Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 2, pp. 405~411

A Study on the Conditional Survival Function with Random Censored Data

Won-Kee Lee[1)] · Myung-Unn Song[2)]

Abstract

In the analysis of cancer data, it is important to make inferences of survival function and to assess the effects of covariates. Cox's proportional hazard model(PHM) and Beran's nonparametric method are generally used to estimate the survival function with covariates.

We adjusted the incomplete survival time using the Buckley and James's(1979) pseudo random variables, and then proposed the estimator for the conditional survival function. Also, we carried out the simulation studies to compare the performances of the proposed method.

Keywords : Random censoring, Covariate, Hazard rate, Conditional survival function

1. 서론

일반적으로 의학자료는 환자의 신체적 조건 또는 병의 진행 정도 등 여러 가지 공변량이 생존시간에 많은 영향을 끼친다. 따라서 이런 공변량들이 주어졌을 때의 조건부 생존함수 추정법들이 많은 연구자에 의해 연구되어 왔다.

가장 대표적인 방법은 Cox 모형(Cox, 1972)으로 공변량에 대해 위험함수들이 비례적이어야 한다는 가정이 필요하므로 흔히 비례위험모형이라고도 한다. Beran(1981)은 비모수적 방법으로 조건부 생존함수에 대한 추정법을 제안하였으며, Cox 모형만큼 널리 사용되고 있지는 않으나 비례위험모형의 가정이 만족되지 않는 경우 하나의 대안이 될 수 있다. Dabrowska(1987)는 Beran의 비모수적 방법으로 구한 조건부 생존함수의 대표본성질을 연구하였고, Kim과 Truong(1998)은 평활(smoothing)을 이용하여 Beran방법을 개선하였다.

한편 주어진 공변량을 이용하여 중도절단된 자료를 개선(update)하는 방법에 대해서도 많은 연구가 이루어져 왔다. 특히 Buckley와 James(1979)는 생존자료 분포가 특

1) 제1저자 : 대구광역시 중구 동인2가 101번지 경북대학교 의과대학 건강증진연구소 연구원
E-mail : wonlee@knu.ac.kr
2) 대구광역시 북구 산격동 1370번지 경북대학교 통계학과

별한 모수모형이 아닌 경우 주어진 공변량들을 이용하여 중도절단된 자료를 개선시키는 방법을 제안하였다. Buckley와 James는 K-M추정량(Kaplan 과 Meier(1958))을 이용하여 중도절단된 생존시간의 조건부 기대수명을 추정하는 방법을 제안하였고, Currie(1996)는 회귀계수 추정량의 극한값에 대한 식을 얻고 분석하였다.

본 연구에서는 조건부 생존함수를 추정하기 위하여 Buckley-James의 방법을 이용하여 중도절단된 자료를 개선하고, 이들 개선된 자료에 Beran의 방법을 적용하여 조건부 생존함수의 추정량을 구하고자 한다. 그리고 기존의 Cox모형과 Beran의 방법으로 얻은 결과들과 비교하고자 한다.

2. 조건부 생존함수의 추정

생존시간 $T_1, T_2, \cdots, T_n$은 서로 독립이면서 동일한 분포 F를 따르고, 중도절단시간 $C_1, C_2, \cdots, C_n$는 마찬가지로 서로 독립이면서 동일한 분포 G를 따른다고 하자. 이들 자료에 대한 공변량은 $Z_1, Z_2, \cdots, Z_n$이며, 여기서 $Z_i=(z_{i1}, z_{i2}, \cdots, z_{ip})^t$이라 하자. 그러면 $\{ (X_i, \delta_i, Z_i), i=1, 2, \cdots, n \}$을 관측하게 되며 여기서 관측시간은 $X_i = T_i \wedge C_i$으로 생존시간과 중도절단시간 중 적은값이며 δ_i는 지시함수로 $\delta_i = I(T_i \le C_i)$이다. 또한 $\{ Z_i \}_{i=1}^n$는 유계이고 주어진 Z_i 에서 T_i 와 C_i 는 조건부 독립이라고 가정한다.

2.1 Cox의 방법

공변량 $Z=z$로 주어졌을 때의 조건부 생존함수 $S(t \mid z)=\Pr(T>t \mid Z=z)$ 를 추정하기 위하여 먼저 Cox 모형을 고려하면

$$\lambda(t \mid \boldsymbol{z}) = \lambda_0(t) \cdot \exp(\boldsymbol{\beta}^t \boldsymbol{z})$$

이며, 여기서 $\lambda_0(t)$는 기저위험함수(baseline hazard function), $\boldsymbol{\beta}$는 $p \times 1$ 회귀벡터이고, $\boldsymbol{z}$는 임의의 한 개체에 대한 공변량으로 $p \times 1$ 벡터이다. Cox 모형에서 회귀계수 β와 기저누적위험함수 $\Lambda_0(t)$의 추정 및 여러 성질들에 대하여는 이미 많은 연구가 이루어져 있으며(Andersen과 Gill(1982) 등), 현재 널리 사용되어 지고 있다.

Cox 모형하에서 추정한 회귀계수 $\hat{\beta}$과 기저누적위험함수 $\hat{\Lambda}_0(t)$를 대입하여 조건부 생존함수를 추정하면 다음과 같다.

$$\hat{S}^{Cox}(t \mid z) = \hat{S}_0(t)^{\exp(\hat{\beta}^t z)}$$

여기서 $\hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}$이다.

2.2 Beran의 방법

두 번째 방법으로 Beran의 비모수적 방법을 고려하기 위해

$$H_1(t \mid z) = \Pr(X > t, \delta = 1 \mid Z = z),\ H_2(t \mid z) = \Pr(X > t \mid Z = z)$$

를 정의하자. 그러면 조건부 누적위험함수 $\Lambda(t \mid z)$와 조건부 생존함수 $S(t \mid z)$는 다음과 같이 표현되어진다.

$$\Lambda(t \mid z) = -\int_0^t \frac{dS(s \mid z)}{S(s- \mid z)} = -\int_0^t \frac{dH_1(s \mid z)}{H_2(s- \mid z)},$$

$$S(t \mid z) = \exp\{-\Lambda^c(t \mid z)\} \prod_{x \le t} \{1 - \Delta\Lambda(s \mid z)\}$$

여기서 누적위험함수의 연속부분과 이산부분의 증가분을 각각 Λ^C, $\Delta\Lambda$로 나타내었다.

또한 $H_1(t \mid z)$과 $H_2(t \mid z)$는 각각

$$H_{1n}(t \mid z) = \sum_{i=1}^{n} I(X_i > t, \delta_i = 1)\, W_i(z),\ H_{2n}(t \mid z) = \sum_{i=1}^{n} I(X_i > t)\, W_i(z)$$

와 같이 경험적으로 추정될 수 있다. 여기서 $W_i(z)$는 단지 공변량 z에만 의존하는 음수가 아닌 가중치로서 커넬과 이웃근접(nearest neighborhood) 가중치가 널리 사용되어진다.

이제 조건부 누적위험함수 $\Lambda(t \mid z)$와 조건부 생존함수 $S(t \mid z)$에 $H_1(t \mid z)$과 $H_2(t \mid z)$의 추정량 $H_{1n}(t \mid z)$과 $H_{2n}(t \mid z)$을 대입하여 Beran이 제안한 $S(t \mid z)$의 추정량 $\hat{S}^B(t \mid z)$를

$$\hat{S}^B(t \mid z) = \prod_{s \le t} [1 - \triangle \widehat{\Lambda(}s \mid z)] = \prod_{s \le t} \left[1 - \frac{H_{1n}(s \mid z) - H_{1n}(s+ \mid z)}{H_{2n}(s \mid z)}\right]$$

으로 얻을 수 있다.

2.3 Buckley-James의 방법

Buckley와 James(1979)는 K-M추정량을 이용하여 중도절단된 생존시간의 조건부 기대수명을 추정하는 방법을 다음과 같이 제안하였다.

먼저 p개의 공변량 Z_i와 생존시간 T_i에 대해 다음과 같은 선형모형을 고려하자.

$$T = Z\beta + \varepsilon,\quad E(\varepsilon) = \alpha 1,\quad \mathrm{Var}(\varepsilon) = \sigma^2 \mathrm{I}$$

여기서 회귀계수 $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$는 $p \times 1$ 벡터이며, ε은 $n \times 1$ 오차벡터로 ε_i의 생존함수는 $S = 1 - F$이며, 평균은 절편회귀계수 α이며, 1은 모두 1인 $n \times 1$ 벡터이다. 이때 $T_i^* = T_i\delta_i + E(T_i \mid T_i > C_i)(1 - \delta_i)$라 두면 $E(T_i^*) = Z\beta$가 되므로, 회귀계수 β를 b로 추정하여, 중도절단된 관측시간 C_i를 $T_i^*(b) = Zb$로 개선하고자 한다. 임의의 주어진 회귀계수 추정치를 $b = (b_1, \cdots, b_p)^T$라 하면, 관측된 잔차의 순서화를 통하여 $e_i(b) = X_i - Z_i^T b = e_{(i)}(b)$을 구하고, 이 잔차에 대한 K-M추정량을 $\hat{S}$이라 하고 이 잔차의 순서에 따라 T_i, δ_i, Z_i를 재배열하면 새로 보완된 반응변수

$T_i^*(\beta)$는

$$T_i^*(\beta) = Z_i^T \beta + \left[e_i(\beta)\delta_i + \left\{ \sum_{k=1}^{n} w_{ik}(\beta) e_k(\beta) \right\} (1-\delta_i) \right]$$

와 같이 구성되어 진다. 여기서

$$w_{ik}(\beta) = \begin{cases} \dfrac{v_k(\beta)\delta_k}{\hat{S}\{e_i(\beta)\}} & \text{만약 } k > i \\ 0 & \text{그 밖에} \end{cases}$$

으로 가중치이며, $v_k(\beta)\delta_k$는 $e_k(\beta)$에 할당된 K-M추정량 $\hat{S}$의 뛈크기(jump size)이다. 그러면 새롭게 만들어진 자료 $T^*(\beta)$를 이용하여 회귀계수 β는 최소제곱법에 의해 $\beta = (Z^T Z)^{-1} Z^T T^*(\beta)$와 같이 새롭게 추정되어 진다.

이러한 절차를 새롭게 추정된 회귀계수가 특정한 값으로 수렴할 때까지 반복하여 회귀계수 추정량 $\hat{\beta}$을 얻을 수 있으며, 회귀계수 β의 초기추정량으로는 관측된 자료인 X_i를 이용한 최소제곱추정량을 많이 사용한다.

이제 조건부 생존함수를 추정하는 세 번째 방법으로 먼저 중도절단된 자료를 완전한 자료 $\{ (T_i^*, Z_i), i = 1, 2, \cdots, n \}$로 개선하고, 이들 개선된 자료를 Beran의 비모수적 방법에 적용하여 조건부 생존함수의 추정량 $\hat{S}^{BJ}(t \mid z)$로 제안한다.

$$\hat{S}^{BJ}(t \mid z) = \prod_{s \le t} \left[1 - \frac{H_n(s \mid z) - H_n(s+ \mid z)}{H_n(s \mid z)} \right]$$

여기서 $H_n(t \mid z) = \sum_{i=1}^{n} I(T_i^* > t)\, W_i(z)$이다.

이는 일표본인 경우에서 중도절단이 없는 경우 생존함수에 대한 K-M추정량이 경험적생존함수와 같아지는 것처럼 공변량이 존재하는 경우에도 중도절단이 없는 경우 Beran의 조건부 생존함수 추정량 $\hat{S}^{BJ}(t \mid z)$은 조건부 가중경험적생존함수와 같아지기 때문이다.

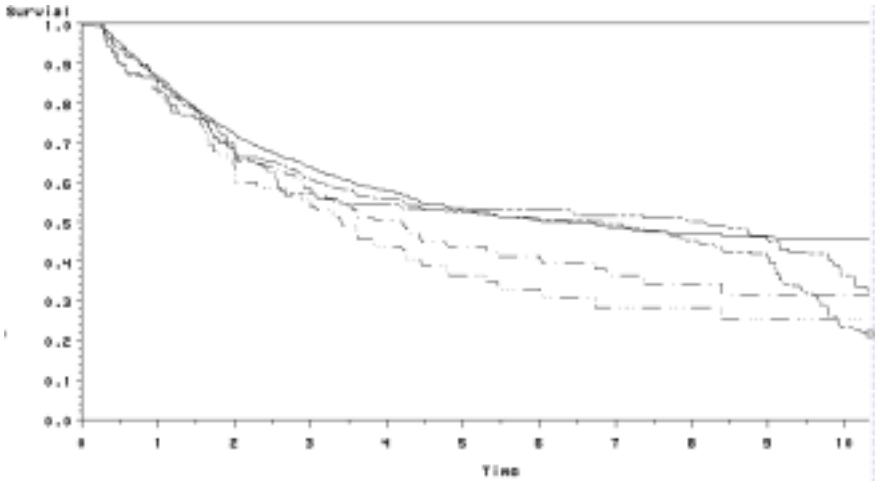
3. 예제

예제 자료는 K대학병원에서 10여 년간 위절제수술을 받았던 환자 1,282명 중 교통사고 등 다른 요인으로 사망한 90명을 제외한 1,192명의 수술 후 생존시간을 관찰한 자료이다. 자료의 일반적 특성에 따른 분포는 남자 773(64.8%)명, 여자는 419(35.2%)명이었고, 평균 연령은 54.37세이었으며, 자료 중 713명(59.8%)이 중도절단되어 관측되었다.

수술 후 생존시간에 유의한 영향을 주리라고 생각되어지는 변수로 Depth, Distant, Size, Borrmann을 고려하였으며, 이들 변수들은 수술 전 각종 검사를 통하여 알 수 있는 변수이다. 여기서 Depth(1,2,3,4)는 종양이 위벽으로부터 얼마나 깊이 위치하고 있는지를 나타내며, Distant(0: no, 1: yes)는 종양이 주변 장기나 임파선 등으로 전이되었는지의 여부를 나타내며, Size는 종양의 장축지름을 측정한 것이며,

Borrmann(0,1,2,3,4)은 종양의 육안적 형태를 나타내는 것이다.

이 자료에 대하여 조건부 생존함수의 추정 결과를 적용시켜 보기 위하여 공변량 값은 적당히 중앙값에 해당하는 값으로 Depth=3, Distance=0, size=50, Borrmann=2를 사용하여 제안된 방법들에 적용한 결과를 <그림 1>에 정리하였다. 그림1에서 _N은 이웃근접가중치를, _K는 커넬가중치를 사용한 것이다.



<그림 1> 주어진 공변량에 대한 조건부생존함수

<그림 1>에서 조건부 생존함수의 추정치들은 시간이 흐름에 따라 제안된 B-J방법의 추정치와 Cox의 추정치는 9년 이후 시점을 제외하고는 유사하게 나타났으나 Beran의 추정치는 이들보다 조금 낮게 추정되었다. 이들 자료가 Cox모형을 따르는 것을 고려하면 제안된 B-J방법이 어느 정도 적절히 조건부 생존함수를 추정한다고 할 수 있다.

4. 모의실험

이 절에서는 조건부 생존함수 추정량들의 효율성을 알아보기 위하여 비례위험모형인 경우와 비례위험모형이 아닌 경우로 나누어 모의실험을 하였으며, 제안된 추정량들의 평균제곱오차(MSE)를 구하여 서로 비교해 보았다.

모의실험에서 생존시간 T는 와이블분포, 중도절단시간의 분포는 지수분포, 공변량 Z의 분포는 $U(0,1)$를 가정하였다. 공변량의 경우 실제 p개를 적용할 수 있으나 모의실험의 경우에는 참모형이 너무 복잡해지므로 공변량이 1개인 경우만 적용하였고, 주어진 공변량의 값은 중간정도가 되게 0.55를 고려하였다.

가정한 모형이 비례위험모형인 경우에는 $\lambda(t \mid z) = e^z$, 비례위험모형이 아닌 경우에는 $\lambda(t \mid z) = e^z I(z \leq 0.5) + 2te^z I(z > 0.5)$을 고려하였다.

먼저 비례위험모형이 만족하는 경우에는 평균제곱오차 값은 거의 비슷하게 나타났

으나 대체로 Cox, Beran, B-J방법 순으로 좋게 나타났다. 전반적으로 표본이 증가하면서 평균제곱오차는 작아지며 중도절단의 비율이 증가하면서 약간 커지는 경향을 보이고 있다. 또한 모든 추정량들은 시간이 증가할수록 추정량의 평균제곱오차가 커짐을 알 수 있다.

비례위험모형이 만족되지 않는 경우에는 평균제곱오차의 크기는 중도절단율에 관계없이 앞쪽에서는 B-J방법을 적용한 경우가 대체로 작게 관찰되다가 뒤쪽으로 갈수록 모든 추정량들의 오차크기는 비슷해 지는 경향을 보였다.

이를 종합하면 비례위험이 타당한 경우에 Cox의 추정량이 다른 추정량에 비해 평균제곱오차측면에서 조금 더 좋음을 알 수 있었으며, 비례위험이 만족되지 않는 모형에서는 전반적으로 Cox의 추정량보다 B-J의 추정량이 더 좋다는 것을 알 수 있다.

<표 1> 주어진 모형하에서의 추청치와 평균제곱오차

			비례위험모형인 경우				비례위험모형 아닌 경우			
	True	TYPE	n=30		n=50		n=30		n=50	
			EST	MSE	EST	MSE	EST	MSE	EST	MSE
CR=10%	0.9	B_N	.899	.003	.897	.003	.804	.015	.808	.013
		B_K	.902	.003	.901	.002	.802	.015	.805	.012
		BJ_N	.900	.003	.897	.003	.807	.014	.813	.012
		BJ_K	.903	.003	.902	.002	.806	.014	.809	.011
		COX	.905	.003	.902	.002	.806	.014	.807	.012
	0.7	B_N	.699	.008	.697	.006	.596	.020	.605	.016
		B_K	.706	.007	.707	.004	.600	.019	.602	.015
		BJ_N	.705	.008	.702	.006	.612	.017	.621	.013
		BJ_K	.712	.007	.712	.004	.615	.016	.618	.012
		COX	.710	.007	.706	.004	.607	.017	.606	.014
	0.5	B_N	.503	.010	.495	.007	.430	.014	.435	.012
		B_K	.514	.009	.512	.005	.438	.013	.439	.009
		BJ_N	.519	.010	.512	.007	.453	.012	.462	.009
		BJ_K	.530	.009	.527	.006	.459	.011	.464	.007
		COX	.513	.009	.505	.005	.445	.012	.441	.009
	0.3	B_N	.307	.008	.301	.007	.276	.009	.278	.007
		B_K	.320	.008	.320	.005	.287	.008	.292	.005
		BJ_N	.326	.010	.319	.008	.289	.010	.295	.008
		BJ_K	.339	.010	.339	.007	.299	.010	.304	.006
		COX	.312	.008	.305	.005	.291	.008	.290	.005
CR=30%	0.9	B_N	.899	.003	.897	.003	.805	.015	.810	.013
		B_K	.902	.003	.901	.002	.804	.015	.805	.012
		BJ_N	.902	.003	.899	.003	.819	.012	.823	.010
		BJ_K	.904	.003	.903	.002	.818	.012	.820	.009
		COX	.905	.003	.902	.002	.809	.014	.808	.012
	0.7	B_N	.700	.009	.697	.006	.598	.021	.606	.017
		B_K	.707	.008	.708	.005	.602	.020	.603	.015
		BJ_N	.720	.008	.716	.006	.653	.011	.655	.009
		BJ_K	.727	.008	.728	.005	.654	.010	.654	.007
		COX	.710	.008	.707	.005	.611	.018	.606	.015
	0.5	B_N	.506	.011	.497	.008	.432	.016	.434	.014
		B_K	.517	.010	.515	.006	.440	.014	.440	.010
		BJ_N	.562	.013	.558	.011	.500	.016	.517	.011
		BJ_K	.572	.014	.572	.010	.504	.015	.516	.009
		COX	.516	.011	.507	.006	.449	.013	.442	.010
	0.3	B_N	.311	.011	.303	.009	.279	.012	.277	.009
		B_K	.325	.010	.325	.007	.290	.010	.294	.006
		BJ_N	.356	.020	.360	.017	.304	.021	.314	.015
		BJ_K	.368	.021	.379	.018	.312	.021	.320	.013
		COX	.315	.010	.307	.006	.299	.010	.294	.006

참고문헌

1. Anderson, P.K., and Gill, R.D.(1982). Cox's Regression Model for Counting Processes: A Large Sample Study, *The Annals of Statistics*, 10, 1100-1120.
2. Beran, R.J.(1981). *Nonparametric Regression with Randomly Censored Data*. Technical Report, University of California, Berkeley.
3. Buckley, J.J. and James, I.R.(1979). Linear Regression with Censored Data. *Biometrika*, 66, 429-436.
4. Cox, D.R.(1972). Regression Models and Life Tables(with Discussion). *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.
5. Currie, I. D.(1996). A Note on Buckley-James Estimators for Censored Data. *Biometrika*, 83, 912-915.
6. Dabrowska, D.M.(1987). Non-paramertic Regression with Censored Survival Time Data. *Scandinavian Journal of Statistics*, 14, 181-198.
7. Kaplan, E.L. and Meier, P.(1958). Nonparametric Estimation from Incomplete Observations. *Journal of American Statistical Association*, 53, 457-481.
8. Kim, H.T. and Truong, Y.K.(1998). Nonparameric Regression Estimates with Censored Data: Local Linear Smoothers and their Applications. *Biometrics*, 54, 1434-1444.

[2004년 3월 접수, 2004년 5월 채택]