The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.6.761
31(6), 761–769

Generating censored data from Cox proportional hazards models

Ji-Hyun Kim[a,1] · Bongseong Kim[a]

[a]Department of Statistics and Actuarial Science, Soongsil University

(Received September 14, 2018; Revised October 19, 2018; Accepted November 2, 2018)

Abstract

Simulations are important for survival analyses that deal with censored data. Cox models are widely used in survival analyses, therefore, we investigate how to generate censored data that can simulate the Cox model. Bender *et al.* (*Statistics in Medicine*, **24**, 1713–1723, 2005) provided a parametric method for generating survival times, but we need to generate censoring times as well as survival times to simulate the censored data. In addition to the parametric method for generating censored data, a nonparametric method is also proposed and applied to a real data set.

Keywords: Cox model, simulation, baseline hazards function, Kaplan-Meier estimator

1. 연구 목적과 필요성

생존분석(survival analysis)의 주요 특징은 완전한 자료가 아닌 중도절단자료(censored data)가 주어진다는 점이다. 중도절단은 주로 제한된 연구기간 때문에 발생한다. 연구시작 시점부터 관심 사건이 발생할 때까지 걸리는 시간을 생존시간(survival times)이라고 할 때, 사건이 발생하지 않은 상태로 연구가 종료된 관측개체에 대해서는 생존시간 대신 중도절단된 값을 관측하게 된다. 이렇게 중도절단된 자료를 우측중도절단자료(right censored data)라고 한다 (좌측중도절단이나 구간중도절단(interval censored)도 있으나 이 논문에서 우측중도절단만을 다루기로 한다). 중도절단된 자료를 완전한 자료인 것으로 간주하고 분석하면 편향된 결과를 얻게 되며, 제외하고 분석하면 유효 표본수가 줄어들어 검정력이나 추정량의 효율성이 떨어지게 된다. 따라서 중도절단된 자료를 분석에 제대로 반영해서 쓰는 것이 중요하다.

우측중도절단자료가 포함된 자료(줄여서 중도절단자료라고 부르기로 함)에 대한 대표적 회귀모형인 Cox 비례위험모형(Cox proportional hazards model, 줄여서 Cox 모형으로 부르기로 함)은 생존시간의 분포 형태에 구애받지 않고 회귀계수 추정을 할 수 있다는 장점 때문에 널리 쓰인다 (Cox, 1972). 한편 생존분석에서 어떤 추정량이나 분석방법의 성능을 알고 싶은데 표본 크기가 작거나 추정량의 성질을 해석적으로 유도하기 힘들면 모의실험이 필요하다. 만약 Cox 모형이 연구에 적절하다면 모의실험을 위해 이 모형을 따르는 자료를 생성할 수 있어야 한다.

[1]Corresponding author: Department of Statistics and Actuarial Science, Soongsil University, Sangdo-Ro 369, Dongjak-Gu, Seoul 06978, Korea. E-mail: jxk61@ssu.ac.kr

Cox 모형으로부터 자료를 생성하는 데에 일반적인 회귀모형과 다른 어려움이 있다. 일반적인 회귀모형은 생성하고자 하는 반응변수 y에 관한 직접적인 모형이므로 모형의 식이 주어지고 오차항의 분포만 결정되면 y를 바로 생성할 수 있다. 하지만 Cox 모형은 y가 아니라 위험함수(hazard function)에 관한 모형이므로 모형의 식이 주어지더라도 y를 바로 생성할 수 없다. 그리고 Cox 모형으로부터 중도절단자료를 생성하기 위해서는 생존시간뿐만 아니라 중도절단시간(censoring times)도 생성해야 한다. 또한 각 관측개체의 특성이 다르므로 하나의 모집단에서 얻어진 표본이 아니라는 점도 일반적인 난수 생성과 다른 점이다.

Moriña와 Navarro (2014)가 만든 survsim이라는 R 패키지를 이용하면 중도절단자료를 생성할 수 있다. 하지만 이 팩키지를 이용하려면 생존시간과 중도절단시간의 분포 종류를 모수와 함께 명시해야 한다. 또한 가속수명(accelerated failure time) 모형을 가정하고 있어 Cox 모형을 따르는 모의자료를 생성할 수 없다.

Cox 모형을 따르는 생존시간을 생성하는 방법을 Bender 등 (2005)이 제안하였다. 하지만 생존분석을 위한 모의실험을 실시하기 위해 이 연구만으로 미흡한 점이 몇 가지 있다. 우선 생존시간에 대한 Cox 모형이 주어졌다고 가정하고 있는데, 모의실험을 위한 모형을 결정하는 방법도 중요하다. 왜냐하면 모의실험에서 쓰는 모형이 실제 상황에서 너무 벗어나면 모의실험의 실용성이 떨어지기 때문이다. 다음으로 Bender 등 (2005)은 생존시간을 생성하는 방법만 제시하였으나 중도절단자료를 생성하려면 중도절단시간도 생성해야 한다. 마지막으로 생존시간의 분포로 모수 분포를 사용하는 모수적 방법만 고려하였으나 비모수적 방법으로 확장하는 구체적 방안에 대한 제시도 필요하다.

본 연구에서는 Bender 등 (2005)의 연구에서 미흡한 점들을 보완하여 Cox 모형을 이용한 모의실험을 통해 생존분석에 관한 연구를 행하고자 하는 연구자들에게 실용적 도움을 주고자 하였다. Cox 모형에서 중도절단자료를 생성하는 방법을 비모수적 방법과 모수적 방법으로 나누어 2절에 서술하고, 3절에서 응용 사례를 제시하였다. 4절에서 중요한 내용을 요약하고 추가적인 사항에 대해 첨언하였다.

2. Cox 모형을 따르는 중도절단자료 생성 방법

생존함수(survival function)를 S, 누적기저위험함수(cumulative baseline hazard function)를 Λ_0, 공변량 벡터를 x로 표기할 때, Cox 모형을

$$S(t|x) = 1 - F(t|x) = \exp\left[-\Lambda_0(t)\exp\left(\beta' x\right)\right]$$

로 표현할 수 있다. U를 0에서 1 사이 균일분포를 따르는 확률변수라고 할 때 위 식으로부터

$$T = \Lambda_0^{-1}\left[-\log(U)\exp\left(-\beta' x\right)\right]$$

로 정의되는 확률변수 T의 생존함수가 $S(t|x)$이며 T는 Cox 모형을 따르는 생존시간이 된다 (Bender 등, 2005). 즉, T의 분포는 Cox 모형이 참일 때 공변량 x를 가진 개체의 생존시간의 분포가 된다.

$\{(Y_i, D_i, x_i),\ i = 1, \ldots, n\}$를 실제로 관측한 중도절단자료라고 하자. 여기서 $Y_i = \min(T_i, C_i)$, $D_i = I(T_i \le C_i)$로서 생존시간 T_i와 중도절단시간 C_i 중에서 먼저 일어난 시간을 관측하게 되며, 중도절단 여부도 관측한다 (D_i가 0이면 중도절단). 모의실험은 특정한 모형을 가정하고 이 모형으로부터 자료를 생성해 이루어진다. 이 때 모형은 비현실적인 상황보다 실제 일어난 상황, 즉 관측한 자료에 근거하여 정하면 더 실용적일 것이다. 관측 중도절단자료 $\{(Y_i, D_i, x_i),\ i = 1, \ldots, n\}$에 Cox 모형을 적용하여 $\hat{\Lambda}_0$, $\hat{\beta}$을 얻은 다음,

$$T_i^* = \hat{\Lambda}_0^{-1}\left[-\log(U)\exp\left(-\hat{\beta}' x_i\right)\right] \tag{2.1}$$

와 같이 생존시간 T_i^*를 생성하면 관측자료에 근거한 Cox 모형에서 자료를 생성할 수 있게 된다. 한편 중도절단시간 C_i^*는 생존시간 T_i^*의 분포와 독립적이며 공변량과 무관하다는 '정보 없는 중도절단(non-informative censoring)'을 가정하여 생성한다. 이로부터 Cox 모형을 따르는 중도절단자료 $\{(Y_i^*, D_i^*, x_i),\ i = 1, \ldots, n\}$를 생성할 수 있다. $Y_i^* = \min(T_i^*, C_i^*)$, $D_i^* = I(T_i^* \leq C_i^*)$이다. T_i^*와 C_i^*를 생성할 때 비모수적 방법과 모수적 방법을 쓸 수 있는데 두 방법을 나누어 설명하고자 한다.

2.1. 비모수적 방법

T_i^*를 비모수적으로 생성한다는 것은 누적기저위험함수 $\Lambda_0(t)$를 비모수적으로 추정하는 것을 의미한다. Breslow (1974) 또는 Nelson-Aalen 추정량이라고 부르는 $\hat{\Lambda}_0(t)$의 식은 다음과 같다.

$$\hat{\Lambda}_0(t) = \sum_{y_{(i)} \leq t} \frac{d_{(i)}}{\sum_{j \in R_{(i)}} e^{\hat{\beta}' x_j}}.$$

위 식에서 $y_{(i)}$와 $d_{(i)}$는 관측시간의 순서통계량 $Y_{(i)}$와 $Y_{(i)}$에 대응되는 $D_{(i)}$의 값을 나타내고, $R_{(i)}$는 $y_{(i)}$에서 정의되는 위험집합(risk set)을 나타낸다. $\hat{\Lambda}_0(t)$는 사건이 일어나는 시점에서만 값이 변하므로 계단함수(step function)이다. 계단함수는 단조증가함수가 아니므로 역함수 값을 구하는 데 어려움이 있다. R의 Hmisc 팩키지 (Harrel, 2017)에 있는 inverseFunction 함수는 내삽을 이용해 계단함수를 단조증가함수로 만들어 역함수 값을 구하는데, 이 함수를 쓰면 $\hat{\Lambda}_0(t)$의 역함수 값을 얻을 수 있다.

$\Lambda_0(t)$를 비모수적으로 추정하면 사건발생 시간의 분포 형태에 대해 가정을 할 필요가 없다는 장점이 있다. 반면에 모수적으로 추정하는 방법과 비교했을 때 단점도 있는데, 먼저 역함수 값을 찾을 때 상대적으로 오래 걸린다. 왜냐하면 다음 절에서 설명할 모수적 방법에서는 $\hat{\Lambda}_0^{-1}(v)$가 v의 식으로 주어지기 때문에 계산만 하면 되지만 비모수적 방법에서는 v에 대응하는 $\hat{\Lambda}_0^{-1}(v)$를 찾아야 하기 때문이다. 모수적 방법에서 난수를 발생하는 데 걸리는 시간은 사건발생 수에 영향을 받지 않지만 비모수적 방법의 경우 사건발생 수가 많아질수록 더 오래 걸린다.

비모수적 방법의 또 다른 단점은 $\hat{\Lambda}_0^{-1}(v)$ 끝 부분에서 어떻게 값을 정의해야 할지가 불분명하다는 점이다. $\hat{\Lambda}_0(t)$는 $d_i = 1$인 관측값에서만 값이 증가하는 계단함수이다. $d_i = 1$인 관측값 중에 제일 큰 값을 $t_{\max}$라고 하고 $v_{\max} = \hat{\Lambda}_0(t_{\max})$라고 했을 때, $v > v_{\max}$에 대해 $\hat{\Lambda}_0^{-1}(v)$를 어떻게 정의해야 할지가 불분명하다. R의 Hmisc 팩키지에 있는 inverseFunction 함수는 $v > v_{\max}$에 대해 $\hat{\Lambda}_0^{-1}(v) = t_{\max}$로 정의한다. 이 정의를 따르게 되면 생성되는 생존시간 T^*의 최대값이 $t_{\max}$보다 클 수 없게 되고, $Y^* = \min(T^*, C^*)$이므로 Y^*도 $t_{\max}$보다 클 수 없게 된다. 만약 $t_{\max} < y_{(n)}$가 되면 생성되는 자료 Y^* 중에 $t_{\max}$보다 큰 값이 없게 되며 $t_{\max}$와 $y_{(n)}$ 사이에 관측된 중도절단된 자료는 아예 생성 불가능하게 된다. 이 문제를 해결하려면 모수적 방법을 쓰거나 $v > v_{\max}$일 때 $\hat{\Lambda}_0^{-1}(v)$를 $y_{(n)}$보다 큰 임의의 값으로 정의하면 된다.

중도절단자료 $\{(Y_i^*, D_i^*, x_i),\ i = 1, \ldots, n\}$를 생성하기 위해 T^*뿐만 아니라 C^*도 생성해야 한다. 생존분석에서 중도절단시간의 분포는 생존시간의 분포와 독립적이며 공변량과 무관하다고 일반적으로 가정한다. 이 '정보 없는 중도절단' 가정에 따라 중도절단시간 C^*를 비모수적으로 생성하려면 d_i의 역할을 바꿔주면 된다. 즉, $d_i^C = 1 - d_i$로 d_i^C를 정의한 다음, 자료 $\{(y_i, d_i^C),\ i = 1, \ldots, n\}$로부터 중도절단시간의 분포 $F^C(t) = P(C \leq t)$에 대한 Kaplan-Meier 추정량 $\hat{F}^C$을 얻을 수 있다. $\hat{F}^C$는 계단함수인데 R의 Hmisc 팩키지에 있는 inverseFunction 함수를 이용하면 분포 $\hat{F}^C$을 따르는 난수 C^*를 생성할 수 있다.

개체 i에 대한 중도절단시간 C_i^*를 랜덤하게 생성하지 않고 상수 c_i로 고정하는 것이 더 적절할 때가 있다. 연구기간이 $(0, T_e)$이고, 오직 연구기간 종료에 의하여 중도절단이 일어난다면 연구 진입 시점이

b_i인 관측개체의 중도절단시간은 $c_i = T_e - b_i$로 결정된다. 만약 모든 개체에 대한 연구가 동시에 시작되어 b_i가 모두 0이라면 모든 개체의 중도절단시간은 동일하게 T_e로 정해진다.

2.2. 모수적 방법

$\Lambda_0(t)$를 모수적으로 추정하는 방법에 대해 알아보자. 기저누적위험함수와 기저생존함수는 $S_0(t) = e^{-\Lambda_0(t)}$로서 일대일 대응관계에 있으므로, 기저생존함수를 모수(parameters)에 의해 결정되는 함수로 가정하고 모수를 추정하면 모수적 방법이 된다.

Bender 등 (2005)은 분포를 결정하는 모수의 값을 알고 있을 때 생존시간을 생성하는 방법을 제시하였다. 모의실험을 위한 모형도 현실에 기반을 둔 모형을 쓰는 것이 좋은데 실제로 관측된 자료로부터 추정한 모형을 모의실험을 위한 모형으로 쓰면 보다 실용적이다. 이 때 모수 추정을 위해 최대가능도(maximum likelihood estimator) 방법을 쓰기로 한다. 공변량 x인 개체의 위험함수를 $\lambda(t|x)$라고 하면 Cox 모형 가정 하에서 $\lambda(t|x) = \lambda_0(t)e^{\beta' x}$, $S(t|x) = [S_0(t)]^{\exp(\beta' x)}$이다. Cox 모형은 $\lambda_0(t)$의 형태에 제한을 두지 않는데 모수모형을 가정하는 경우 가능도함수가 어떻게 되는지 유도해보자. 확률밀도함수를 $f(t|x)$라고 하면 위험함수의 정의에 의해 $f(t|x) = \lambda(t|x)S(t|x)$이다. 관측된 중도절단자료를 $\{(y_i, d_i, x_i),\ i = 1, \ldots, n\}$라고 할 때, '정보 없는 중도절단'을 가정하면 가능도함수를 다음과 같이 표현할 수 있다.

$$\begin{aligned}\prod_{i=1}^{n}\left[f(y_i|x_i)^{d_i}S(y_i|x_i)^{1-d_i}\right] &= \left[\prod_{i:d_i=1}^{n}\lambda(y_i|x_i)\right]\left[\prod_{i=1}^{n}S(y_i|x_i)\right] \\ &= \left[\prod_{i:d_i=1}^{n}\lambda_0(y_i)e^{\beta' x_i}\right]\left[\prod_{i=1}^{n}(S_0(y_i))^{\exp(\beta' x_i)}\right] \qquad (2.2)\end{aligned}$$

기저분포(기저위험함수 λ_0와 기저생존함수 S_0)는 분포모수 θ에 의하여 그 형태가 결정된다 (θ는 벡터일 수 있다). 여기서 부분가능도(partial likelihood)와 완전가능도(full likelihood)를 구분할 필요가 있다. Cox 모형에서 부분가능도로 추정한 회귀계수는 기저위험함수(baseline hazard function) λ_0 형태에 무관하지만 (Cox 1972), 위 식과 같은 완전가능도로 추정한 회귀계수 값은 기저위험함수의 형태에 따라 달라진다. 왜냐하면 완전가능도 (2.2)에서 λ_0 또는 S_0를 결정하는 기저분포모수 θ와 회귀모수 β를 동시에 추정하기 때문이다. 분포모수를 추정할 때 회귀계수도 같이 추정하면 자료에 더 적합한 모수적 비례위험모형을 찾아줄 수 있을 것이다. 하지만 공변량의 수가 많거나 표본크기가 크지 않으면 추정이 어렵거나 불안정해진다. 회귀계수 추정이 우리의 목적이 아니고 모의자료 생성이 목적이므로 완전가능도에서 분포모수 θ에만 관심을 갖기로 한다. 즉, 관측자료 $\{(y_i, d_i, x_i),\ i = 1, \ldots, n\}$에 먼저 Cox 모형을 적용해서 얻은 회귀계수추정량 $\hat{\beta}$를 β로 간주하고, 가능도 식 (2.2)에서 기저분포모수 θ만 추정하기로 한다. 또는 β를 자료에서 구하지 않고 모의실험의 목적에 맞는 값으로 연구자가 직접 지정할 수도 있다. 모수적 방법에 의해 생성된 자료도 비례위험모형을 따르므로 Cox 모형의 특수한 경우로 간주할 수 있다.

만약 모수적 기저분포로 Weibull 분포를 가정한다면, $\lambda_0(t) = \lambda\nu t^{\nu-1}$, $t > 0$이고 $S_0(t) = e^{-\lambda t^{\nu}}$로서 $\theta = (\nu, \lambda)$이다. Cox 모형의 부분가능도를 최대화하는 추정량을 $\hat{\beta}$라고 하고, 가능도 (2.2)를 최대화하는 추정량을 $\hat{\theta} = (\hat{\nu}, \hat{\lambda})$라고 하면, $\hat{\Lambda}_0^{-1}(v) = (v/\hat{\lambda})^{1/\hat{\nu}}$이고

$$T^* = \left(-\frac{\log(U)\exp\left(-\hat{\beta}' x\right)}{\hat{\lambda}}\right)^{\frac{1}{\hat{\nu}}} \qquad (2.3)$$

Table 2.1. Execution time taken to generate one survival time (in microseconds)

	n: 자료의 크기				
	100	1000	10000	100000	1000000
비모수적 방법	71	124	501	5552	101163
모수적 방법	3	3	3	5	5

임을 식 (2.1)로부터 알 수 있다. 모수적 기저분포로 Gompertz 분포를 가정하는 경우 Bender 등 (2005)에 있는 표 I과 II를 참조하면 된다 (Bender 등 (2005)의 표 I에서 $\lambda_0(t) = e^{\alpha t}$는 $\lambda_0(t) = \lambda e^{\alpha t}$로 수정해야 함).

중도절단자료를 생성하려면 생존시간 T^*과 함께 중도절단시간 C^*도 생성해야 한다. 이 때 '정보 없는 중도절단'을 가정한다. $d_i^C = 1 - d_i$로 정의하면 자료 $\{(y_i, d_i^C),\ i = 1, \ldots, n\}$에 관한 가능도를 다음과 같이 표현할 수 있다.

$$\prod_{i=1}^{n} \left[f^C(y_i)^{d_i^C} S^C(y_i)^{1-d_i^C} \right] = \left[\prod_{i:d_i^C=1}^{n} \lambda^C(y_i) \right] \left[\prod_{i=1}^{n} S^C(y_i) \right].$$

위 가능도를 최대화하는 중도절단시간의 분포 f^C 또는 S^C에 관한 모수 추정량 $\hat{\theta}^C$로부터 $\hat{S}^C$를 얻을 수 있고, $\hat{S}^C$의 역함수로부터 C^*를 생성할 수 있다. Weibull 분포를 가정하는 경우 $\hat{\theta}^C = (\hat{\nu}^C, \hat{\lambda}^C)$라고 하면

$$C^* = \left(-\frac{\log(U)}{\hat{\lambda}^C} \right)^{\frac{1}{\hat{\nu}^C}}$$

이다.

모수적 방법과 비모수적 방법의 계산 시간을 비교해보고자 한다. Weibull 분포를 가정한 모수적 방법으로 T^*를 한 개 생성할 때 걸리는 시간은 식 (2.3)에서 공변량을 무시한다면 $(-\log(U)/\hat{\lambda})^{1/\hat{\nu}}$를 계산할 때 걸리는 시간과 같으므로 표본 크기에 무관하다 (분포모수 $\hat{\lambda}$, $\hat{\nu}$는 이미 추정되었다고 가정한다). 반면에 비모수적 방법을 적용한다면 식 (2.1)에서 $\hat{\Lambda}_0^{-1}[-\log(U)]$를 구해야 하는데, 표본 크기가 커질수록, 보다 정확하게는 서로 다른 생존시간 개수가 많아질수록 역함수 값을 찾는 데 시간이 더 걸리게 된다. 정량적으로 비교해보기 위해 R의 microbenchmark 팩키지 (Mersmann, 2018)에 있는 microbenchmark 함수를 이용하였다. 한 개의 생존시간 T^*를 생성하는 데 걸리는 시간을 두 방법에 대해 각각 측정하여 보았다. microbenchmark 함수는 지정한 연산을 100번 수행한 다음 소요되는 시간의 평균, 중앙값, 최대값 등을 구해준다. Table 2.1에 평균을 정리하였다. 예상대로 모수적 방법은 자료의 크기에 영향을 받지 않지만 비모수적 방법은 그렇지 않음을 알 수 있다.

모수적 방법이 빠르지만 모수적 방법에서 가정한 기저분포가 관측자료에 가까운 분포인지를 판단하는 데 어려움이 있다. 기저분포는 공변량 값이 0인 개체들의 분포를 나타낸다고 볼 수 있는데 관측자료에 이런 개체에 대한 자료는 없거나 드물다. 따라서 특정한 형태의 모수적 분포가 관측자료에 적합한지를 판단하기가 어렵다. 모의실험을 위한 자료이므로 문제가 안 될 수도 있지만, 가능하면 실제로 관측된 자료에 가까운 기저분포를 적용하고 싶다면 모수적 방법보다 비모수적 방법을 적용하는 것이 좋다.

3. 모의자료 생성 예

실제 사례를 들어 Cox 모형을 따르는 모의자료를 생성하는 비모수적 방법과 모수적 방법을 설명하고자 한다. 생성된 모의자료를 이용해서 무엇을 할 수 있는지에 대해서도 살펴본다.

Table 3.1. Cox model for the observed data

	$\hat{\beta}$	$\exp(\hat{\beta})$	$\hat{\beta}$의 표준오차	p-값 (H_0: $\beta = 0$)
CLINIC2	−0.9318	0.3939	0.2183	1.98e−05
PRISON	0.2869	1.3323	0.1683	0.08826
DOSE2	−0.5255	0.5913	0.1772	0.00301
DOSE3	−1.5472	0.2128	0.3079	5.02e−07

Table 3.2. The mean and 95% confidence interval of the regression coefficient for CLINIC2, obtained from 400 iterations of simulation

		관측자료에 적용한 Cox 모형에서 자료생성	DOSE3 계수를 1.5배 강화한 모형에서 자료생성
비모수적 방법	DOSE 고려	−0.9297 (−0.9510, −0.9084)	−0.9382 (−0.9604, −0.9160)
	DOSE 누락	−0.9842 (−1.003, −0.9648)	−1.0071 (−1.0261, −0.9882)
모수적 방법 (Weibull 분포)	DOSE 고려	−0.9501 (−0.9731, −0.9271)	−0.9452 (−0.9669, −0.9234)
	DOSE 누락	−0.9892 (−1.0097, −0.9686)	−0.9939 (−1.0116, −0.9761)

Caplehorn과 Bell (1991)은 메타돈(methadone)이라는 약물이 헤로인 중독 치료에 미치는 효과를 연구하였는데 자료 일부가 공개되어 있다. 238명의 헤로인 중독자를 두 병원으로 나누어 치료 받게 하고 입원부터 퇴원까지의 시간을 관측하였다. 정해진 연구기간이 종료되기까지 퇴원하지 않은 환자의 시간은 중도절단된 자료이다. 환자가 입원 전에 교도소에 수감된 적이 있는가를 나타내는 이항형 변수(PRISON)와 환자에게 투여된 메타돈의 최댓값을 세 개의 범주로 구분한 변수(DOSE), 그리고 환자가 입원한 병원을 나타내는 변수(CLINIC)를 공변량으로 관측하였다. 우리의 목적은 Cox 모형을 따르는 모의자료 생성을 설명하고자 하는 것인데, 메타돈의 효과보다 두 병원의 차이에 관심을 두고자 한다. 두 병원은 의료 인력이나 설비, 입원 환자의 증상 정도에 차이가 있을 수 있겠지만, 무엇보다 메타돈 투여량에 뚜렷한 차이가 있음을 관측 자료에서 알 수 있다. 병원 1보다 병원 2가 보다 많은 양의 메타돈을 투여하였는데, 메타돈 투여량을 보정했을 때와 하지 않았을 때 두 병원의 차이를 나타내는 회귀계수가 어떻게 달라지는지를 알아보자.

238명의 자료에 대해 Cox 모형을 적합시킨 결과는 Table 3.1과 같다. Table 3.1에서 CLINIC2는 병원 2를 나타내는 가변수로서 병원 1이 기준범주이다. 메타돈 최대투여량이 하루 60 밀리그램 미만이면 1, 60 이상 80 미만이면 2, 80 이상인 경우 3으로 나타낸 범주형 변수가 DOSE인데, 첫 번째 범주 1을 기준 범주로 하고 2와 3인 경우를 나타내는 가변수를 각각 DOSE2와 DOSE3으로 표기하였다. 위 모형을 참 모형으로 가정하고 위 모형을 만족하는 모의자료를 생성할 수 있다. 관측자료를 $\{(y_i, d_i, x_i),\ i = 1, \ldots, 238\}$라고 하고 관측자료에 적합시킨 Cox 모형으로부터 생성한 모의자료를 $\{(y_i^*, d_i^*, x_i), i = 1, \ldots, 238\}$라고 하자. 두 자료의 중요한 차이점은 관측자료의 참 모형은 알 수 없으나 모의자료 $\{(y_i^*, d_i^*, x_i),\ i = 1, \ldots, 238\}$의 참 모형은 알고 있다는 점이다. 관측자료는 모의자료 생성을 위한 참 모형을 정의하기 위해 이용하였다. 모의자료의 참 모형을 알고 있으므로, 참 모형과 반복생성한 모의자료를 이용하면 메타돈 투여량 보정여부에 따라 두 병원의 차이가 어떻게 달리 추정되는가를 알 수 있다.

모의자료를 생성한 참 모형으로부터 메타돈 투여량에 따라 생존시간(입원부터 퇴원까지의 시간)의 분포

가 달라진다는 것을 알고 있다. 만약 메타돈 투여량을 무시하고 두 병원의 차이를 추정하게 되면 편향된 결과를 얻게 될 것인데 이 편향의 크기가 얼마나 될까를 알고 싶다고 하자. 비모수적 방법으로 모의자료를 생성해서 메타돈 투여량 DOSE를 무시하고 CLINIC과 PRISON 두 공변량만 고려한 잘못된 모형을 적합시키면 CLINIC2 회귀계수의 평균값이 약 -0.9842가 된다 (평균값을 구하기 위해 모의실험을 400번 반복했다). DOSE를 포함시킨 제대로 된 모형을 적합시켰을 때 얻게 되는 회귀계수의 평균값은 약 -0.9297이다. 따라서 편향의 크기는 $-0.9842-(-0.9297)=-0.0545$로서, 약 5.8% $(=0.0545/0.9297)$의 상대적 편향이 발생한다. 위험비(hazard ratio)로 나타내면 병원 2에 입원한 환자는 병원 1에 입원한 환자에 비해 퇴원 위험률(hazard rate)이 $e^{-0.9297}\doteq 0.3947$배이다. 만약 메타돈 투여량을 무시하고 비교하면 $e^{-0.9842}\doteq 0.3737$배로 잘못 추정하게 된다. 실용적으로 큰 차이가 아닐 수 있지만 통계적으로 유의한 차이임을 400번의 모의실험 결과로 알 수 있다. 만약 메타돈 투여량이 입원기간에 미치는 영향이 더 큰 경우에 이를 무시하면 편향의 크기가 어떻게 달라지는가를 알아보자. 모의실험의 장점은 참 모형을 우리가 원하는 대로 변형시켜볼 수 있다는 점이다. DOSE2 회귀계수는 그대로 두고 DOSE3 회귀계수를 -1.5472에서 -2.3208로 1.5배 강화해 준 모형에서 모의자료를 생성하여 CLINIC2 계수 변화를 살펴보았다 (Table 3.2). DOSE를 무시한 경우와 제대로 고려한 경우를 비교해 보면, 더 중요해진 변수를 무시했으므로 편향이 좀 더 커짐을 알 수 있다.

모수적 방법을 적용하면 계수의 크기는 조금씩 달라지지만 DOSE3 계수를 바꿔줬을 때 편향이 확대되는 정성적 성질은 같았다 (Table 3.2). 가정한 모수적 분포가 참 분포에 가깝다면 비모수적 방법과 모수적 방법에 별 차이가 없을 것이다.

4. 요약과 첨언

실제 자료에 근거해서 Cox 모형을 만든 후 그 모형을 따르는 중도절단자료를 생성하는 방법을 제시하였다. 모수적 방법뿐만 아니라 비모수적 방법을 같이 제시하였다. 비모수적 방법을 적용하면 자료 크기가 커질수록 계산 속도가 느려진다는 단점이 있는 반면에 실제 자료가 갖는 분포에 더 가까운 모의자료를 생성할 수 있다는 장점이 있다.

Kropko와 Harden (2018)이 만든 coxed라는 R 패키지의 sim.survdata 함수를 이용하면 Cox 모형을 따르는 중도절단자료를 생성할 수 있다. 이 함수를 이용하려면 기저위험함수의 식을 사용자가 직접 정의하거나 자동 설정된 기저위험함수를 써야 한다. 본 연구에서 제안한 비모수적 방법은 실제자료를 이용해서 누적기저위험함수를 추정하므로 sim.survdata 함수가 적용하는 방법과 다르다. 모수적 방법의 경우 분포모수의 최대가능도추정량을 실제자료로부터 먼저 추정한 다음 추정된 최대가능도추정량으로 표현한 기저위험함수를 sim.survdata 함수의 입력함수로 정의한다면 sim.survdata 함수가 적용하는 방법과 같아질 것이다.

Cox 모형에서 모의자료를 생성할 수 있다면 생존분석에서 해석적 방법의 한계를 벗어날 수 있어 유용하게 쓸 수 있을 것이다. 예를 들어 관측된 중도절단자료에 대해 인과추론(causal inference)을 하고자 할 때, 관심 있는 평균처리효과(average treatment effect)를 추정하기 위해 몇 가지 다른 방법을 적용해볼 수 있는데, 다양한 상황에서 여러 방법의 성능을 비교하고 싶을 때 모의자료를 생성할 수 있다면 유용할 것이다.

본 연구에서 공변량 생성에 대해서는 다루지 않았으나 공변량도 필요하다면 생성할 수 있다. 예를 들어 관측자료보다 더 많은 모의자료를 추출하고자 할 때 새로운 개체가 필요하므로 생존시간뿐만 아니라 공변량 생성도 필요하다. 공변량은 중도절단자료가 아니므로 일반적인 난수생성 방법을 적용하면 된다. 다만, 회귀분석에서 추정은 공변량 값이 관측된 값으로 고정되었다고 가정하고 이루어진다는 점에 유의

할 필요가 있다.

Cox 모형은 공변량이나 회귀계수가 시간에 의존하지 않는다는 것을 가정한다. 본 연구에서 제안한 방법도 이런 가정 하에서 적용할 수 있으며, 시간에 의존하는 공변량(time-dependent covariates)이나 시간에 의존하는 회귀계수(time-dependent coefficients)의 경우 적용할 수 없다.

References

Bender, R., Augustin, T., and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models, *Statistics in Medicine*, **24**, 1713–1723.

Breslow, N. (1974). Covariance analysis of censored survival data, *Biometrics*, **30**, 89–99.

Caplehorn, J. R. and Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment, *The Medical Journal of Australia*, **154**, 195–199.

Cox, D. R. (1972). Regression models and life tables (with discussion), *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Harrell Jr, F. with contributions from Charles Dupont and many others (2017). Hmisc: Harrell Miscellaneous. R package version 4.0-3. https://CRAN.R-project.org/package=Hmisc

Kropko, J. and Harden, J. (2018). coxed: Duration-Based Quantities of Interest for the Cox Proportional Hazards Model. R package version 0.2.0. https://CRAN.R-project.org/package=coxed

Mersmann, O. (2018). microbenchmark: Accurate Timing Functions. R package version 1.4-4. https://CRAN.R-project.org/package=microbenchmark

Moriña, D. and Navarro, A. (2014). The R package survsim for the simulation of simple and complex survival data, *Journal of Statistical Software*, **59**, 1–20.

Cox 비례위험모형을 따르는 중도절단자료 생성

김지현[a,1] · 김봉성[a]

[a]숭실대학교 정보통계보험수리학과

(2018년 9월 14일 접수, 2018년 10월 19일 수정, 2018년 11월 2일 채택)

요 약

통계학 연구에 모의실험이 중요하게 쓰이며 중도절단자료를 다루는 생존분석에서도 마찬가지다. 생존분석에서 Cox 모형이 널리 쓰이는데, Cox 모형을 따르는 중도절단자료를 생성하는 방법에 대해 살펴보았다. Bender 등 (*Statistics in Medicine*, **24**, 1713–1723, 2005)은 생존시간을 생성하는 모수적 방법을 제시하였으나 생존시간뿐만 아니라 중도절단시간도 생성해야 중도절단자료를 얻게 된다. 중도절단자료를 생성하기 위한 모수적 방법과 함께 비모수적 방법도 제시하였으며 실제 자료에도 적용해 보았다.

주요용어: Cox 모형, 모의실험, 기저위험함수, Kaplan-Meier 추정량

[1]교신저자: (06978) 서울시 동작구 상도로 369, 숭실대학교 정보통계보험수리학과. E-mail: jxk61@ssu.ac.kr