12-7-2005

Inferences in Censored Cost Regression Models with Empirical Likelihood

Xiao-Hua Zhou
University of Washington, azhou@u.washington.edu

Gengsheng Qin
Georgia State University, gqin@gsu.edu

Huazhen Lin
University of Washington, huazhen.lin@med.va.gov

Gang Li
University of California at Los Angeles, vli@ucla.edu

Suggested Citation

1 Introduction

Prospective payment models, such as capitation, have a long history in the financing of private and public sector health care. Capitated payments are set at the expected total cost of a patient, which payors (e.g., employers or Medicare) and recipients (e.g., health plans or Medicare HMOs) find mutually acceptable if those payments equal actual average cost (Maciejewski et al, 2005). Problems arise if prospective payment models do not account for distributional aspects of costs that can lead to significant deviations from actual average costs, especially for particular populations or groups. Hence, it is important to accurately predict the expected total cost of a patient over a certain time period $[0, \tau]$ after adjusting for patients' characteristics. If one is interested in median regression with censored cost data, see the paper by Bang and Tsiatis (2002).

One of the main features in the distribution of health care costs that can impede reliable prediction is its skewness due to the small percentage of patients who invariably incur extremely high costs relative to most patients. Recent efforts have yielded new statistical analysis methods that can adjust for the special features in the distribution of health care costs (Zhou et al, 1997; Zhou and Tu, 1999; Zhou et al, 2001).

Censoring can also be a major issue in estimating the average lifetime cost or cost in a certain time period. Censoring occurs when the complete costs of some subjects in the certain time period for some subjects are not available because the subjects are lost to follow-up before the end of the study. If we only include uncensored subjects in analysis, we may underes-

2

timate the average cost. This is because subjects who survive a long time are likely to be censored and not included in analysis while subjects who die early are likely to be uncensored and included in analysis. Subjects who die shortly after entering a study often use the fewest resources whereas those who remain alive for a long time are most likely to be expensive (Etzioni et al, 1999).

The main challenge in the analysis of censored cost data is that the total cost at the time of censoring is not independent of the total cost at the time of death, even if the time of death and time of censoring are independent. Hence, standard survival analysis techniques (e.g Cox models), which assume independent censoring, cannot be directly used for the analysis of censored costs data by treating censored costs as censored survival times. For estimating the average cost of censored cost data without covariates, Lin (1997), Bang and Tiastis (2000), and Jiang and Zhou (2004) proposed several appropriate methods. For estimating the average cost of censored cost data with covariates, Lin (2000a, 2000b, 2003) proposed several regression models.

The existing regression models for incomplete cost data focus mostly on finding consistent and asymptotically normal estimators for the individual components of the vector of regression parameters (β). However, the expected total cost of a patient with a vector of given covariates is a complicated function of β.

Although it is possible to derive confidence intervals for the expected total cost of a patient over a certain period based on the asymptotical nor-

3
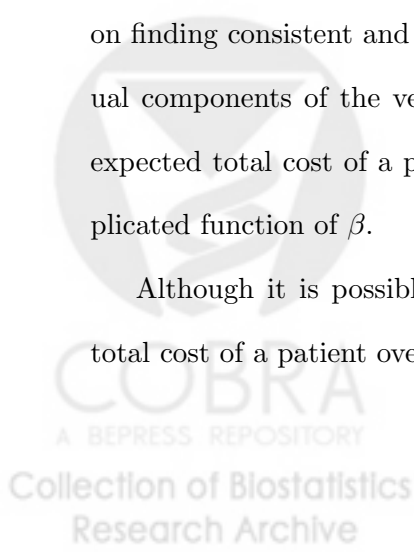
mality of the estimator for β, there may be several problems with these kinds of the confidence intervals. First, the normal approximation can have poor coverage accuracy if the distribution of data is skewed, which is common for cost data. Second, as it is well known in the literature, the confidence region of the multi-dimensional parameter β, based on the normal approximation, can have poor coverage accuracy even if the coverage probabilities for the univariate components of β are close to the nominal level.

Empirical likelihood (EL) methods are popular non-parametric methods for constructing confidence intervals and bands. As previously demonstrated (Owen, 2001), an EL method has several advantages over the normal approximation method in constructing confidence bands and intervals. First, EL methods do not assume a symmetrical shape, as is assumed in the normal approximation method; instead its shape is determined by data, and the EL regions are Bartlett correctable in most cases (DiCiccio et al., 1991). Hence, the EL-based method is especially suitable for skewed data. Second, EL methods allow for confidence band construction without an information/variance estimator. Third, the EL methods allow us to employ likelihood methods without having to pick a parametric family for the data. In this paper, we develop a new EL based confidence region for β and intervals for the expected total cost over the period $[0, \tau]$.

2 Data Setup and regression models

In this section, we use regression models for the mean of cost over the period $[0, \tau]$ and for the survival time. We follow the notation used in Lin (2003).

4
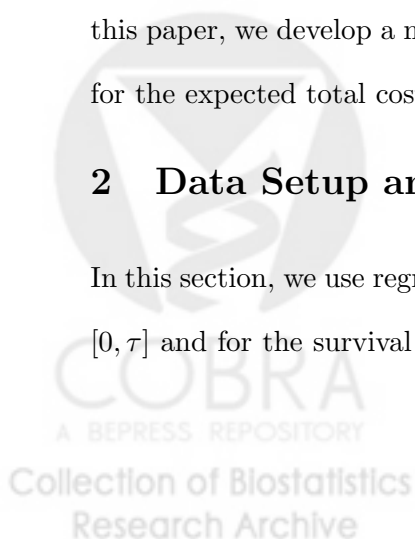
For patient i $(i = 1, ..., n)$, let $Y_i(t)$ be the total cost of the patient up to time t. We can only observe $Y_i(t)$ at a finite number of possible time points, $t_0, \ldots, t_K = \tau$. Let y_{ki} be the total cost over the kth $(k = 1, ..., K)$ interval $[t_{k-1}, t_k)$, where $t_0 = 0$ and $t_K = \tau$. That is, $y_{ki} = Y_i(t_k) - Y_i(t_{k-1})$. Then, the total cost accumulated by the patient i over the entire interval $[0, \tau)$ is $Y_i = \sum_{k=0}^{K} y_{ki}$. Let T_i and C_i be the survival and censoring times of patient i, respectively. Let $\mathbf{Z}_i(t)$ be the $p \times 1$ vector of potentially time-dependent covariates for patient i. Denote $\mathbf{Z}_{ki}$ to be the value of $\mathbf{Z}_i(t)$ when t is in the kth interval. Since we take a position that no additional cost can be accumulated after death, we have $Y_i(t) = Y_i(t \wedge T_i)$. To model the effect of covariates Z on the marginal distribution of y_{ki}, we use the same model as in Lin (2003),

$$E(y_{ki}|\mathbf{Z}_{ki}) = g(\beta'\mathbf{Z}_{ki}), k = 1, ..., K; i = 1, ..., n, \qquad (1)$$

where g is some link function. This model includes both the previously proposed linear regression model and the proportional mean model for censored medical cost (Lin, 2000a, Lin, 2000b).

3   An Existing Estimation Procedure

In the presence of censoring, not all the y_{ki}'s are observable. Let $T_{ki}^* = min(t_k, T_i)$, $\delta_{ki}^* = I(T_{ki}^* \le C_i)$, $X_i = min(T_i, C_i)$, and $\delta_i = I(T_i \le C_i)$. So, y_{ki} is observable if and only if $\delta_{ki}^* = 1$. Define $\mathbf{F}_i = \{I(T_i \le t), Y_i(t), \bar{\mathbf{L}}_\mathbf{i}(t)\}$, where $\bar{\mathbf{L}}_\mathbf{i}(t)$ represents all the measured covariate processes, and $\bar{\mathbf{H}}(t) = \{\mathbf{H}(s) : s \le t\}$ for any process $\mathbf{H}(.)$. Let $G(t \mid \bar{\mathbf{F}}_i) = P(C_i > t \mid \bar{\mathbf{F}}_i(T_i))$. Lin

5

(2003) proposed the following generalized estimating equation for β:

$$\widehat{U}(\beta) \equiv \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\delta_{ki}^{*}}{\widehat{G}(T_{ki}^{*} \mid \bar{\mathbf{F}}_i)} h(\mathbf{Z}_{ki}; \beta)(y_{ki} - g(\beta' \mathbf{Z}_{ki})) \mathbf{Z}_{ki} = 0,$$

where $h(\mathbf{Z}_{ki}; \beta)$ is a given scalar function. From the theory of estimating equations, we know that an optimal choice of $h(Z_{ik}; \beta)$ is given by

$$h(Z_{ik}; \beta) = g^{(1)}(\beta' Z_{ik}) / var(y_{ki}).$$

However, since $var(y_{ki})$ is unknown, we let the weight function $h(Z_{ik}; \beta)$ be 1 in the analysis presented in this article, although more general choices are possible. Misspecification of the weight function will not affect the consistency of the resulting estimator, only the efficiency. Here $\widehat{G}(\cdot \mid \mathbf{F}_i)$ is a consistent estimator of $G(\cdot \mid \mathbf{F}_i)$. In the case of completely random censoring, we may set $\widehat{G}(\cdot \mid \bar{\mathbf{F}})$ to be the Kaplan-Meier estimator $\widehat{G}(\cdot)$ for the common survival function of C_i. Otherwise, we take $\widehat{G}(\cdot \mid \mathbf{F}_i)$ to be the Breslow (1972) estimator, defined by

$$\widehat{G}(\cdot \mid \bar{\mathbf{F}}_i) = \exp\left[-\sum_{j=1}^{n} \frac{\bar{\delta}_i I(X_j < t) e^{\widehat{\gamma}' \mathbf{W}_i(X_j)}}{S^{(0)}(X_j; \widehat{\gamma})} \right],$$

where $\mathbf{W}_i(t)$ is a vector of known function of $\mathbf{F}_i$, $\bar{\delta}_i = 1 - \delta_i$, and $\widehat{\gamma}$ is the maximum partial likelihood estimator of the regression parameters in the proportional hazards model (Cox, 1972)

$$\lambda(t \mid \bar{\mathbf{F}}_i) = \lambda_0(t) e^{\widehat{\gamma}' \mathbf{W}_i(t)}, \quad i = 1, ..., n,$$

and

$$S^{(\rho)}(t; \gamma) = \sum_{i=1}^{n} I(X_i \geq t) e^{\gamma' \mathbf{W}_i^{\otimes \rho}(t)}, \quad \rho = 0, 1, 2.$$

Here and in the sequel, we adopt the notation: $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{aa}'$.

The solution $\widehat{\beta}$ to the above estimating equation is defined as an estimator of β. Lin (2003) obtained the limiting distribution of $\widehat{\beta}$:

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{\mathcal{L}} N(0, A^{-1}VA^{-1}), \tag{2}$$

where $A = -\lim_{n\to\infty} n^{-1}E(\frac{\partial \widehat{U}(\beta)}{\partial \beta})$, and V is given by (5) on Page 9 when we discuss our EL method.

4 Empirical likelihood confidence region for β

In this section we propose EL-based confidence region for β. Let

$$\mathbf{D}_i = \sum_{k=1}^{K} \frac{\delta_{ki}^*}{G(T_{ki}^* \mid \bar{\mathbf{F}}_i)} h(\mathbf{Z}_{ki}; \beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}$$

and

$$\widehat{\mathbf{D}}_i = \sum_{k=1}^{K} \frac{\delta_{ki}^*}{\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i)} h(\mathbf{Z}_{ki}; \beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}.$$

First consider the testing problem,

$$H_0 : \beta = \beta_0 \quad vs. \quad H_1 : \beta \neq \beta_0.$$

Since $E(\mathbf{D}_i) = 0$ for all $i = 1, ..., n$, the problem of testing whether β_0 is the true parameter of β is equivalent to testing whether $EU(\beta_0) = 0$, where $U(\beta_0) = \sum_{i=1}^{n} \mathbf{D}_i$.

This can be done by using Owen's EL method (1990, 1991). Let $p = (p_1, \cdots, p_n)$ be a probability vector, i.e., $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for all i. Then, the empirical likelihood, evaluated at the true parameter value β_0, is

7

defined by

$$\widetilde{L}(\beta_0) = \sup\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \mathbf{D}_i = 0\right\}.$$

Since $\mathbf{D}_i$'s depend on $G(\cdot \mid \bar{\mathbf{F}}_i)$, which is unknown, replacing $\mathbf{D}_i$ by $\widehat{\mathbf{D}}_i$, we obtain the estimated empirical likelihood for β_0:

$$L(\beta_0) = \sup\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \widehat{\mathbf{D}}_i = 0\right\}.$$

Then, by the Lagrange multiplier, we can easily get

$$p_i = \frac{1}{n}\left\{1 + \lambda'\widehat{\mathbf{D}}_i\right\}^{-1}, \qquad i = 1, \cdots, n,$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)'$ is the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\mathbf{D}}_i}{1 + \lambda'\widehat{\mathbf{D}}_i} = 0. \tag{3}$$

Note that $\prod_{i=1}^{n} p_i$, subject to $\sum_{i=1}^{n} p_i = 1$, attains its maximum n^{-n} at $p_i = n^{-1}$. So we define the empirical likelihood ratio at β_0 by

$$R(\beta_0) = \prod_{i=1}^{n}(np_i) = \prod_{i=1}^{n}\{1 + \lambda'\widehat{\mathbf{D}}_i\}^{-1}.$$

Therefore, the corresponding empirical log-likelihood ratio can be defined as

$$l(\beta_0) = -2\log R(\beta_0) = 2\sum_{i=1}^{n}\log\{1 + \lambda'\widehat{\mathbf{D}}_i\}, \tag{4}$$

where $\lambda = (\lambda_1, \cdots, \lambda_p)'$ is the solution to Equation (3).

Before introducing the main theorem, we need some additional notation. If censoring occurs in a completely random fashion, we define

$$\eta_i = \int_{o}^{\infty} \mathbf{q}(t)dM_i(t),$$

8

where

$$M_i(t) = \bar{\delta}_i I(X_i \le t) - \int_0^t I(X_i \ge x)\lambda(x)dx,$$

$$\lambda(x) = -\frac{d\log G(x)}{dx}, \text{ and}$$

$$\mathbf{q}(t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^* I(T_{ki}^* > t)}{G(T_{ki}^* \mid \bar{\mathbf{F}}_i)P(X_i \ge t)}h(\mathbf{Z}_{ki};\beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}.$$

Otherwise, we define

$$\eta_i = \int_0^\infty \left[\mathbf{q(t)} + \mathbf{b}\Omega^{-1}\left(\mathbf{W}_i(t) - \bar{\mathbf{w}}(t)\right)\right]dM_i(t),$$

where

$$M_i(t) = \bar{\delta}_i I(X_i \le t) - \int_0^t I(X_i \ge x)e^{\gamma'\mathbf{W}_i(t)}\lambda_0(x)dx,$$

$$\mathbf{q}(t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^* I(T_{ki}^* > t)e^{\gamma'\mathbf{W}_i(t)}}{\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i)s^{(0)}(t)}h(\mathbf{Z}_{ki};\beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki},$$

$$\mathbf{s}^{(\rho)}(t) = \lim_{n\to\infty} n^{-1}\mathbf{S}^{(\rho)}(t) \quad (\rho = 0, 1, 2),$$

$$\mathbf{b} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^*}{\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}\mathbf{r}'(T_{ki}^*;\mathbf{W}_i),$$

$$\mathbf{r}(t;\mathbf{W}) = \int_0^t e^{\gamma'\mathbf{W}(x)}\left[\mathbf{W}(x) - \bar{\mathbf{w}}(x)\right]\lambda_0(x)dx,$$

$$\bar{\mathbf{w}}(t) = \mathbf{s}^{(1)}(t)/s^{(0)}(t),$$

$$\Omega = \int_0^\infty \left[\mathbf{s}^{(2)}(t)/s^{(0)}(t) - \bar{\mathbf{w}}^{\otimes 2}(t)\right]s^{(0)}(t)\lambda_0(t)dt.$$

Let

$$V_1 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\mathbf{D}_i^{\otimes 2} \text{ and } \qquad V = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left(\mathbf{D}_i + \eta_i\right)^{\otimes 2}. \qquad (5)$$

9

The following conditions are needed in this paper:

C1. $\mathbf{q}(t) < \infty$ and $\mathbf{s}^{(\rho)}(t) < \infty$ $(\rho = 0, 1, 2)$ for every t.

   C2. $\|\mathbf{b}\| < \infty$ and $\|\Omega\| < \infty$.

   C3. V_1 and V are positive definite matrix.

   C4. $\max_{k,i} \left\| \frac{\delta_{ki}^*}{G(T_{ki}^* | \bar{\mathbf{F}}_i)} h(\mathbf{Z}_{ki}; \beta)(y_{ki} - g(\beta' \mathbf{Z}_{ki})) \mathbf{Z}_{ki} \right\| = o_p(n^{1/2})$.

   Theorem 1. *Assume the conditions* **C1-C4** *hold. If* β_0 *is the true value of* β, *then* $l(\beta_0)$ *has the asymptotical distribution as a weighted sum of independent chi-square random variables with 1 degree of freedom; that is,*

$$l(\beta_0) \xrightarrow{\mathcal{L}} l_1 \chi_{1,1}^2 + \cdots + l_p \chi_{p,1}^2,$$

where $\chi_{i,1}^2$*'s, for* $1 \leq i \leq p$, *are independent chi-square random variables with one degree of freedom, and the weights* l_i, $1 \leq i \leq p$, *are the eigenvalues of* $V_1^{-1} V$.

We provide a proof of Theorem 1 in the Appendix. In order to apply Theorem 1, we first need to estimate the weights l_i, $1 \leq i \leq p$. To estimate the weights, we define

$$\widehat{\eta}_i = \bar{\delta}_i \mathbf{Q}(X_i) - \sum_{j=1}^n \frac{\bar{\delta}_j I(X_j \leq X_i) \mathbf{Q}(X_j)}{\sum_{l=1}^n I(X_l \leq X_j)}, \text{ if } \widehat{G} \text{ is the Kaplan-Meier estimator,}$$

where

$$\mathbf{Q}(t) = \sum_{i=1}^n \sum_{k=1}^K \frac{\delta_{ki}^* I(T_{ki}^* > t) h(\mathbf{Z}_{ki}; \widehat{\beta})(y_{ki} - g(\widehat{\beta}' \mathbf{Z}_{ki})) \mathbf{Z}_{ki}}{\widehat{G}(T_{ki}^*)} \Big/ \sum_{j=1}^n I(X_j \geq t)$$

or

$$\widehat{\eta}_i = \bar{\delta}_i \mathbf{N}_i(X_i) - \sum_{j=1}^n \frac{\bar{\delta}_j I(X_j \leq X_i) e^{\widehat{\gamma}' \mathbf{W}_i(X_j)} \mathbf{N}_i(X_j)}{S^{(0)}(X_j; \widehat{\gamma})}, \text{ if } \widehat{G} \text{ is the Breslow estimator.}$$

10

Here

$$
\mathbf{N}_i(t) = \widetilde{\mathbf{Q}}(t) + \mathbf{B}\widehat{\Omega}^{-1}\left[\mathbf{W}_i(t) - \mathbf{S}^{(1)}(t;\widehat{\gamma})/S^{(0)}(t;\widehat{\gamma})\right],
$$

$$
\widetilde{\mathbf{Q}}(t) = \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^{*}I(T_{ki}^{*}>t)e^{\widehat{\gamma}'\mathbf{W}_i(t)}}{\widehat{G}(T_{ki}^{*}\mid\bar{\mathbf{F}}_i)S^{(0)}(t;\widehat{\gamma})}h(\mathbf{Z}_{ki};\widehat{\beta})(y_{ki}-g(\widehat{\beta}'\mathbf{Z}_{ki}))\mathbf{Z}_{ki},
$$

$$
\mathbf{B} = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{\delta_{ki}^{*}}{\widehat{G}(T_{ki}^{*}\mid\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\widehat{\beta})(y_{ki}-g(\widehat{\beta}'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}\mathbf{R}'(T_{ki}^{*};\mathbf{W}_i),
$$

$$
\mathbf{R}(t;\mathbf{W}) = \sum_{i=1}^{n}\bar{\delta}_{i}I(X_i<t)e^{\widehat{\gamma}'\mathbf{W}(X_i)}\left[\mathbf{W}(X_i)-\frac{\mathbf{S}^{(1)}(X_i;\widehat{\gamma})}{S^{(0)}(X_i;\widehat{\gamma})}\right]\Big/S^{(0)}(X_i;\widehat{\gamma})\,,\text{ and}
$$

$$
\widehat{\Omega} = \sum_{i=1}^{n}\bar{\delta}_{i}\left[\frac{\mathbf{S}^{(2)}(X_i;\widehat{\gamma})}{S^{(0)}(X_i;\widehat{\gamma})}-\frac{S^{(1)}(X_i;\widehat{\gamma})^{\otimes2}}{S^{(0)}(X_i;\widehat{\gamma})^{2}}\right].
$$

Then we can consistently estimate V_1 and V by

$$
\widehat{V}_1 = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{D}}_i^{\otimes2}, \tag{6}
$$

$$
\widehat{V} = n^{-1}\sum_{i=1}^{n}\left(\widetilde{\mathbf{D}}_i+\widehat{\eta}_i\right)^{\otimes2}, \tag{7}
$$

respectively, where

$$
\widetilde{\mathbf{D}}_i = \sum_{k=1}^{K}\frac{\delta_{ki}^{*}}{\widehat{G}(T_{ki}^{*}\mid\bar{\mathbf{F}}_i)}h(\mathbf{Z}_{ki};\widehat{\beta})(y_{ki}-g(\widehat{\beta}'\mathbf{Z}_{ki}))\mathbf{Z}_{ki}.
$$

Here the estimator $\widehat{V}$ of V is the same as the one given in Lin (2003). Hence l_i, $1 \le i \le p$ can be consistently estimated by the eigenvalues $\widehat{l}_i$'s of $\widehat{V}_1^{-1}\widehat{V}$.

Confidence regions for β can be constructed as follows. Let

$$
R_\alpha(\beta) = \{\beta : l(\beta) \le c_\alpha\}, \tag{8}
$$

where c_α is the Monte Carlo approximation to the $(1-\alpha)th$ quantile of the weighted chi-square distribution $l_1\chi_{1,1}^2 + \cdots + l_p\chi_{p,1}^2$. Then from the

11

earlier discussion, $R_\alpha(\beta)$ gives an approximate confidence region of β with asymptotically correct coverage probability $1 - \alpha$, i.e.,

$$P(\beta_0 \in R_\alpha(\beta)) = 1 - \alpha + o(1).$$

Note that Monte Carlo simulation is needed to calculate the critical value c_α in (8). This can be done by first generating a large number of realizations of $\widehat{l}_1\chi^2_{1,1} + \cdots + \widehat{l}_p\chi^2_{p,1}$ and then taking c_α to be the $(1-\alpha)$-th sample quantile.

Next we describe another method for constructing a confidence region of β without resorting to Monte Carlo simulation. Define

$$r_n(\beta) = \frac{tr(\widehat{V}^{-1}S_n)}{tr(V_{1n}^{-1}S_n)},$$

where

$$V_{1n} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{D}}_i\widehat{\mathbf{D}}_i', \quad S_n = \Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\mathbf{D}}_i\Big)\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\mathbf{D}}_i\Big)',$$

and $\widehat{V}$ is defined by equation (7) in Section 4. Then, by examining the proof of Theorem 1 (see Appendix), we have

$$r_n(\beta)l(\beta) \xrightarrow{\mathcal{L}} r(\beta)\sum_{i=1}^{p}l_i\chi^2_{i,1}, \quad \text{as } n \to \infty,$$

where $r(\beta) = p/tr(V_1^{-1}V)$ with $tr(\cdot)$ denoting the trace operator. Rao and Scott (1981) showed that the distribution of $r(\beta)\sum_{i=1}^{p}l_i\chi^2_{i,1}$ could be approximated by the standard χ^2_p distribution. Therefore, an approximate $1 - \alpha$ confidence region of β_0 can be constructed as follows:

$$\{\beta: \quad r_n(\beta)l(\beta) \le \chi^2_p(\alpha)\}, \tag{9}$$

where $\chi^2_p(\alpha)$ is the $(1-\alpha)$-th quantile of the standard χ^2_p distribution. It is worth noting that the adjustment factor $r_n(\beta)$ can be motivated from the

fact that $r(\beta)$ can be rewritten as $r(\beta) = tr(V^{-1}V)/tr(V_1^{-1}V)$, replacing V^{-1}, V_1^{-1} and V by $\widehat{V}^{-1}$, V_{1n}^{-1} and S_n respectively leads to $r_n(\beta)$.

Before we end this section, we remark that when there is no censoring in the observations, $\eta_i = 0$ for $i = 1, ..., n$, and $l(\beta_0) \xrightarrow{\mathcal{L}} \chi_p^2$. So Theorem 1 reduces to the Wilks' theorem in the context of generalized linear regression models.

5 Empirical likelihood based intervals for the expected total costs

Let z_{k0} and y_{k0} be the covariate value and the total cost of a patient at the kth interval $[t_k, t_{k+1})$, where $k = 1, \ldots, K$. Then, the total cost of this patient over the entire interval $[0, \tau)$ is $Y_0 = \sum_{k=1}^{K} y_{k0}$. We want to construct a confidence interval for $u_0 = \sum_{k=1}^{K} E(y_{k0} \mid z_{k0})$. Based on the assumed generalized linear model in Section 2, we obtain an expression for u_0 as follows:

$$u_0 = \sum_{k=1}^{K} g(\beta' z_{k0}). \tag{10}$$

Let R be the $(1-\alpha)100\%$ empirical likelihood based confidence region for β, as defined in (9). Then, we can obtain a confidence interval for the expected cost u_0 of a patient with $z = (z_{01}, \ldots, z_{0K})'$ as follows:

$$\{\mu(z) = \sum_{k=1}^{K} g(\beta' z_{k0}) : \beta \in R\}. \tag{11}$$

This confidence interval has the coverage probability that is greater than or equal to $1 - \alpha$ with the equality achieved when $g(.)$ is an one-one function.

13

6 Numerical studies

We carried out two simulation studies to compare the finite-sample properties of our proposed method with the method of Lin (2003). Since the confidence interval for the expected cost u_0 is determined by the confidence region of β, in the simulation studies we focus on the coverage accuracy of the confidence regions of β.

In the first simulation, we adopt a similar parameter set-up as in Lin (2003). Survival and censoring times are generated from the exponential distribution with mean m and uniform $(0, c)$ distribution, respectively. The combinations of $(m, c) = (5, 40), (5, 20)$, and $(10, 20)$ yield the mean censored rate of approximately 12.6%, 24.4%, and 43.2%, respectively. We divide the entire study period into three equally spaced intervals. We set

$$
\begin{aligned}
y_{ki} &= \Big[I(k = 1)u_i^d + I(T_i > t_k)(\epsilon_i + u_{ki}) \\
&\quad + I(t_{k-1} < T_i \le t_k)\{(\epsilon_i + u_{ki})(T_i - t_{k-1}) + u_i^f\} \Big] \exp(\xi Z_i)
\end{aligned}
$$

for $k = 1, 2, 3; i = 1, \cdots, n$, where $\epsilon_i, u_{ki}, u_i^d$ and u_i^f are independent random variables with uniform distributions. Specifically, ϵ_i and u_{ki} have the uniform $(0, 1)$ distribution, u_i^d and u_i^f have the uniform $(0, 5)$ and $(0, 10)$ distributions, respectively. This scheme creates J-shaped time patterns. For the same subject, the costs in different intervals share a common random effect and are thus positively correlated. It is easy to see that the cost data satisfy

$$
E[y_{ki}|Z_i] = \mu_k \exp\{\xi Z_i\}
$$

14

So, $\beta = (\xi, \mu_1, \mu_2, \mu_3)$, μ_k is the mean cost in time $(k-1, k]$ for the subject with the covariate $Z = 0$. We choose two different sets of values for (m, u_1, u_2, u_3): $(5, 4.313, 1.484, 1.215)$, and $(10, 3.928, 1.292, 1.1689)$. We set Z to be a treatment indicator with $n/2$ subjects in each of the two groups and ξ to be 1. We choose $n = 100, 200$ and 500 as in Lin (2003). We summarize the results from 500 repetitions in Table 1 along with the coverage accuracy of the confidence regions for β using our method and the normal approximation method based on Lin's approach. Our results for ξ are very similar to those reported in Lin (2003), and hence are not reported in Table 1 as our focus is on β.

Table 1 goes here

In Table 1, EL.CP is the coverage probability of the 95% nominal level confidence region for β, based on the empirical likelihood method. The CP is the coverage probability of the 95% nominal level confidence region for β, based on the normal approximation of $\widehat{\beta}$, given in Lin (2003), and defined by

$$n(\widehat{\beta} - \beta)^T (\widehat{A}^{-1} \widehat{V} \widehat{A}^{-1})^{-1} (\widehat{\beta} - \beta) \leq \chi_p^2(\alpha), \qquad (12)$$

where $\widehat{A}$ and $\widehat{V}$ are consistent estimators of A and V respectively (see also (2) in Section 3). From Table 1 we see that both the empirical likelihood and normal approximation methods yield the confidence regions for β that are close to the nominal level, and the empirical likelihood method is slightly better than the normal approximation method for heavy censoring.

15

Since generated cost observations in Table 1 above are from some uniform distributions, the resulting cost data have an approximately normal distribution. In fact, simulation studies done in Lin's papers (2000a, 2000b, 2003) assumed that cost data followed a normal distribution. However, as we know from the literature (Zhou et al, 1997; Jiang and Zhou, 2004), cost data are not normally distributed but instead are skewed. In the second simulation study, we generate cost data from a skewed distribution. This simulation study is similar to the first one, except that covariates are generated from a normal distribution $N(\nu, \sigma^2)$, where $\nu = 2$, σ is chosen to be 1 and 2, and the coefficient ξ was chosen to be 0.1, 0.2, 0.4, and 0.6. Under this setup, the distribution of the total medical cost of a patient becomes more skewed as σ and ξ increase.

The results with a fixed sample size of 100 from 500 repetitions are summarized in Table 2. With lightly skewed cost data, the improvement in the coverage accuracy of the empirical likelihood based confidence region is minimal compared to the one based on the normal approximation confidence region. But when the skewness increases, the improvement is noticeably significant, and the coverage probability of the empirical likelihood based confidence region is much closer to the nominal level than the normal approximation confidence region.

Table 2 goes here

Numerical studies are also conducted at a larger sample size under the same simulation scheme as in Table 2. Table 3 shows a comparison of the two types of confidence regions with $n = 400$. As the sample size increases, the

16
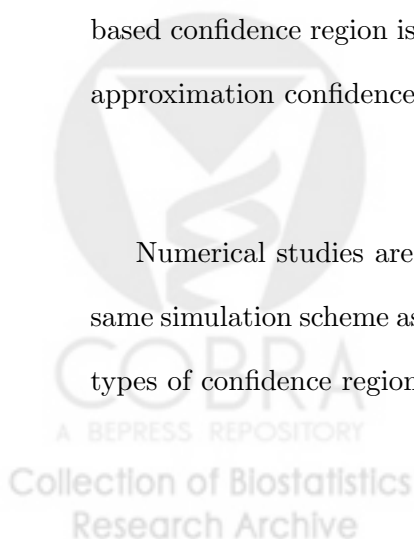
performance of both types of confidence regions improve; however, the coverage probabilities from the normal approximation approach are still much lower than the nominal level when the cost distribution is severely skewed. The empirical confidence region has the better and more robust performance than the normal approximation approach for all the cases considered here.

Table 3 goes here

In summary, the accuracy of the EL-based confidence regions and the normal approximation based regions for β are close when data is less skewed. When cost data are highly skewed, which are likely to occur in practice, the EL-based confidence regions greatly outperform the normal approximation method although there is still room for further improvement.

7 A real data example

To illustrate the application of our methodology, we use the same SEER Medicare database as in Lin (2003). Our data consist of 985 and 2647 patients diagnosed with regional and distant stages of epithelial ovarian cancer, respectively. The data on survival time and monthly medical expenditures are available from 1983 to 1990. The subjects who were still alive at the end of 1990 are censored. There is no voluntary loss to follow-up in this study, so that censoring, which is solely caused by limited study duration, can be regarded as completely random. Thus, the proposed methods with $\widehat{G}$ as the Kaplan-Meier estimator can be used. Since most of the patients did not survive to the 7th year, we confine our attention to the first 6 years after the diagnosis. The focus of our analysis is to provide a confidence interval

17

for the expected total cost of a patient during the first 6 years after the first diagnosis of cancer, using the given covariates of the patient.

From Figure 4 in Lin (2003), we see that the effects of the stages on the cost are not constant over time on either an additive or multiplicative scale. So, we compute the expected total cost on $[0, \tau]$ separately for regional and distant groups. To illustrate the proposed methodology, we also include a continuous covariate Z, the time of the first diagnosis, in the model, where $Z = 0$ corresponds to a new cancer patient. We are interested in constructing a confidence interval for the expected total cost over $[0, \tau]$ for a patient with $Z = z$, where $\tau = 72$ months. Let Y_0 be the total cost over $[0, \tau]$ of a patient with $Z = z_0$. Then, we like to construct a 95% confidence interval for $u_0 = E(Y_0 \mid Z = z_0)$.

Let y_{ki} denote the total cost over the kth month for patient i, $k = 1, \ldots, \tau = 72$, and let Z_i be the value of Z for the ith patient. We fit a separate generalized linear model for patients with regional stages and patients with distant stages, respectively. The fitted model has the following general form:

$$E(y_{ki}|Z_i) = \mu_k \exp\{\xi Z_i\},$$

where $k = 1, \ldots, \tau = 72$. Since there is no closed form for the confidence interval of the expected total cost when the empirical likelihood method is used, we propose a numerical method to determine the EL-based confidence interval. Note that the expected total cost over $[0, \tau]$ is $\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\}$ when $Z = z$ and that the univariate empirical likelihood confidence region is

18

always an interval. Let $\beta = (\xi, \mu_1, \cdots, \mu_\tau)^T$, and R be the 95% confidence region for β. Then, we can write the EL-based confidence interval for the expect total cost on $(0, \tau)$ as (q_0, q_1), where

$$q_0 = \min\left\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : \beta \in R\right\},$$

$$\text{and } q_1 = \max\left\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : \beta \in R\right\}.$$

From (8), we know that we can write q_0 and q_1 as

$$q_0 = \min\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : l(\beta) = c, 0 \leq c \leq c_\alpha\}$$

$$\approx \min\{\cup_{i=1}^{N}\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : l(\beta) = c_i\}\} \text{ for large N,}$$

$$q_1 = \max\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : l(\beta) = c, 0 \leq c \leq c_\alpha\}$$

$$\approx \max\{\cup_{i=1}^{N}\{\sum_{k=1}^{\tau} \mu_k \exp\{\xi z\} : l(\beta) = c_i\}\} \text{ for large N,}$$

where $\{c_1, \ldots, c_N\}$ is a random sample of size N generated from the uniform distribution over $[0, c_\alpha]$. Therefore, for estimating q_0 and q_1, we need to solve the equation $l(\beta) = c$ for any $c \in [0, c_\alpha]$. Tian et al. (2003) proposed a numerical algorithm for a similar problem, but their method requires an initial approximation solution for the equation $l(\beta) = c$ which is difficult to obtain in our case. Therefore, we propose a nonparametric technique to solve $l(\beta) = c$. First, we note that it is feasible to compute $l(\beta)$ for any given β and that R may be approximated by R_0, which is defined by

$$R_0 = \{\beta : \hat{\mu}_k - 1.96\hat{\sigma}_k \leq \mu_k \leq \hat{\mu}_k + 1.96\hat{\sigma}_k, k = 1, \cdots, \tau,$$

$$\text{and } \hat{\xi} - 1.96\hat{\sigma} \leq \xi \leq \hat{\xi} + 1.96\hat{\sigma}\},$$

19

where $\hat{\sigma}_k$ is the estimator of the standard error of $\hat{\mu}_k, k = 1, \cdots, \tau$ and $\hat{\sigma}$ is the estimator of the standard error of $\hat{\xi}$. By generating J vectors $\beta^{(j)}, j = 1, \cdots, J$ uniformly over R_0 that satisfy $l(\beta)^{(j)} \leq c_\alpha$, we can estimate β which satisfies $l(\beta) = c$ for any given $c \in [0, c_\alpha]$ by a smoothing technique (for example, local linear or spline) based on data $(\beta^{(j)}, l(\beta^{(j)})), j = 1, \cdots, J$, where the value of J depends on the number of parameters. In our example, for $\tau = 12$, since the number of the parameters is 13, we take $J = 1000$ because the results do not change significantly as J is chosen to be greater than 10000. Similarly, for $\tau = 24$, we choose $J = 2000$, and for $\tau = 72$, we choose $J = 20000$.

In Tables 4 and 5, we report the 95% confidence interval for $u_0 = E(Y_0 \mid Z = z_0)$ when $z_0 = 0$. The EL-based confidence interval is wider than the interval based on the normal approximation. The result is consistent with our simulation results which have shown that the normal approximation interval has a coverage probability that is lower than the nominal level while the EL based interval has a coverage probability that is close to the nominal level.

Tables 4 and 5 go here

8   Discussion

In this paper we develop an empirical likelihood (EL) based interval estimation method for the expected total cost of a patient with given covariates over a certain period when costs of some patients were censored. The issue of correctly predicting such an expected cost has important implications

20
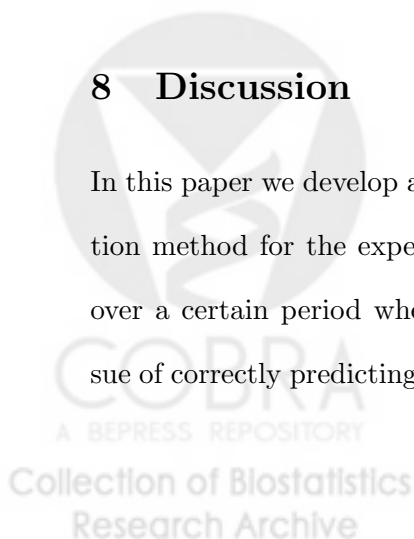
in health economics, especially in prospective payment systems. We have also developed the underlying asymptotic theory for the proposed EL-based method and conducted a simulation study to compare its performance with the existing method in finite-sample sizes. Our simulation results show that the proposed EL-based method performs equally well with the existing method when cost data are not so skewed, and outperforms the existing method when cost data are moderately or highly skewed in terms of coverage accuracy in almost all cases.

Since in almost all cost studies, cost data are skewed, and many of them have the skewness of greater than 1.0 and a sample size between 100 and 4000 (see Katon et al (2004) and Liu et al (2003)), we believe that our new method has more practical relevance that the existing method.

The EL have better coverage probability than the direct normal approximation, which is a phenomenon happened in many applications of EL methods. For example, see Qin and Jing (2001); Qin and Tsao (2003); Li and Wang(2003); Wang, Linton, and Hrdle(2004) among others. The future research will be in the direction of finding the edgeworth expansion for the coverage probability of EL intervals, which may shed some light on why EL method having better coverage accuracy than the direct normal approximation.

As noticed by the referee, the EL confidence intervals can have poor coverage, which occur when the data is seriously skewed. A future research direction is to see whether we can obtain better intervals if we can find a transformation that can transform the original data into less skewed or

21

almost symmetric data before we apply our EL method.

<div align="center">Acknowledgements</div>

References

Bang, H. and Tsiatis, A.A. (2000). Estimating medical costs with censored data. *Biometrika*, **87**, 329-343.

Bang, H. and Tsiatis, A.A. (2002). Median regression with censored cost data. *Biometrics*, **58**, 643-649.

DiCiccio, T. J., Hall, P., and Romano, J. (1991). Empirical likelihood is Barlett-correctable. *Annals of Statistics*, **19**, 1053-1061.

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

Etzioni, R. D., Feuer, E. J., Sullivan, S. D., Lin D. Y., Hu, C., and Ramsey, R. D. (1999). On the use of survival analysis techniques to estimate medical care costs. *Journal of Health Economics*, **18**, 365-380.

Jiang, H. and Zhou, X. H. (2004). Comparison of several independent population means when their sample contain log-normal and possibly zero observations. *Statistics in Medicine,* in press.

Katon WJ, Von Korff M, Lin EH, Simon G, Ludman E, Russo J, Ciechanowski P, Walker E, and Bush T. (2004). The Pathways Study: a randomized trial of collaborative care in patients with diabetes and depression. Arch Gen Psychiatry;61:1042-9.

Li, G. and Wang, Q. (2003). Empirical likelihood regression analysis for right censored data. *Statistica Sinica*, **13**, 51-68.

Liu CF, Hedrick SC, Chaney EF, Heagerty P, Felker B, Hasenberg N, Fihn S, and Katon W. (2003). Cost-effectiveness of collaborative care for depression in a primary care veteran population. Psychiatr Serv. ;54(5):698-704.

Lin, D. Y. (2000a). Linear regression of censored medical costs. *Biostatistics*, **1**, 35-47.

Lin, D. Y. (2000b). Proportional means regression for censored medical costs. *Biometrics*, **56**, 775-778.

Lin, D. Y. (2003). Regression analysis of incomplete medical cost data. *Statistics in Medicine*, **15**, 1181-1200.

Lin, D. Y., Fleming, T. R. and Wei, L. J. (1994). Confidence bands for survival curves under proportional hazards model. *Biometrika*, **80**, 573-581.

Maciejewski, M. L., Zhou, X. H., Fortney, J. C., and Burgess, J. F. (2005). Alternative Methods for Modelling Heteroscedastic Non-normally Dis-

23

tributed Costs. **Submitted**.

Owen, A. (1990) Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Owen, A. (1991). Empirical likelihood for linear models. *Ann. Statist.* **19**, 1725-1747.

Owen, A. (2001). Empirical Likelihood. *Chapman & Hall/CRC.* New York.

Qin, G. S., and Jing, B. Y. (2001). Censored partial linear models and empirical likelihood. *J. Multivariate Analysis*, **78**, 37-61.

Rao, J.N.K. and Scott, A.J. (1981). The Analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fits and independence in two-way tables. *J. Amer. Statist. Assoc.*, 76, 221-230.

Tian,L., Liu, J., Zhao, M., and Wei, L.J.(2003). Statistical inferences based on non-smooth estimating functions. Harvard University Biostatistics Working Paper Series.

Wang, Q; Linton, O; Hrdle, W. (2004). Semiparametric regression analysis with missing response at random. *J. Amer. Statist. Assoc.*, 99, 334-345.

Zhou, X. H. and Tu, W. (1999). Comparison of several independent population means when their sample contain log-normal and possibly zero observations. *Biometrics,* 55, 645-651.

Zhou, X. H., Gao, S., and Hui, S. L. (1997). Methods for comparing the means of two independent log-normal samples . *Biometrics,* 53, 1129-1135.

Zhou, X. H., Stroupe, K. T., and Tierney, W. M. (2001). Regression analysis of health care charge data with heteroscedasticity. *JRSS Series C,* 50,

24

303-312.

Appendix. Proof of Theorem 1

We need a few lemmas for proving Theorem 1.

Lemma 1. (See Lin, 2003)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i \xrightarrow{\mathcal{L}} N(0, V)$$

Lemma 2. We have the following properties for $\widehat{\mathbf{D}}_i$.

$$(i) \quad \max_i \|\widehat{\mathbf{D}}_i\| = o_p(n^{1/2}) \qquad (ii) \quad \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i \widehat{\mathbf{D}}_i' \xrightarrow{p} V_1.$$

Proof of the Lemma 2.

(i). From the condition **C4**, we have

$$\max_i \|\mathbf{D}_i\| = o_p(n^{1/2}).$$

Using the uniform consistency of Kaplan-Meier estimator and Breslow estimator, we get

$$
\begin{aligned}
\widehat{\mathbf{D}}_i - \mathbf{D}_i &= \sum_{k=1}^{K} \frac{\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i) - G(T_{ki}^* \mid \bar{\mathbf{F}}_i)}{G(T_{ki}^* \mid \bar{\mathbf{F}}_i)\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i)} \delta_{ki}^* h(\mathbf{Z}_{ki}; \beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))\mathbf{Z}_{ki} \quad (13) \\
&= o_p(1)
\end{aligned}
$$

uniformly for $i = 1, ..., n$. So,

$$\max_i \|\widehat{\mathbf{D}}_i\| \leq \max_i \|\widehat{\mathbf{D}}_i - \mathbf{D}_i\| + \max_i \|\mathbf{D}_i\| = o_p(n^{1/2})$$

25

(ii) Let $\widetilde{V}_1 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i\mathbf{D}_i'$. Note that $V_{1n} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{D}}_i\widehat{\mathbf{D}}_i'$. For any $\mathbf{a} \in \mathbf{R}^p$, we have the following decomposition:

$$
\begin{aligned}
\mathbf{a}'\left(V_{1n} - \widetilde{V}_1\right)\mathbf{a} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{a}'(\widehat{\mathbf{D}}_i - \mathbf{D}_i)\right)^2 + \frac{2}{n}\sum_{i=1}^{n}(\mathbf{a}'\mathbf{D}_i)\left(\mathbf{a}'(\widehat{\mathbf{D}}_i - \mathbf{D}_i)\right) \\
&\leq \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}|\mathbf{a}'(\widehat{\mathbf{D}}_i - \mathbf{D}_i)|\right)\left(\frac{1}{\sqrt{n}}\max_i|\mathbf{a}'(\widehat{\mathbf{D}}_i - \mathbf{D}_i)| + \frac{2}{\sqrt{n}}\max_i|\mathbf{a}'\mathbf{D}_i|\right) \\
&\equiv J_0(J_1 + 2J_2).
\end{aligned}
\tag{14}
$$

From the proof of (i), we obtain that $J_1 = o_p(1)$ and $J_2 = o_p(1)$. Now let's look at the term J_0. If the Kaplan-Meier estimator $\widehat{G}$ is used as the estimator of G, using (13) and the following martingale representation for $\widehat{G}$,

$$
\frac{n^{1/2}(G(t) - \widehat{G}(t))}{G(t)} = n^{-1/2}\sum_{j=1}^{n}\int_0^t \frac{dM_j(x)}{P(X_j \geq x)} + o_p(1),
$$

we have

$$
\begin{aligned}
J_0 &= \left|n^{-1/2}\sum_{j=1}^{n}\int_0^{\infty} q_1(t)dM_j(t)\right| + o_p(1) \\
&= O_p(1) + o_p(1) = O_p(1),
\end{aligned}
$$

where

$$
q_1(t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\left|\sum_{k=1}^{K}\frac{\delta_{ki}^* I(T_{ki}^* > t)}{\widehat{G}(T_{ki}^* \mid \bar{\mathbf{F}}_i)P(X_i \geq t)}h(\mathbf{Z}_{ki};\beta)(y_{ki} - g(\beta'\mathbf{Z}_{ki}))(\mathbf{a}'\mathbf{Z}_{ki})\right|
$$

Similarly, if we use the Breslow estimator $\widehat{G}(t \mid \bar{\mathbf{F}})$ of $G(t \mid \bar{\mathbf{F}})$, using (13) and the following representation due to Lin, Fleming and Wei (1994), we obtain that

$$
\begin{aligned}
\frac{n^{1/2}\left(G(t \mid \bar{\mathbf{F}}) - \widehat{G}(t \mid \bar{\mathbf{F}})\right)}{G(t \mid \bar{\mathbf{F}})} &= n^{-1/2}\sum_{j=1}^{n}\int_0^t \frac{e^{\gamma'\mathbf{W}(x)}dM_j(x)}{s^{(0)}(x)} \\
&\quad + \mathbf{r}'(t;\mathbf{W})\Omega^{-1}n^{-1/2}\sum_{j=1}^{n}\int_0^{\infty}\left[\mathbf{W}_j(x) - \bar{\mathbf{w}}(x)\right]dM_j(x) + o_p(1).
\end{aligned}
$$

26

Hence we can also get $J_0 = O_p(1)$. Therefore $V_{1n} = \widetilde{V}_1 + o_p(1)$, and Lemma 2(ii) is thus proved.

Proof of Theorem 1. Applying Taylor's expansion to (4), we get

$$l(\beta_0) = 2\sum_{i=1}^n \log\{1 + \lambda'\widehat{\mathbf{D}}_i\} = 2\sum_{i=1}^n \left(\lambda'\widehat{\mathbf{D}}_i - \frac{1}{2}(\lambda'\widehat{\mathbf{D}}_i)^2\right) + r_n, \qquad (15)$$

where

$$|r_n| \leq C\sum_{i=1}^n (\lambda'\widehat{\mathbf{D}}_i)^3 \quad \text{in probability.}$$

Write $\lambda = \kappa\theta$, where $\kappa \geq 0$ and $\|\theta\| = 1$. From the proof of Lemma 2(ii), we get

$$\theta'V_{1n}\theta = \theta'\widetilde{V}_1\theta + o_p(1).$$

Then, using Lemma 1, Lemma 2(ii), and the argument similar to the one in Owen (1990), we can show that

$$\|\lambda\| = O_p(n^{-1/2}). \qquad (16)$$

Hence, using (16) and Lemma 2 together we obtain

$$|r_n| \leq C\|\lambda\|^3 \max_{1 \leq i \leq n} \|\widehat{\mathbf{D}}_i\| \sum_{i=1}^n \|\widehat{\mathbf{D}}_i\|^2 = o_p(1). \qquad (17)$$

Note that

$$\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\mathbf{D}}_i}{1 + \lambda'\widehat{\mathbf{D}}_i} = \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{D}}_i \left[1 - \lambda'\widehat{\mathbf{D}}_i + \frac{(\lambda'\widehat{\mathbf{D}}_i)^2}{1 + \lambda'\widehat{\mathbf{D}}_i}\right]$$

$$= \frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{D}}_i - \left(\frac{1}{n}\sum_{i=1}^n \widehat{\mathbf{D}}_i\widehat{\mathbf{D}}_i'\right)\lambda + \frac{1}{n}\sum_{i=1}^n \frac{\widehat{\mathbf{D}}_i\left(\lambda'\widehat{\mathbf{D}}_i\right)^2}{1 + \lambda'\widehat{\mathbf{D}}_i}.$$

27

From (3), (16), and Lemma 2, it follows that

$$\lambda = \left(\sum_{i=1}^{n} \widehat{\mathbf{D}}_i \widehat{\mathbf{D}}_i'\right)^{-1} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i + o_p(n^{-1/2}). \tag{18}$$

Again by (3), we get that

$$0 = \sum_{i=1}^{n} \frac{\lambda' \widehat{\mathbf{D}}_i}{1 + \lambda' \widehat{\mathbf{D}}_i} = \sum_{i=1}^{n} (\lambda' \widehat{\mathbf{D}}_i) - \sum_{i=1}^{n} (\lambda' \widehat{\mathbf{D}}_i)^2 + \frac{1}{n} \sum_{i=1}^{n} \frac{(\lambda \widehat{\mathbf{D}}_i')^3}{1 + \lambda' \widehat{\mathbf{D}}_i}. \tag{19}$$

By (16) and Lemma 2, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(\lambda' \widehat{\mathbf{D}}_i)^3}{1 + \lambda' \widehat{\mathbf{D}}_i} = o_p(1). \tag{20}$$

From (19) and (20), we get

$$\sum_{i=1}^{n} \lambda' \widehat{\mathbf{D}}_i = \sum_{i=1}^{n} (\lambda' \widehat{\mathbf{D}}_i)^2 + o_p(1). \tag{21}$$

By (15), (17), (18) and (21), we get

$$
\begin{aligned}
l(\beta_0) &= \sum_{i=1}^{n} \lambda' \widehat{\mathbf{D}}_i \widehat{\mathbf{D}}_i' \lambda + o_p(1) \\
&= \left(n^{-1/2} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i\right)' \left(n^{-1} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i \widehat{\mathbf{D}}_i'\right)^{-1} \left(n^{-1/2} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i\right) + o_p(1) \\
&= \left(V^{-1/2} n^{-1/2} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i\right)' \left(V^{1/2} V_1^{-1} V^{1/2}\right) \left(V^{-1/2} n^{-1/2} \sum_{i=1}^{n} \widehat{\mathbf{D}}_i\right) + o_p(1).
\end{aligned}
$$

Then Theorem 1 directly follows from Lemma 1, Lemma 2(ii) and Lemma 5 in Qin and Jing (2001).

28

Table 1: Coverage accuracy of confidence regions for β with the symmetric distribution

m	c	n	censored rate	CP	EL.CP
5	40	100	0.126	0.916	0.913
		200		0.922	0.920
		500		0.942	0.932
5	20	100	0.244	0.902	0.911
		200		0.920	0.918
		500		0.938	0.938
10	20	100	0.432	0.916	0.929
		200		0.928	0.936
		500		0.938	0.938

29

Table 2: Simulation results for the asymmetric distribution($n = 100$)

		censored				β	
m	c	rate	σ	ξ	Skewness	CP	EL.CP
5	40	0.1256	1	0.1	0.7841	0.9128	0.9226
			2	0.1	0.9763	0.8887	0.9085
			1	0.2	0.9763	0.8800	0.8972
			2	0.2	1.5151	0.7816	0.8360
			1	0.4	1.5151	0.7800	0.8283
			2	0.4	2.7259	0.5000	0.7419
			1	0.6	2.1101	0.6480	0.7445
			2	0.6	3.9317	0.2773	0.6721
5	20	0.2444	1	0.1	0.7841	0.9063	0.9163
			2	0.1	0.9763	0.8864	0.9106
			1	0.2	0.9763	0.8760	0.8994
			2	0.2	1.5151	0.7711	0.8421
			1	0.4	1.5151	0.7680	0.8347
			2	0.4	2.7259	0.4900	0.7425
			1	0.6	2.1101	0.6600	0.7404
			2	0.6	3.9317	0.2872	0.6755
10	20	0.4318	1	0.1	1.0155	0.9047	0.9165
			2	0.1	1.1760	0.8763	0.8925
			1	0.2	1.1760	0.8700	0.8880
			2	0.2	1.6308	0.7856	0.8193
			1	0.4	1.6308	0.7840	0.8096
			2	0.4	2.7321	0.5140	0.7379
			1	0.6	2.1592	0.6460	0.7264
			2	0.6	3.9001	0.2990	0.6862

30

Table 3: Simulation results for the asymmetric distribution($n = 400$)

		censored				θ	
m	c	rate	σ	ξ	Skewness	CP	EL.CP
5	40	0.1247	1	0.1	0.8221	0.948	0.950
			2	0.1	1.0491	0.952	0.952
			1	0.2	1.0491	0.952	0.952
			2	0.2	1.7364	0.920	0.920
			1	0.4	1.7364	0.920	0.920
			2	0.4	3.6446	0.700	0.832
			1	0.6	2.6122	0.840	0.878
			2	0.6	5.9674	0.458	0.734
5	20	0.2455	1	0.1	0.8221	0.936	0.940
			2	0.1	1.0491	0.942	0.946
			1	0.2	1.0491	0.942	0.946
			2	0.2	1.7364	0.932	0.932
			1	0.4	1.7364	0.932	0.928
			2	0.4	3.6446	0.696	0.830
			1	0.6	2.6122	0.846	0.886
			2	0.6	5.9674	0.470	0.734
10	20	0.4334	1	0.1	1.0477	0.932	0.936
			2	0.1	1.2323	0.924	0.932
			1	0.2	1.2323	0.924	0.932
			2	0.2	1.8215	0.916	0.926
			1	0.4	1.8215	0.916	0.926
			2	0.4	3.5986	0.724	0.829
			1	0.6	2.6179	0.842	0.869
			2	0.6	5.9025	0.450	0.723

Table 4: The average cost for the regional-stage patients in the first 6 years

τ	average cost	95%CI(normal)	95%CI(EL)
12	31638.17	[30325.31, 32951.03]	[28928.03, 35801.20]
24	45321.87	[43049.58, 47594.16]	[40075.41, 51216.07]
36	56053.82	[52619.14, 59488.51]	[49916.64, 60339.97]
48	63734.03	[59266.65, 68201.42]	[56401.28, 74029.19]
60	71861.62	[66095.95, 77627.29]	[63872.28, 84207.23]
72	77967.85	[71267.34, 84668.37]	[70927.91, 89366.97]

31

Table 5: The average cost for the distant-stage patients in the first 6 years

τ	average cost	95%CI(normal)	95%CI(EL)
12	38028.34	[37007.67, 39049.00]	[35195.77, 41972.41]
24	56373.66	[54557.11, 58190.21]	[51378.09, 62906.75]
36	70895.30	[68057.08, 73733.51]	[66108.69, 76880.35]
48	82330.58	[78034.04, 86627.12]	[76459.93, 88756.02]
60	92018.35	[86056.14, 97980.55]	[84334.12, 100357.07]
72	99249.84	[91981.43, 106518.24]	[91162.97, 109599.23]

32