Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 1, pp. 97 ~ 105

A Nonparametric Additive Risk Model Based on Splines

Cheolyong Park[1)]

Abstract

We consider a nonparametric additive risk model that is based on splines. This model consists of both purely and smoothly nonparametric components. As an estimation method of this model, we use the weighted least square estimation by Huffer and McKeague (1991). We provide an illustrative example as well as a simulation study that compares the performance of our method with the ordinary least square method.

Keywords : Nonparametric Additive Risk Model, Weighted Least Square

1. 머리말

McKeague와 Sasieni (1994)에 의해 제안된 부분모수적 가법위험모형(partly parametric additive risk model)은 Lin과 Yang (1994)에 의해 고려된 Cox (1972)의 비례위험모형의 가법모형을 포함하는 준모수 모형(semiparametric model)으로 t시점의 위험함수(hazard function)가

$$\lambda(t\,|\,x,z) = \alpha(t)^T x + \beta^T z$$

형태로 주어진다. 단 여기서 x,z는 q,p 차원의 공변량, $\alpha(\cdot)$은 미지의 비모수함수이며 β는 미지의 모수이다. Park (2006)은 위의 모형을 확장하여 미지의 모수 부분이 $\beta(t)=\beta \cdot f(t)$ 형태로서 $f(\cdot)$는 알려진 함수로 주어지는 일반화 부분모수적 가법위험모형을 고려하였다.

이 논문에서는 위의 모형을 더욱 확장하여 $\beta(t)$가 부드러운(smooth) 비모수 함수로 주어지며 따라서 t시점의 위험함수(hazard function)가

1) Associate Professor, Department of Statistics, Keimyung University, Taegu 704-701
E-mail : cypark1@kmu.ac.kr

$$\lambda(t\,|\,x,z)=\alpha(t)^{T}x+\beta(t)^{T}z \tag{1.1}$$

로 주어지는 모형을 고려한다. 따라서 이 모형은 Aalen (1980)의 순수 비모수 부분인 $\alpha(t)^{T}x$와 부드러운 비모수 부분인 $\beta(t)^{T}z$의 가법모형인 것이다. 이 논문에서 제안하고 있는 것은 부드러운 비모수 함수인 $\beta(t)$를 3차 스플라인 함수(cubic spline function)로 근사화시켜 일반화 부분모수적 가법위험모형에서 적용했던 기법들을 그대로 적용하는 것이다. 구체적으로 이 논문에서 고려하는 형태는 $\beta(t)=Bb$이다. 여기서 $B=(\beta_{ij})$는 $p\times m$인 모수 행렬이고 $b=(B_1(t),B_2(t),\cdots,B_m(t))^{T}$로서 $B_1(t),B_2(t),\cdots,B_m(t)$는 3차 B-스플라인의 기저(B-spline basis)이다. 이 논문에서는 기저의 숫자 선택에 대한 문제는 고려하지 않고 있으며 너무 커지도 않고 너무 작지도 않다고 자의적으로 판단하는 $m=8$을 사용하였다.

앞에서 고려된 모형에서 추정의 초점은 모수인 B 및 (순수) 비모수 부분의 누적위험함수(cumulative hazard function)

$$A(t)=\int_0^t a(s)\,ds$$

이다. 이 논문에서는 Huffer와 McKeague (1991)에서 그룹화 자료(grouped data)의 위험함수의 추정에 사용하였던 조각별 고정위험(piecewise constant risk)을 사용하여 모수 B와 위험함수의 추정에 적용하려고 한다. 구체적으로 주어진 추적기간(follow-up period)을 작은 구간으로 분할하고 각 구간 안에서는 위험이 동일하다는 가정 하에서 가중최소제곱법(weighted least squares method)에 의해 모수 B와 구간별 위험함수를 추정하는 방법이다. 이 방법은 실제로 적용하기 아주 쉬운 방법이며 또한 계산이 용이한 보통최소제곱법(ordinary least square method)에 비해 효율성이 좋은 것으로 Huffer와 McKeague (1991)와 Park (2006)의 연구에서 밝혀졌기 때문에 사용되었다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 Huffer와 McKeague (1991)에서 제시되었고 Park (2006)에서 약간 변형한 가중최소제곱법을 간략히 소개하고 각 공변량에 대응되는 누적 위험함수의 분산추정량을 제공한다. 3절에서는 실제 적용 예제와 모의실험 결과를 제공한다. 먼저 (1.1)의 모형에서 생성된 하나의 표본에 대해 이 방법을 적용하였을 때 신뢰구간이 모집단의 누적 위험함수를 포함시키는지 여부를 알아보는 실제 적용 예제를 제공한다. 다음에 (1.1)의 모형에서 난수를 5000 번 생성하여 보통최소제곱법과 이 논문에서 제안하는 가중최소제곱법을 평균제곱오차(MSE; mean square error)를 통해 비교하는 모의실험 결과를 제공한다. 이 모의실험에서는 보통최소제곱법과 이 논문의 가중최소제곱법에 의해 누적 위험함수의 95% 신뢰구간을 계산하였을 때 포함비율이 명목 포함확률(coverage probability)인 95%에 가까운지 알아보는 실험을 병행하였다.

2. 가중최소제곱법과 모수의 분산추정량

이 절에서는 Huffer와 McKeague (1991)에서 제시되었고 Park (2006)에서 약간 변형된 가중최소제곱법을 간략하게 소개하고 각 공변량에 대응되는 누적 위험함수의 분산추정량을 제공한다. 구체적으로 이 방법을 소개하기 전에 여러 가지 표기법을 정의

하도록 하겠다. 표본크기는 n으로 나타내고, 각 개체 $i=1,2,\cdots,n$에 대해 U_i는 고장시간(failure time), V_i는 중도절단시간(censoring time), $T_i=\min(U_i,V_i)$ 및 $\delta_i=I(U_i \leqq V_i)$로 정의한다. 또한 추적기간(follow-up period)은 $[0,T]$이고 x_i,z_i는 각각 q,p차원의 공변량이다. 그리고 (T_i,δ_i,x_i,z_i)는 (1.1)의 조건부 위험함수를 가지는 (T,δ,x,z)의 서로 독립이고 동일분포를 가지는 확률표본이라고 가정한다.

이 논문에서 사용하는 가중최소제곱법은 집단화된 자료에 적용되는 방법이다. 따라서 (1.1)의 위험함수를 바로 사용하지 않고 추적기간을 분할하여 각 세부구간에서 고정위험을 가지는 조각별 고정위험(piecewise constant risk) 방법을 사용한다. $I_1,I_2,\cdots,I_d$는 추적시간 $[0,T]$의 분할로서 각각의 길이가 $l_1,l_2,\cdots,l_d$이며 시간순서에 따라 배열되어 있다고 가정한다. 그러면 위험함수 (1.1)은 다음과 같이 나타낼 수 있다.

$$\lambda(t\,|\,x,z)=\sum_{j=1}^{q}\left(\sum_{r=1}^{d}\alpha_{rj}I(t\in I_r)\right)x_j+\sum_{j=1}^{p}\sum_{k=1}^{m}\left[\beta_{kj}\left(\sum_{r=1}^{d}b_{rk}I(t\in I_r)\right)\right]z_j$$

여기서 b_{rk}는 구간 I_r의 중간점에서 계산된 B-스플라인 기저 $B_k(t)$의 값을 사용한다. 또한 $i=1,2,\cdots,n$에 대해 위험함수 (1-1)은

$$\lambda(t\,|\,x,z)=\sum_{j=1}^{q}\left(\sum_{r=1}^{d}\alpha_{rj}I(t\in I_r)\right)x_{ij}+\sum_{j=1}^{p}\sum_{k=1}^{m}\left[\beta_{kj}\left(\sum_{r=1}^{d}b_{rk}I(t\in I_r)\right)\right]z_{ij}$$

라고 표기하고 $\lambda_{ir}\equiv\lambda_i(t)I(t\in I_r)\ \forall\ t\in I_r$라고 정의한다.

여기서 관심의 대상이 되는 모수는 $\theta=(\alpha_{11},\alpha_{12},\cdots,\alpha_{dq},\beta_{11},\beta_{12},\cdots,\beta_{pm})$로서 총 $dq+mp$ 개이다. 이 모수를 조각별 고정위험 방법에 의해 추정하는 방법을 간략히 설명하도록 하겠다. 위험함수가 $\lambda(t)$이면 확률밀도함수는

$$\lambda(t)\exp\left(-\int_0^t\lambda(s)ds\right)$$

로 주어지기 때문에 표본 $(T_i,\delta_i,x_i,z_i),i=1,2,\cdots,n$의 로그우도는

$$\begin{aligned}l(\theta)&=\sum_{i=1}^{n}\delta_i\log\lambda_i(T_i)-\sum_{i=1}^{n}\int_0^{T_i}\lambda_i(t)dt\\&=\sum_{i=1}^{n}\sum_{r=1}^{d}\delta_{ir}\log\lambda_{ir}-\sum_{i=1}^{n}\sum_{r=1}^{d}T_{ir}\,\lambda_{ir}\end{aligned}$$

가 된다. 여기서

$$\delta_{ir}=\delta_i\,I(T_i\in I_r),\ T_{ir}=\int_{I_k}I(T_i\geqq t)dt$$

이다. 따라서

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_{r=1}^{d} \left(\sum_{i=1}^{n} \delta_{ir} \psi_{ir} / \lambda_{ir} - \sum_{i=1}^{n} T_{ir} \psi_{ir} \right)$$

가 성립된다. 여기서 $\psi_{ir} = \partial \lambda_{ir} / \partial \theta$이며 열벡터로 놓는다. 그러면 $\lambda_{ir} = \psi_{ir}^T \theta$가 성립하는 것을 쉽게 알 수 있다. 따라서 가중치를 $w_{ir} = 1/\lambda_{ir}$로 놓게 되면 앞의 식은

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_{r=1}^{d} \left(\sum_{i=1}^{n} w_{ir} \delta_{ir} \psi_{ir} - \sum_{i=1}^{n} w_{ir} T_{ir} \psi_{ir} \psi_{ir}^T \theta \right)$$

가 된다. 따라서

$$C = \sum_{r=1}^{d} \sum_{i=1}^{n} w_{ir} \delta_{ir} \psi_{ir}, \quad D = \sum_{r=1}^{d} \sum_{i=1}^{n} w_{ir} \delta_{ir} \psi_{ir} \psi_{ir}^T$$

라 놓으면 $\partial l(\theta)/\partial \theta = 0$는 $C = D\theta$로 표현되기 때문에 $\theta = D^{-1} C$로서 해를 구할 수 있는 것이다.

가중치 w_{ir}가 알려진 경우 앞에서 설명된 방법에 의해 바로 모수 추정량을 구할 수 있지만 가중치가 미지의 위험함수 λ_{ir}의 역수로 표현되기 때문에 이 방법을 바로 적용할 수는 없다. 따라서 보통최소제곱법, 즉 모든 가중치를 1로 놓고 $\tilde{\theta} = D^{-1} C$를 구해 새로운 가중치 $\hat{w}_{ir} = 1/(\psi_{ir}^T \tilde{\theta})$를 구한다. 이 가중치를 대입하여 다시 $\hat{\theta} = \hat{D}^{-1} \hat{C}$를 구할 수 있다. 이 과정을 반복적으로 적용하면 최적의 추정량을 구할 수 있다.

이 논문에서는 기본적으로 Huffer와 McKeague (1991)의 방법을 이용하여 모수 추정량을 구하도록 한다. 구체적으로 앞의 과정을 두 번 반복 계산하여 모수 추정량을 구하게 된다. 또한 새로운 가중치 계산에서 $\tilde{\theta}$ 대신에 평활 추정량 θ^*을 사용하도록 한다. 구체적으로 $\tilde{a}_{rj}$ 대신에 전후 구간에서 계산된 $\tilde{\alpha}_{ij}, \forall i \neq r$ 값들을 구간의 길이에 비례하게 가중 평균한 α_{ir}^*를 사용하는 것이다. 이 방법은 Huffer와 McKeague (1991)에서 과거의 구간만 사용하여 예측가능한(predictable) 추정량을 사용한 것과 약간의 차이가 있다. 이 방법을 사용하는 이유는 (1.1) 모형에서는 각 구간의 모수 추정량이 다른 구간의 관찰값에도 의존하여 예측가능한 추정량을 만들 수 없으며 또한 과거의 자료뿐만 미래의 자료로 이용하면 효율성을 높일 수 있기 때문이다. 구체적으로 누적 관측 고장시간이 최소 K가 넘는 최초의 전후 구간 n_K개를 이용할 것이며, 누적 관측 고장시간이 K가 넘지 않는 전반부의 구간에는 가중치 1을 그냥 사용한다. 구체적으로 이 논문에서는 K로 30을 사용한다. 또한 McKeague와 Sasieni (1994)의 제안에 따라 $\lambda_{ir}^* = \psi_{ir}^T \theta^*$에 대해 추가적인 평활을 하겠다. 구체적으로 λ_{ir}^* 추정과정에서 음수나 0에 가까운 값이 발생하면 그 구간 r에서 관측된 모든 고장시간의 λ_{jr}^*을 평균한 값의 0.25배가 되도록 조정하였고, 또한 누적 관측 고장시간이 K가 넘지 않는 전반부의 구간에는 각 구간에서 관측된 모든 고장시간의 λ_{jr}^*을 평균한 값을 사용하였다.

모수의 추정량 $\hat{\theta} = \hat{D}^{-1} \hat{C}$의 분산 추정량으로는 Huffer와 McKeague (1991)에서 사

용한 것과 같이 다음의 세 가지 방법을 사용한다.

$$\widehat{Var}_1(\hat{\theta}) = \hat{D}^{-1}\hat{Q}\hat{D}^{-1}$$
$$\widehat{Var}_2(\hat{\theta}) = \hat{D}^{-1}$$
$$\widehat{Var}_3(\hat{\theta}) = \hat{D}^{-1}\hat{H}\hat{D}^{-1}$$

위의 공식을 이용하면 우리가 관심을 가지는 선형추정량 $\psi^T\hat{\theta}$의 분산 추정량을 쉽게 계산할 수 있다. 구체적으로 우리가 주로 관심을 가지는 선형추정량은 순수 비모수 부분의 공변량 x_j에 대응되는 누적 위험함수 $\sum_{r=1}^{d} l_r \hat{\alpha}_{rj}$와 부드러운 비모수 부분의 공변량 z_j에 대응되는 누적 위험함수 $\sum_{r=1}^{d} l_r \sum_{k=1}^{m} \hat{\beta}_{kj} b_{rk}$ 등이 있다. 실제로 3절에서 이러한 누적 위험함수에 대한 신뢰구간 계산에 이 분산 추정량 공식을 이용하게 된다.

3. 적용 예제와 모의실험

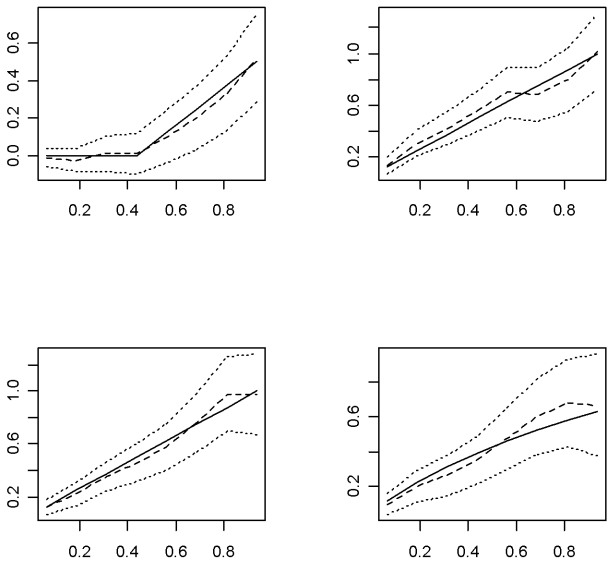
이 절에서는 (1.1)의 모형에서 생성되는 자료에 대한 적용 예제와 모의실험 결과를 제시한다. 먼저 (1.1)의 모형에서 하나의 표본을 뽑아 각 공변량에 대응되는 추정 누적 위험함수의 평균값과 신뢰구간을 계산하였을 때 모집단의 누적 위험함수를 포함시키는지 알아보는 실제 적용 예제를 제시한다. 다음으로 (1.1)의 모형에서 5000개의 표본을 뽑아 보통최소제곱법과 이 논문의 가중최소제곱법에 의해 각 공변량에 대응되는 추정 누적 위험함수를 계산하여 평균제곱오차(mean squares error)를 비교하는 모의실험 결과를 제공한다. 이 모의실험에서는 보통최소제곱법과 이 논문의 가중최소제곱법에 의해 누적 위험함수의 95% 신뢰구간을 계산하였을 때 포함비율이 명목 포함확률(coverage probability)인 95%에 얼마나 가까운지 알아보는 실험을 병행하였다.

이 절의 적용 예제와 모의실험의 난수 생성에 사용되는 고장시간의 위험함수는 다음과 같다.

$$\lambda(t\,|\,x,z) = I(t \geqq 0.5)x_1 + t\,x_1 + t\,z_1 + e^{-t}z_2 \tag{3.1}$$

여기서 x_1, x_2, z_1, z_2은 아래와 위에서 각각 1%씩 절사한 평균이 1/2인 지수분포에서 생성된 확률표본이다. 중도절단시간은 평균이 1/0.3인 지수분포에서 생성하였다. 추적기간은 $[0,1]$, 표본크기는 1000이며 그룹화된 자료를 만들 때 $d=8$의 균등분할을 사용하였다.

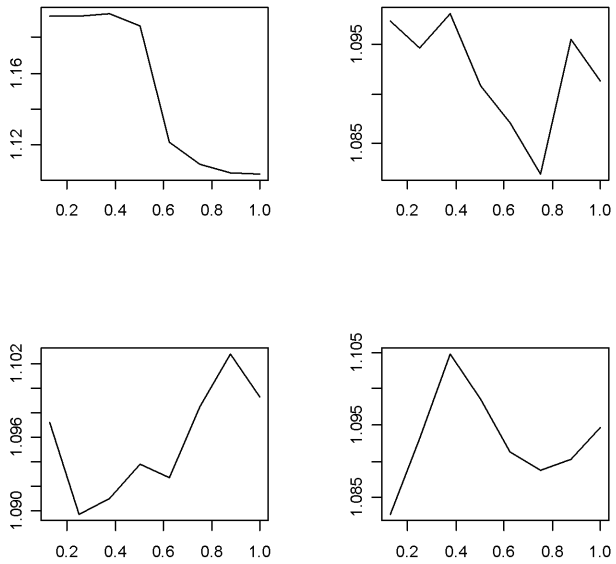
먼저 이 논문의 가중최소제곱법의 적용 예제를 살펴보도록 하겠다. 먼저 (3.1)의 위험함수를 가지는 모형에서 표본크기 1000인 표본을 하나 생성하였다. 이 표본에 대해 각 공변량에 대응되는 균등분할 구간의 추정 누적 위험함수를 계산하여 표본평균값과 95% 신뢰구간을 계산하여 모집단의 누적 위험함수와 함께 그림으로 나타낸 것이 <그림 1>이다.



<그림 1> 예제 자료의 모집단 누적위험, 표본평균값 및 신뢰구간

구체적으로 <그림 1>에는 x_1, x_2, z_1, z_2에 대응되는 네 개의 누적 위험함수 그림들이 첫 번째 행에 순수 비모수 부분, 두 번째 행에 부드러운 비모수 부분에 해당되는 그림들이 순서대로 나열되어 있다. 참고로 x_2, z_1, z_2에 대응되는 모집단의 누적 위험함수는 쉽게 $t, t, 1-e^{-t}$로 계산되는 것을 알 수 있다. 각 그림에는 모집단의 값은 실선, 표본평균값은 굵은 점선 그리고 $\widehat{Var}_1(\hat{\theta})$에 의한 95% 신뢰구간이 가는 점선으로 표시되어 있다. 이 예제 표본에서는 모집단의 값이 8개 구간 모두 95% 신뢰구간 안에 잘 포함되어 있음을 알 수 있다.

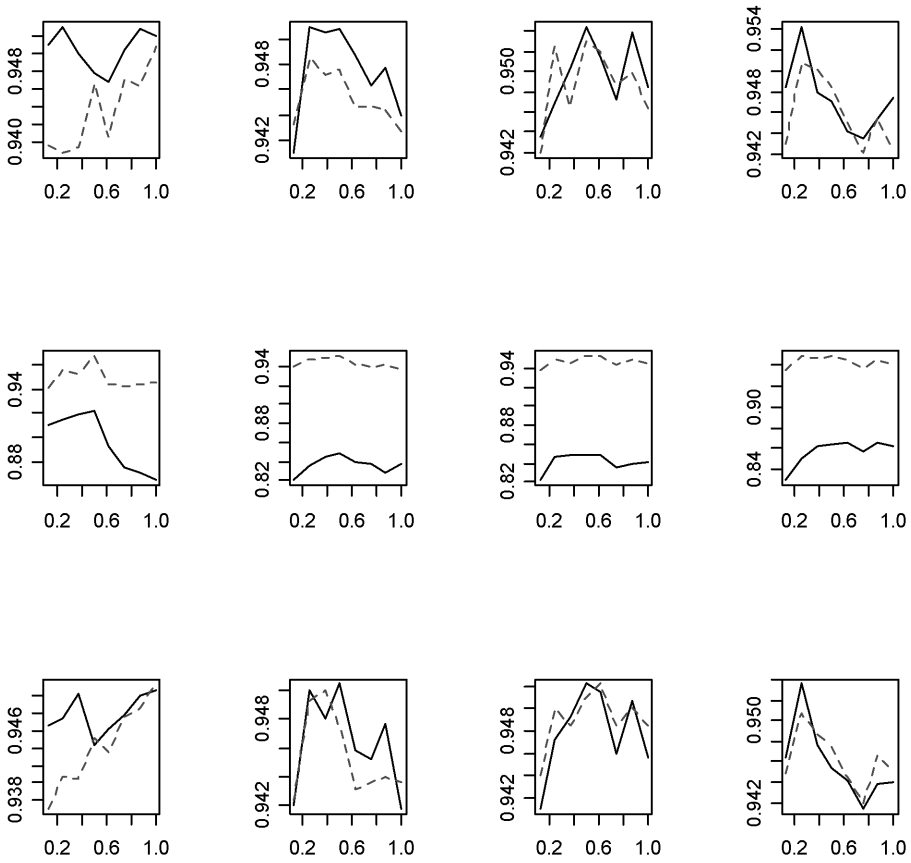
이 논문에서 제안한 방법의 성능을 알아보기 위하여 모의실험을 실행하였다. 모의실험은 기본적으로 적용 예제에서와 똑 같은 실험방법을 사용한다. 다만 표본을 5000개 반복해서 뽑아서 보통최소제곱법과 이 논문의 가중최소제곱법에 의해 추정된 누적위험함수의 평균제곱오차인 MSE를 계산하여 이것을 비교하고자 한다. 보통최소제곱법의 평균제곱오차를 이 논문의 방법에 의한 평균제곱오차로 나눈 값인 MSE_{OLS}/MSE_{WLS}를 그림으로 나타난 것이 <그림 2>이다.



<그림 2> 보통최소제곱법과 이 논문의 가중최소제곱법의 MSE_{OLS}/MSE_{WLS}

구체적으로 <그림 2>에는 x_1, x_2, z_1, z_2에 대응되는 네 개의 평균제곱오차의 비가 첫 번째 행에 순수 비모수 부분의 두 개 그림, 두 번째 행에 부드러운 비모수 부분의 두 개 그림 순서대로 나열되어 있다. 네 개의 그림 모두에서 이 논문의 방법에 의한 평균제곱오차가 훨씬 더 작은 것으로 나타났다. 구체적으로 이 논문에서 제안한 방법이 보통최소제곱법에 비해 최소 8% 최대 19%까지 효율이 높은 것으로 나타나 아주 우수한 성능을 보여주는 것을 알 수 있다.

이 모의실험에서 앞 절에서 제시된 분산 추정량의 정확성을 알아보는 실험도 동시에 수행하였다. 구체적으로 누적 위험함수의 95% 신뢰구간을 계산하여 모집단의 누적 위험함수를 포함하는 포함비율을 계산하여 명목상의 포함확률 95%와 비교하는 것이다. 세 가지 분산 추정법에 의한 포함비율을 그림으로 나타낸 것이 <그림 3>이다.



<그림 3> 세 가지 분산 추정법에 의한 누적 위험함수의 95% 신뢰구간의 포함비율

<그림 3>에서 그림의 배치는 다음과 같다. 각 행은 세 가지 방법에 대응된다. 즉 첫 번째 행은 $\widehat{Var}_1(\hat{\theta})$에 의한 방법, 두세 번째 행은 각각 $\widehat{Var}_2(\hat{\theta}), \widehat{Var}_2(\hat{\theta})$에 의한 방법에 대응되는 그림이다. 그리고 각 열은 네 가지 공변량에 대응된다. 즉 x_1, x_2, z_1, z_2에 대응되는 누적 위험함수의 포함비율이 네 개의 열에 순서대로 배치되어 있다. 그리고 각 그림에서 실선은 보통최소제곱법에 의한 포함비율이고 점선은 이 논문에서 제안한 가중최소제곱방법에 의한 포함비율이다.

이 그림으로부터 다음과 같은 특징을 요약할 수 있을 것이다. 첫 번째 변량 x_1 혹은 두 번째 분산 추정방법을 제외하면 보통최소제곱법과 이 논문의 가중최소제곱법의 포함비율이 비슷하며 또한 대체로 명목확률인 95%에 근접하고 있다. 두 방법 사이에 차이가 발생하는 부분은 크게 두 가지로 요약할 수 있을 것이다. 먼저 이 논문에서 제안한 방법은 첫 번째 변량 x_1에 대응되는 전반부 구간에서 포함비율이 95% 보다 약간 떨어지는 현상이 첫 번째, 세 번째 분산 추정방법에서 발견되고 있다. 또한 보통

최소제곱법은 두 번째 분산 추정방법에서 전반적으로 큰 문제점을 보여주고 있어 보통최소제곱법과 두 번째 분산 추정방법의 동시 사용을 피하는 것이 좋을 듯하다.

참고문헌

1. Aalen, O. O. (1980). A model for nonparametric regression analysis of counting processes. In *Mathematical Statistics and Probability Theory*, Lecture Notes in Statistics, 2, Ed. W. Klonecki, A. Kozek and J. Rosinski, pp. 1-25. Springer-Verlag, New York.
2. Huffer, F. W. and McKeague, I. W. (1991). Weighted least squares estimation for Aalen's additive risk model. *Journal of American Statistical Association*, 86, 38-53.
3. Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81, 61-71.
4. McKeague, I. W. and Sasieni, P. D. (1994). A partly parametric additive risk model. *Biometrika*, 81, 501-514.
5. Park, C. (2006). A generalized partly-parametric additive risk model. *Journal of the Korean Data & Information Science Society*, 17, 401-409.

[2007년 1월 접수, 2007년 2월 채택]