Far East J. Theo. Stat. **16**(1) (2005), 105-112
(A special volume on Biostatistics and related areas)

EMPIRICAL LIKELIHOOD INFERENCE FOR THE ADDITIVE-MULTIPLICATIVE HAZARD MODEL

YICHUAN ZHAO and ZHIHENG XU

Department of Mathematics and Statistics
Georgia State University
Atlanta, GA 30303, U. S. A.
e-mail: yzhao@mathstat.gsu.edu

Winship Cancer Institute and Department of Biostatistics
Emory University
Atlanta, GA 30323, U. S. A.

Abstract

Cox's proportional hazards model has so far been the most popular model for the regression analysis of censored survival data. This model has also been extended to the general additive-multiplicative hazard model. In this paper, we apply empirical likelihood ratio method to this model and derive the limiting distribution of the empirical likelihood ratio. Based on the result, we construct a confidence region for the regression parameter. Simulation studies are conducted to evaluate the performance of the proposed method.

1. Introduction

Recently, survival analysis has received a lot of attention and experienced tremendous growth in medical studies. We are interested in the regression model for the survival rate incorporating information from the covariates which allows us to quantify the relationship between the failure time and a set of explanatory variables. Cox [1] proportional

2000 Mathematics Subject Classification: Primary 62G10; Secondary 62G05.

Key words and phrases: counting process, martingale, regression, right censoring.

Received February 21, 2005

hazards model has been considered as a major tool for regression analysis of survival data. Under this model, the covariates $Z(\cdot)$ have multiplicative effects on the hazard function. Cox [1, 2] introduced a semiparametric approach for the proportional hazards model which has contributed tremendously to clinical trial studies.

However, the proportional hazards assumption may not be appropriate for some data analyses in medical research, the additive risk model [3] provides a useful alternative to Cox's proportional hazards model for studying the association between failure time and risk factors through hazard functions. This valuable model for the analysis of survival data seems simple and easy to interpret for medical researchers. In some applications, we should consider models which allow certain covariates to have both the additive and multiplicative effects. To enhance the modeling capability, a general additive-multiplicative hazard model [4, 5] has been proposed:

$$\lambda(t; Z) = g\{\beta_0^T W(t)\} + \lambda_0(t) h\{\gamma_0^T X(t)\}, \tag{1.1}$$

where $Z = (W^T, X^T)^T$ is a p-vector of covariates, $\theta_0 = (\beta_0^T, \gamma_0^T)^T$ is a p-vector of unknown regression parameters, g and h are known link functions and λ_0 is an unspecified baseline hazard function.

Empirical likelihood (EL) method, first introduced by Thomas and Grunkemeier [9], is a powerful nonparametric method. They applied the method to survival probabilities estimated by the Kaplan-Meier curve which can be obtained as a nonparametric maximum likelihood estimate of the survival function from censored data. Owen [6, 7] introduced empirical likelihood confidence regions for the mean of a random vector based on independent and identical complete data. Since then, the empirical likelihood has been widely applied to statistical models to make inference. Recent work includes Qin and Jing [8], Zhao and Hsu [10], among others.

In the present article, we focus on model (1.1), make full use of the estimating function provided by Lin and Ying [4], and find one tractable likelihood-ratio based confidence region for the unknown regression parameters. Our approach does not require to estimate the limiting

covariance matrices; instead one carries out a constrained maximization of the empirical likelihood, which can be done reliably by Newton-Raphson method. Moreover, the EL confidence region is adapted to the data set and reflects the nature of the underlying data and hence give a more representative way to make inferences about the parameter of interest. The proposed confidence region and main asymptotic result are presented in Section 2. In Section 3, we conduct simulation to investigate the performance of the empirical likelihood method in terms of coverage probability. Proof is given in the Appendix.

2. Main Results

2.1. Preliminaries

Let T denote the failure time and C denote the censoring time. Suppose that data consist of n independent samples of (X_i, δ_i, Z_i), where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \le C_i)$. Let $N_i(t) = \delta_i I(X_i \le t)$ be a counting process for the ith subject, which indicates that the failure time of the ith subject is observed up to time t. Let $Y_i(t) = I(X_i \ge t)$ denote the predictable indicator process indicating whether or not the ith subject is at risk just before time t. $Z_i(\cdot)$ is the covariate process for the ith subject. Let τ satisfy $P(X_i > \tau) > 0$.

We denote

$$M_i(\theta, t) = N_i(t) - \int_0^t Y_i(s)[g\{\beta^T W_i(s)\} + h\{\gamma^T X_i(s)\}\lambda_0(s)]ds.$$

Let $D_i(\theta, t)$ be a p-dimensional predictable process, which is a smooth function of Z_i and θ not involving λ_0. Lin and Ying [4, 5] proposed the following estimating function:

$$\begin{aligned} S(\theta) &= \sum_{i=1}^{n} \int_0^{\tau} D_i(\theta, s)[dN_i(s) - Y_i(s)g\{\beta^T W_i(s)\}ds \\ &\qquad - Y_i(s)h\{\gamma^T X_i(s)\}d\hat{\Lambda}_0(\theta, s)] \\ &= \sum_{i=1}^{n} \int_0^{\tau} \{D_i(\theta, s) - \overline{D}(\theta, s)\}[dN_i(s) - Y_i(s)g\{\beta^T W_i(s)\}ds] \end{aligned}$$

$$= \sum_{i=1}^{n} \int_0^{\tau} \{D_i(\theta, s) - \overline{D}(\theta, s)\} dM_i(\theta, s), \tag{2.1}$$

with

$$\overline{D}(\theta, t) = \frac{\sum_{i=1}^{n} Y_i(t) h\{\gamma^T X_i(t)\} D_i(\theta, t)}{\sum_{i=1}^{n} Y_i(t) h\{\gamma^T X_i(t)\}}.$$

2.2. EL confidence region

Now consider empirical likelihood approach. It is clear that $E[S(\theta_0)] = 0$ from the estimating equation (2.1) since the $M_i(\theta_0, t)$ is a martingale. For $1 \le i \le n$, we define

$$W_{n,i} = \int_0^{\tau} \left(D_i(\theta_0, t) - \frac{\hat{\alpha}_1(t)}{\hat{\alpha}_0(t)} \right) dM_i(\theta_0, t), \tag{2.2}$$

where $\hat{\alpha}_r(t) = n^{-1} \sum_i D_i^{\otimes r}(\theta_0, t) Y_i(t) h(\gamma_0^T X_i(t))$ with $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$. Therefore, an empirical likelihood at the true value θ_0 is given by

$$L(\theta_0) = \sup\left\{ \prod_{i=1}^{n} p_i : \sum p_i = 1, p_i \ge 0, \sum_{j=1}^{n} p_j W_{n,j} = 0 \right\}.$$

Thus, the empirical likelihood ratio at the θ_0 is defined by

$$R(\theta_0) = \sup\left\{ \prod_{i=1}^{n} n p_i : \sum p_i = 1, p_i \ge 0, \sum_{j=1}^{n} p_j W_{n,j} = 0 \right\}.$$

By using Lagrange multipliers, we have

$$-2 \log R(\theta_0) = 2 \sum_{i=1}^{n} \log(1 + \lambda^T W_{n,i}),$$

where $\lambda = (\lambda_1, ..., \lambda_p)^T$ satisfies the equation

$$\frac{1}{n} \sum_{i=1}^{n} \frac{W_{n,i}}{1 + \lambda^T W_{n,i}} = 0.$$

Denote

$$\Gamma = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\{D_i(\theta_0, t) - \overline{D}(\theta_0, t)\}^{\otimes 2} dN_i(t). \tag{2.3}$$

Assume that Γ is positive definite. Suppose that $g(t)$, $h(t)$, and $\lambda_0(t)$ are continuous and covariate vector Z_i is time-invariant and bounded. Now, we state our main result and explain how it can be used to construct confidence region for θ.

Theorem 2.1. *Under above conditions, we have* $-2\log R(\theta_0)$ *converges to chi-squared distribution with p degrees of freedom.*

Then, an asymptotic $100(1-\alpha)\%$ confidence region for θ is given by

$$R = \{\theta : -2\log R(\theta) \le \chi_p^2(\alpha)\}, \tag{2.4}$$

where $\chi_p^2(\alpha)$ is the upper α-quantile of the chi-squared distribution with p degrees of freedom.

3. Simulation Study

In this section we investigate the performance of the proposed empirical likelihood (EL) confidence region in terms of coverage probability.

Let $g(x) = x$ and $h(x) = e^x$. Consider a simple general additive-multiplicative model with $\beta_0 = 1$ and $\gamma_0 = 1$. The model is

$$\lambda(t \mid Z) = W(t) + \lambda_0 e^{X(t)}, \tag{3.1}$$

where $Z(t) = (W(t), X(t))^T$.

Let T denote failure time, C denote the censoring time. Censoring rate is approximately 30%, 50%, and 70%, respectively. The sample size n is chosen to be 50, 70, and 100, respectively. W and X are drawn from uniform distribution $U(0, 1)$. Suppose W and X are independent uniform variables, and n random samples are selected for W_i and X_i,

respectively. To simulate failure time T_i, we assume model (3.1) in which $\lambda_0 = 1$ has an Exponential (1) distribution. Censoring time C_i's are drawn from the uniform distribution $U(0, k)$, where k is chosen to get a desired censoring rate which is chosen to be 30%, 50%, and 70%, respectively. Finally, simulated observations from the above general additive-multiplicative model are (Y_i, Z_i, δ_i) for $i = 1, ..., n$, where $Y_i = \min(T_i, C_i)$, $Z_i = (W_i, X_i)$, and $\delta_i = I(T_i \leq C_i)$.

Such simulation is repeated 1,000 times to generate simulated data. Then, the coverage probabilities for the empirical likelihood methods based on these 1,000 simulated data sets are simply the proportions of these data sets which satisfy the inequalities (2.4). The nominal confidence level $1 - \alpha$ is taken to be 0.90, 0.95, and 0.99, respectively. The simulation results are presented in Table 3.1.

Table 3.1. EL coverage probabilities for $\theta = (\beta, \gamma)^T$ at three significant levels

Censoring Rate	Sample Size (n)	$1 - \alpha = 90\%$	$1 - \alpha = 95\%$	$1 - \alpha = 99\%$
30%	50	0.858	0.923	0.952
	70	0.869	0.932	0.969
	100	0.874	0.940	0.975
50%	50	0.861	0.909	0.973
	70	0.862	0.932	0.973
	100	0.884	0.933	0.976
70%	50	0.846	0.910	0.975
	70	0.863	0.932	0.978
	100	0.881	0.942	0.984

From Table 3.1, we make the following observations.

1. Generally, at each nominal level, the coverage accuracy for the empirical likelihood method increases as the sample size n increases.

2. The empirical likelihood method works well at three different censoring rates.

3. The coverage probabilities of empirical likelihood method appear to be close to the nominal levels especially with moderate sample size $(n = 100)$.

Thus, the proposed approach can be applied to this model and leads to reasonable results.

Appendix

Proof of Theorem 2.1. Let

$$W_i = \int_0^\tau \left\{ D_i(\theta_0, t) - \frac{\alpha_1(t)}{\alpha_0(t)} \right\} dM_i(\theta_0, t),$$

where

$$\alpha_r(t) = E(D_i^{\otimes r}(\theta_0, t) Y_i(t) h(\gamma_0^T X_i(t))), \quad r = 0, 1.$$

For any $a \in R^p$, we have $\sup_{0 \le t \le \tau} E[n(\hat{\alpha}_0(t) - \alpha_0(t))^2] = 0(1)$, and $\sup_{0 \le t \le \tau} E[n(a^T \hat{\alpha}_0(t) - a^T \alpha_0(t))^2] = 0(1)$. Following the proof of Lemma 2 of Qin and Jing [8], we have for each i

$$E(a^T(W_{n,i} - W_i))^2 = \int_0^\tau E\left(\frac{a^T \hat{\alpha}_1(t)}{\hat{\alpha}_0(t)} - \frac{a^T \alpha_1(t)}{\alpha_0(t)} \right)^2 Y_i(t) [g\{\beta_0^T W_i(t)\} + h\{\gamma_0^T X_i(t)\} \lambda_0(t)] dt$$

$$\le M \int_0^\tau E\left(\frac{a^T \hat{\alpha}_1(t)}{\hat{\alpha}_0(t)} - \frac{a^T \alpha_1(t)}{\alpha_0(t)} \right)^2 dt$$

$$= 0(1),$$

where in the second step we use the fact $\lambda_0(t)$, $g(t)$, and $h(t)$ are continuous and Z_i are bounded and $0 \le Y_i(t)[g\{\beta_0^T W_i(t)\} + h\{\gamma_0^T X_i(t)\} \lambda_0(t)] \le M$, here M is a constant.

Applying the same argument as that in Qin and Jing [8], we get

$$\sum_{i=1}^{n} W_{n,i} W_{n,i}^T / n \xrightarrow{P} \Gamma \quad \text{and} \quad \max_{1 \le i \le n} \| W_{n,i} \| = o_P(n^{1/2}).$$

From Lin and Ying [4], it is obvious that $n^{-1/2} \sum_{i=1}^{n} W_{n,i} \xrightarrow{D} N(0, \Gamma)$. Combining these results, we conclude Theorem 2.1 by the argument of Owen [7].

References

[1] D. R. Cox, Regression models and life-tables (with discussion), J. Roy. Statist. Soc. Ser. B 34 (1972), 187-220.

[2] D. R. Cox, Partial likelihood, Biometrika 62 (1975), 269-276.

[3] D. Y. Lin and Z. Ying, Semiparametric analysis of the additive risk model, Biometrika 81 (1994), 61-71.

[4] D. Y. Lin and Z. Ying, Semiparametric analysis of general additive-multiplicative hazard models for counting processes, Ann. Statist. 23 (1995), 1712-1734.

[5] D. Y. Lin and Z. Ying, Additive hazards regression models for survival data, Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis, D. Y. Lin and T. R. Fleming, eds., Springer-Verlag, New York, 1997, pp. 185-198.

[6] A. Owen, Empirical likelihood ratio confidence intervals for a single functional, Biometrika 75 (1988), 237-249.

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

[8] G. Qin and B. Y. Jing, Empirical likelihood for Cox regression model under random censorship, Comm. Statist. Simulation Comput. 30 (2001), 79-90.

[9] D. R. Thomas and G. L. Grunkemeier, Confidence interval estimation for survival probabilities for censored data, J. Amer. Statist. Assoc. 70 (1975), 865-871.

[10] Y. Zhao and Y. S. Hsu, Semiparametric analysis for additive risk model via empirical likelihood, Comm. Statist. Simulation Comput. 34 (2005), 135-143.

■