CONFIDENCE INTERVALS FOR SURVIVAL QUANTILES IN THE COX REGRESSION MODEL

by

Tze Leung Lai
Zheng Su

December 2005

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065
CONFIDENCE INTERVALS FOR SURVIVAL QUANTILES IN THE COX REGRESSION MODEL

by

Tze Leung Lai
Department of Statistics
Stanford University

Zheng Su
Department of Health Research & Policy
Stanford University

Technical Reports No. 2005-27
December 2005

This research was supported in part by the National Science Foundation grant DMS-0305749 and National Institute of Health 5R01-CA088890

Department of Statistics
STANFORD UNIVERSITY
Stanford, California 94305-4065

http://www-stat.stanford.edu
Confidence intervals for survival quantiles
in the Cox regression model

TZE LEUNG LAI*

Department of Statistics, Stanford University, Stanford, CA 94305, USA
lait@stat.stanford.edu

ZHENG SU

Division of Biostatistics, Department of Health Research & Policy,
Stanford University, Stanford, CA 94305, USA
zhengsu@stanford.edu

SUMMARY

Median survival times and their associated confidence intervals are often used to summarize the survival outcome of a group of patients in clinical trials with failure-time endpoints. Although there is an extensive literature on this topic for the case in which the patients come from a homogeneous population, few papers have dealt with the case in which covariates are present as in the proportional hazards model. In this paper we propose a new approach to this problem and demonstrate its advantages over existing methods, not only for the proportional hazards model but also for the widely studied cases where covariates are absent and where there is no censoring. As an illustration, we apply it to the Stanford Heart Transplant data. Asymptotic theory and simulation studies show that the proposed method indeed yields confidence intervals and bands with accurate coverage errors.

Keywords: Bootstrap; Median survival; Proportional hazards model; Test-based confidence intervals.

* To whom correspondence should be addressed.
1. INTRODUCTION

The proportional hazards model of Cox (1972) is a log-linear regression model that relates the cumulative hazard function $\Lambda(t|x)$ associated with a covariate vector $x$ to a baseline hazard function $\Lambda(t)$ via

$$\Lambda(t|x) = \Lambda(t) \exp(\beta^T x).$$

(1.1)

Based on a sample consisting of $n$ observations $(\tilde{t}_i, \delta_i, x_i)$, where $\tilde{t}_i = \min(t_i, c_i)$ and $\delta_i = I(t_i \leq c_i)$ is the indicator of whether the actual failure time $t_i$ is observed or is censored by $c_i$, the estimate $\hat{\beta}$ of $\beta$ is the maximizer of the partial likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T x_i - \log \left( \sum_{j: t_j \geq \tilde{t}_i} \exp(\beta^T x_j) \right) \right\}.$$

(1.2)

Confidence regions for $\beta$ can be constructed by using the asymptotic normality of $(-\ell(\hat{\beta}))^{-1/2}(\hat{\beta} - \beta)$ or the limiting $\chi^2$ distribution of $\ell(\hat{\beta}) - \ell(\beta)$. In many applications, it is useful to estimate also the median survival time given a subject’s covariate vector. In particular, by combining $\hat{\beta}$ with Breslow’s (1974) estimate $\hat{\Lambda}$ of the baseline cumulative hazard function, Miller and Halpern (1982) used the median of the distribution function $1 - \exp\{-\hat{\Lambda}(\cdot)e^{\beta^T x}\}$ to estimate median survival, given a subject’s age that forms the covariate vector $x = (\text{age}, \text{age}^2)$, from the Stanford Heart Transplant data. Dabrowska and Doksum (1987) and Burr and Doss (1993) subsequently studied the problem of constructing confidence intervals for the median survival time given a subject’s covariates. Letting $\xi_p(x)$ denote the $p$th quantile of the failure time distribution for a given covariate vector $x$ (so that $p = \frac{1}{2}$ corresponds to the median) and $\hat{\xi}_p(x)$ be the $p$th quantile of the preceding estimated distribution function, their approach is based on the approximate normality of $\{\hat{\xi}_p(x) - \xi_p(x)\}/\hat{se}_p(x)$, where $\hat{se}_p(x)$ denotes the estimated standard error of $\hat{\xi}_p(x)$.

A major difficulty with this approach for sample sizes commonly encountered in practice lies in $\hat{se}_p(x)$. The variance of the limiting normal distribution of $\sqrt{n}(\hat{\xi}_p(x) - \xi_p(x))$ involves the baseline hazard function $\lambda(t) = (d/dt)\Lambda(t)$. Although Dabrowska and Doksum (1987, p.802) cite Tsiatis (1981) and Andersen and Gill (1982) in claiming consistency of their proposed estimator of the limiting variance, Tsiatis, Anderson and Gill have only established consistency for Breslow’s estimate of $\Lambda$ but not of the derivative $\lambda$. Burr and Doss (1993) make use of kernel smoothing of $\hat{\Lambda}$ to estimate $\lambda$, and instead of applying the large-sample theory of $\{\hat{\xi}_p(x) - \xi_p(x)\}/\hat{se}_p(x)$ directly to construct confidence intervals for
\( \xi_p(x) \), they use it to provide a theoretical justification of the bootstrap-t method to construct confidence intervals. However, as pointed out by Efron and Tibshirani (1993, Section 12.6), the bootstrap-t method requires stable estimates of standard errors for it to work well in practice. Therefore the difficulties in estimating the standard error of \( \hat{\xi}_p(x) \) also cause difficulties with bootstrap-t confidence intervals for \( \xi_p(x) \).

In fact, even without censoring and covariate effects so that the problem reduces to that of confidence intervals for the \( p \)th quantile \( \xi_p \) of a distribution function based on a sample of independent and identically distributed survival times \( t_1, \ldots, t_n \) with common density function \( f \) that has a consistent kernel estimator \( \hat{f} \), the limiting normal distribution of

\[
\hat{f}(\hat{\xi}_p)(n/[p(1-p)])^{1/2}(\hat{\xi}_p - \xi_p)
\]

is seldom used in constructing confidence intervals for \( \xi_p \). Besides issues with finite-sample performance of the density estimator \( \hat{f}(\hat{\xi}_p) \), the adequacy of the linear approximation \( f(\xi_p)(\hat{\xi}_p - \xi_p) \) to \( F(\hat{\xi}_p) - F(\xi_p) \) used to derive the asymptotic normality of \( \hat{\xi}_p - \xi_p \) (where \( F \) is the distribution function whose derivative is \( f \)) is problematic when \( \hat{\xi}_p \) is not sufficiently near \( \xi_p \). Instead, a standard nonparametric confidence interval is of the form \( t_{(k_1)} < \xi_p < t_{(k_2)} \), where the \( t_{(i)} \) denote the order statistics of the sample and \( k_1 \) and \( k_2 \) are respectively the smallest and largest integers such that

\[
P\{t_{(k_1)} \leq \xi_p < t_{(k_2)}\} = P\{k_1 \leq \text{Bi}(n, p) < k_2\} \geq 1 - 2\alpha,
\]

the lower bound of (1.4) may not be attainable because of the discreteness of the binomial distribution \( \text{Bi}(n, p) \). As shown by Efron (1979) and Chen and Hall (1993, p.1169) respectively, bootstrap percentile confidence intervals and empirical likelihood confidence intervals (obtained by inverting empirical likelihood ratio tests) for \( \xi_p \) are of this form; see also Efron and Tibshirani (1993, p.174). Chen and Hall (1993) also showed that the inability of (1.4) to attain \( 1 - 2\alpha \) with an \( O(n^{-1}) \) error due to the discreteness of the binomial distribution can be overcome by using a smoothed version of empirical likelihood. An alternative method to achieve a coverage probability of \( 1 - 2\alpha + O(n^{-1}) \) was proposed by Beran and Hall (1993) who used convex combinations of sample quantiles to develop interpolated confidence intervals. Recently Ho and Lee (2005) made use of smoothed bootstrap iterations to achieve more accurate one-sided coverage errors of the bootstrap percentile interval. Their method, however, is very computationally intensive and involves an additional layer of bootstrapping to determine the bandwidth used to smooth the empirical distribution.

For censored survival data without covariates, Li, Hollander, McKeague and Yang (1996) made use of empirical likelihood to construct confidence bands for \( \xi_p \), jointly in \( p_1 \leq p \leq p_2 \).
Their results on coverage probabilities are based on weak convergence and do not provide convergence rates of the kind in Chen and Hall (1993). They have, however, not smoothed the empirical likelihood function, nor have they compared the empirical likelihood approach with other test-based methods to construct confidence intervals for $\xi_p$ when the $t_i$ are subject to censoring. These alternative test-based intervals date back to Brookmeyer and Crowley (1982) who invert a generalized sign test, leading to an approximate $1 - 2\alpha$ confidence interval of the form

$$\{t : |\hat{S}(t) - 1/2| \leq z_{1-\alpha} \hat{\sigma}(t)\}$$

(1.5)

for the median $\xi_{1/2}$, where $\hat{S}(t)$ is the Kaplan-Meier estimator of the survival function, $\hat{\sigma}(t)$ is the estimated standard error of $\hat{S}(t)$ and $z_q$ denotes the $q$th quantile of the standard normal distribution. Instead of using the normal approximation, Strawderman, Parzen and Wells (1997) used Edgeworth expansions for the Studentized cumulative hazard function to derive more accurate test-based confidence limits for $\xi_p$.

In this paper we develop a new method to construct confidence intervals and confidence bands for the quantile $\xi_p(x)$ in the proportional hazards model (1.1). Unlike the methods of Dabrowska and Doksum (1987) and Burr and Doss (1993) that use $\{\xi_p(x) - \xi_p(x)\}/\hat{\sigma}_p$ as an approximate pivot, we use a test-based approach, using $\hat{\Lambda}(t|x)$ to test if $\Lambda(t|x) = \log(p^{-1})$, where $\hat{\Lambda}(t|x) = \hat{\Lambda}(t) \exp(\hat{\beta}^T x)$ and $\hat{\Lambda}(t)$ is Breslow’s estimator of the baseline cumulative hazard function $\Lambda(t)$. Instead of using the normal approximation or its second-order refinement as in Strawderman, Parzen and Wells (1997) to find the quantiles of the test statistic, we use the bootstrap method to evaluate the quantiles of an approximate pivot obtained by Studentizing the test statistic. The details are described in Section 2 which also extends this approach to confidence bands for $\xi_p(x)$, jointly in $x$ belonging to some given set $\mathcal{K}$. The advantages of the proposed procedure are demonstrated in the asymptotic theory in Section 2 and the simulation studies in Section 3. In Section 4 we apply the proposed methods to construct confidence intervals and confidence bands for median survival given a patient’s covariates from the Stanford Heart Transplant data, and compare our results with those of Burr and Doss (1993). Section 5 concludes with some remarks and further discussion.

2. METHODOLOGY

2.1 A new test-based bootstrap confidence interval

In this section we propose a new test-based confidence interval for the $p$th quantile $\xi_p(x)$ given a subject’s covariate vector $x$ in the Cox model, and provide an associated algorithm
for computing the endpoints of the interval. An obvious generalization of the Brookmeyer-Crowley confidence interval (1.5) for $\xi_{i/2}$ to $\xi_p(x)$ in the Cox model is
\[ \{ t : |\widehat{S}(t|x) - (1 - p)| \leq z_{1-a} \widehat{\sigma}(t|x) \}, \]
where $\widehat{\sigma}(t|x)$ is the asymptotic variance of
\[ \widehat{S}(t|x) = \exp\{-\widehat{\lambda}(t)e^{\beta^T x}\}, \]
in which $\widehat{\beta}$ is the maximizer of (1.2) and $\widehat{\Lambda}$ is Breslow's (1974) estimate of the baseline cumulative hazard function based on $(\check{t}_i, \delta_i, x_i), 1 \leq i \leq n$. The asymptotic variance formula was derived by Tsiatis (1981) using the delta method; see (2.6) below for its consistent estimate $\check{\sigma}(t|x)$. Note that this asymptotic variance is a nonlinear function of the asymptotic covariance matrix of $(\widehat{\Lambda}(t) - \Lambda(t), (\widehat{\beta} - \beta)^T x)$. Although $\widehat{S}(t|x)$ takes values in $[0, 1]$, $(\widehat{\beta} - \beta)^T x$ does not have such constraints and its variance in finite samples can be substantial. Moreover, the normal approximation to $|\widehat{S}(t|x) - S(t|x)|/\check{\sigma}(t|x)$ used in (2.1) may be inadequate when the sample size is not large enough; in particular, its symmetry about $S(t|x)$ fails to incorporate skewness that is especially relevant for censored data.

Instead of using $\widehat{S}(t|x) - (1 - p)$ as the test statistic, we use the logarithmic transformation to transform it into $\widehat{\lambda}(t|x) - \log(1 - p)^{-1}$. An advantage of this transformation is that unlike $\widehat{S}(t|x)$, $\widehat{\lambda}(t|x)$ is no longer constrained to belong to $[0, 1]$ and therefore the variability due to $(\widehat{\beta} - \beta)^T x$ in its asymptotic variance formula can be more compatible with its magnitude. Another advantage is that the asymptotic variance of $\widehat{S}(t|x)$ involves further linear approximation around $\check{\lambda}(t|x)$. In fact, after deriving the asymptotic variance $\nu(t|x)$ of
\[ \check{\lambda}(t|x) = \check{\lambda}(t) \exp(\beta^T x) \]
from the asymptotic covariance matrix of $(\check{\Lambda}(t) - \Lambda(t), (\check{\beta} - \beta)^T x)$, Tsiatis (1981) used it to derive the asymptotic variance of $\widehat{S}(t|x)$ via the nonlinear transformation $\widehat{S}(t|x) = e^{-\check{\lambda}(t|x)}$.

Letting $x_i = (x_{i1}, \ldots, x_{ik})^T$ and $a = (a_1, \ldots, a_k)^T$, define
\[ W(t) = \sum \exp(\beta^T x_j), \quad W(t) = \sum x_{ji} \exp(\beta^T x_j), \]
\[ Q_i(t, a) = \sum \delta_i (W_i(\check{t}_i)/W(\check{t}_i) - a_i)/W(\check{t}_i), \]
and $Q(t, a) = (Q_1(t, a), \ldots, Q_k(t, a))^T$. Replacing the unknown parameters in $\nu(t|x)$ by their consistent estimates yields
\[ \check{\sigma}(t|x) = e^{2\beta^T x} \left\{ \sum \delta_i/W^2(\check{t}_i) + (Q(t, x))^T (-\check{\beta})^{-1} Q(t, x) \right\}, \]

4
which in turn yields
\[ \vartheta^2(t|x) = (\tilde{S}(t|x))^2 \nu(t|x), \]
by applying the delta method to the transformation \( \tilde{S}(t|x) = e^{-\tilde{\lambda}(t|x)} \), see Tsiatis (1981).

Instead of the normal quantiles \( z_{1-\alpha} \) and \( z_{\alpha} = -z_{1-\alpha} \) used in (2.1), we approximate the \( \alpha \)th and \( (1-\alpha) \)th quantiles, denoted by \( c_{\alpha}(t) \) and \( c_{1-\alpha}(t) \), by the quantiles \( \tilde{c}_{\alpha}(t) \) and \( \tilde{c}_{1-\alpha}(t) \) of the bootstrap distribution of \( \{ \tilde{\lambda}(t|x) - \Lambda(t|x) \} / \tilde{\nu}^{1/2}(t|x) \). For fixed \( x \), the cumulative hazard function \( \tilde{\lambda} \) is a step function with jumps at the uncensored observations \( \tilde{t}_i \) with \( \delta_i = 1 \), and so is the function \( \tilde{\nu} \). The jumps at the uncensored \( \tilde{t}_i \)'s also cause discontinuities of \( \tilde{c}_q \) at these points. Let \( \tilde{\Lambda}(\cdot|x) - \log(1-p)^{-1} / \tilde{\nu}^{1/2}(t|x) \) denote the modification of \( [\tilde{\Lambda}(\cdot|x) - \log(1-p)^{-1}] / \tilde{\nu}^{1/2}(t|x) \) that linearly interpolates between the corresponding values at two adjacent uncensored \( \tilde{t}_i \)'s, and define \( \tilde{c}_q(\cdot) \) similarly as the interpolated variants of \( \tilde{c}_q(\cdot) \). Define the confidence set
\[ T = \{ t : \tilde{c}_{\alpha}(t) \leq [\tilde{\Lambda}(t|x) - \log(1-p)^{-1}] / \tilde{\nu}^{1/2}(t|x) \leq \tilde{c}_{1-\alpha}(t) \} \]
for the \( p \)th quantile \( \xi_p(x) \) at a given covariate vector \( x \). Although the set (2.7) may not be an interval, it often suffices to give only the upper and lower limits of the confidence set. We next describe an algorithm to compute the upper and lower limits of (2.7), thereby obtaining a confidence interval.

Suppose the covariates \( x_i \) are independent and identically distributed, as is often the case in randomized clinical trials. Then the bootstrap distribution of the asymptotic pivot \( \{ \tilde{\Lambda}(t|x) - \Lambda(t|x) \} / (\tilde{\nu}(t|x))^{1/2} \) can be evaluated by resampling from \( \{(\tilde{t}_i, \delta_i, x_i) : 1 \leq i \leq n\} \) to obtain \( B \) bootstrap samples \( \{(\tilde{t}_i^b, \delta_i^b, x_i^b) : 1 \leq i \leq n, 1 \leq b \leq B\} \). At each given value of \( t \) that will be specified below, \( \omega_{\tilde{\alpha}}(t) := \{ \tilde{\Lambda}_{\tilde{\alpha}}(t|x) - \tilde{\Lambda}(t|x) \} / (\tilde{\nu}_{\tilde{\alpha}}(t|x))^{1/2} \) is computed from the \( b \)th bootstrap sample, and the \( \alpha \)th and \( (1-\alpha) \)th quantiles of \( \{ \omega_{\tilde{\alpha}}(t), \ldots, \omega_{\tilde{\alpha}}(t) \} \) are computed to yield \( \tilde{c}_{\alpha}(t) \) and \( \tilde{c}_{1-\alpha}(t) \). We can use the following iterative procedure to choose the values of \( t \), belonging to the ordered set \( U \) of uncensored \( \tilde{t}_i \)'s, at which \( \tilde{c}_{\alpha}(t) \) or \( \tilde{c}_{1-\alpha}(t) \) is computed. For definiteness, we consider \( \tilde{c}_{\alpha}(t) \). The objective of the iterative procedure is to solve the equation \( g(t) = 0 \), where
\[ g(t) = (\tilde{\Lambda}(t|x) - \log(1-p)^{-1}) / \tilde{\nu}^{1/2}(t|x) - \tilde{c}_\alpha(t). \]
Let \( a \) be the smallest and \( b \) be the largest element of \( U \). With \( g(a_1) < 0 \) and \( g(b_1) > 0 \), we can use the bisection method, applied to the set of ranks of the elements of the ordered set \( U \), to find two adjacent elements of \( U \) where \( g \) changes sign. Then we linearly interpolate between these two points to find the solution of \( g(t) = 0 \).
Burr and Doss (1993) use another resampling scheme under the assumption that the censoring variables $c_i$ have the same distribution function $C$. Let $\hat{C}$ be the Kaplan-Meier estimate of $C$. A bootstrap sample is of the form \{(\hat{t}_i^*, \delta_i^*, x_i) : 1 \leq i \leq n\}, where $\hat{t}_i^* = \min(t_i^*, c_i^*)$ and $\delta_i^* = I_{t_i^* \leq c_i^*}$, in which $c_i^*$ is generated from $\hat{C}$ and $t_i^*$ is generated from $\tilde{S}(\cdot|x_i)$ independently of $c_i^*$. This resampling scheme does not need the $x_i$ to be identically distributed but assumes the $c_i$ to be identically distributed instead.

### 2.2 Asymptotic theory

When there are no covariates, Lai and Wang (1993) have derived Edgeworth expansions for the sampling distribution and also for the bootstrap distribution of $(\Lambda(t) - \lambda(t))/\sqrt{\hat{V}}(t)$. In the Cox regression model, Gu (1992) has developed an Edgeworth expansion for $(-\hat{L}(\hat{\beta}))^{1/2}(\hat{\beta} - \beta)$ and also for its bootstrap counterpart under certain regularity conditions in the case $k = 1$; his arguments can be readily extended to multidimensional covariates. Combining these results yields an Edgeworth expansion for $(\Lambda(t|x) - \lambda(t|x))/\sqrt{\hat{V}}(t|x)$ and its bootstrap counterpart, from which it follows that

$$
\tilde{\xi}_q(t) - c_q(t) = o_p(n^{-1/2}) \tag{2.9}
$$

for every fixed $t$ and $0 < q < 1$ (in particular, for $q = \alpha$ and $q = 1 - \alpha$). We now show that similar to smoothing used by Chen and Hall (1993) and Ho and Lee (2005), the linear interpolation in (2.7) is effective in improving the coverage accuracy of the test-based confidence intervals.

For notational simplicity, write $\xi_p$ instead of $\xi_p(x)$ and note that $\Lambda(\xi_p|x) = \log(1 - p)^{-1}$. As in Burr and Doss (1993), Chen and Hall (1993) and Ho and Lee (2005), assume that the baseline cumulative hazard function $\Lambda$ has a continuous positive derivative $\lambda$ in addition to the regularity conditions assumed by Gu (1992). The censoring variables $c_i$ are assumed to have a common continuously differentiable distribution function. Then it can be shown that

$$
\Lambda(t|x) - \tilde{\Lambda}(t|x) = o_p(n^{-1/2}), \quad \hat{v}(t|x) - v(t|x) = o_p(n^{-1/2}), \quad \tilde{\xi}_q(t) - \tilde{c}_q(t) = o_p(n^{-1/2}).
$$

Combining this with (2.9) and using an argument similar to that of Hall (1992, Section 5.3) then yield

$$
P\{\xi_p \in T\} = P\{\tilde{\xi}_a(\xi_p) \leq [\tilde{\Lambda}(\xi_p|x) - \Lambda(\xi_p|x)]/\sqrt{\hat{V}}(\xi_p|x) \leq \tilde{c}_{1-a}(\xi_p)\}
\leq 1 - 2\alpha + o(n^{-1/2}). \tag{2.10}
$$

Without introducing interpolation, a test-based confidence set more closely related to (2.1) is

$$
\{t : \tilde{c}_a(t) \leq [\tilde{\Lambda}(t|x) - \log(1 - p)^{-1}]/\sqrt{\hat{V}}(t|x) \leq \tilde{c}_{1-a}(t)\}. \tag{2.11}
$$
However, the preceding argument can only show that $P\{\xi_p(x) \in (2.11)\} = 1 - 2\alpha + O(n^{-1/2})$. Moreover, the $O(n^{-1/2})$ error term cannot be sharpened to $o(n^{-1/2})$, as was shown by Chen and Hall (1993, p.1169) for the uncensored case without covariates.

2.3 Extension to confidence bands

Let $K$ be a compact subset of the covariate space. Noting that $\{\sqrt{n}(\hat{\xi}_p(x) - \xi_p(x)) : x \in K\}$ converges weakly to a Gaussian process indexed by $x \in K$ as $n \to \infty$, Burr and Doss (1993) used

$$\sqrt{n} \max_{x \in K} |\hat{\xi}_p(x) - \xi_p(x)| / \hat{se}_p(x)$$

(2.12)

as an approximate pivot to construct bootstrap confidence bands for $\{\xi_p(x), x \in K\}$. We can also modify the approach in Section 2.1 to construct test-based bootstrap confidence bands as follows. Let $d_\alpha$ denote the $\alpha$-th quantile of the distribution of

$$\min_{x \in K} \{\tilde{\Lambda}(\xi_p(x)|x) - \log(1 - p)^{-1} / \tilde{\sigma}^2(\xi_p(x)|x),$$

(2.13)

noting that $\Lambda(\xi_p(x)|x) = \log(1 - p)^{-1}$. We can estimate $d_\alpha$ by the $\alpha$-th quantile $\tilde{d}_\alpha$ of the bootstrap distribution of $\min_{x \in K} [\tilde{\Lambda}^*(\xi_p(x)|x) - \tilde{\Lambda}(\xi_p(x)|x)] / (\tilde{\sigma}^*(\tilde{\xi}_p(x)|x))^{1/2}$; see the last two paragraphs of Section 2.1. Similarly use the bootstrap quantile $\tilde{d}_{1-\alpha}$ to estimate the $(1-\alpha)$-th quantile $\tilde{d}_{1-\alpha}$ of

$$\max_{x \in K} [\tilde{\Lambda}(\xi_p(x)|x) - \log(1 - p)^{-1} / \tilde{\sigma}^2(\xi_p(x)|x).$$

(2.14)

With the same notation as that in (2.7), let $T_x = \{t : \tilde{d}_\alpha \leq \tilde{\Lambda}(t|x) - \log(1 - p)^{-1} / \tilde{\sigma}^2(\xi_p(x)|x) \leq \tilde{d}_{1-\alpha}\}$. Then $\{T_x, x \in K\}$ is a confidence band for $\{\xi_p(x), x \in K\}$ satisfying

$$P\{\xi_p(x) \in T_x \text{ for all } x \in K\}$$

$$= P\{\tilde{d}_\alpha \leq \min_{x \in K} [\tilde{\Lambda}(\xi_p(x)|x) - \log(1 - p)^{-1} / \tilde{\sigma}^2(\xi_p(x)|x),$$

$$\tilde{d}_{1-\alpha} \geq \max_{x \in K} [\tilde{\Lambda}(\xi_p(x)|x) - \log(1 - p)^{-1} / \tilde{\sigma}^2(\xi_p(x)|x)]\} = 1 - 2\alpha + O(n^{-1/2}),$$

(2.15)

under the same regularity conditions as those for (2.10). In Section 4 we apply this method to construct confidence bands for median survival from the Stanford Heart Transplant data and compare them with those constructed by the Burr-Doss method.

3. SIMULATION STUDIES

This section contains simulation studies of the coverage errors of the test-based confidence set (2.7) with $p = \frac{1}{2}$ and compares them with those of Dabrowska and Doksum (1987) and Burr and Doss (1993). While Dabrowska and Doksum have described their procedure
explicitly for us to implement in the comparative study in Section 3.2, Burr and Doss (1993, p.1333) "use the biweight kernel and choose bin width subjectively" in the numerical studies of their procedure, for which there are many possible choices of the kernel and the smoothing parameter in estimating $\lambda$. To simplify matters, we assume $\lambda$ to be known in the estimation of $\hat{\sigma}_p(x)$ (see the first two paragraphs of Section 1) for their bootstrap-$t$ confidence intervals; this circumvents issues concerning how the bandwidth and kernel should be chosen for their procedure to compare with ours which does not require kernel smoothing. In Section 3.1 we simplify the simulation study even further by considering the case in which covariates are absent so that the problem reduces to interval estimation of the median based on a sample of independent and identically distributed $t_i$ when there is no censoring, or on $(l_i, \delta_i)$ when there are censoring variables $c_i$.

3.1 Case without covariates

We first consider the case where there is no censoring. In this case, a test-based confidence set of the type (2.7) can be re-expressed in the form

$$\{t : \tilde{\omega}_q(t) \leq \sqrt{n}[\tilde{F}(t) - 1/2]/[\tilde{F}(t)(1 - \tilde{F}(t))]^{1/2} \leq \tilde{\omega}_{1-\alpha}(t)\},$$

(3.1)

where $w_q(t)$ denotes the $q$th quantile of the Studentized variate $\sqrt{n}[\tilde{F}(t) - F(t)]/[\tilde{F}(t)(1 - \tilde{F}(t))]^{1/2}$, $\tilde{\omega}_q(t)$ denotes the estimate of $w_q(t)$ via bootstrap resampling, $\tilde{\omega}_q(t)$ denotes the linear interpolation of $\tilde{\omega}_q(t_{(k)})$ and $\tilde{\omega}_q(t_{(k+1)})$ when $t_{(k)} = t < t_{(k+1)}$, and $[\tilde{F}(t) - 1/2]/[\tilde{F}(t)(1 - \tilde{F}(t))]^{1/2}$ denotes a similar interpolated version of $[\tilde{F}(t) - 1/2]/[\tilde{F}(t)(1 - \tilde{F}(t))]^{1/2}$. The coverage errors of (3.1) are given in Table 1a,b for $n = 30, 100$ and for the three distributions considered in the simulation study of Ho and Lee (2005): standard normal $F$, double exponential with density function $f(x) = e^{-|x|}/2$, and lognormal $F$ which is the distribution function of $\exp(N(0,1))$. As in Ho and Lee, $\alpha = 5\%$ and each result is based on 1000 simulations. Moreover, 1000 bootstrap samples are used to compute the bootstrap quantiles.

INSERT TABLE 1 ABOUT HERE

Following Burr and Doss (1993), a bootstrap-$t$ confidence interval in the present setting without censoring and covariates uses $\{\text{med}(\tilde{F}) - \text{med}(F))]/\hat{\sigma}$ as an approximate pivot and bootstrap resampling to estimate its $\alpha$th and $(1-\alpha)$th quantiles. As pointed out in (1.3), the asymptotic standard error is $\{2\sqrt{n}f(\tilde{\xi}_{1/2})\}^{-1}$ and its estimate requires a density estimator $\hat{f}$. To simply matters in comparing this method with (3.1), we consider a more favorable version of the method that can use the true $f$ to define $\hat{\sigma} = \{2\sqrt{n}f(\tilde{\xi}_{1/2})\}^{-1}$. Parts (a) and
(b) of Table 1 consider these bootstrap-t (abbreviated by Boot-t) confidence intervals. Also given for comparison are the coverage errors, taken from Table 3 of Ho and Lee (2005), of the confidence intervals of Ho and Lee, Beran and Hall (1993, abbreviated by BeHa) and Chen and Hall (1993) who construct the confidence intervals by using smoothed empirical likelihood (abbreviated by SmoEL). The results of Table 1a,b show that (3.1) and the Beran-Hall confidence limits have coverage errors that are close to the nominal value of 5%. The smoothed empirical likelihood and Ho-Lee confidence limits also perform well except for a couple of cases. In contrast, the bootstrap-t confidence intervals based on the approximate pivot \{\text{med}(\hat{F}) - \text{med}(F)\}/\hat{s}_e\ have coverage errors that are markedly different from 5% in most cases.

Parts (c) and (d) of Table 1 consider the case where there are independent censoring variables \(c_i\) that have a common distribution \(C\) which is assumed to be exponential with intensity parameter \(2/3\) or 1. The observations are \((\tilde{t}_i, \delta_i), i = 1, \ldots, n\). Let \(\hat{\sigma}(t)\) be the estimated standard error of \(\hat{S}(t)\) given by Greenwood’s formula; see Andersen et al. (1993, p.258). As in (2.7) and (3.1), we use \([\hat{S}(t) - 1/2]/\hat{\sigma}(t)\) and \(\tilde{w}_q(t)\) to denote the interpolated versions of \([\hat{S}(t) - 1/2]/\hat{\sigma}(t)\) and \(\tilde{w}_q(t)\). Since we do not have to estimate \(\beta\) and adjust for the variability of \(\hat{\beta} - \beta\) here, we can use \(\hat{S}(t) - 1/2\) as in the Brookmeyer-Crowley interval (1.5) (abbreviated by BrCr) instead of transforming it to \(\hat{\lambda}(t) - \log 2\) as in (2.7). This leads to a test-based confidence set of the form

\[\{t : \tilde{w}_\alpha(t) \leq [\hat{S}(t) - 1/2]/\hat{\sigma}(t) \leq \tilde{w}_{1-\alpha}(t)\}.\] (3.2)

Besides (1.5) and (3.2), parts (c) and (d) of Table 1 also consider the bootstrap-t confidence intervals (abbreviated by Boot-t) of the type considered by Burr and Doss (1993). Since there are no covariates, the asymptotic standard error of \(\hat{\xi}_{1/2} - \xi_{1/2}\) is simpler than that in their Theorem 1. It is equal to \(\sigma(\xi_{1/2})/f(\xi_{1/2})\), where \(\sigma^2(t)\) is the asymptotic variance of \(\hat{S}(t)\). Assuming \(f\) to be known, the estimated standard error of \(\hat{\xi}_{1/2} - \xi_{1/2}\) is \(\hat{s}_e = \hat{\sigma}(\hat{\xi}_{1/2})/f(\hat{\xi}_{1/2})\), where \(\hat{\sigma}(t)\) is the estimated standard error of \(\hat{S}(t)\) given by Greenwood’s formula; see Andersen et al. (1993, pp.257,258,276 which assume \(t_i\) to be nonnegative and use \(\sigma^2(t)\) to denote the asymptotic variance of \(\hat{S}(t)/S(t)\) instead).

Strawderman, Parzen and Wells (1997) gave a review of confidence intervals for \(\xi_p\) based on censored observations in the earlier literature and proposed a new test-based confidence interval for \(\xi_p\), which uses an Edgeworth expansion for \([\hat{\lambda}(t) - \Lambda(t)]/\sqrt{\hat{\xi}^2(t)}\) to improve the normal approximation and which they denote by \(I_2\). Their simulation study shows that \(I_2\), which Table 1 refers to as SPW (abbreviation for the authors), “is superior to all others considered
in terms of maintaining coverage accuracy." The results in Table 1c,d, however, contain cases where SPW has worse coverage accuracy than the Brookmeyer-Crowley confidence limits, and show that (3.2) has better coverage accuracy. Parts (c) and (d) of Table 1 also show that the bootstrap-confidence intervals based on the approximate pivot \((\hat{\xi}_{1/2} - \xi_{1/2})/\hat{se}\) have poor coverage accuracy when \(F\) is double exponential or lognormal.

3.2 Two covariates one of which is treatment indicator

Consider a clinical trial which has a duration of 5.5 years and in which \(n = 100\) patients enter uniformly during a 3-year accrual period and are randomized to treatment or control with probability \(1/2\). The patient's failure time (measured from time of entry) is assumed to follow the proportional hazards model (1.1) in which \(x = (x_1, x_2)\), where \(x_1\) is the treatment indicator (taking the value 1 if the treatment is received and 0 otherwise) and \(x_2\) is a patient characteristic, which is assumed to be uniformly distributed on \((0, 1)\). Suppose the unknown parameters of (1.1) take the values \(\beta_2 = 1/2\), \(e^{\beta_1} = 1, 2/3\) or 1/2, and \(\lambda(t) = t/3\) (so the baseline distribution is exponential with 3 years as its mean). Coverage errors of the confidence set (2.7) for the median survival given the covariate \(x_2\) for the treatment group \((x_1 = 1)\) and for the control group \((x_1 = 0)\) are given in Table 2 for various values of \(x_2\). As in Table 1, each result is based on 1000 simulations, and 1000 bootstrap samples are used to compute the bootstrap quantiles. Also given for comparison are the coverage errors (in parentheses) of the confidence intervals of Dabrowska and Doksum (1993); see the second paragraph of Section 1. Table 2 shows that the coverage errors of (2.7) are close to \(\alpha = 5\%\) but those of Dabrowska-Doksum differ markedly from 5%.

**INSERT TABLE 2 ABOUT HERE**

Besides the confidence set (2.7), this simulation study also considers the coverage errors of confidence bands of the type in Section 2.3. For the control group, we choose \(K = \{(0, x_2) : 0.25 \leq x_2 \leq 0.75\}\) and the confidence band \(\{T_2, x \in K\}\) has coverage error 11.2\% for \(\beta_1 = 0\), 11.7\% for \(\beta_1 = \log 2/3\), and 9.3\% for \(\beta_1 = \log 1/2\), in close agreement with the nominal coverage error of 10\%. The corresponding values for the treatment group (with \(K = \{(1, x_2) : 0.25 \leq x_2 \leq 0.75\}\)) are 9.2\% for \(\beta_1 = 0\), 10.7\% for \(\beta_1 = \log 2/3\), and 9.2\% for \(\beta_1 = \log 1/2\). Figure 1 plots the 90\% confidence bands for the median survival times of the treatment and control groups based on one simulated data set from this randomized clinical trial with 100 subjects.

**INSERT FIGURE 1 ABOUT HERE**
4. APPLICATION TO STANFORD HEART TRANSPLANT DATA

In this section we illustrate the methods in Sections 2.1 and 2.3 for constructing confidence intervals and bands on the 1980 version of the Stanford Heart Transplant (SHT) data as given in Miller and Halpern (1982), who fitted a proportional hazards regression model to the data involving 152 patients that had survived at least 10 days, and who chose quadratic regression of \( \log_{10} \) (survival time in days) on age (in years) as the predictor variable for the final model. Burr and Doss (1993, p.1338) have applied their bootstrap-\( t \) method to construct confidence intervals and bands for median survival (days in the \( \log_{10} \) scale) from these data. In addition, they have also used the limiting Gaussian process for \( \{ \hat{\xi}_{1/2}(x) - \xi_{1/2}(x) \}/\hat{\sigma}_{1/2}(x) \) to construct confidence intervals and confidence bands for \( \xi_{1/2}(x) \), simulating the Gaussian process to determine the half-width of the band. Their results at 38.5 and 48.7 years of age for 95% simulated process (SP) and bootstrap-\( t \) (Boot) confidence bands and intervals are included in Table 3. Also given in Table 3 are the test-based (Tb\( \hat{A} \)) confidence intervals and bands in Sections 2.1 and 2.3 that use \( \{ \hat{\Lambda}(t|x) - \log(1 - p)^{-1} \}/\hat{\sigma}(t|x) \) as the test statistic. Instead of \( \hat{\Lambda}(t|x) \), an alternative is to use \( \{ \hat{S}(t|x) - (1 - p) \}/\hat{\sigma}(t|x) \) as pointed out in the first paragraph of Section 2.1, and its associated test-based (Tb\( \hat{S} \)) confidence intervals and bands are also given in Table 3 for comparison. The largest uncensored lifetime in the SHT data is 7.9 years. In view of the interpolation scheme used to evaluate the upper limits of the confidence intervals and bands in Section 2.1 and 2.3, the value of the upper limit is undetermined if it exceeds the largest uncensored observation. This explains the entry 7.9+ (which means exceeding 7.9 years) for the upper limit of the confidence band Tb\( \hat{A} \) or Tb\( \hat{S} \) in Table 3, which shows Tb\( \hat{A} \) to yield somewhat shorter confidence intervals and bands than Tb\( \hat{S} \). Moreover, Tb\( \hat{A} \) yields markedly shorter confidence bands than SP and Boot. Figure 2 plots the entire confidence band for Tb\( \hat{A} \); note that the upper band ends at \( \log_{10}(2878 \) days).

INSERT TABLE 3 and Figure 2 ABOUT HERE

5. DISCUSSION

An important ingredient in the test-based confidence intervals/bands developed herein for survival quantiles in the Cox regression model is the interpolation scheme that can produce effects similar to those of smoothing in the smoothed empirical likelihood approach of Chen and Hall (1993) and the iterated smoothed bootstrap confidence intervals of Ho and Lee (2005) for population quantiles (without covariates and censoring). On the other hand, the interpolation scheme does not involve smoothing whose delicate accompanying issues like bandwidth selection are a major source of difficulty in the alternative approach based
on \(\{\hat{\xi}_p(x) - \xi_p(x)\}/\hat{\sigma}_p\) that has been considered by Dabrowska and Doksum (1987) and Burr and Doss (1993). Another useful ingredient is to work with \(\hat{\Lambda}(t|x) - \log(1 - p)^{-1}\), instead of \(\hat{S}(t|x) - (1 - p)\) that has been used by Brookmeyer and Crowley (1982) and subsequent authors for the case without covariates. In the presence of covariates, there is additional variability due to the estimation of the regression parameter \(\beta\) and it is useful to transform \(\hat{S}(t|x)\), which is constrained to belong to \([0, 1]\), to the unconstrained \(\hat{\Lambda}(t|x) - \log(1 - p)^{-1}\). This transformation often leads to shorter confidence intervals and bands. A third important ingredient of our approach is the use of bootstrap quantiles to approximate the quantiles of \(\{\hat{\Lambda}(t|x) - \Lambda(t|x)\}/\hat{\nu}(x)\), instead of using the normal approximation (or Edgeworth expansions) as in previous works on test-based confidence intervals for \(\xi_p\) (in the absence of covariates) from censored survival data.

ACKNOWLEDGEMENT

Lai’s research was supported by the National Science Foundation and the National Cancer Institute. Su’s research was supported by the Cancer Center at Stanford University.

REFERENCES


Fig 1. 90% confidence bands for the median survival time. The solid curve represents the true median survival time (in years) as a function of the covariate $x_2$.

Fig 2. 95% confidence bands for the median survival time of the Stanford Heart Transplant data. The solid curve represents the estimate of median survival time (days in the log_{10} scale) as a function of age (in years).
Table 1. *Coverage errors* (in %) of confidence intervals for median of three distributions.

<table>
<thead>
<tr>
<th>Interval</th>
<th>Normal</th>
<th></th>
<th>Double Exponential</th>
<th></th>
<th>Lognormal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td>Total</td>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>(3.1)</td>
<td></td>
<td>4.5</td>
<td>5.0</td>
<td>9.5</td>
<td>5.0</td>
<td>4.4</td>
</tr>
<tr>
<td>BeHa</td>
<td></td>
<td>5.1</td>
<td>5.0</td>
<td>10.1</td>
<td>5.9</td>
<td>4.9</td>
</tr>
<tr>
<td>SmoEL</td>
<td></td>
<td>4.9</td>
<td>5.7</td>
<td>10.6</td>
<td>6.2</td>
<td>5.7</td>
</tr>
<tr>
<td>Ho-Lee</td>
<td></td>
<td>3.7</td>
<td>7.3</td>
<td>11.0</td>
<td>4.0</td>
<td>5.4</td>
</tr>
<tr>
<td>Boot-t</td>
<td></td>
<td>5.4</td>
<td>5.5</td>
<td>10.9</td>
<td>2.8</td>
<td>3.0</td>
</tr>
<tr>
<td>(b) n = 100, no censoring</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3.1)</td>
<td></td>
<td>5.0</td>
<td>5.2</td>
<td>10.2</td>
<td>5.1</td>
<td>4.5</td>
</tr>
<tr>
<td>BeHa</td>
<td></td>
<td>4.6</td>
<td>5.0</td>
<td>9.6</td>
<td>4.6</td>
<td>4.3</td>
</tr>
<tr>
<td>SmoEL</td>
<td></td>
<td>4.9</td>
<td>5.6</td>
<td>10.5</td>
<td>5.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Ho-Lee</td>
<td></td>
<td>5.1</td>
<td>6.0</td>
<td>11.1</td>
<td>4.2</td>
<td>5.3</td>
</tr>
<tr>
<td>Boot-t</td>
<td></td>
<td>4.5</td>
<td>6.1</td>
<td>10.6</td>
<td>2.9</td>
<td>3.6</td>
</tr>
<tr>
<td>(c) n = 30, $C \sim \text{Exp}(2/3)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3.2)</td>
<td></td>
<td>5.3</td>
<td>5.0</td>
<td>10.3</td>
<td>5.1</td>
<td>5.3</td>
</tr>
<tr>
<td>BrCr</td>
<td></td>
<td>5.3</td>
<td>6.2</td>
<td>11.5</td>
<td>4.7</td>
<td>4.6</td>
</tr>
<tr>
<td>SPW</td>
<td></td>
<td>5.3</td>
<td>4.2</td>
<td>9.5</td>
<td>5.5</td>
<td>4.8</td>
</tr>
<tr>
<td>Boot-t</td>
<td></td>
<td>5.2</td>
<td>3.6</td>
<td>8.8</td>
<td>2.6</td>
<td>1.6</td>
</tr>
<tr>
<td>(d) n = 100, $C \sim \text{Exp}(1)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3.2)</td>
<td></td>
<td>5.6</td>
<td>4.5</td>
<td>10.1</td>
<td>4.4</td>
<td>5.7</td>
</tr>
<tr>
<td>BrCr</td>
<td></td>
<td>4.2</td>
<td>4.7</td>
<td>8.9</td>
<td>4.4</td>
<td>4.8</td>
</tr>
<tr>
<td>SPW</td>
<td></td>
<td>6.4</td>
<td>4.5</td>
<td>10.9</td>
<td>6.2</td>
<td>4.8</td>
</tr>
<tr>
<td>Boot-t</td>
<td></td>
<td>4.5</td>
<td>5.9</td>
<td>10.4</td>
<td>3.4</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Table 2. Coverage errors (in %) of (2.7) and of the Dabrowska-Doksum confidence intervals (in parentheses) for median survival in Cox model.

<table>
<thead>
<tr>
<th>$x_2$</th>
<th>$\beta_1 = 0$</th>
<th>$\beta_1 = \log \frac{2}{3}$</th>
<th>$\beta_1 = \log \frac{1}{2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lower</td>
<td>Upper</td>
<td>Total</td>
</tr>
<tr>
<td>Treatment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>5.6</td>
<td>5.4</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td>(3.4)</td>
<td>(2.9)</td>
<td>(6.3)</td>
</tr>
<tr>
<td>.5</td>
<td>5.5</td>
<td>4.9</td>
<td>10.4</td>
</tr>
<tr>
<td></td>
<td>(2.7)</td>
<td>(4.5)</td>
<td>(7.2)</td>
</tr>
<tr>
<td>.75</td>
<td>5.2</td>
<td>4.1</td>
<td>9.3</td>
</tr>
<tr>
<td></td>
<td>(2.2)</td>
<td>(4.7)</td>
<td>(6.9)</td>
</tr>
<tr>
<td>1</td>
<td>4.8</td>
<td>5.4</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>(2.2)</td>
<td>(4.3)</td>
<td>(6.5)</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.25</td>
<td>4.5</td>
<td>5.6</td>
<td>10.1</td>
</tr>
<tr>
<td></td>
<td>(2.8)</td>
<td>(4.3)</td>
<td>(7.1)</td>
</tr>
<tr>
<td>.5</td>
<td>4.2</td>
<td>4.8</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>(3.2)</td>
<td>(4.3)</td>
<td>(7.5)</td>
</tr>
<tr>
<td>.75</td>
<td>5.4</td>
<td>5.6</td>
<td>11.0</td>
</tr>
<tr>
<td></td>
<td>(2.3)</td>
<td>(3.5)</td>
<td>(5.8)</td>
</tr>
<tr>
<td>1</td>
<td>5.2</td>
<td>4.2</td>
<td>9.4</td>
</tr>
<tr>
<td></td>
<td>(2.7)</td>
<td>(3.6)</td>
<td>(6.3)</td>
</tr>
</tbody>
</table>

Table 3. 95% confidence intervals and bands for median survival (in years) from SHT data.

<table>
<thead>
<tr>
<th>Age=38.5</th>
<th>Age=48.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP</td>
<td>Boot</td>
</tr>
<tr>
<td>Interval</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>2.4</td>
</tr>
<tr>
<td>U</td>
<td>6.7</td>
</tr>
<tr>
<td>Band</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>1.6</td>
</tr>
<tr>
<td>U</td>
<td>9.7</td>
</tr>
</tbody>
</table>