Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.4 pp. 975~982

Nonparametric Estimation of the Bivariate Survival Function under Koziol-Green Model I[1)]

Choon-Mo Ahn[2)] · Sang-Gue Park[3)]

Abstract

In this paper we considered the problem of estimating the bivariate survival distribution of the random vector (X, Y) when Y may be subject to random censoring but X is always uncensored. Adapting conditional Koziol-Green model, simplified estimator for bivariate survival function is proposed. We perform simulation to compare the proposed estimator with popular estimators and discussed the performance of it.

Keywords : Conditional Survival Function; Koziol-Green Censorship Model; Nearest Neighbor Estimates;

1. Introduction

Statistical inferences of the bivariate survival distribution under random censoring has received considerable attention over the past 20 years (See Wang and Wells(1997) for review). Here we consider the estimation problem of the bivariate survival function of (X, Y) when Y may be subject to random censoring C, but X is always uncensored. This is a frequent case in the linear regression model where the covariate X is uncensored and the response variable Y is subject to random censoring. The estimating procedure for survival function under this type of data has been considered by many authors, Stute (1993), Akritas (1994), De Uña-Álvarez and González-Manteiga(1998), Subramanian (2000) etc.

---

1) This paper is supported in part by Korea Research Foundation 2001-015-DP0071.

2) Senior researcher, Information Technology Management Research Group, ETRI, Daejon, 305-350
E-mail : cmahn@etri.re.kr

3) Professor, Department of Statistics, Chung-Ang University, Seoul, 156-756, Korea

Koziol and Green (1976) introduced the appealing and useful survival model in which there exists a positive constant β such that $P(C > t) = P(Y > t)^{\beta}$. β is called censoring parameter. They showed this pattern of censorship often does occur in clinical trials, and that a study of this specific model is worthwhile. The case $\beta = 0$ corresponds to no censoring, and the expected number of the censored observation increases as the β increases. Also it is well known fact that $\alpha = P(Y \leq C) = 1/(1+\beta)$.

There has been many works for Koziol-Green model. See Csörgő and Horváth (1981), Cheng and Lin (1987), Gijbels and Veraverbeke (1989), Stute (1992), De Uña-Álvarez and González-Manteiga (1998, 1999), Gather and Pawlitschko (1998), Braekers and Veraverbeke (2001). Especially, Cheng and Lin (1987) studied the maximum likelihood estimator of it and compared it with the asymptotic efficiency of Kaplan and Meier (1958) estimator. Recently Subramanian (2000) discussed the efficient estimation of the regression coefficients and cumulative baseline hazard function under Cox (1972) model and conditional Koziol-Green model assumption. That is, if we denote $S_{C|X}$ and $S_{Y|X}$ as conditional survival function of C and Y given X, it is assumed that for covariate $X = x$,

$$S_{C|X}(t \mid x) = (S_{Y|X}(t \mid x))^{\beta}, \ \alpha > 0. \tag{1}$$

Our aim in this paper is to propose, under the model assumption (1), intuitive and computationally easy survival function. In section 2, we propose a new estimator and show asymptotic properties of it. In section 3, simulation studies of propose estimator are performed and compared with Akritas' one.

2. Estimation and asymptotic properties

Consider a sequence of independent and identically distributed random vectors (Y_i, C_i, X_i), $i = 1, \cdots, n$, such that given X_i, Y_i and C_i are independent. The observed data are of the form (Z_i, δ_i, X_i), $i = 1, \cdots, n$, where $Z_i = Y_i \wedge C_i = \min(Y_i, C_i)$ and $\delta_i = I(Y_i = Z_i)$.

We know that $S(x, y)$ can be represented as

$$S(x, y) = \int P(Y > y \mid X = t) I(t > x) dG(t) \tag{2}$$

where $S(x, y) = P(X > x, Y > y)$, $G(t) = P(X \leq t)$. Therefore, one could estimate (2) by proposing the estimator of $S(y \mid t) = P(Y > y \mid X = t)$. Akritas (1994) was consider the nearest type estimator of $S(y|t)$, that is,

$$\hat{S}(y|t)=\prod_{Z_i\le y,\delta_i=1}\left\{1-\frac{K((\hat{G}(t)-\hat{G}(X_i))/a_n)}{\sum_{j=1}^{n}I(Z_j>Z_i-)K((\hat{G}(t)-\hat{G}(X_j))/a_n)}\right\}$$

where a_n is a deterministic sequence of real numbers converging to zero, $K(u)=0.5I(-1<u<1)$ and $\hat{G}$ is the empirical distribution function for X_i ($i=1,2,\cdots,n$). Nearest-type estimator use smoothing technique and does not have explicit choice mechanism of smoothing parameter a_n. Also under partial Koziol-Green model, estimating procedures are considered by Gather and Pawlitschko(1998), Braekers and Veraverbeke (2001).

For proposing simpler estimator, we consider estimates $\hat{S}(y|x)$ that are based on Koziol-Green model assumption in (1). Under Koziol-Green assumption, the conditional survival function $P(Y>y|X=t)$ can be represented by the observed sample Z as

$$P(Y>y|X=t)=P(Z>y|X=t)^{1/(\beta+1)}.$$

Therefore, for estimating $S(x,y)$, we use the relation (3).

$$S(x,y)=\int P(Z>y|X=t)^{\alpha}I(t>x)dG(t) \tag{3}$$

where $\alpha=1/(1+\beta)$.

Since Z is uncensored observation, the simplest estimator of $P(Z>y|X=t)$ is

$$\hat{P}(Z>y|X=t)=\hat{G}_2(dt)/\hat{G}(dt), \tag{4}$$

where $\hat{G}(t)=n^{-1}\sum_{i=1}^{n}I(X_i\le t)$ and $\hat{G}_2(t)=n^{-1}\sum_{i=1}^{n}I(X_i\le t,Z_i>y)$. As in the univariate case, α can be estimated by $\hat{\alpha}$, where

$$\hat{\alpha}=\hat{P}(Y\le C)=n^{-1}\sum_{i=1}^{n}I(Y_i\le C_i)$$

Using the assumption (1) and the relation (2), the natural estimator of $S(x,y)$ is

$$\begin{aligned}\hat{S}(x,y)&=\int\hat{P}(Z>y|X=t)^{\hat{\alpha}}I(t>x)\hat{G}(dt)\\&=\int[\hat{G}_2(dt)/G(dt)]^{\hat{\alpha}}I(t>x)\hat{G}(dt)\\&=n^{-1}\sum_{i=1}^{n}\{\hat{G}_2(dX_i)/\hat{G}(dX_i)\}^{\hat{\alpha}}I(X_i>x).\end{aligned} \tag{5}$$

The proposed estimator is very simple to compute and does not use smoothing technique. In section 3, we show that the proposed one perform well when X is recorded by discrete fashion.

Now we show the asymptotic property of proposed estimator. For notational

simplicity, we denote $P(Z>y|X=t)$ as $S(y|t)$. Using bivariate Taylor expansion of $\hat{S}(y|t)^{\hat{\alpha}}$ around the point $(\alpha, S(y|t))$, we have

$$\begin{aligned}\hat{S}(y|t)^{\hat{\alpha}} &= S(y|t)^{\alpha}+\alpha S(y|t)^{\alpha-1}[\hat{S}(y|t)-S(y|t)] \\ &\quad + S(y|t)^{\alpha}\log S(y|t)[\hat{\alpha}-\alpha]+R_n\end{aligned} \tag{6}$$

where

$$\begin{aligned}R_n &= \frac{1}{2}\bar{\alpha}(\bar{\alpha}-1)\,\bar{S}(y|t)^{\bar{\alpha}-2}(\hat{S}(y|t)-S(y|t))^2 \\ &\quad +\frac{1}{2}\bar{S}(y|t)^{\bar{\alpha}}(\log\bar{S}(y|t))^2(\hat{\alpha}-\alpha)^2 \\ &\quad +[\bar{S}(y|t)^{\bar{\alpha}-1}+\bar{\alpha}\bar{S}(y|t)^{\alpha-1}\log\bar{S}(y|t)](\hat{S}(y|t)-S(y|t))(\hat{\alpha}-\alpha)\end{aligned} \tag{7}$$

with $\bar{\alpha}$ between α and $\hat{\alpha}$, and $\bar{S}(y|t)$ between $S(y|t)$ and $\hat{S}(y|t)$. Under this decomposition, we can get asymptotic property of the proposed estimator. The proposed estimator has very similar form of De Uña-Álvarez and González-Manteiga (1998)'s one. So the their proof procedure can be applied for proving Theorem 1.

Theorem 1. Assume that $\int S(y|t)^{\beta-1}I(t>x)dG(t)<\infty$. Then we have

$$\sqrt{n}\{\hat{S}(x,y)-S(x,y)\}\rightarrow N(0,\sigma^2). \tag{8}$$

Proof. Using decomposition (6), we have

$$\begin{aligned}\sqrt{n}(\hat{S}(x,y)-S(x,y)) &= \sqrt{n}\int S(y|t)^{\alpha}I(t>x)[\hat{G}(dt)-G(dt)] \\ &\quad +\sqrt{n}\int S(y|t)^{\alpha}\log S(y|t)[\hat{\alpha}-\alpha]\hat{G}(dt) \\ &\quad +\sqrt{n}\int \alpha S(y|t)^{\alpha-1}[\hat{S}(y|t)-S(y|t)]I(t>x)\hat{G}(dt) \\ &\quad +\sqrt{n}\int R_n(t,y)I(t>x)\hat{G}(dt)\end{aligned}$$

Assume for a moment that we can establish the relation

$$\sqrt{n}\int R_n(t,y)I(t>x)\hat{G}(dt)=o_P(1).$$

First note that for each $\varepsilon>0$, there exists a constant $M=M(\varepsilon)$ satisfying $|\log u|\le Mu^{-\varepsilon}$, $0<u\le 1$. Using this inequality, we have

$$\sqrt{n}\int S(y|t)^{\alpha}\log S(y|t)I(t>x)[\hat{\alpha}-\alpha](\hat{G}(dt)-G(dt))=o_P(1) \tag{9}$$

This can be done by

$$|\text{left of (9)}|\le\sqrt{n}|\hat{\alpha}-\alpha|\cdot M\left|\int S(y|t)^{\alpha-\varepsilon}I(t>x)(\hat{G}(dt)-G(dt))\right|.$$

By the CLT, $\sqrt{n}|\hat{\alpha}-\alpha|=O_P(1)$ and from the assumption of theorem, we have $\int S(y|t)^{\alpha-\varepsilon}I(t>x)G(dt)<\infty$. Therefore, the SLLN can be applied to (9) to the

second absolute term. This leads to $o_P(1)$.

Similarly, we have

$$\sqrt{n}\int \alpha S(y|t)^{\alpha-1}[\hat{S}(y|t)-S(y|t)]I(t>x)(\hat{G}(dt)-G(dt))=o_P(1) \tag{10}$$

This result is based on

$$|\text{left term of (10)}|$$
$$\leq \sqrt{n}\ \sup|\hat{S}(y|t)-S(y|t)|\cdot\left|\int S(y|t)^{\alpha-1}I(t>x)[\hat{G}(dt)-G(dt)]\right|$$

Since $\hat{S}(y|t)$ is a quotient of empirical distributions, we can also have

$$\sqrt{n}\sup|\hat{S}(y|t)-S(y|t)|=O_P(1).$$

Also by assumption, applying SLLN to second absolute value term, result can be satisfied.

Note that $\sqrt{n}[\hat{S}(y|t)-S(y|t)]$ can be represented asymptotically equivalent to

$$\sqrt{n}[-\hat{G}(dt)G_2(dt)/G(dt)^2+\hat{G_2}(dt)/G(dt)]+O_p(n^{-1/4}(\log n)^{3/4})$$

similar to Lo and Singh (1985)′ representation (p.461).

Therefore, $\sqrt{n}[\hat{S}(x,y)-S(x,y)]$ is asymptotically the sum of iid random variables. By the assumption of theorem, (8) can be achieved by applying CLT. ■

3. Numerical Study and Discussion

We conduct a simulation study to compare the performance of both estimators, the proposed estimator and Akritas one. We calculated the survival probabilities and mean squared error. In simulation, the pairs of failure times were distributed according to the bivariate exponential model

$$S(x,y)=(e^{x/\theta}+e^{y/\theta}-1)^{-\theta}$$

of Clayton (1978), with $\theta=0.25$, represents fairly strong positive dependence, with $\theta\to\infty$ giving independence and $\theta\to 0$ giving maximal dependence. As the same in Prentice and Cai (1992), Clayton model failure time (x,y) were obtained from uniform (0,1) variates (u,v) using the transformation

$$y=-\log(1-v),\quad x=\theta\log\{(1-a)+a(1-u)^{-(1+\theta)^{-1}}\}.$$

Also we assume that X is observed as 0.1 step. Censoring variables are exponential distribution with mean 5. This censoring mechanism gives censoring rate about 20%. We should note that Akritas method takes a considerable time to obtain the estimates. The simulation results are based on 1,000 simulation runs.

[Table 1] Simulated survival probabilities and its mean squared errors of Clayton Model ($\theta = 0.25$, $n = 30$)

X \ Y	0	0.2231	0.5108	0.9163
0	1	0.8*	0.6	0.4
	1 (-)	0.786** (0.0053)	0.565 (0.0090)	0.349 (0.0105)
	1 (-)	0.819*** (0.0061)	0.616 (0.0104)	0.392 (0.0113)
0.2231	0.8	0.72	0.575	0.396
	0.823 (0.0054)	0.707 (0.0065)	0.543 (0.0086)	0.346 (0.0102)
	0.826 (0.0054)	0.702 (0.0087)	0.552 (0.0104)	0.370 (0.0105)
0.5108	0.6	0.575	0.513	0.384
	0.613 (0.0080)	0.564 (0.0080)	0.478 (0.0094)	0.345 (0.0099)
	0.613 (0.0080)	0.542 (0.0103)	0.445 (0.0143)	0.315 (0.0133)
0.9163	0.4	0.393	0.384	0.337
	0.427 (0.0088)	0.401 (0.0081)	0.366 (0.0085)	0.312 (0.0095)
	0.414 (0.0088)	0.380 (0.0086)	0.325 (0.0120)	0.239 (0.0173)

*) true survival probability of Clayton model

**) mean of estimated probability by proposed estimator and MSE

***) mean of estimated probability by Akritas′ estimator with bandwidth 0.5 and MSE

In Table 1, we can see that the marginal distribution of Y, when X=0, the estimator given by Akritas′ shows better performance in mean sense. However, at the all other points except X=0, the proposed estimator shows better performance in estimating probability and mean squared error sense. Also, when X=0, the mean squared error of proposed one shows smaller value. Also the computing time of the proposed estimator is much less compared to the Akritas′ nearest type estimator.

We proposed a relatively simple estimator for the bivariate survival function comparing with any other kinds in the literature, in the hope that this would offer an insight for analyzing more complicate types of models for the conditional censoring distribution. Like Koziol and Green(1976)′s suggestion in the univariate case, this pattern of censorship in the bivariate case could occur and a study this specific model might be worthwhile. Some researches should be done about the

meanings and the practical applications of this kind of generalized Koziol-Green assumption and need to be considered a goodness of fit testing procedures for it.

Appendix 1

In this section, we want to prove

$$\sqrt{n}\int R_n(t,y)\,I(t>x)\,\widehat{G}(dt)=o_P(1).$$

The proof procedures are very similar to De Uña-Álvarez and González-Manteiga (1998). Note that $\widehat{S}(y|t)$ is asymptotically the sum of two empirical distribution $\widehat{G}$ and $\widehat{G}_2$, the lemma 3.1 through 3.4 in latter paper can be applied to $\widehat{S}(y|t)$.

First we show that

$$\sqrt{n}\int \overline{S}(y|t)^{\overline{\alpha}}(\log \overline{S}(y|t))^2(\widehat{\alpha}-\alpha)^2 I(t>x)\,\widehat{G}(dt)=o_P(1). \tag{A1}$$

(A1) can be bounded for each $\eta>0$ and $M=M(\eta)$ as

$$\begin{aligned}(A1) &\le \sqrt{n}|\widehat{\alpha}-\alpha|^2 M\int \overline{S}(y|t)^{\overline{\alpha}-2\eta} I(t>x)\,\widehat{G}(dt)\\ &\le \sqrt{n}|\widehat{\alpha}-\alpha|^2 M\int (S(y|t)^{\overline{\alpha}-2\eta}+\widehat{S}(y|t)^{\overline{\alpha}-2\eta}) I(t>x)\,\widehat{G}(dt)\\ &\le \sqrt{n}|\widehat{\alpha}-\alpha|^2 M\int (1+(\widehat{S}(y|t)/S(y|t))^{\overline{\alpha}-2\eta}) S(y|t)^{\alpha-2\eta-\varepsilon} I(t>x)\,\widehat{G}(dt)\end{aligned}$$

where we use inequality $\overline{S}(y|t)^{\overline{\alpha}-2\varepsilon}\le S(y|t)^{\overline{\alpha}-2\varepsilon}+\widehat{S}(y|t)^{\overline{\alpha}-2\varepsilon}$. It is known that $\sqrt{n}|\widehat{\alpha}-\alpha|^2=o(1)$ a.s. and using property Lemma 3.2 in De Uña-Álvarez and González-Manteiga (1998), it can be shown (A1).

Secondly, let us show that

$$\sqrt{n}\int \overline{S}(y|t)^{\overline{\alpha}-1}(1+\overline{\alpha}\log \overline{S}(y|t))[\widehat{S}(y|t)-S(y|t)][\widehat{\alpha}-\alpha] I(t>x)\,\widehat{G}(dt)=o_P(1) \tag{A2}$$

Similar to (A1) derivation, we have

$$\begin{aligned}|(A2)| &\le \sqrt{n}|\widehat{\alpha}-\alpha| M\int |\widehat{S}(y|t)-S(y|t)|\,\overline{S}(y|t)^{\overline{\alpha}-1-\eta} I(t>x)\,\widehat{G}(dt)\\ &\le \sqrt{n}|\widehat{\alpha}-\alpha| M\int |\widehat{S}(y|t)-S(y|t)|[S(y|t)^{\overline{\alpha}-1-\eta}+\widehat{S}(y|t)^{\overline{\alpha}-1-\eta}] I(t>x)\,\widehat{G}(dt)\\ &\le \sqrt{n}|\widehat{\alpha}-\alpha| M\int \frac{|\widehat{S}(y|t)-S(y|t)|}{S(y|t)^{1-\varepsilon}}\left(1+\left(\frac{\widehat{S}(y|t)}{S(y|t)}\right)^{\overline{\alpha}-1-\eta}\right) S(y|t)^{\alpha-\eta-2\varepsilon} I(t>x)\,\widehat{G}(dt)\end{aligned}$$

Using Lemmma 3.2 and 3.3 of De Uña-Álvarez and González-Manteiga (1998) and $\sqrt{n}|\widehat{\alpha}-\alpha|=O_P(1)$, we can derive (A2).

Finally, we show

$$\sqrt{n}\int \overline{\alpha}(\overline{\alpha}-1)\,\overline{S}(y|t)^{\overline{\alpha}-2}(\widehat{S}(y|t)-S(y|t)^2 I(t>x)\,\widehat{G}(dt)=o_P(1) \tag{A3}$$

For $0<\gamma<1$, (A3) can be bounded as

$$|(A3)| \le \sqrt{n} \sup \frac{|\widehat{S}(y|t) - S(y|t)|}{S(y|t)^{\gamma}} \cdot \sup \frac{|\widehat{S}(y|t) - S(y|t)|}{S(y|t)^{2\gamma}}$$
$$\times \int S(y|t)^{3\gamma} \left(1 + \left(\frac{\widehat{S}(y|t)}{S(y|t)}\right)\right)^{\alpha - 2} S(y|t)^{\alpha - 2 - \varepsilon + 3\gamma} I(t > x) \widehat{G}(dt)$$

If we choose $\gamma > (1+\varepsilon)/3$, by Lemma 3.3 and 3.4, (A3) can be derived. ■

References

1. Akritas, M.G.(1994). Nearest neighbor estimation of a bivariate distribution under random censoring, *Annals of Statistics*, 22, 1299-1327.
2. Braekers, R. and Veraverbeke, N.(2001). The partial Koziol-Green model with covariates. *Journal of Statistical Planning and Inference*, 92, 55-71.
3. Cheng, P.E. and Lin, G.D.(1987). Maximum likelihood estimation of a survival function under the Koziol-Green Proportional hazard model. *Statistics & Probability Letters,* 5, 75-80.
4. Cox, D.R.(1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, 34, 187-220.
5. Gather, U. and Pawlitschko, J.(1998). Estimating the survival function under a generalized Koziol-Green model with partially informative censoring. *Metrika*, 48, 189-207.
6. Gijbels I. and Veraverbeke, N.(1989). Quantile estimation in the proportional hazard model of random censorship. *Communication in Statistics- Theory and Methods*, 18, 1645-1663.
7. Koziol, J.A. and Green, S.B.(1976). A Cramer-von Mises statistic for randomly censored data. *Biometrika*, 63, 465-475.
8. Stute, W.(1992). Strong consistency under the Koziol-Green Model. *Statistics and Probability Letters*. 14, 313-320.
9. Subramanian, S.(2000). Efficient estimation of regression coefficients and baseline hazard under proportionality of conditional hazard. *Journal of Statistical Planning of Inference,* 84, 81-94.
10. De Uña-Álvarez, J. and González-Manteiga, W. (1998). Distributional Convergence under proportional censorship when covariables are present. *Statistics and Probability Letters*, 39, 305-315.
11. De Uña-Álvarez, J. and González-Manteiga, W. (1999). Strong consistency under proportional censorship when covariables are present. *Statistics and Probability Letters*, 42, 283-292.
12. Wang, W. and Wells, M.T.(1997). Nonparametric estimators of the bivariate survival function under simplifies censoring conditions. *Biometrika*, 84, 863-880.

[received date : Aug. 2003, accepted date : Nov. 2003]