(u)\,du. \tag{8}$$

From the covariance function, we known that $W(t)$ has independent increments.

This weak convergence can be heuristically interpreted as meaning that the likelihood of the empirical process $M(t; \beta_0)$ is approximately equal to that based on the Gaussian martingale $W(t)$ on $(-\infty, T^*]$. This idea will be

employed in our EPA approach to construct an approximated likelihood function.

To set up our EPA inferences for $\beta$, we need the local asymptotic linearity uniform in $t \in (-\infty, T^*]$ given in the next theorem. This theorem is an extended version of Theorem 1 of Ying (1993). It follows directly from Lemmas 1, 3, 4, and 5 of Ying (1993) and the stochastic equicontinuity of the process $M(t; \beta_0)$ ensured by its weak convergence. Therefore, its proof is omitted here.

THEOREM 3.1. *Under conditions* (A)–(F), *the empirical process* $M(t; \beta)$ *is uniformly asymptotically linear on* $(-\infty, T^*]$ *in the sense that for every sequence* $d_n > 0$ *with* $d_n \to 0$ *a.s.*,

$$\sup\left\{|M(t;\beta) - M(t;\beta_0) - \sqrt{N}(\beta - \beta_0)g_0(t)|/\left(1 + \sqrt{N}|\beta - \beta_0|\right)\right\} = o(1) \quad a.s., \tag{9}$$

*where the* "sup" *is taken over* $-\infty \le t \le T^*$, $|\beta - \beta_0| \le d_n$ *and* $g_0(t) = \int_{-\infty}^{t} A(u; \beta_0)\lambda'(u)\,du$.

*Furthermore, for every sequence of* $\tilde{d}_n > 0$ *with* $\tilde{d}_n \to 0$ *a.s.*,

$$\sup\left\{|M(s;\beta) - M(t;\beta_0) - \sqrt{N}(\beta - \beta_0)g_0(t)|/\left(1 + \sqrt{N}|\beta - \beta_0|\right)\right\} = o(1) \quad a.s., \tag{10}$$

*where the "sup" is taken over* $|\beta - \beta_0| \le d_n$, $|s - t| \le \tilde{d}_n$, $-\infty \le t, s \le T^*$.

Theorem 3.1 is an improved version of Theorems 3.1 and 3.2 of Tsiatis (1990). And the use of the preceeding asymptotic linearities is twofold: (1) to motivate the EPA estimate of $\beta$ as a generalized partial likelihood estimator (see Remark 1); (2) to derive the consistency and asymptotic normality for our EPA estimator.

To define our EPA estimator, we shall use the following vector notation: let $-\infty < t_1 < t_2 \cdots < t_k\ (= T^*)$,

$$\mathbf{t} = (t_1, \ldots, t_k)^T,$$

$$U(\mathbf{t}) = \left(U(t_1), \ldots, U(t_k)\right)^T,$$

$$\Delta U(\mathbf{t}) = (\Delta_1 U, \ldots, \Delta_k U)^T, \qquad \Delta_i U = U(t_i) - U(t_{i-1}),$$

where $U$ can be any function or empirical process in $t$. For example, $M(\mathbf{t}; \beta) = (M(t_1; \beta), \ldots, M(t_k; \beta))^T$.

By ignoring the approximation error in (9), we have a system of $k$ regression equations as

$$M(\mathbf{t}; \beta) = M(\mathbf{t}; \beta_0) + N^{1/2}(\beta - \beta_0)g_0(\mathbf{t})$$

or, equivalently,

$$\Delta M(\mathbf{t}; \beta) = \Delta M(\mathbf{t}; \beta_0) + N^{1/2}(\beta - \beta_0)\Delta g_0(\mathbf{t}). \tag{11}$$

From the weak convergence of $M(t; \beta_0)$, we have

$$\Delta M(\mathbf{t}; \beta_0) \approx_d N\left(0, \Sigma_k^0\right), \tag{12}$$

where $\Sigma_k^0$ is $k \times k$ diagonal matrix with diagonal vector $\Delta H(\mathbf{t})$. Hence we have a regression setup in (11) with approximated normal errors.

Furthermore, we need an initial estimator for $\beta$. It is desirable that this estimator is not only root-$N$ consistent, but also robust to parameter $\lambda(u)$ and its estimates. Here we propose the following initial estimator $\hat{\beta}_0$, say,

$$\hat{\beta}_0 = \arg\inf_{\beta} [\Delta M(\mathbf{t}, \beta)]^T [\Delta M(\mathbf{t}, \beta)].$$

Clearly, $\hat{\beta}_0$ is robust in the previous sense, since its definition is independent of $\lambda$. And its root-$N$ consistency is ensured by Theorem 3.1. One other choice of initial estimator is the one proposed in Fygenson and Ritov (1994).

With $\hat{\beta}_0$, the function $H(t)$ can be estimated consistently by

$$\hat{H}(t) = \int_{-\infty}^{t} \hat{A}\left(u, \hat{\beta}_0\right) d\hat{\Lambda}(u), \tag{13}$$

where $\hat{A}$ is the empirical sum of squares of $Z$'s that are at risk on a time scale of $\{X_i - \hat{\beta}_0 Z_i\}$, that is,

$$\hat{A}(u, \beta) = \frac{1}{N} \sum_{i=1}^{N} \left\{Z_i - \bar{Z}(u; \beta)\right\}^2 Y_i(u + \beta Z_i)$$

and $\hat{\Lambda}$ is the Nelson estimate of the cumulative hazard function $\Lambda(t) = \int_{-\infty}^{t} \lambda(u)\, du$ based on calculated residuals $\{X_i - \hat{\beta}_0 Z_i\}$, that is,

$$\hat{\Lambda}(t) = \sum_{i=1}^{N} \int_{-\epsilon}^{t} \frac{dN_i\left(u + \hat{\beta}_0 Z_i\right)}{\sum_{j=1}^{N} Y_j\left(u + \hat{\beta}_0 Z_j\right)}.$$

With $\hat{H}(t)$ given in (13), our EPA estimator of $\beta$ is defined as

$$\begin{aligned} \hat{\beta}_k &= \arg\inf_{\beta} [\Delta M(\mathbf{t}, \beta)]^T \left(\hat{\Sigma}_k^0\right)^{-1} \Delta M(\mathbf{t}, \beta) \\ &= \arg\inf_{\beta} L_k(\beta), \quad \text{say,} \end{aligned}$$

where $\hat{\Sigma}_k^0$ is the plug in estimate of $\Sigma_k^0$.

In practice, we might like to iterate the preceding construction once or twice for better finite sample properties. We call these iterative estimates EPA estimators as well.

It is noted that the log-likelihood function based on (12) is equal to $(-2)\{\log|\Sigma_k^0| + L_k(\beta)\}$. Furthermore, by using the asymptotic linearity in (11) and taking the derivative of $L_k(\beta)$ with respect to $\beta$, we arrive at the unfeasible generalized least squares (GLS) estimator

$$\hat{\beta}_* = \text{the solution of } \Delta g_0(\mathbf{t}) \left(\Sigma_k^0\right)^{-1} \Delta M(\mathbf{t}, \beta) = 0. \tag{14}$$

It is clear that both estimates $\hat{\beta}_k$ and $\hat{\beta}_*$ are asymptotically equivalent. Therefore, we have the following approximated equation:

$$N^{1/2}\left(\hat{\beta} - \beta_0\right) = \left\{ \sum_{i=1}^{k} (\Delta_i g_0)^2 / \Delta_i H \right\}^{-1} \sum_{i=1}^{k} \Delta_i g_0 \{\Delta_i M(t; \beta)\} / \Delta_i H. \tag{15}$$

The variance of this normalized error $N^{1/2}(\hat{\beta} - \beta_0)$ is denoted by $\sigma_k^2$ and

$$\sigma_k^{-2} = \sum_{i=1}^{N} \frac{(\Delta_i g_0)^2}{\Delta_i H} \tag{16}$$

$$= \sum_{i=1}^{k} \frac{\left(\int_{t_{i-1}}^{t_i} A(u, \beta_0)\lambda'(u)\, du\right)^2}{\int_{t_{i-1}}^{t_i} A(u, \beta_0)\lambda(u)\, du}$$

$$\approx \int_{-\infty}^{T^*} A(u, \beta_0) \frac{(\lambda'(u))^2}{\lambda(u)}\, du. \tag{17}$$

The last approximation holds when the vector $\mathbf{t}$ becomes dense on $(-\infty, T^*]$ at a slow enough rate as $N \to \infty$.

The preceding argument gives a sketchy proof of the following theorem which summarizes the asymptotic properties of our EPA estimators.

THEOREM 3.2. *Under the conditions of Theorem* 3.1 *and with a fixed choice of* $\mathbf{t}$, *we have*

$$\sqrt{N}\left(\hat{\beta}_k - \beta\right) \approx_d N\left(0, \sigma_k^2\right),$$

*and an approximate confidence region for* $\beta$ *is given as*

$$\Omega_\alpha(\beta_0) = \left\{\beta \mid L_k(\beta) \le \chi_k^2(\alpha)\right\},$$

*where* $\chi_k^2(\alpha)$ *is the specified* $\alpha$*-percentage point of the chi-square distribution with degrees of freedom* $k$.

It should be emphasized here again that the point as well as the interval estimations derived from the EPA make no use of nonparametric smoothed estimated of $\lambda(u)$ or its derivative.

From (17) we see that this EPA estimator can perform very nearly as well as the optimal estimator in Tsiatis' class, that is, the estimator solving equation $S_N(W_N, \beta) = 0$ with $W_N(u) = \lambda'(u)/\lambda(u)$. Furthermore, if the set of cutoff points becomes dense on the support of $F$ and the truncation $T^*$ is adaptively chosen to tend to $\infty$ at a slow enough rate as sample size $N$ tends to $\infty$, then the asymptotic variance of our EPA estimator given in Theorem 3.2 will achieve the Fisher information bound given in Ritov and Wellner (1988). The issue of the asymptotic efficiency of the EPA estimators will be further discussed in Section 6.

A consistent estimate of $\sigma_k^{-2}$ can be constructed as follows:

$$\sigma_k^{-2} = \sum_{i=1}^{N} \frac{(\Delta_i \hat{g}_0)^2}{\Delta_i \hat{H}},$$

where

$$\Delta_i \hat{g}_0 = \int_{t_{i-1}}^{t_i} \hat{A}\left(u, \hat{\beta}_k\right)\hat{\lambda}'(u)\, du$$

and $\hat{\lambda}'$ is the kernel smoothed estimate of the derivative of the hazard function proposed in Section 3.3 of Ramlau-Hansen (1983). For example,

$$\hat{\lambda}'(t) = \frac{1}{b_N^2} \int_0^1 K_B'(u)\, d\hat{\Lambda}(u),$$

where $K_B'(u)$ is the first derivative of the biweight kernel $K_B(u) = 15/16 \cdot (1 - u^2)^2$, $-1 \le u \le 1$, and the bandwidth $b_N = b_0 N^{-1/5}$. A suitable constant $b_0$ can be found according to the stability of the estimate $\hat{g}_0(t)$ shown on the graphic display. See also Section 4.2 of Andersen, Borgan, Gill and Keiding (1992) for a discussion of the optimal bandwidth.

REMARK 1 (Generalized partial likelihood function). It shall be noted here that, instead of using the maximum likelihood approach based on the full likelihood function of the whole process $W(t) + N^{1/2}(\beta - \beta_0)g_0(t)$, $t \in (\infty, T^*]$, we use only its finite-dimensional approximation. One reason is that the full likelihood function of this continuous-time process is, in general, very difficult to compute explicitly, while its finite-dimensional approximation is much more feasible. Slud (1992) gave a very nice discussion on this issue and termed the latter likelihood as the generalized partial likelihood function. Therefore, based on (12), our EPA estimator is also asymptotically equivalent to the maximum generalized partial likelihood estimator. The other reason is that the (strong or weak) approximation of the empirical process $M(t; \beta_0)$ or $M(t; \beta)$ usually involves a remainder term of small order $N$. This also prevents us from using the full likelihood. See Section 6 for further discussion.

**4. Testing the semiparametric model assumption.** In this section, we briefly discuss how a Pearson chi-square statistic can be easily constructed as a by-product of the EPA for testing the semiparametric model assumption in (1). This convenient feature is shared with the minimum chi-square estimate advocated by Joseph Berkson in parametric models [see Berkson (1980) and the references therein]. However, in the recent literature related to semiparametric models, most existing approaches are only devoted to estimating rather than goodness-of-fit testing.

Recall the definition of the empirical process $M(t; \beta)$:

$$M(t; \beta) = N^{-1/2} \sum_{i=1}^{N} \int_{-\infty}^{t} dN_i(u + \beta Z_i)\{Z_i - \bar{Z}(u; \beta)\}, \qquad t \le T^*. \tag{18}$$

It is noted that this empirical process involves only $\beta$, not the unknown baseline hazard function $\lambda(u)$. Therefore, it is a natural basis for testing the semiparametric model assumption in (1).

Here we propose to use $L_k(\hat{\beta}_k)$ as a Pearson chi-square statistic for testing the goodness-of-fit of model (1). The asymptotic distribution of this testing

statistic is derived from the following approximations:

$$L_k(\beta_0) \approx L_k\big(\hat{\beta}_k\big) + \big(\hat{\beta}_k - \beta_0\big)^2 \big[\Delta \hat{g}_0(\mathbf{t})\big]^T \big(\hat{\Sigma}_k^0\big)^{-1} \big[\Delta \hat{g}_0(\mathbf{t})\big],$$

$$0 \approx \big[\Delta \hat{g}_0(\mathbf{t})\big]^T \big(\hat{\Sigma}_k^0\big)^{-1} \Delta M\big(\mathbf{t}, \hat{\beta}_k\big).$$

The next theorem summarizes that this statistic is approximately distributed as $\chi^2_{k-1}$.

THEOREM 4.1. *Under the conditions of Theorem* 3.2, *let* $\hat{\beta}_k$ *be the minimizer of* $L_k(\beta)$. *Then* $L_k(\hat{\beta})$ *is approximately distributed as the chi-square with degrees of freedom* $k - 1$, *that is,* $\chi^2_{k-1}$.

This simple construction of the Pearson chi-square statistic for testing the semiparametric model is an important advantage of our EPA. It should have some practical value in applied statistics.

**5. Extensions to multiple covariates.** Up to now, we assumed that $\beta$ is a scale parameter. In this section, we extend the EPA to multiple regression models.

For ease of exposition, we will keep the time scale $t \in R^1$ and let $v$, $s \in R^p$. And we take $\mathbf{U}(t) = \mathbf{U}(v)$ with $v = (t, \ldots, t)^T$ for any $p$-dimensional process $\mathbf{U}(v)$ and $(s \wedge v) = (s_1 \wedge v_1, \ldots, s_p \wedge v_p)$.

Suppose that the underlying responses $T_i$ are linearly related to covariates $Z_i \in R^p$. That is, $\beta \in R^p$ and

$$T_i = \beta^T Z_i + e_i.$$

As discussed in Section 3, we construct a system of $p$ empirical processes as follows. Let

$$\mathbf{M}(t; \beta) = \big(M_1(t; \beta), \ldots, M_p(t; \beta)\big)^T,$$

where

$$M_j(t; \beta) = N^{-1/2} \sum_{i=1}^{N} \int_{-\infty}^{t} R_{ij}(u, \beta)\, dN_i(u + \beta Z_i),$$

with

$$R_{ij}(u, \beta) = Z_{ij} - \bar{Z}_j(u, \beta), \qquad j = 1, \ldots, p,$$

$$\bar{Y}(u, \beta) = \sum_{i=1}^{N} Y_i\big(u + \beta^T Z_i\big),$$

$$\bar{\bar{Z}}_j(u, \beta) = \sum_{i=1}^{N} Z_{ij} Y_i\big(u + \beta^T Z_i\big) / \bar{Y}(u, \beta).$$

Then, by using Rebolledo's central limit theorem, the system of empirical processes $\mathbf{M}(v; \beta_0)$ converges weakly to a $p$-variate Gaussian martingale $\mathbf{W}(v)$, say, that is,

$$\mathbf{M}(v; \beta) \to_d \mathbf{W}(v) \quad \text{as } N \to \infty \text{ in } D\big((-\infty, t^*]^p\big).$$

The Gaussian martingale $\mathbf{W}(v)$ has its $p \times p$-matrix of covariance functions

$$\operatorname{cov}(\mathbf{W}(s), \mathbf{W}(v)) = H(s \wedge v),$$

with

$$H_{l,m}(t) = \int_{-\infty}^{t} A_{lm}(u, \beta)\lambda(u)\, du, \qquad l, m = 1, \ldots, p,$$

and $A_{lm}(u, \beta)$ defined as the limit of

$$\frac{1}{N}\sum_{i=1}^{N} R_{il}(u, \beta)R_{im}(u, \beta)Y_i\big(u + \beta^T Z_i\big).$$

Similarly, the local asymptotic linearity uniform in time scale $t$ given in Theorem 3.1 can be established here and applied to ensure the $\sqrt{N}$-consistency of the initial estimates and the asymptotic normality of our EPA estimators, which are likewise constructed as in Section 3. For testing the semiparametric model assumption of this multiple regression model, the Pearson chi-square statistic can be likewise derived as in Section 4. Results similar to Theorems 3.2 and 4.1 can also be established.

**6. A simulation study and discussion.** In this section, we discuss several issues related to the EPA and its further applications. First, the numerical method used in this paper to find our EPA estimates is the grid search. The feasibility of this simple method is very much enhanced by modern computer technologies, such as a graphics display, among others, and by our numerical experience which indicates the approximated likelihood function $L_k(\beta)$ in Section 3 having a unique minimum with very high probability if $k \geq 2$.

The next issue is how to choose $k$. We focus on this issue in the following way: if $k$ is allowed to depend on $N$, what is the fastest rate for $k$ such that the result in Theorem 3.2 remains valid? To facilitate the discussion of this issue, we assume the following two strong approximations hold without proof.

First, in view of the strong approximation results of product-limit and empirical cumulative hazard processes given in Theorems 1 and 2 in Burke, Csörgő and Horváth (1988) and the results in Section 3 of Koning (1994), we expect that, under suitable tail conditions on $F$ and $G_i$ [see condition A3 of Hsieh (1996a)], the following strong approximation of the empirical process $M_1(t, \beta_0)$ also holds:

$$M_1(t, \beta_0) = W(t) + R_N^{(1)}(t) \quad \text{a.s.}, \qquad t \in (-\infty, T^*], \tag{19}$$

where $R_N^{(1)}(t) = O(N^{-1/4} \log N)$.

This strong approximation is similar to the one for the classic empirical process derived in Kolmós, Major and Tusnády (1975). A more precise description should be based on Dudley's almost sure representation and described in terms of the sequence of Gaussian processes on a new probability space [see Pollard (1990)].

Second, the refinement on Theorem 1 in Ying (1993) should lead to the following strong approximation: for $\beta$ is a small neighborhood of $\beta_0$,

$$M_1(t, \beta) = M_1(t, \beta_0) + (\beta - \beta_0) g_0(t) + R_N^{(2)}(t) \quad \text{a.s.}, \tag{20}$$

where $g_0(t)$ is defined in Section 3, and $R_N^{(2)}(t)$ has the same order as $R_N^{(1)}(t)$.

With (19) and (20), we further obtain a signal plus noise model with resolution error:

$$M_1(t, \beta) = \tilde{W}(t) + \beta g_0(t) + R_N^*(t) \quad \text{a.s.}, \qquad t \in (-\infty, T^*], \tag{21}$$

where $\tilde{W}$ is a Gaussian martingale with covariance function $\text{cov}(\tilde{W}(t), \tilde{W}(s)) = \hat{H}(t \wedge s)$, and $R_N^*(t) = O(N^{-1/4} \log N)$ is called the resolution error which dominates the fastest rate of $k$ allowed.

By some standard calculations based on (21), it can be found that a chosen rate of $k$ such that the result of Theorem 3.2 remains valid has to satisfy

$$k^2/N \to 0 \quad \text{as } N \to \infty.$$

Otherwise the bias term incurred by our EPA estimate will overwhelm the approximating normal component and make the result of Theorem 3.2 invalid. A similar condition is required in Portnoy (1988) for a valid normal approximation in a model with the number of parameters increasing with the sample size. Furthermore, under the preceding condition and taking $T^* = O(\log N)$, we can show that the EPA estimators achieve the semiparametric Fisher information bound given in Ritov and Wellner (1988) [see Theorem 3.3 in Hsieh (1996a)].

Although the previous condition gives $N^{-1/2}$ as an upper bound on the rate of $k$, this asymptotic result is still not practical enough to lead to a choice of $k$. Here we suggest a practical algorithm for choosing a $k$ and a $k$-vector of cutoff time points $\mathbf{t}$.

The algorithm is:

1. Choose a suitable truncation point $T^*$ and a time point $t_0$ near the median of uncensored $Y_i$. Take $\mathbf{t} = \{t_0\}$.
2. Calculate the initial estimate $\hat{\beta}_0$ with $\mathbf{t}$ and then the Nelson estimate $\hat{\Lambda}(u)$ and its corresponding estimate $\hat{F}(u) = 1 - \exp\{-\hat{\Lambda}(u)\}$.
3. Choose a $k$ and a new $\mathbf{t}$ such that the $k$-vector consists of time points where $\hat{F}(u)$ has equal increments $1/\{k + 1\}$ and each cell contains at least 20 uncensored data.
4. Iterate steps 2 and 3 once.
5. Calculate the EPA estimate and iterate it once.

In Table 1, we report results from a small simulation study estimating the regression parameter $\beta$ in the homoscedastic model (1) using Tsiatis', Fygenson and Ritov's and our own EPA estimates. Our EPA estimates $\hat{\beta}_i$, $i = 3, 5, 7$ are respectively constructed with $k = 3$, 5 and 7 and time points where $\hat{F}(u)$ has equal increments $1/\{k + 1\}$. For convenience, Fygenson and Ritov's estimate $\hat{\beta}^{\text{FR}}$ is also used as the initial estimate in the constructions of our EPA estimates. Three of Tsiatis' estimates $\hat{\beta}_i^T$, $i = 0, 1, \infty$, say, are con-

TABLE 1
*Simulation study results: estimates of the regression parameters $\beta$ (with MSEs in parentheses) for cases with log-Weibull, normal, Cauchy and $F_{\mathrm{MIX}}$ errors using the EPA, Tsiatis' and Fygenson and Rotov's estimators*

| Estimate | log-Weibull | Normal | Cauchy | $F_{\mathrm{MIX}}$ |
|---|---|---|---|---|
| Uncensored % | 90.8 | 85.7 | 70.6 | 100 |
| $\hat{\beta}_3$ | 1.995486 (0.001198) | 1.988992 (0.018708) | 1.980305 (0.047785) | 1.987473 (0.079400) |
| $\hat{\beta}_5$ | 1.998961 (0.001225) | 1.994882 (0.019408) | 1.980805 (0.046526) | 1.974928 (0.098660) |
| $\hat{\beta}_7$ | 1.998716 (0.001176) | 1.994432 (0.020245) | 1.979625 (0.042797) | 1.972414 0.100563) |
| $\overline{\beta}_\infty^T$ | 2.000154 (0.000648) | 2.003696 (0.017670) | 1.973794 (0.066995) | 1.984986 (0.298983) |
| $\overline{\beta}_1^T$ | 2.000154 (0.000648) | 1.999696 (0.017955) | 1.984619 (0.053678) | 2.043327 (0.132160) |
| $\hat{\beta}_0^T$ | 1.999989 (0.000666) | 1.997916 (0.017716) | 1.957784 (0.062479) | 2.021653 (0.088977) |
| $\hat{\beta}^{\mathrm{FR}}$ | 1.999306 (0.000943) | 1.999367 (0.013923) | 1.981315 (0.053956) | 1.991312 (0.099489) |

structed from log-rank estimating equations with truncation at times 0, 1 and $\infty$. These three estimators of Tsiatis' family will numerically bring out the effect of local confounding in the fourth case considered later where the baseline hazard function goes up and down.

This simulation study consists of 200 repetitions of the log-Weibull, normal and Cauchy cases, and consists of 1000 repetitions of the fourth case in which the baseline hazard function is defined as

$$\lambda_{\mathrm{MIX}}(t) = \begin{cases} \exp t/(1+\exp t), & \text{if } t \le 0, \\ 1/2(1+t), & \text{if } t > 0. \end{cases}$$

Let $F_{\mathrm{MIX}}$ denote its corresponding distribution. This hazard function has its maximum at time 0 and $F_{\mathrm{MIX}}(0) = 1/2$.

In each repetition, we have sample size $N = 100$, $\beta_0 = 2$ and the $Z_i$ are generated from Uniform$(0, 3)$. For the first three cases, the error variables $\varepsilon_i$ are distributed as log-Weibull$(2, 2)$, $N(0, 1)$ and Cauchy$(0, 1)$. In each case, the censoring variables $C_i$ are independently generated according to the distribution of $c + 2Z_i + \varepsilon_i$, where the constant $c$ is chosen to adjust the censoring proportions. In the fourth case no censoring is considered.

From Table 1, we can see that our EPA estimators are comparable with Tsiatis' in the log-Weibull and normal cases, while performing better in the Cauchy case. It is noted that $\hat{\beta}_\infty^T$ is the optimal estimator in the log-Weibull case. For the fourth case, we can clearly see the effect of local confounding on

$\hat{\beta}_1^T$ and the even more signficant effect on $\hat{\beta}_\infty^T$ [the original log-rank estimator used in Ying (1993)].

We then briefly comment on the issue of the finite truncation $T^*$. In view of redistribute-to-the-right, the estimate $\hat{F}$, given in step 3 of the preceding algorithm seems able to indicate which $T^*$ could be a reasonable choice. Furthermore, the finite truncation issue is in part due to the martingale structure employed here. In fact, we can make statistical inferences under the setting considered here without applying the martingale structure, and then this issue disappears or is relieved to some extent. For example, Hsieh and Hsu (1996) developed the EPA approach employing the theory of classic empirical processes and having Brownian bridges as limiting Gaussian processes under the same models (1), but with complete data. There is no need to impose a finite truncation $T^*$ there.

Further applications of the EPA approach are for regression models with complicated incomplete data, such as data from biased sampling. However, in general, it is not likely to have martingale structures with these types of data. The EPA together with the method of sieves, remains applicable. In a separate report, our EPA is shown to have several advantages over the inference based on maximizing the profiled likelihood obtained by plugging in a solution of the self-consistent equation of $F$.

The EPA approach can also be applied to the accelerated failure model with time-dependent covariates. In a separate report, the author discussed the local confounding issue pertaining to the estimating equation proposed in Robins and Tsiatis (1992) and derived an asymptotically efficient estimator via the EPA based on a system of two empirical processes.

At the end of the paper, it is worth mentioning that, in a separate report Hsieh (1996b), the EPA is applied to a censored regression model with heteroscedasticity, that is,

$$T_i = \beta^T Z_i + \sigma_i e_i, \tag{22}$$

where $\sigma_i$ expresses the heteroscedasticity, for example, $\sigma_i = \exp\{\eta V_i\}$ [see Bickel (1978)], and the covariate $V_i$ is assumed to be linearly independent of $Z_i$.

These types of models are important in making predictions. A reason is that if we ignore the heteroscedasticity and apply the original log-rank estimating equation, then biased estimations of $\beta$ will result. The related two-sample problem was previously discussed in Hsieh (1996a) based on an empirical $Q$–$Q$ plot. Furthermore, there is no corresponding version of Fygenson and Ritov's estimator (1994) in the model (22), since the presence of heteroscedasticity will make the Hodges–Lehmann-type estimators invalid.

**Acknowledgments.** The author wishes to thank Professor L. D. Brown and the referee for helpful remarks and suggestions that led to a considerable improvement in presentation.

## REFERENCES


ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (1992). *Statistical Models Based on Counting Processes*. Springer, New York.

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

BERKSON, J. (1980). Minimum chi-square, not maximum likelihood! *Ann. Statist.* **8** 547–469.

BICKEL, P. J. (1978). Using residuals robustly. I: tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266–291.

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

BURKE, M. D., CSÖRGŐ, S. and HORVÁTH, L. (1988). A correction to and improvement of strong approximations of some biometric estimates under random censorship. *Probab. Theory Related Fields* **79** 51–57.

CARROLL, R. J. and RUPPERT, D. (1988). *Transformation and Weighting in Regression*. Chapman and Hall, London.

COX, D. R. and OAKES, D. (1984). *Analysis of Survival Data*. Chapman and Hall, London.

FYGENSON, F. and RITOV, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.* **22** 732–746.

HSIEH, F. (1996a). Empirical process approach in two-sample location-scale model with censored data. *Ann. Statist.* **24** 2705–2719.

HSIEH, F. (1996b). Empirical process approach (EPA) in heteroscedastic regression models with right-censored data. Unpublished manuscript.

HSIEH, F. and HSU, C. (1996). Empirical process approach in heteroscedastic linear model with symmetric error. Unpublished manuscript.

HSIEH, F. and TURNBULL, B. W. (1996). The empirical process approach for general semiparametric regression models: theory and applications. Unpublished lecture notes.

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

KOLMÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sum of independent R.V.'s and sample D.F. I. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.

KONING, A. (1994). Approximation of the basic martingale. *Ann. Statist.* **22** 565–579.

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

LAI, T. L. and YING, Z. (1992). Linear rank statistics in regression analysis with censored or truncated data. *J. Multivariate Anal.* **40** 13–45.

LOUIS, T. A. (1981). Nonparametric analysis of an accelerated failure time model. *Biometrika* **68** 381–390.

MILLER, R. (1976). Least squares regression with censored data. *Biometrika* **63** 449–464.

POLLARD, D. (1990). *Empirical Processes*: *Theory and Applications*. IMS, Hayward, CA.

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential family when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366.

RAMLAU-HANSEN, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11** 453–466.

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

RITOV, Y. and WELLNER, J. A. (1988). Censoring, martingales and the Cox model. *Contemp. Math.* **80** 191–220.

ROBINS, J. and TSIATIS, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika* **79** 311–319.

SLUD, E. V. (1992). Partial likelihood of continuous-time stochastic processes. *Scand. J. Statist.* **19** 97–109.

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

WEI, L. J. and GAIL, M. H. (1983). Nonparametric estimation for a scale-change with censored observations. *J. Amer. Statist. Assoc.* **78** 382–388.

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

DEPARTMENT OF MATHEMATICS
NATIONAL TAIWAN UNIVERSITY
TAIPEI
TAIWAN
E-MAIL: fushing@math.ntu.tw