Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 327～334

Estimation in a Two-Sample Proportional Odds Model[1)]

Ju-Sung Kim[2)] · Min-Ja Seo[3)]

Abstract

In this paper we propose a new estimator of relative odds ratio in the two-sample case of proportional odds model. Also, we show that the estimator is consistent and asymptotically normal. The efficiency of the proposed is assessed through a simulation study.

Keywords : 마팅게일, 비례오즈모형, 생존함수, 오즈비

1. 머리말

비록 Cox의 비례위험모형이 생존시간과 공변량들의 관계를 연구하는데 사용되어지는 가장 일반적인 접근법이기는 하나, 자료가 잘못 적합 되었을 경우 즉 시간이 지남에 따라 위험비가 비례하지 않고 상수 1로 수렴하는 경우에 대안 모형이 필요하다. 이표본의 경우에 비례오즈모형이 시간이 지남에 따라 위험비가 1로 수렴하는 자료를 적합 시키는데 유용하다. 위험비가 비례하지 않고 수렴하는 경우는 시간이 지남에 따라 공변량 효과가 사라지는 상황에서 발생할 수 있다. $S_i(t)$가 군 $i(i=1,\ 2)$에서 개체에 대한 실제 생존함수를 나타낸다고 했을 때, 군 i에서 생존시간이 t보다 더 큰 개체에 대한 실제 오즈는 $\phi_i(t)=\dfrac{S_i(t)}{1-S_i(t)}(i=1,\ 2)$로 나타내어지고 모든 $t>0$와 어떤 $\theta>0$에 대해서 식 $\phi_2(t)=\theta\phi_1(t)$이 성립할 때 두 군은 비례오즈모형을 만족한다고 말한다. 만일 λ_i가 $S_i(t)$에 대응되는 위험함수이면 비례오즈모형 하에서 $\dfrac{\lambda_2(t)}{\lambda_1(t)}=\dfrac{1}{1+(\theta-1)S_1(t)}$이다. 따라서, $t\rightarrow\infty$일때 $\dfrac{\lambda_2(t)}{\lambda_1(t)}\rightarrow 1$ 이 된다.

---

1) 이 논문은 2004년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었음.
2) 청주시 흥덕구 개신동 산 48번지 충북대학교 정보통계학과 교수
E-mail : kimjs@chungbuk.ac.kr
3) 청주시 흥덕구 개신동 산 48번지 충북대학교 정보통계학과

Bennett(1983b)에 의해 생존분석에 도입된 비례오즈모형은 비례위험모형의 대안으로 각광을 받고 있다. 비례오즈모형에 대한 많은 연구가 이루어져 왔는데, 그중 몇몇을 소개하면 Song Yang과 Prentice(1999)는 우측 중도절단 된 생존시간을 갖는 비례오즈회귀모형의 적합을 위해 가중경험오즈함수를 이용하여 회귀추정량을 얻었다. 또 ,Huang과 Rossini(1997)는 구간 중도절단 자료를 갖는 비례오즈사망시간회귀모형에 대한 추정을 고려하였으며, Wu(1995)는 이표본 비례오즈모형에서 실제모수 추정을, Murphy등(1997)은 우측 중도절단 자료를 갖는 경우, Bennett(1983b)의 준모수적 최우추정량(semiparametric maximum likelihood estimator) 방법이 회귀계수에 대한 효율적인 추정량을 제공한다는 사실을 보였다. Dabrowska-Doksum(1988)는 이표본 일반화 오즈비 모형에서 스코어함수를 이용하여 상대 오즈비를 추정하였다. 본 논문에서, 우리는 중도 절단 되지 않은 자료를 갖는 비례오즈모형 하에서 상대 오즈비를 추정하는 새로운 방법을 제안하려고 하여, 2절에서는 새로운 추정량을 제안하고 이 제안된 추정량의 일치성과 근사적 정규성을 마팅게일 방법을 이용하여 보인다. 또한 3절에서는 우리가 제안한 추정량에 대한 유효성을 로그-로지스틱분포를 사용하여 모의실험을 통해 제안한 추정량과 Dabrowska-Doksum의 추정량에 대한 각각의 MSE(Mean Square Error)를 비교한다.

2.추정량의 제안과 정규 근사

생존시간을 나타내는 독립된 양의 확률변수들 $X_{ij}(\ i=1,\ 2, j=1,\cdots,n_i)$ 가 연속생존함수 $S_i(t)=pr(X_{ij}>t)$, 분포함수 $F_i(t)=1-S_i(t)$, 비례오즈함수 $\phi_i(t)=\dfrac{S_i(t)}{1-S_i(t)}$ 누적위험함수 $\Lambda_i(t)=-\log S_i(t)$ $(i=1,\ 2)$를 갖는다고 하자. 또한 생존시간 X_{ij}는 중도절단 되지 않았다고 가정하자. 1절에서 언급한 바와 같이, 비례오즈모형은 위험비가 시간이 무한이 증가함에 따라 $(t\to\infty)$ 1로 수렴하는 경우에 대단한 가치를 갖는다. 만일 두 군이 비례오즈모형을 만족한다면 모든 $t>0$와 어떤 $\theta>0$ 에 대해서 식 $\phi_2(t)=\theta\phi_1(t)$이 성립한다. 따라서

$$\frac{S_2(t)}{S_2(t)-1}=\theta\frac{S_1(t)}{S_1(t)-1} \tag{2-1}$$

이다.

한편 (2-1)에서 $d(-\phi_2(t))=\dfrac{d(-\phi_2(t))}{dt}=\theta\dfrac{-dS_1(t)}{[S_1(t)-1]^2}$ 이고 $\int_0^\infty -dS_1(t)=1$이므로 $\int_0^\infty(1-S_1(t))^2 d(-\phi_2(t))=\theta$가 성립한다. 따라서 $\hat{\theta}=\int_0^\infty(1-\hat{S}_1(t))^2 d(-\hat{\phi}_2(t))$ 의하여 θ를 추정할 수 있다. 여기서

$-\widehat{\phi}_2(t)=\dfrac{\widehat{S}_2(t)}{\widehat{S}_2(t)-1}$ 이고 $\widehat{S}_i(t)(i=1,\ 2)$는 $S_i(t)$에 대한 경험 생존함수이다.
따라서 우리는
$\widehat{\theta}=\int_0^{\infty}(1-\widehat{S}_1(t))^2 d(-\widehat{\phi}_2(t))$을 $\theta=\int_0^{\infty}(1-S_1(t))^2 d(-\phi_2(t))$에 대한 추정량으로 제안한다.

이제, 아래의 정리에 나열된 이 추정량이 갖는 중요한 특성을 증명하자.

정리 $n=n_1+n_2$라 하고, $n_i\rightarrow\infty$ $(i=1,\ 2)$함에 따라 $\dfrac{n}{n_i}=\rho_i$이라고 가정하자.
그러면
(1) $\widehat{\theta}$는 θ의 일치추정량이다.
(2) $\sqrt{n}(\widehat{\theta}-\theta)$의 근사분포는 평균이 0, 분산 σ^2이 아래와 같은 정규분포를 따른다.

여기서, $\sigma^2=12\rho_1\theta^2\int_0^{\infty}S_1(t)^2 dF_1(t)+\rho_2\theta^4\int_0^{\infty}\left(\dfrac{S_1(t)}{S_2(t)}\right)^4(2S_2(t)+1)dF_2(t).$

증명: (1) $\widehat{S}_i(t)$이 $S_i(t)$의 일치추정량이라는 사실로부터 $\widehat{\theta}$이 θ의 일치추정량이 됨은 자명하다.[Serfling, 1980 pp.24 Theorem].

(2) $N_i=\sum_{j=1}^{n_i}I\{X_{ij}\le t\}$ $(i=1,\ 2)$를 군 i에서의 t 또는 그 이전에 사망자의 총수라 하고 $Y_i=\sum_{j=1}^{n_i}I\{X_{ij}\ge t\}$를 군 i에서의 위험과정 상태의 총수라 하자. 여기서 I는 지표함수를 나타낸다. 과정(process) $M_i=N_i-\int Y_i d\Lambda_i$는 잘 알려진 이차적분가능 마팅게일이며 평균이 0이고 예측 가능한 분산과정 $\int_0^t(1-\triangle\Lambda_i)dF_i=\int_0^t(1-F_i)d\Lambda_i$을 갖는 가우스 과정 M_i^*에 수렴한다. 여기서 Λ_i는 군 i에서의 누적위험함수를 표시 한다. 기호 $\xrightarrow{p}$ 와 $\xrightarrow{d}$ 는 확률수렴과 분포수렴을 각각 나타낸다. 이제 아래의 식을 살펴보면

$$\begin{aligned}\sqrt{n}(\widehat{\theta}-\theta)&=\sqrt{n}\left[\int_0^{\infty}(1-\widehat{S}_1(t))^2 d(-\widehat{\phi}_2(t))-\int_0^{\infty}(1-S_1(t))^2 d(-\phi_2(t))\right]\\&=\sqrt{n}\int_0^{\infty}[(1-\widehat{S}_1(t))^2-(1-S_1(t))^2]d(-\phi_2(t))\\&\quad+\sqrt{n}\int_0^{\infty}(1-\widehat{S}_1(t))^2 d(\phi_2(t)-\widehat{\phi}_2(t))\end{aligned}$$

위 식에서 앞부분을 Z_1, 뒷부분을 Z_2라 하자. 그러면

$$
\begin{aligned}
Z_1 &= \sqrt{n}\int_0^{\infty}\left[(1-\widehat{S}_1(t))^2-(1-S_1(t))^2\right]d(-\phi_2(t)) \\
&= \sqrt{n}\int_0^{\infty}(S_1(t)-\widehat{S}_1(t))(2-S_1(t)-\widehat{S}_1(t))d(-\phi_2(t)) \\
&= \sqrt{\frac{n}{n_1}}\int_0^{\infty}\sqrt{n_1}\,\frac{S_1(t)-\widehat{S}_1(t)}{S_1(t)}S_1(t)(2-S_1(t)-\widehat{S}_1(t))d(-\phi_2(t)) \\
&= \sqrt{\frac{n}{n_1}}\int_0^{\infty}\frac{\widehat{S}_1(u_-)}{S_1(u)}\,\frac{dM_1(u)}{1-H(u_-)}\,S_1(t)(2-\widehat{S}_1(t)-S_1(t))d(-\phi_2(t))
\end{aligned}
$$

여기서 $1-H(u_-)=\dfrac{Y_1(u)}{n_1}$ 이다.

$\widehat{S}_1 \xrightarrow{p} S_1$이고 $1-H_1(u_-) \xrightarrow{p} 1-F_1(u)=S_1(u)$이며 $M_1^*(u)$이 $M_1(u)$의 평균 0을 갖는 정규극한과정이기 때문에 Slutsky 정리와 Mann -Wald정리에 의해

$$
\begin{aligned}
Z_1 &\xrightarrow{d} 2\sqrt{\rho_1}\int_0^{\infty}\int_0^t dM_1^*(u)(1-S_1(t))d(-\phi_2(t)) \\
&= 2\sqrt{\rho_1}\int_0^{\infty}M_1^*(t)(1-S_1(t))d(-\phi_2(t)) \\
&= 2\sqrt{\rho_1}\int_0^{\infty}\phi_2(t)d[(1-S_1(t))M_1^*(t)] \text{이다}
\end{aligned}
$$

그리고 $2\sqrt{\rho_1}\int_0^{\infty}\phi_2(t)d[(1-S_1(t))M_1^*(t)]$ 의 분산은

$$
\begin{aligned}
\sigma_1^2 &= 4\rho_1\int_0^{\infty}(\phi_2(t))^2 d[(1-S_1(t))^2\int_0^t dF_1(u)] \\
&= 4\rho_1\int_0^{\infty}\left(\frac{S_2(t)}{1-S_2(t)}\right)^2 d[(1-S_1(t))^3] \\
&= 12\rho_1\theta^2\int_0^{\infty}S_1^2(t)dF_1(t)
\end{aligned}
$$

이다.

따라서 Z_1의 극한분포는 평균이 0이고 분산 $\sigma_1^2=12\rho_1\theta^2\int_0^{\infty}S_1^2(t)dF_1(t)$을 갖는 정규분포를 따른다.

같은 방법으로 우리는 Z_2의 극한분포를 구할 수 있다.

$$Z_2 = \sqrt{n}\int_0^{\infty}(1-\widehat{S}_1(t))^3\, d(\phi_2(t)-\widehat{\phi}_2(t))$$

에 대해서

$$\begin{aligned}
\phi_2(t)-\widehat{\phi}_2(t) &= \frac{S_2(t)}{1-S_2(t)} - \frac{\widehat{S}_2(t)}{1-\widehat{S}_2(t)} \\
&= \frac{\phi_2(t)}{1-\widehat{S}_2(t)}\frac{S_2(t)-\widehat{S}_2(t)}{S_2(t)} \\
&= \frac{\phi_2(t)}{1-\widehat{S}_2(t)}\frac{1}{\sqrt{n_2}}\int_0^t \frac{\widehat{S}_2(u_-)}{S_2(u)}\frac{dM_2(u)}{1-H_2(u_-)}
\end{aligned}$$

여기서 $1-H_2(u_-)=\dfrac{Y_2(u)}{n_2}$ 이며 $\widehat{S}_2 \xrightarrow{p} S_2$ 이고 $1-H_2(u_-)\xrightarrow{p} 1-F_2(u)=S_2(u)$ 이다. $M_2^*(u)$이 $M_2(u)$의 극한과정이기 때문에

$$Z_2 \xrightarrow{p} \sqrt{\rho_2}\int_0^{\infty}(1-S_1(t))^2 d\left(\frac{S_2(t)}{(1-S_2(t))^2}\int_0^t \frac{1}{S_2(u)}\, dM_2^*(u)\right)$$

이다.

그리고 $M_2^*(t)$이 정규과정이므로 Slutsky정리와 Mann-Wald정리에 의해 Z_2의 극한분포는 평균 0 이고 분산

$$\begin{aligned}
\sigma_2^2 &= \rho_2\int_0^{\infty}(1-S_1(t))^4 d\left(\frac{S_2^2(t)}{(1-S_2(t))^4}\int_0^t \frac{1}{S_2^2(u)}\, d\int_0^u dF_2(s)\right) \\
&= \rho_2\theta^4\int_0^{\infty}\left(\frac{S_1(t)}{S_2(t)}\right)^4 (2S_2(t)+1)dF_2(t)
\end{aligned}$$

인 정규분포를 따른다.

$M_1^*(t)$와 $M_2^*(t)$이 서로 독립이라는 사실로부터 $\sqrt{n}(\widehat{\theta}-\theta)=Z_1+Z_2$의 극한분포는 평균 0 이고 분산

$$\sigma^2 = 12\rho_1\theta^2\int_0^{\infty}S_1(t)^2 dF_1(t) + \rho_2\theta^4\int_0^{\infty}\left(\frac{S_1(t)}{S_2(t)}\right)^4(2S_2(t)+1)dF_2(t)$$

인 정규분포를 따른다.

3. 모의실험결과

이 절에서 우리는 제안된 추정량의 근사적 결과에 대한 평가를 위한 모의실험 연구 결과를 MSE를 가지고 Dabrowska- Doxsum의 추정량과 비교하였다. 모의실험은 로그-로지스틱 분포에서 이루어졌는데, 그 이유는 이 모형이 와이블 모형과 지수모형처럼 생존함수와 위험함수에 대한 용이한 대수적 표현을 갖는 이점 때문이다. 로그-로지스틱 분포에서 생존함수와 분포함수 그리고 오즈함수는 아래와 같이 주어진다.

$$S(\lambda,\ \gamma) = 1/[1+(t^{-\lambda}e^{-\gamma})^{-1}],\quad \lambda > 0,\ -\infty < \gamma < \infty,\ t > 0$$

$$F(\lambda,\ \gamma) = \frac{1}{1+t^{-\lambda}e^{-\gamma}}$$

$$\phi(\lambda,\ \gamma) = (1+t^{-\lambda}e^{-\gamma})/[1+(t^{-\lambda}e^{-\gamma})^{-1}]$$

여기서 λ는 위치모수(location parameter)이고, γ는 척도모수(scale parameter)이다. 두 표본과 관련된 모의실험을 위하여 λ는 동일한 것으로 간주할 것이다. s번째 표본에 대한 위험함수는 $h_s(t) = \lambda/[t(1+t^{-\lambda}e^{-\gamma_s})]$이고, 두 표본 s 와 r에 대한 위험비 $h_s(t)/h_r(t)=$ $(1+t^{-\lambda}e^{-\gamma_r})/(1+t^{-\lambda}e^{-\gamma_s})$이고 그 위험 비는 시간이 커짐에 따라 1로 수렴한다. 따라서 비례오즈모형을 따른다는 가설 하에서 로그-로지스틱분포에서 모의실험을 실행할 수 있다. 먼저 두 모집단의 분포를 각각 로그-로지스틱분포에서 실제 오즈비 θ의 여러 값들에 대해, 첫 번째 집단의 자료 n_1과 두 번째 집단의 자료 n_2로 하여 (n_1, n_2)를 (10, 10), (10, 20), (30, 30), (50, 50), (50, 60), (100, 120)씩 10000번 발생시켜 몬테카를로 방법을 사용하여 상대오즈 비를 추정하였다. θ에 대한 각각의 추정치 $\hat{\theta}$에 대한 MSE (mean squared error)를 제안한 추정치와 Dabrowska-Doxsum의 추정치에 대하여 구하고 <표-1>에 나타내어 비교하였다.

<표-1>을 보면 θ가 e^{-2}, e^{-1}, e^{0}, $\sqrt{e}$, e일 경우에는 표본의 크기에 관계없이 본 논문에서 제안한 추정치에 대한 MSE가 Dabrowska- Doxsum의 추정치에 대한 MSE 보다 작음을 알 수 있다. 이것은 본 논문에서 제안한 추정치가 더 효율적임을 의미한다. 그러나 θ가 e^2, e^3일 때는 표본의 크기가 $(n_1, n_2) \ge (50, 50)$ 인 경우에 한하여 본 논문에서 제안한 추정치에 대한 MSE가 더 적음을 알 수 있다. 이것은 위의 경우에 본 논문에서 제안한 추정치가 표본이 클수록 효율적이라는 것을 의미한다. 앞으로 비례오즈모형 하에서 완비자료에 대한 상대오즈 비 추정에서와 마찬가지로 중도 절단된 자료의 경우에 대한 상대오즈 비 추정과 관련된 연구도 대단히 유용할 것으로 생각한다.

<표-1>비례오즈모형에서의 오즈비 θ의 추정치 $\hat{\theta}$에 대한 MSE(mean squared error)

표본의 크기	실제 오즈 비	MSE		표본의 크기	실제 오즈 비	MSE	
(n_1, n_2)	θ	제안된 추정치	Dabrowska-Doxsum의 추정치	(n_1, n_2)	θ	제안된 추정치	Dabrowska-Doxsum의 추정치
(10,10)	e^{-2}	2.234475	9.3768437	(50,50)	e^{-2}	0.9915872	2.4728555
	e^{-1}	2.0239183	5.5229297		e^{-1}	0.3849301	1.0558413
	$e^0=1$	0.9459186	5.5332015		$e^0=1$	0.1891787	0.3074017
	$\sqrt{e}$	1.2611676	4.6993762		$\sqrt{e}$	0.4752366	0.5926444
	e	3.3014934	5.779668		e	2.4908526	2.6324642
	e^2	40.336951	32.941631		e^2	38.548883	38.758442
	e^3	359.1251	323.56549		e^3	351.70312	355.8503
(10, 20)	e^{-2}	2.3392444	10.051982	(50,60)	e^{-2}	0.9558953	2.2645271
	e^{-1}	2.2203729	4.2729116		e^{-1}	0.3925656	0.9860724
	$e^0=1$	0.952780	3.882438		$e^0=1$	0.1676463	0.2723049
	$\sqrt{e}$	1.4629075	3.3901161		$\sqrt{e}$	0.444311	0.5901122
	e	3.2097224	4.6880146		e	2.528403	2.6069307
	e^2	38.516059	34.576475		e^2	38.276585	38.844595
	e^3	349.79063	337.48746		e^3	349.52232	356.60688
(30,30)	e^{-2}	1.2596789	3.6177286	(100,120)	e^{-2}	0.8134715	1.5159569
	e^{-1}	0.541423	1.574743		e^{-1}	0.2938143	0.7282953
	$e^0=1$	0.3187047	0.8298428		$e^0=1$	0.0684861	0.0911494
	$\sqrt{e}$	0.708447	0.8200524		$\sqrt{e}$	0.285136	0.6139063
	e	2.5788262	2.7406448		e	2.5616693	2.7209612
	e^2	39.379009	37.342202		e^2	37.385741	39.571129
	e^3	354.1513	350.6143		e^3	344.12274	359.78621

참고문헌

1. Bennett, S. (1983a). Log-Logistic Regression Models for Survival data, *Applied. statist.*, 32, No. 2, 165-171.
2. Bennett, S. (1983b). Analysis of Survival Data by the Proportional Odds Model, *Statistics in Medicine*, 2. 273-277.
3. Dorota M. Daborowska and Kjell A Doksum (1988). Estimation and

Testing in a Two-sample Generalized odds-Rate Model, *Journal of the American Statistical Association*, 83, No. 403, 744-749.

4. Galen R. Shorack. and Jon A. Weller (1986). *Empirical Processes with Applications to Statistics*, John Willey & Sons, Inc. 1-150, 258-333.
5. Jean-Yves Dauxois and Syed N. U. A. Kirmani (2003). Testing the Proportional Odds Model under Random Censoring, *Biometrika*, 90, 4, 913-922.
6. Jian Huang and A. J. Rossini (1997). Sieve Estimation for the Proportional-Odds Failure-Time Regression Model with The Interval Censoring, *Journal of the American Statistical Association*, 92. 960-967.
7. Per Krag Andersen, ∅rnulf Borgan, Niels Keiding and Richard D. Gill (1992). *Statistical Models Based on Counting Processes*, Springer Series in Statistics.
8. S. A. Murphy, A. J. Rossini and A. W Van Der Vaart (1997). Maximum Likelihood Estimation in the Proportional Odds Model, *Journal of the American Statistical Association*, 92, 968-976.
9. S. N. U. A. Kirmani and Ramesh C. Gupta (2001). On the Proportional Odds Model in Survival Analysis, *Ann. Inst. Statist*. Math. 53, 2, 203-216.
10. Song Yong and Ross L. Prentice (1999). Semiparametric Inference in the Proportional Odds Regression Model, *Journal of the American Statistical Association*, 94, 125-136.
11. Robert J. Serfling (1980). *Approximation theorems of mathematical statistics*, John Willy & Sons Inc. 1-54.

[2005년 3월 접수, 2005년 5월 채택]