Youngnam Statistical Letters
June 1993, Vol. 4 , 1-11

Comparison of Survival Function Estimators for the Cox's Regression Model using Bootstrap Method [1]

Young Joon Cha[2]

ABSTRACT

The Cox's regression model is frequently used for covariate effects in survival data analysis. But, much of the statistical work has focused on asymptotic behavior, so the small sample evaluation has been neglected. In this paper, we compare the small or moderate sample performances of the survival function estimators for the Cox's regression model using bootstrap method. The smoothed PL type estimator and the Link estimator are slightly better than corresponding the PL type estimator and the Nelson type estimator in the sense of the achieved error rates.

1. Introduction

The Cox's regression model plays a significant role in survival analysis with covariate effects. The model assumes that the hazard function $\lambda(t;z)$ at time t for an individual with a column vector of covariates, z, is given by

$$\lambda(t;z) = \exp(\beta' z)\lambda_0(t), \tag{1.1}$$

[1] This paper was supported by Non Directed Research Fund, Korea Research Foundation, 1991-1992.
[2] Department of Statistics, Andong National University, Andong, 760-749, Korea.

where β is a column vector of unknown parameters and $\lambda_0(t)$ is the hazard for a base level individual that is left unspecified.

A rigorous development of the asymptotic properties of the Cox's regression model has been given by Breslow(1974), Kalbfleisch and Prentice(1980), Tsiatis(1981), Andersen and Gill(1982), Andersen and Borgan(1985), and Arjas and Haara(1988). As seen in the preceeding paragraphs, much of the statistical work for this model has focused on asymptotic behavior, so the small-sample properties have been neglected. A small-smaple estimation evaluation using bootstrap method, of particular interest to the applied statistician, is presented in Section 2. An example is illustrated in Section 3 using experimental Leukemia data set. Finally Disscusion is presented in Section 4.

2. Small-Sample Estimation Performance

The observed data from n individuals consist of the vector (t_i, δ_i, z_i), where t_i is the observed death or censoring time; δ_i is censoring indicator equal to 1 if a death is oberved and equal to 0 if the observation is censord ; and $z_i = (z_{i1}, \cdots, z_{ip})'$ is a p-dimensional vector of covariates. $R(t_{(i)})$ will be used to represent the risk set at the i-th observation time $t_{(i)}$. From equation (1.1), the survival function for the Cox's regression model can be written by

$$\begin{aligned} S(t;z) &= \exp\left\{-\int_0^t \lambda(u;z)\,du\right\} \\ &= \exp\left\{-\exp(\beta' z)\int_0^t \lambda_0(u)du\right\} \\ &= \left\{\exp\left(-\int_0^t \lambda_0(u)du\right)\right\}^{\exp(\beta' z)} \\ &= S_0(t)^{\exp(\beta' z)}, \end{aligned} \qquad (2.1)$$

where

$$S_0(t) = \exp\left\{-\int_0^t \lambda_0(u)du\right\}.$$

Therefore, we turn our attention first to the question of estimating the unknown survival function. To estimate $S_0(t)$ nonparametrically, we turn to another formula, which estimates the cumulative hazard function $\Lambda_0(t)$, equal to the negative log of $S_0(t)$, as

$$\widehat{\Lambda}_0(t) = \sum_{i;t_{(i)} \le t} \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)}, \tag{2.2}$$

where $\widehat{\beta}$ represents the vector of estimated regression coefficients, and the first summation sign refers to the sum of all terms in the following brackets at times up to and including time $t_{(i)}$. And $m_{(i)}$ are the multiplicities of the death times. The advantage of this formulation is that it accomodates tied death times as well as single death time. Sometimes we call this estimator the Breslow estimator.

From equations (2.1) and (2.2), the estimated survival function for an individual with covariate vector z may be written

$$\begin{aligned} \widehat{S}_{NC}(t;z) &= \{\widehat{S}_0(t)\}^{\exp(\widehat{\beta}' z)} \\ &= \{\exp(-\widehat{\Lambda}_0(t))\}^{\exp(\widehat{\beta}' z)} \\ &= \exp\{-\exp(\widehat{\beta}' z) \sum_{i;t_{(i)} \le t} \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)}\}\ . \end{aligned} \tag{2.3}$$

Note that $\widehat{S}_{NC}$ adjusted for a set of measured covariates under the Cox's model corresponds to the Nelson estimato, and reduces to the Nelson estimator when $\widehat{\beta} = 0$. Also the estimator $\widehat{S}_{NC}$ will, in general, form a step function whose value drops at the next observed death time.

Link (1984) has modified this step function, equation (2.3),to a smoothed curve by linearly interpolating between observed times. When the possibility of ties is included, the survival function may be estimated by

$$\widehat{S}_{LINK}(t;z) = \exp\{-\exp(\widehat{\beta}' z)\tilde{\Lambda}_0(t)\}, \tag{2.4}$$

where

$$\begin{aligned} \tilde{\Lambda}_0(t) &= \sum_{i=1}^{l} (t_i - t_{i-1})\widehat{\lambda}_i \ + \ (t - t_l)\widehat{\lambda}_{l+1} \\ &= \sum_{i=1}^{l} \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)} + \frac{t - t_{(l)}}{t_{(l+1)} - t_{(l)}} \frac{m_{(l+1)}}{\sum_{j \in R(t_{(l+1)})} \exp(\widehat{\beta}' z_j)}, \end{aligned}$$

l is such that $t_{(l)} < t$ and $t_{(l+1)} \geq t$, and

$$\widehat{\lambda}_i = \frac{m_{(i)}}{(t_i - t_{i-1}) \sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)} \ .$$

Next, we consider the survival function estimator of the PL type in stead of the Nelson type under the Cox's regression model as follows :

$$\widehat{S}_{PLC}(t;z) = \left\{ \prod_{i;t_{(i)} \leq t} \left(1 - \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)}\right)\right\}^{\exp(\widehat{\beta}' z)}, \qquad (2.5)$$

where the outer product is over the true death time $t_{(i)}$ less than or equal to t.

In the similar method, we have the smoothed PL type estimator as follows:

$$\widehat{S}_{SPLC}(t;z) = \left\{ \prod_{i=1}^{l} \left(1 - \frac{m_{(i)}}{\sum_{j \in R(t_{(i)})} \exp(\widehat{\beta}' z_j)}\right) \cdot \right.$$

$$\left. \left(1 - \frac{t - t_{(l)}}{t_{(l+1)} - t_{(l)}} \frac{m_{(l+1)}}{\sum_{j \in R(t_{(l+1)})} \exp(\widehat{\beta}' z_j)}\right)\right\}^{\exp(\widehat{\beta}' z)} \qquad (2.6)$$

In this formula, we consider the proportion of time between t and last observed death time prior to time t, which is ignored by $\widehat{S}_{PLC}(t;z)$.

To carefully study the small or moderate sample evaluation of mentioned survival function estimators, we take coverage probabilities as the measure of the performance. A method for calculating confidence intervals particularly suited to Monte Carlo simulations has been used by Efron(1981,1985) for censored survival data. Efron discussed its theoretical justification for the case in which the bootstrap distribution is not found by Monte Carlo methods. In particular, nonparmetric Monte Carlo confidence intervals may be useful when analytic intervals are unobtainable. The Monte Carlo simulations by the bootstrap method were carried out to determine the effects of varing the lifetime distributions, censoring distributions, covariates, nominal levels of confidence, and sample sizes under the Cox's model. Achieved error rates for the estimators were determined for all combinations of the following situations with 100 times replications and 1000 times trials :

(1) survival distributions(exponential regression model, Weibull regression model with decreasing failure rate, and Weibull regression model with increasing failure rate) ;

(2) sample sizes n= 10, 20, 30, 50 ;

(3) censoring levels(uncensored case, 10%, 20%, 50%) ;

(4) covariates (0,1) ;

(5) nominal levels of confidence : 0.99, 0.95, 0.90 ;

(6) time t : $S(t;z) = 0.1(0.1)0.9$.

Note that the parameters of the censoring distributions are calculated by using the numerical integration DCADRE subroutine of IMSL, so that the proportion of censoring, $P(Y > C; z)$, is equal to 0%, 10%, 30%, or 50% approximately.

Table 1 - 2 illustrates the following general findings :

(1) The Link estimator has more accurate coverage probabilities than do the Nelson type estimator. This superiority is due to smoothness, that is, the piecewise linear curve obtained by connecting linearly the start point of successive steps of the Nelson type estimator.

(2) To be completely satisfactory, confidence intervals should fall above and below the true parameter value an approximately equal number of times. But this is a more difficult task than to achieve an acceptable total error rates. Almost all the intervals produce unbalanced number of failures.

(3) Fleming and Harrington(1984) found that the Nelson estimator is more efficient than the PL estimator for t satisfying $S(t) \geq 0.20$. In this respect, we found the similar thing under the Cox's regession models.

(4) The smoothed PL type estimator and the Link estimator have the similar achieved error rates as the sample size is larger. But the Link estimator is slightly better than the smoothed PL type estimator for the small or moderate sample size.

(5) The smoothed PL type estimator are better than the PL type estimator under the Cox's regression model in terms of the achieved error rates.

3. An Example

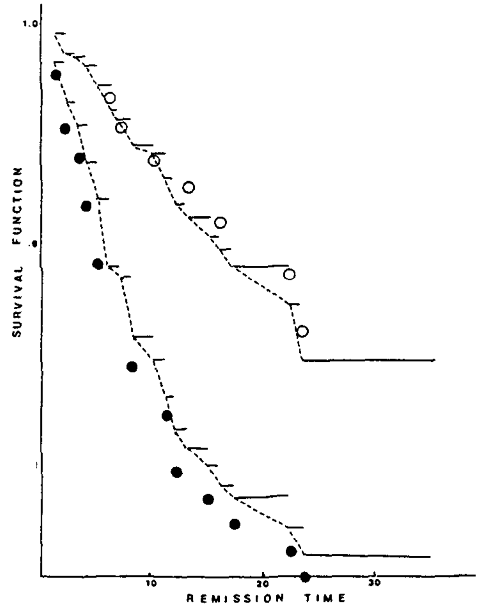
Gehan and others have discussed the results of a clinical trial reported by Freireich et al. in which the drug 6-MP was compared to a placebo with respect to the ability to maintain remission in acute leukemia patients. The 42 patients who entered the study were randomized into two groups, one group given the placebo (lengths of remission: 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23) and the other the drug 6-MP(lengths of remission: 6,6,6,6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22,23, 25^*, 32^*, 32^*, 34^*, 35^*, * denote censored data).

We consider graphical presentation to display $\widehat{S}_{NC}$, $\widehat{S}_{LINK}$, $\widehat{S}_{PLC}$, and $\widehat{S}_{SPLC}$. Clear departures of the survival function estimators under the Cox's model from the PL eswimator would be shown by the following figures. Finally, $\widehat{S}_{NC}$, $\widehat{S}_{LINK}$, $\widehat{S}_{PLC}$, $\widehat{S}_{SPLC}$, and $\widehat{S}_{PL}$ are displayed by Figure 1 and Figure 2.

4. Discussion

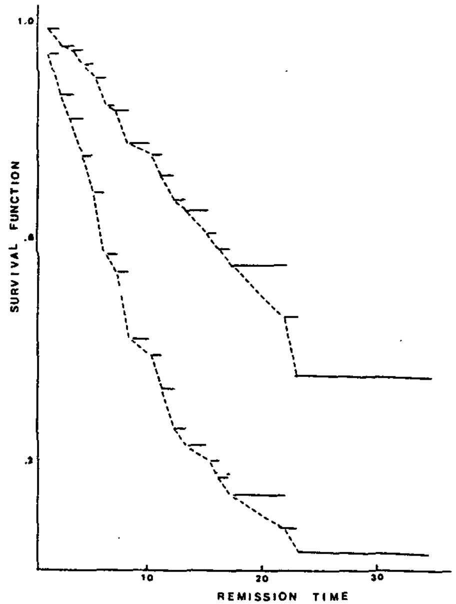
On the basis of overall performance, The Nelson type estimator and Link estimator are preferred for the small samples in the sense of the achieved error rates with comparing to the PL type estimator and the smoothed PL type estimator,respectively, when the true survival probability is at least 0.2. This result is particularly enticing for the Cox's regression models with covariates, in which one or more groups of patients are often followed for a limited period of time after onset of disease or initiation of treatment, since one is quite certain that the true survival function is much larger than 0.2 over the entire interval on which it is being estimated. Moreover, the smoothed PL type estimator and the Link estimator are slightly better than corresponding the PL type estimator and the Nelson type estimator.

Figure 1. Graphic presentation of $\widehat{S}_{NC}$, $\widehat{S}_{LINK}$, and $\widehat{S}_{PL}$



PL estimator : ● , sample 0 (control) ; ○ , sample 1 (drug 6-MP).
For clarity, PL estimators are indicated by the left ends of the defining horizontal lines. Under the Cox's model, —— , $\widehat{S}_{NC}$; ...... , $\widehat{S}_{LINK}$.

Figure 2. Graphic presentation of $\widehat{S}_{PLC}$ and $\widehat{S}_{SPLC}$



——, $\widehat{S}_{PLC}$; .... , $\widehat{S}_{SPLC}$.

Table 1. The distribution of the achieved error rates above and below the true value of the survival function for confidence intervals with nominal level 90%(Exponential survival and uncensoring case)

sample size = 10

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.027	0.345	0.046	0.160	0.087	0.110	0.238	0.058	0.726	0.009
LINK	0.040	0.341	0.082	0.094	0.109	0.070	0.248	0.034	0.727	0.005
PLC	0.030	0.343	0.054	0.135	0.100	0.085	0.249	0.043	0.726	0.006
SPLC	0.043	0.340	0.090	0.087	0.127	0.057	0.254	0.023	0.728	0.004

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.032	0.376	0.042	0.159	0.083	0.106	0.217	0.051	0.633	0.014
LINK	0.041	0.374	0.061	0.105	0.103	0.067	0.240	0.033	0.649	0.010
PLC	0.035	0.375	0.047	0.148	0.089	0.093	0.234	0.037	0.637	0.013
SPLC	0.044	0.372	0.073	0.096	0.123	0.053	0.260	0.026	0.657	0.006

sample size = 20

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.027	0.161	0.031	0.124	0.055	0.084	0.097	0.064	0.518	0.015
LINK	0.046	0.131	0.054	0.086	0.075	0.067	0.110	0.048	0.520	0.013
PLC	0.029	0.156	0.039	0.113	0.070	0.076	0.109	0.051	0.521	0.011
SPLC	0.051	0.130	0.062	0.082	0.084	0.054	0.126	0.041	0.524	0.008

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.018	0.176	0.036	0.128	0.047	0.101	0.092	0.063	0.413	0.007
LINK	0.042	0.159	0.049	0.101	0.067	0.074	0.108	0.042	0.429	0.006
PLC	0.018	0.168	0.038	0.120	0.060	0.090	0.106	0.047	0.426	0.006
SPLC	0.044	0.156	0.057	0.090	0.078	0.069	0.126	0.029	0.443	0.001

sample size = 30

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.029	0.144	0.042	0.124	0.048	0.103	0.068	0.085	0.347	0.027
LINK	0.046	0.096	0.056	0.093	0.059	0.089	0.082	0.072	0.350	0.024
PLC	0.031	0.136	0.044	0.117	0.057	0.093	0.082	0.068	0.351	0.020
SPLC	0.052	0.095	0.060	0.084	0.065	0.077	0.094	0.057	0.355	0.018

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.028	0.171	0.045	0.125	0.049	0.109	0.064	0.070	0.277	0.016
LINK	0.048	0.113	0.055	0.098	0.056	0.084	0.074	0.054	0.287	0.014
PLC	0.029	0.168	0.047	0.118	0.053	0.095	0.073	0.060	0.295	0.011
SPLC	0.054	0.110	0.061	0.091	0.067	0.076	0.088	0.045	0.306	0.005

Table 2. The distribution of the achieved error rates above and below the true value of the survival function for confidence intervals with nominal level 90%(Survival and censoring are Weibull with IFR and uniform, respectively)

sample size = 10

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.000	0.616	0.012	0.369	0.048	0.177	0.195	0.060	0.592	0.005
LINK	0.000	0.511	0.014	0.281	0.065	0.118	0.207	0.034	0.622	0.001
PLC	0.000	0.570	0.013	0.324	0.054	0.142	0.207	0.041	0.612	0.002
SPLC	0.001	0.490	0.019	0.251	0.080	0.093	0.222	0.023	0.638	0.001

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.001	0.630	0.011	0.405	0.056	0.227	0.172	0.078	0.533	0.015
LINK	0.002	0.532	0.014	0.330	0.062	0.152	0.193	0.044	0.574	0.008
PLC	0.001	0.604	0.013	0.385	0.062	0.191	0.195	0.047	0.562	0.007
SPLC	0.002	0.514	0.016	0.301	0.077	0.128	0.223	0.027	0.619	0.005

sample size = 20

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.000	0.753	0.002	0.454	0.031	0.198	0.099	0.050	0.401	0.004
LINK	0.000	0.687	0.005	0.399	0.040	0.165	0.117	0.037	0.419	0.002
PLC	0.000	0.733	0.004	0.428	0.035	0.172	0.117	0.036	0.414	0.002
SPLC	0.000	0.666	0.006	0.359	0.041	0.139	0.141	0.025	0.445	0.002

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.001	0.725	0.004	0.452	0.021	0.206	0.101	0.069	0.400	0.006
LINK	0.001	0.659	0.006	0.380	0.027	0.177	0.116	0.053	0.424	0.004
PLC	0.001	0.709	0.005	0.422	0.023	0.185	0.118	0.046	0.445	0.003
SPLC	0.001	0.649	0.007	0.357	0.031	0.156	0.141	0.035	0.480	0.002

sample size = 30

S(t;0)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.000	0.828	0.001	0.532	0.013	0.228	0.106	0.049	0.363	0.004
LINK	0.000	0.779	0.001	0.479	0.015	0.200	0.124	0.041	0.397	0.003
PLC	0.000	0.815	0.001	0.511	0.016	0.212	0.129	0.039	0.417	0.002
SPLC	0.000	0.764	0.001	0.456	0.022	0.178	0.142	0.035	0.449	0.000

S(t;1)	0.9		0.7		0.5		0.3		0.1	
	Below	Above	Below	Above	Below	Above	Below	Above	Below	Above
NC	0.000	0.818	0.003	0.509	0.020	0.215	0.102	0.056	0.348	0.003
LINK	0.000	0.778	0.005	0.449	0.022	0.179	0.115	0.042	0.372	0.002
PLC	0.000	0.810	0.005	0.485	0.022	0.192	0.118	0.040	0.402	0.002
SPLC	0.000	0.769	0.006	0.436	0.023	0.160	0.135	0.030	0.425	0.001

References

1. Andersen, P.K. and Gill, R.D.(1982), Cox's Regression Model for Counting Process : a Large Sample Study. *Ann.Statist.*, 10, 1100-1120.

2. Arjas, E. and Haara, P.(1988), A Note on the Asymptotic Normality in the Cox Regression Model. *Ann.Statist.*, 16, 1133-1140.

3. Bie, O., Borgan, Ø., and Liestøl,K.(1986), Confidence Intervals and Confidence Bands for the Cumulative Hazard Rate Function and Their Small Sample Properties. *Institute of Informatics, Univ. of Oslo Tech. Report.*

4. Breslow, N. and Crowley, J.(1974), A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Ann.Statist.*, 2, 437-453.

5. Buckland, S.T.(1984), Monte Carlo Confidence Intervals. *Biometrics*, 40, 811-817.

6. Cox, D.R.(1972), Regression Models and Life-Tables (with discussion), *J.Roy.Statist.Soc.,B*, 34, 187-220.

7. Efron, B.(1981), Censored Data and the Bootstrap. *J.Amer.Statist.Assoc.*, 76, 312-319.

8. Efron, B.(1985), Bootstrap Confidence Intervals for a Class of Parametric Problems. *Biometrika,* 72, 45-58.

9. Fleming, T.R. and Harrington, D.P.(1984), Nonparametric Estimation of the Survival Distribution in Censored Data. *Commun.Statist.- Theor.Meth.,* 2469-2486.

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

11. Link, C.L.(1984), Confidence Intervals for the Survival Function using Cox's Proportional-Hazard Model with Covariates. *Biometrics,* 40, 601-610.

12. Tsiatis, A.A.(1981), A Large Sample Stwdy of Cox's Regression Model. *Ann.Statist.*, 9,93-108.