The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.7.1231
29(7), 1231–1246

Estimation methods and interpretation of competing risk regression models

Mijeong Kim[a,1]

[a]Department of Statistics, Ewha Womans University

(Received July 18, 2016; Revised September 3, 2016; Accepted October 7, 2016)

Abstract

Cause-specific hazard model (Prentice *et al.*, 1978) and subdistribution hazard model (Fine and Gray, 1999) are mostly used for the right censored survival data with competing risks. Some other models for survival data with competing risks have been subsequently introduced; however, those models have not been popularly used because the models cannot provide reliable statistical estimation methods or those are overly difficult to compute. We introduce simple and reliable competing risk regression models which have been recently proposed as well as compare their methodologies. We show how to use SAS and R for the data with competing risks. In addition, we analyze survival data with two competing risks using five different models.

Keywords: proportional odds models, competing risks, cumulative incidence function, subdistribution

1. 서론

t 시점에서의 위험함수(hazard function) $h(t)$는 다음과 같이 정의할 수 있다.

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

$F(t) = P(T < t)$이고, $f(t) = F'(t)$이다. $h(t)$는 t 시점에 장비가 고장나는 비율, 사망하는 개체의 비율 등에 적용될 수 있다. 이러한 위험 함수에 대해 Cox (1972)는 위험에 영향을 줄 수 있는 공변량을 고려하여 다음과 같은 비례 위험 모형(proportional hazard model)을 제안했다.

$$\lambda(t; Z) = \lambda_0(t) \exp\left(Z^{\mathrm{T}}\beta\right). \tag{1.1}$$

이 때, $\lambda_0(t)$는 t 시점의 기저위험률(baseline hazard)이고, $\lambda(t; Z)$는 공변량의 함수 $\exp(Z^{\mathrm{T}}\beta)$가 $\lambda_0(t)$에 비례하는 식으로 표현된다. Cox 모형은 경쟁위험(competing risk)을 고려하지 않은 위험함수에 대한 초기 모형이라고 할 수 있다. 현재까지도 경쟁 위험이 존재하는 생존 자료(survival data)에 대해서 위험함수 및 누적 발생률(cumulative incidence)에 대한 많은 연구가 Cox 모형 (1.1)을 발전시킨 형태로 이루어지고 있다. 위험함수와 누적 발생률을 설명하는 모형에 대한 접근 방법으로 모수

This research is supported by grants from Ewha Womans University.
[1]Department of Statistics, Ewha Womans University, 52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Korea. E-mail: m.kim@ewha.ac.kr

모형(parametric model), 준모수 모형(semiparametric model) 이외에도 비모수 방법(nonparametric model)이 있다. 비모수 방법으로는 Gray (1988), Lin (1997)에서 제시된 방법을 통해 누적발생함수를 추정할 수 있다. 비모수 방법은 자료의 패턴을 보는 면에서 유용할 수 있으나, 모수를 통해 설명할 수 있는 부분이 없다는 면에서 해석 측면에서 한계가 있다. 이 논문에서는 경쟁 위험에 대해 준모수 방법으로 접근한 몇가지 회귀모형에 대해 소개하고자 한다. 비교적 자주 이용되는 경쟁 위험 회귀 모형은 Prentice 등 (1978)의 cause-specific 위험 모형과 Fine과 Gray (1999)의 subdistribution을 이용한 비례 위험 모형으로써, Cox 비례 위험 모형 (1.1)을 발전시킨 모형이다. 이 두 모형은 위험 함수로부터 누적 발생률을 모형화하였다. Scheike 등 (2008)의 이항 회귀 모형(direct binomial model)은 위험 함수를 통하지 않고 누적 발생률 함수에 대한 모형을 만드는 방법이다. Scheike 등 (2008)의 이항 회귀 모형은 누적 발생률 함수에 다양한 형태의 모델링이 가능하다. Gerds 등 (2012)의 절대 위험 회귀 모형(absolute risk regression model)은 누적 발생률 함수에 대한 회귀 계수에 대한 설명력이 의미가 있는 모형이다. 마지막으로, Eriksson 등 (2015)가 제안한 비례 오즈 모형(proportional odds model)에 대한 새로운 추정방법을 소개하고자 한다

2. 경쟁 위험 회귀 모형

2.1. Cause-specific 위험 모형

우선 위험 집합(risk set), 경쟁 위험(competing risk)의 개념을 이해할 필요가 있다. t 시점에서의 위험 집합(risk set)은 t 시간까지 고장(failure)이 일어나지도 중도 절단(censored)도 일어나지 않은 상태로써 추후에 위험에 노출될 가능성이 있는 개체들의 집합을 뜻한다. K개의 각각 다른 종류의 위험이 있는 경우, 각각의 원인을 '원인 1, 원인 2, ..., 원인 K'라고 하고, 원인 j로 인한 위험이 발생한 사건을 사건 j ($j = 1, \ldots, K$) (event j)라고 하자. 서로 다른 위험은 서로에게 경쟁 위험(competing risk)의 관계에 있다. 예를 들어, 골수 이식으로 인한 사망률을 연구한다고 가정하자. 골수 이식을 받은 사람 중에 골수 이식으로 인해 사망한 사람, 백혈병의 재발로 인한 사람이 있다면, 골수 이식을 원인 1, 백혈병의 재발을 원인 2로 보는 것이 합리적이다. 이 때 백혈병의 재발은 골수 이식의 경쟁 위험이라고 한다. 이미 원인 1이 아닌 다른 위험, 예를 들어 원인 2로 인한 위험이 있었다면, 그 개체는 원인 1의 위험 집합에는 포함되지 않고, 중도 절단된 것과 같이 취급한다. 연속 시간을 가정한다면, 원인 j로 인한 cause-specific hazard는 다음과 같다.

$$h_j(t) = \lim_{\Delta t \to 0} \frac{P(t < T \le t + \Delta t, \epsilon = j | T > t)}{\Delta t} = \frac{f_j^*(t)}{S(t)}.$$

이 때, $F_j^*(t) = P(T < t, \epsilon = j)$, $S(t) = P(T > t) = \exp\{-\int_0^t \sum_{k=1}^K h_k(u)du\}$이고 $f_j^*(t) = \partial F_j^*(t)/\partial t$이다. $\int_t f_j^*(t)dt < 1$이므로, 부적절 분포(improper distribution)이고, 부적절 분포를 표시하기 위해 윗첨자 *을 붙인다. Cause-specific 위험에 근거한 비례 위험 모형은 다음과 같다.

$$h_j(t; Z) = h_{j0}(t) \exp\left(Z^{\mathrm{T}} \beta_j\right), \quad j = 1, \ldots, K. \tag{2.1}$$

Holt (1978)는 식 (2.1)로부터 다음과 같이 β_j에 대한 편우도 함수(partial likelihood function)를 구하였다.

$$\prod_{j=1}^{K} \left[\prod_{\nu=1}^{d_j} \frac{\exp\left\{z_{j(\nu)}^{\mathrm{T}} \beta_j\right\}}{\sum_{l \in R\{t_{j(\nu)}\}} \exp\left\{z_l^{\mathrm{T}} \beta_j\right\}} \right]. \tag{2.2}$$

이 때, $t_{j(\nu)}$ $(\nu = 1, \ldots, d_j)$는 원인 j로 인한 d_j개의 고장, $R\{t_{j(\nu)}\}$는 $t_{j(\nu)}$ 시간까지의 위험 집합, $z_{j(\nu)}$는 시간 $t_{j(\nu)}$에 해당하는 공변량이다. 식 (2.2)을 최대로 만드는 β_j를 찾음으로써 추정치를 구할 수 있다.

2.2. Fine과 Gray (1999) 모형

Subdistribution 위험은 cause-specific 위험과 다르게 정의된 위험 집합에 기초하여 정의된다. t 시점에서의 원인 1에 대한 위험 집합은 t 시점 전에 사건 1과 중도 절단을 경험하지 않은 모든 개체들이 포함된다. 예를 들어, t 시점 전에 이미 원인 2로 인한 위험에 노출되어서 t 시점에서의 상태를 알 수 없을 경우에도 원인 1의 위험을 겪을 가능성이 있다고 가정하고 그 개체를 원인 1에 대한 위험 집합에 포함시킨다. 사건 j에 대한 subdistribution 위험 모형은 다음과 같다.

$$\lambda_j(t; Z) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\{t \le T \le t + \Delta t, \epsilon = j | T \ge t \cup (T \le t \cap \epsilon \ne j), Z\} = \frac{f_j^*(t; Z)}{1 - F_j^*(t; Z)}.$$

이 때, $T^* = I(\epsilon = j) \times T + \{1 - I(\epsilon = j)\} \times \infty$라고 하면, $t < \infty$일 때 T^*의 분포함수는 $F_j^*(t; Z)$이고 $f_j^*(t; Z) = \partial F_j^*(t; Z)/\partial t$이다. $t = \infty$이면, $\Pr(T^* = \infty; Z) = \Pr(T < \infty, \epsilon \ne j; Z) = 1 - F_j(\infty; Z)$이다. 이제부터는 편의상 사건 1에 대한 위험 모형을 설명하고자 한다. Fine과 Gray (1999)는 Cox 모형 (1.1)에 근거하여, 경쟁 위험을 고려한 비례 subdistribution 위험 모형을 다음과 같이 정의하였다.

$$\lambda_1(t; Z) = \lambda_{10}(t) \exp\left(Z^{\mathrm{T}}\beta\right). \tag{2.3}$$

여기서 $\lambda_{10}(t)$는 사건 1에 대한 subdistribution 기저 위험률로써 단조 증가하는 임의의 양의 함수를 가정한다. Fine과 Gray (1999)에서는 식 (2.3)을 로그 변환하고, $g(x) = \log\{-\log(1 - x)\}$를 이용하여 다음과 같은 모형을 제안했다.

$$g(F_1^*(t; Z)) = \log\{\Lambda_{10}(t)\} + Z^{\mathrm{T}}\beta. \tag{2.4}$$

Robins와 Rotnitzky (1995)가 소개한 중도절단 가중치에 대한 역확률 inverse probability of censoring weighting(IPCW)을 이용하여 불완비자료(incomplete data)에 대한 추정이 가능하게 하였다. 중도 절단된 시점 C가 시간 T, 위험 발생의 원인 ϵ, 공변량 Z와 독립이라는 가정하에, $P(C \ge t) = G(t)$라고 하자. 특정 원인으로 인한 위험이 발생한 시간(failure time)을 T^*이라고 하자. 중도 절단이 있는 경우, T^*은 관찰되지 않는다. 따라서, 실제로 관찰된 시간을 T라고 한다면, $T = T^* \wedge C$로 표시할 수 있다. 중도 절단된 경우에는 위험 발생의 원인(cause of failue) ϵ는 관찰되지 않는다. 중도 절단의 지표로서 $\Delta = I(T^* \le C)$를 이용한다. Δ는 중도 절단 되지 않을 경우 1, 중도 절단되었을 때에는 0의 값을 갖는다. t 시점에서 개체 i $(i = 1, \ldots, n)$의 vital 상태 $r_i(t)$를 $I(C_i \ge T_i^* \wedge t)$로 계산한다. 중도절단 가중치 $w_i(t, G)$는 다음과 같다.

$$w_i(t, G) = \frac{r_i(t)G(t-)}{G\{(T_i \wedge t)-\}}, \tag{2.5}$$

여기서 T_i^*, ϵ_i가 주어졌을 때 $r_i(t)/G\{(T_i \wedge t)-\}$의 조건부 기대값은 1이다. G를 추정하는 방법으로 Kaplan과 Meier (1958) 방법이 주로 쓰인다. 가중치 $w_i(t)$를 적용하여 편우도 함수를 만든 후 다음과 같은 추정 방정식이 얻어진다.

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i(s) - \frac{\sum_j w_j(s) Y_j(s) Z_j(s) \exp\left\{Z_j^{\mathrm{T}}(s)\beta\right\}}{\sum_j w_j(s) Y_j(s) \exp\left\{Z_j^{\mathrm{T}}(s)\beta\right\}} \right\} w_i(s) dN_i(s).$$

이 때, $N_i(t) = I(T_i \le t, \epsilon_i = 1)$이고 $Y_i(t) = 1 - N_i(t-)$이다. 위의 편우도 함수를 최대로 하는 추정치를 구할 수 있다. Fine과 Gray 방법은 경쟁 위험 모형으로 자주 쓰이는 방법이지만, 식 (2.4)에서 $F_1^*(t;Z)$는 cause-specific 위험과 관련된 분포가 아니라, subdistribution에 근거한 분포이기 때문에 위험률 $\lambda_1(t;Z)$을 해석하기가 쉽지 않다 (Fine 2001). 반면에, subdistribution에 근거한 오즈비 $g(F_1^*(t;Z)) = \text{logit}(F_1^*(t;Z)) = \log\{F_1^*(t;Z)/(1-F_1^*(t;Z))\}$는 사건 1로 인한 위험과 그 사건의 여집합의 확률의 비로 해석이 되기 때문에 이해하기가 더 수월할 수 있다.

2.3. 이항 회귀 모형

경쟁 위험이 있는 데이터를 다룰 때, 특정 사건에 기인한 위험률, 누적 발생 확률 또는 생존 함수에 관심이 있는 경우가 있다. 앞서 설명한 비례 위험 모형은 시간에 따른 변수의 효과가 일정하다는 가정 하에 만들어진 모형이다. 그러나 누적 발생 모형에서 시간에 따라 다른 효과를 가진 변수가 존재할 가능성이 있으며, 특정 변수는 위험률에는 큰 영향이 있지만, 누적발생 확률 관점에서는 상태적으로 효과가 작을 수도 있다. 이런 면에서 위험률로부터 누적 발생 확률 모형을 만들어내는 것이 합리적이지 않을 수 있고, 시간에 따른 변수의 효과를 고려해야하는 필요가 생긴다. Scheike 등 (2008)에서 비례 위험 모형으로부터 모형 변환을 거치지 않고, 누적 발생 확률 함수를 모델링하는 이항 회귀 모형(direct binomial model)을 고안했다. 이 장에서도 사건 1에 초점을 맞추어서 설명하겠다. Cox 모형 (1.1)은 효과가 시간에 따라 변하지 않는 모형이지만, 이 모형에서는 다음과 같이 시간에 따라 변하는 효과 $\eta(t)$를 고려한다.

$$P_1^{\eta}(t;X_i) = E\{N_i(t)|X_i\} = h\left\{X_i^{\mathrm{T}}\eta(t)\right\}. \tag{2.6}$$

이 때, $N_i(t)$는 2.2장에서 정의한 것과 같이 $N_i(t) = I(T_i \le t, \epsilon_i = 1)$이고, 원인 1에 해당하므로 $P_1^{\eta}(t;X_i)$에 아래첨자 1을 붙였다. 연결 함수 h의 형태는 알고 있다고 가정하고 다양한 형태의 양의 증가함수를 적용할 수 있다. 절단된 데이터에 $\Delta_i N_i(t)/G(T_i|X_i)$ 만큼의 가중치를 부여할 경우, 다음 식을 만족한다.

$$E\left\{\frac{\Delta_i N_i(t)}{G(T_i|X_i)}\right\} = E\left[E\left\{\frac{\Delta_i N_i(t)}{G(T_i|X_i)}\middle|T_i,\epsilon_i,X_i\right\}\right] = E\{N_i(t)|X_i\} = P_1^{\eta}(t;X_i),$$

여기서 $G(T_i|X_i)$은 Kaplan-Meier 방법으로 추정하고, 모형 (2.6)에서 $\eta(t)$는 다음과 같은 점수 함수 $U_n(\eta,\widehat{G})(t) = 0$를 통해 추정할 수 있다. 즉, 각각의 t에서 $U_n(\eta,\widehat{G})(t) = 0$을 찾는 방식이다.

$$U_n\left(\eta,\widehat{G}\right)(t) = \sum_{i=1}^{n} D_{\eta}^{T}(t,X_i)w(t,X_i)\left\{\frac{\Delta_i N_i(t)}{\widehat{G}(T_i|X_i)} - P_1^{\eta}(t,X_i)\right\},$$

여기서 $D_{\eta}^{T}(t,X_i) = \partial P_1^{\eta}(t,X_i)/\partial\eta(t)$이고, $w(t,x)$는 가중치이다.

한편, Scheike 등 (2008)은 모형 (2.6)의 확장된 모형으로써 다음과 같이 두 가지 종류의 준모수 모형을 고려하였다.

$$P_1^{\eta,\gamma}(t;X_i,Z_i) = h\left\{X_i^{\mathrm{T}}\eta(t)g(\gamma,Z_i,t)\right\}, \tag{2.7}$$

$$P_1^{\eta,\gamma}(t;X_i,Z_i) = h\left\{X_i^{\mathrm{T}}\eta(t) + g(\gamma,Z_i,t)\right\}. \tag{2.8}$$

이 때, Z_i는 시간과 무관하게 일정한 효과를 갖는 공변량이고, 연결 함수 g는 비모수 모형 또는 모수 모형이 될 수 있다. g는 비모수 모형으로써 상수 또는 조각별 상수(piecewise-constant) 함수를 이용할 수

있다. 모형 (2.7)을 승법 준모수 모형(multiplicative semiparametric model)이라 하고, 모형 (2.8)을 가법 준모수 모형(additive semiparametric model)이라고 한다. 모형 (2.7)과 (2.8)은 h 함수의 선택에 따라 기존 모형으로 변환할 수 있다. subdistribution 방법으로 위험 집합을 구하고, $h(y) = 1 - \exp(-y)$, $x_i = 1$로 두면, 식 (2.7)을 통해 Fine과 Gray 모형 (2.3)이 된다.

$(\widehat{\eta}, \widehat{\gamma})$는 $U_n(\eta, \gamma, \widehat{G})(t) = \{U_n^1(\eta, \gamma, \widehat{G})(t), U_n^2(\eta, \gamma, \widehat{G})(t)\} = 0$을 통해 구할 수 있다. 이 때, $U_n(\eta, \gamma, \widehat{G})(t)$는 다음과 같다.

$$U_n^1\left(\eta, \gamma, \widehat{G}\right)(t) = \sum_{i=1}^{n} D_\eta^T(t, X_i, Z_i) w(t, X_i, Z_i) \left\{ \frac{\Delta_i N_i(t)}{\widehat{G}(T_i|\boldsymbol{X}_i, Z_i)} - P_1^{\eta,\gamma}(t, X_i, Z_i) \right\},$$

$$U_n^2\left(\eta, \gamma, \widehat{G}\right)(t) = \sum_{i=1}^{n} \int_a^\tau D_\gamma^T(t, X_i, Z_i) w(t, X_i, Z_i) \left\{ \frac{\Delta_i N_i(t)}{\widehat{G}(T_i|\boldsymbol{X}_i, Z_i)} - P_1^{\eta,\gamma}(t, X_i, Z_i) \right\} dt.$$

$D_\eta^T(t, X_i, Z_i) = \partial P_1^{\eta,\gamma}(t, X_i, Z_i)/\partial\eta(t)$, $D_\gamma^T(t, X_i, Z_i) = \partial P_1^{\eta,\gamma}(t, X_i, Z_i)/\partial\eta(t)$이고, $w(t, X_i, Z_i)$는 가중치이다. 이 연구에서는 중도 절단된 데이터가 있는 경우에 G를 추정하는 데 있어서 Kaplan Meier 방법보다 더 효율적인 방법을 제안했다. 또한 회귀 계수의 추정치가 일치성과 효율성을 보임을 증명했다.

2.4. 절대 위험 회귀 모형

비례 위험 모형 cause-specific 위험 모형과 subdistribution 위험 모형은 위험률에 초점을 맞추어 모형이 만들어진다. 위험률 함수로부터 누적 발생 확률을 만들기 위해서는 식 (2.4)와 같은 모형의 변환이 필요한데, 위험률 함수로부터 변환된 누적 발생함수는 연결 함수(link function)에 따라 누적 발생 확률의 예측의 정확성이 달라지고, 회귀 계수도 다르게 해석된다. Gerds 등 (2012)은 절대 위험 회귀 모형(absolute risk regression model)을 제안했다. 이 모형은 이항 회귀 모형처럼 비례 위험 모형에 근거하지 않고, 직접 누적 발생 확률 함수를 만드는 방법을 이용한다. 절대 위험 회귀 모형은 변수값이 한 단위 변화했을 때 누적 발생 확률이 얼마나 달라지는지 설명이 가능한 모형이다. K개의 경쟁 위험이 있는 경우, 사건 1의 누적 발생률 함수는 다음과 같다.

$$F_1(t; Z) = F_{1,0}(t) \exp\left(Z^{\mathrm{T}}\beta\right), \tag{2.9}$$

여기서 $F_{1,0}$은 $Z = 0$일 때의 누적 발생률 함수이다. 공변량 Z는 p차원이라고 가정한다. 다음 식으로부터 Z_k가 한단위 증가할 때의 누적 확률의 변화가 $\exp(\beta_k)$임을 알 수 있다.

$$\frac{F_1^{(ARR)}(t, Z_1 = z_1, \ldots, Z_k = z_k + 1, \ldots, Z_p = z_p)}{F_1^{(ARR)}(t, Z_1 = z_1, \ldots, Z_k = z_k, \ldots, Z_p = z_p)} = \exp(\beta_k).$$

이 때, $k = 1, \ldots, p$이다. 다음 준모수 모형을 생각해 보자.

$$g\{F_1(t; Z)\} = \beta_0(t) + \beta_1 Z_1 + \cdots + \beta_p Z_p \tag{2.10}$$

이고 g는 미분 가능한 함수로써 그 형태를 알고 있다고 가정한다. $g(p) = \log(p)$, $\beta_0(t) = \log F_{1,0}(t)$이면, 식 (2.9)는 식 (2.10)과 같게 된다. 절대 위험 회귀 모형에서도 Fine과 Gray의 모형과 유사하게 중도 절단된 데이터에 대해서는 IPCW 방법을 이용한다. 모형 (2.10)에 대해서 일반화 추정방정식(generalized estimation equation)에 대한 근을 찾는 방식으로 공변량의 효과를 추정한다. 이 모형은 특정 사건에 대한 예측 확률이 1을 초과할 수 있다는 단점이 있다.

2.5. 비례 오즈 모형

Fine (2001)은 subdistribution에 기초한 위험률 함수가 직관적인 해석이 어렵기 때문에, subdistribution에 바탕을 둔 비례 오즈 모형(proportional odds model)이 더 선호된다는 언급을 했다. Fine (2001)에 따라 비례 오즈 모형을 만들 수 있으나, 소프트웨어로 구현되어 있지 않아 쉽게 사용하기 어려운 점이 있다. Eriksson 등 (2015)는 비례 오즈 모형에 대한 새로운 추정방법을 제안했다. K개의 경쟁위험이 존재하는 상황에서 t 시점까지 원인 j $(j = 1, \ldots, K)$로 인한 사망하는 확률을 $F_j(t) = P(T^* \leq t, \epsilon = j)$이라고 하자. 이 장에서도 사건 1에 초점을 맞춰서 설명하겠다. 이 때 T^*은 사망하는 시간, ϵ은 사망의 원인이다. T^*의 정의는 2.2장과 같다. 즉, T^*은 관측되지 않을 수 있는 값이고, 실제 관측 값 $T = T^* \wedge C$이다. 따라서 $F_1(t)$는 subdistribution 위험 집합에 근거한 누적 확률이다. 식 (2.4)에서 $g(x) = \text{logit}(x) = \log\{x/(1-x)\}$로 두면, 다음과 같은 비례오즈 모형이 정의된다.

$$\text{logit}(F_1(t; Z)) = \log\left\{\frac{F_1(t; Z)}{1 - F_1(t; Z)}\right\} = \log(\Lambda_{10}(t)) + Z^{\mathrm{T}}\beta, \tag{2.11}$$

누적 기저위험률 $\Lambda_{10}(t)$는 임의의 단조 증가하는 양의 함수(positive monotone increasing function)이고, $\Lambda_{10}(0) = 0$이다. Z는 누적 발생률에 영향을 미칠 가능성이 있는 공변량이다. Fine과 Gray (1999) 모형은 subdistribution에 근거한 사건에 대한 위험률에 대한 모형인데, cause-specific에 근거하지 않았기 때문에 위험률을 이해하기가 다소 어렵지만, 반면에 이 모형은 subdistribution에 근거하여 사건 1이 일어날 확률과 그 여집합에 대한 확률의 비에 대한 함수로 설명할 수 있으므로, 이해하기가 더 수월하다. 이 모형으로부터 다음과 같은 누적발생률 함수(cumulative incidence function)가 생성된다.

$$F_1(t; Z) = \frac{\Lambda_{10}(t)\exp\left(Z^{\mathrm{T}}\beta\right)}{1 + \Lambda_{10}(t)\exp\left(Z^{\mathrm{T}}\beta\right)}.$$

원인 1로 인한 subdistribution 위험 함수는 다음과 같다.

$$\frac{\partial}{\partial t}\log(1 - F_1(t; Z)) = \frac{1}{e^{Z^{\mathrm{T}}\beta + \Lambda_{10}(t-)}}\lambda(t).$$

이 때, $\lambda(t)$는 $\Lambda_{10}(t)$의 미분값이다. 중도 절단되는 사건은 독립적으로 일어나고, $P(C > t) = G(t)$이다. i번째 개체의 관측치는 $T_i = T_i^* \wedge C_i, \Delta_i, \Delta_i\epsilon_i, Z_i$로 표시하고, C_i는 i번째 개체가 중도 절단된 시점, Δ_i는 i번째 개체의 중도 절단 여부, Z_i는 i번째 개체의 공변량이다. 셈과정(counting process) $N_i(t)$는 2.2장에서 정의된대로 $N_i(t) = I(T_i^* \leq t, \epsilon_i = 1)$이고, t 시점 이전에 개체 i가 원인 1로 인해 위험이 발생했는지의 여부를 뜻한다. 위험에 노출될 가능성에 대한 지표(modified at-risk indicator)를 $Y_i(t) = 1 - N_i(t-)$로 표시된다. t 시점 직전까지 원인 1로 인한 위험이 없었다면 t 시점에 위험에 처할 가능성이 있으므로 1이 되고, 즉 개체 i는 위험 집합에 포함된다. 중도 절단이 전혀 없는 데이터의 경우, 다음과 같은 compensated 셈 과정을 생각할 수 있다.

$$M_i(t) = N_i(t) - \int_0^t Y_i(u)\alpha(\Lambda_{10}(u-), \beta, Z_i)d\Lambda_{10}(u), \tag{2.12}$$

이 때, $\alpha(\Lambda_{10}, \beta, Z) = (e^{-Z^{\mathrm{T}}\beta} + \Lambda_{10})^{-1}$이고, $M_i(t)$는 마팅게일(martingale)이 된다. 위의 식은 Chen 등 (2002)과 유사한 방법으로 도출되었다. 중도 절단이 있다면, N_i와 Y_i는 알 수 없게 된다. 따라서 중도 절단이 있는 경우에는 식 (2.12)을 이용한 추정이 불가능해진다. Eriksson 등 (2015)에서는 중도 절단된 데이터를 처리하기 위해 Fine과 Grary (1999)가 도입한 IPCW 가중치를 이용하거나 Kaplan-Meier 방법을 이용한다. 2.2장에서 언급한 vital 상태 $r_i(t)$를 이용하면, 중도 절단된 데이터에서 $r_i(t)N_i$와 $r_i(t)Y_i$는 구할 수 있게 된다. Fine과 Grary (1999)에서 중도 절단된 데이터에 가중치

(2.5)를 부여한 것처럼 Eriksson 등 (2015)에서도 가중치 (2.5)를 식 (2.12)에 도입하였다. 마팅게일의 성질을 이용하면, β와 $\Lambda_{10}(\cdot)$에 대한 다음과 같은 추정 방정식(estimating equation)을 얻는다.

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau} Z_i w_i(t,G)\left\{dN_i(t) - Y_i(t)\alpha(\Lambda_{10}(t-),\beta,Z_i)d\Lambda_{10}(t)\right\} = 0,$$

$$n^{-1}\sum_{i=1}^{n}\int_0^{\tau} w_i(t,G)\left\{dN_i(t) - Y_i(t)\alpha(\Lambda_{10}(t-),\beta,Z_i)d\Lambda_{10}(t)\right\} = 0,$$

여기서 τ는 추적 관찰한 가장 마지막 시간이고, $\forall t \in [0,\tau]$이다.

3. 경쟁 위험 회귀 모형의 이용

3.1. 각 방법의 장점 및 단점

지금까지 몇 가지 경쟁위험 모형과 각각의 추정 방법에 대해서 간략하게 설명하였다. 2장에 소개된 경쟁 위험 모형은 도입된 시간 순서대로 설명되어 있다. 각각의 모형은 앞선 모형의 단점을 개선하고자 하는 노력에 의해 도입되었다. 어떤한 모형이 절대적으로 우위에 있다기보다는 모든 모형이 장단점이 있기 때문에, 목적에 따라 또는 데이터에 따라 적합한 모형을 선택해서 쓸 수 있다면 좋을 것이다. Cause-specific 위험과 subdistribution 위험은 Cox의 비례 위험 모형에 근거한 모형으로 위험 집합을 어떻게 정하는가에 달려있다. Lau 등 (2009)에 따르면, 이미 일어난 사건의 원인을 찾는데 관심이 있는 병인학(etiology)에서는 경쟁 원인으로 인한 위험을 중도 절단한 cause-specific 위험 모형이 적합하고, 예측 측면에서는 경쟁 위험으로 인한 위험을 겪은 개체 또한 위험 집합에 포함시키는 subdistribution 위험 모형이 더 설명력이 있다.

이항 회귀 모형의 장점은 변수의 시간에 따른 영향력이 다르게 조절할 수 있는 모형이라는 점이다. 특정 변수가 누적 확률에 시간에 따라 다른 영향으로 준다는 사전 지식이 있거나 가능성이 있다면 이항 회귀 모형을 쓰는 것이 좋을 것이다. 이항 회귀 모형은 연결 함수의 선택에 따라 Fine과 Gray 모형과 비례 오즈 모형 등 다양한 모형을 만들 수 있는데, 이렇게 모형이 유연하기 때문에 cause-specific 위험 모형이나 subdistribution 위험 모형보다 활용 범위가 넓다는 점에서는 장점이지만, 연결 함수를 잘 선택하지 못했을 때 설명력이나 예측력의 문제가 생길 수 있는 단점도 있다. 절대 위험 모형은 회귀 계수가 설명력을 갖도록 만들어졌다. 공변량이 한 단위 변화할 때, 변화하는 확률을 알고자 하는 것이 주된 관심사라면 절대 위험 모형이 유용하게 쓰일 수 있다. 해석 측면에서는 절대 위험 모형이 비례 오즈 모형보다 더 이해하기가 쉬울 수 있다. 그러나 절대 위험 모형의 누적 발생률은 1을 초과할 가능성도 있다는 단점이 있다. Eriksson 등 (2015)의 비례 오즈 모형은 subdistribution 위험 집합을 이용한다는 점에서는 Fine과 Gray (1999) 모형과 공통점이 있지만, 비례 오즈 모형이 비례 위험 모형보다 좀 더 직관적으로 이해가 된다는 점에서 발전된 모형이다. 비례 오즈 모형의 해석 측면에서의 장점 때문에 Eriksson 등 (2015) 외에도 많은 학자들이 비례 오즈 모형에 대한 연구를 하였지만 Fine과 Gray (1999) 방법만큼 효율적(efficient)이지 않으나, Eriksson 등 (2015) 의 방법은 Fine과 Gray (1999) 방법만큼 효율적인 추정치를 제공한다.

3.2. 모형의 비교

각각의 모형이 어떠한 관계에 있는지 설명하고자 한다. 첫 번째로 cause-specific 위험 모형과 Fine과 Gray 모형으로 어떠한 관계에 있는지, 두 번째로 Fine과 Gray 모형과 이항 회귀 모형, 세 번째로 이항

회귀 모형과 비례 오즈 모형의 비교에 대한 언급을 하고자 한다. 절대 위험 모형은 다른 네 가지 모형으로는 변환할 수 없으므로, 절대 위험 모형은 다루지 않도록 한다.

3.2.1. Cause-specific 위험 모형과 Fine과 Gray 모형 Gail (2005)에는 상대 위험(relative hazard)을 주어진 시간에서 두 개의 위험률의 비율이라고 정의한다. Lau 등 (2009)에서는 cause-specific 위험 모형과 subdistribution 위험 모형을 상대 위험 관점에서 비교하였다. 공변량 $Z = 1$을 노출, $Z = 0$을 노출되지 않은 경우라고 하자. Cause-specific 위험 모형의 식 (2.1)에서 원인 j에 대해서 $\exp(\beta_j)$는 Z가 한 단위 증가할 때(노출되었을 때) 변화하는 위험의 정도로 해석할 수 있다. 식 (2.1)의 β_j를 $\beta_{\text{CS}j}$라고 하자. cause-specific 위험 모형에서 변수 j에 해당하는 상대 위험 $\text{CSRH}_j = \exp(\beta_{\text{CS}j})$가 된다. 이와 유사하게, Fine과 Gray 모형에서는 식 (2.3)에서 원인 j에 대해서 $\exp(\beta_j)$는 Z가 한 단위 증가할 때(노출되었을 때) 변화하는 위험의 정도로 설명된다. 식 (2.3)에서 β_j를 $\beta_{\text{SD}j}$라고 하면, subdistribution 위험 모형에서 변수 j에 해당하는 상대 위험 $\text{SDRH}_j = \exp(\beta_{\text{SD}j})$가 된다. 경쟁 위험이 두 가지인 경우($j = 1, 2$)를 생각해보자. Lau 등 (2009)에서는 Beyersmann 등 (2007)에 따라 다음과 같은 관계식을 유도하였다. 사건 1에 초점을 맞추어서 설명하고자 한다.

$$\text{CSRH}_{j=1}(t) = \frac{\left\{1 + \dfrac{P(T \le t, j = 2, Z = 1)}{P(T > t, Z = 1)}\right\}}{\left\{1 + \dfrac{P(T \le t, j = 2, Z = 0)}{P(T > t, Z = 0)}\right\}} \text{SDRH}_{j=1}(t).$$

cause-specific 위험 집합과 subdistribution 위험 집합이 다르므로 일반적으로 CSRH_j와 SDRH_j는 다른 값을 갖는다. 노출이 각각의 사건과 독립적으로 일어난다고 가정한다.

- $\text{CSRH}_1 < 1$, $\text{CSRH}_2 < 1$인 경우
  노출되었을 때, 사건 1에 대한 cause-specific 위험률과 사건 2에 대한 cause-specific 위험률이 동시에 감소하는 경우이다. 사건 2에 대한 cause-specific 위험률이 감소하는 것은 사건 1의 subdistribution 위험 집합의 크기가 감소하는 것을 의미하므로, $\text{SDRH}_1 > \text{CSRH}_1$이 된다.
- $\text{CSRH}_1 < 1$, $\text{CSRH}_2 > 1$인 경우
  노출되었을 때, 사건 1에 대한 cause-specific 위험률은 감소하고 사건 2에 대한 cause-specific 위험률은 증가하는 경우이다. 사건 2에 대한 cause-specific 위험률이 증가하는 것은 사건 1의 subdistribution 위험 집합의 크기가 증가하는 것을 의미하므로, $\text{SDRH}_1 < \text{CSRH}_1$이 된다.
- $\text{CSRH}_1 > 1$, $\text{CSRH}_2 < 1$인 경우
  노출되었을 때, 사건 1에 대한 cause-specific 위험률은 증가하고, 사건 2에 대한 cause-specific 위험률은 감소하는 경우이다. 사건 2에 대한 cause-specific 위험률이 감소하는 것은 사건 1의 subdistribution 위험 집합의 크기가 감소하는 것을 의미하므로, $\text{SDRH}_1 > \text{CSRH}_1$이 된다.
- $\text{CSRH}_1 > 1$, $\text{CSRH}_2 > 1$인 경우
  노출되었을 때, 사건 1에 대한 cause-specific 위험률과 사건 2에 대한 cause-specific 위험률이 동시에 증가하는 경우이다. 사건 2에 대한 cause-specific 위험률이 증가하는 것은 subdistribution 위험 집합의 크기가 증가하는 것을 의미하므로, $\text{SDRH}_1 < \text{CSRH}_1$이 된다.

3.2.2. Fine과 Gray 모형과 이항 회귀 모형 식 (2.7) 또는 식 (2.8)를 변형하여 Fine과 Gray 모형으로 만들 수 있다. Fine과 Gray 모형에 해당하는 이항 회귀 모형은 다음과 같다.

$$\log[-\log\{1 - F_1(t; Z)\}] = \log\{\Lambda_{10}(t)\} + Z^{\mathrm{T}}\beta.$$

Table 3.1. Software commands for competing risk regressions

Models	SAS	R
Cause-specific	PROC PHREG	package:riskRegression (CSC)
Fine-Gray	PROC PHREG	package:cmprsk (crr) package:riskRegression (FGR)
Direct binomial		package:timereg (comp.risk)
Absolute risk		package:riskRegression (ARR)
Proportional odds		package:timereg (prop.odds.subdist)

Fine과 Gray 방법은 편우도 함수를 통해 추정 방정식을 구한 후 근을 구하고, 이항 회귀 모형은 준모수 모형의 효율적 추정 방법으로 접근한다는 점에서 차이가 있다. 두 방법 모두 중도 절단에 대한 분포를 추정해야한다는 단점이 있다. Scheike 등 (2008)에서는 시뮬레이션을 통해서 대부분의 경우, Fine과 Gray 방법과 비교해서 회귀 계수에 대한 편차와 표준 오차가 큰 차이가 나지 않음을 보였다. 그러나 Fine과 Gray 방법을 이용하면 예측치의 분산이 시간에 따라서 증가하는 경우가 있는데, 이항 회귀 모형의 경우에는 시간에 따라 일정한 분산의 패턴을 보이는 점에서 이항 회귀 모형의 장점을 설명하였다.

3.2.3. 이항 회귀 모형과 비례 오즈 모형 이항 회귀 모형 (2.6)은 비례 오즈 모형 (2.11)과 같은 형태로 만들 수 있다.

$$\text{logit}(F_1(t;Z)) = \log(\Lambda_{10}(t)) + Z^{\mathrm{T}}\beta.$$

같은 모형에 대한 추정 방정식이 다르다. 이항 회귀 모형은 준모수 모형의 효율적 추정 방법을 이용했으나, Eriksson 등 (2015)의 비례 오즈 모형에서는 Chen 등 (2002)에서와 같이 $M_i(t)$을 마팅게일 과정으로 고려하여 추정 방정식에 도입했다. Eriksson 등 (2015)에서 Eriksson 등 (2015)의 방법과 Scheike 등 (2008)의 이항 회귀 모형을 시뮬레이션을 통해 비교한 결과, Eriksson 등 (2015)의 방법이 Scheike 등 (2008)의 방법에 비해 추정치 편의(bias)와 분산이 모두 작음을 보였다.

3.3. SAS와 R 활용

Cause-specific 위험 모형과 Fine과 Gray의 subdistribution 위험 모형은 SAS와 R로 구현되어 있고, 비교적 최근에 소개된 모형인 이항 회귀 모형과, 절대 위험 모형, Eriksson 등 (2015)의 비례 오즈 모형은 R에서만 구현되어 있다. Table 3.1에 각 소프트웨어에 해당하는 명령어를 간단하게 기재하였다. 옵션 지정하는 방식에 대해서는 자세히 언급하지 않았다.

두 가지 경쟁 위험이 있고 두 개의 공변량를 모형에 포함시키는 경우를 생각해보자. 변수 time, type, $Z1$, $Z2$가 포함된 데이터를 고려한다. time은 사건이 일어난 시간이다. type을 다음 세 가지로 고려한다. type = 0은 중도 절단된 데이터, type = 1은 경쟁 위험 1, type = 2는 경쟁 위험 2이다. $Z1$, $Z2$는 공변량에 해당한다.

3.3.1. Cause-specific 모형 다음은 SAS와 R로 사건 1에 대한 회귀 모형을 추정한 것이다.

- SAS

```
PROC PHREG data=data_name;
model time*type(0,2) = Z1 Z2;
run;
```

Cause-specific 모형에서는 관심 있는 사건이 아닌 경우, 중도 절단된 것으로 간주한다. 위의 코드에서 $(0, 2)$ 부분은 type $= 0$, 2는 중도 절단되었다는 것을 표시한 것이다. 사건 2의 회귀 모형에 대해서는 이 부분만 $(0, 1)$로 바꾸면 된다.

- R (package:riskRegression)

```
library(riskRegression)
cs1<-CSC(Hist(time,type)~Z1+Z2,data=data_name, cause = 1)
print(cs1)
```

관심있는 사건은 cause $= 1$로 지정한다.

3.3.2. Fine과 Gray 모형

- SAS

```
PROC PHREG data=data_name;
model time*type(0) = Z1 Z2 /eventcode = 1;
run;
```

Fine과 Gray 모형에서는 관심있는 사건을 eventcode $= 1$로 지정해준다. 사건 2에 대해서는 이 부분만 eventcode $= 2$로 바꾸면 된다.

- R (pakage:cmprsk)

```
library(cmprsk)
attach(data_name)
fg1<-crr(time,type,cbind(Z1,Z2),faildcode=1,cencode=0)
summary(fg1)
detach()
```

관심있는 사건은 faildcode $= 1$로 지정하고, 중도 절단된 데이터는 type $= 0$이므로 cencode $= 0$으로 지정한다.

- R (package:riskRegression)

```
library(riskRegression)
fg2<-FGR(Hist(time,type)~Z1+Z2,data=data_name, cause = 1)
summary(fg2)
```

관심있는 사건은 cause $= 1$로 지정한다.

3.3.3. 이항 회귀 모형

- R (package:timereg)

```
library(timereg)
add1<-comp.risk(Event(time,type)~Z1+Z2,data=data_name,cause=1,
                resample.iid=1,n.sim=100,model="additive")
summary(add1)
```

```
add2<-comp.risk(Event(time,type)~const(Z1)+const(Z1),data=data_name,cause=1,
                resample.iid=1,n.sim=100,model="additive")
summary(add2)
```

add1은 시간에 따른 공변량의 효과를 고려한 회귀식이고, add2는 공변량의 효과가 시간에 따라 일정한 회귀식이다. const(·)로 지정함으로써 시간의 따른 효과를 일정하게 하도록 지정할 수 있다. 식 (2.7)과 식 (2.8)은 다음과 같이 좀 더 일반적인 식으로 전환이 가능하다.

$$P(T < t, \text{cause}=1|x, z) = P_1(t, x, z) = h(g(t, x, z)).$$

위의 이항 회귀 모형에서 에서 연결함수 h와 g에 대한 지정을 해야하는데, 위의 코드에서 model = "additive"이 그 역할을 한다. 이 경우 $h = 1 - \exp(-x)$이고, $g(t, x, z) = x^{\mathrm{T}}\alpha(t) + (z^{\mathrm{T}}\beta)t$이다. 이 때, $\alpha(t)$는 데이터에 근거한 시뮬레이션을 통해 비모수 방법으로 만들어지는 함수로써 데이터 분석 전에는 알 수 없는 함수이다. resample.iid = 1, n.sim = 100은 시뮬레이션과 관련된 옵션이다. 다른 연결 함수 지정에 대해서는 Scheike와 Zhang (2011)을 참조하기 바란다.

3.3.4. 절대 위험 회귀 모형

- R (package:riskRegression)

```
library(riskRegression)
arr<-ARR(Hist(time,type)~Z1+Z2,data=data_name,cause=1)
summary(arr)
```

원인이 1인 사건을 기준으로 분석을 할 경우, cause = 1로 지정한다.

3.3.5. 비례 오즈 모형

- R (package:timereg)

```
library(timereg)
po <- prop.odds.subdist(Event(time,type)~Z1+Z2,data=data_name,cause=1,
                        cens.model="KM",detail=0,n.sim=1000)
summary(po)
```

원인이 1인 사건을 기준으로 분석을 할 경우, cause = 1로 지정한다. Eriksson 등 (2015)의 비례 오즈 모형은 중도 절단된 데이터의 분포를 지정해야 한다. cens.model = "KM"이 이에 대한 옵션이다. 여기에서는 Kaplan-Mier 방법을 이용했다. detail은 시뮬레이션의 결과 출력에 대한 옵션으로 detail = 0은 iteration 결과를 출력하지 않고 detail = 1로 지정하면 iteration 결과가 출력된다. n.sim = 1000은 시뮬레이션 최대 횟수를 지정하는 옵션이다.

4. 자료 분석

이 장에서는 Lau 등 (2009)에 소개된 Women's Interagency HIV Study(WIHS) 데이터를 다루었다. WIHS는 1993년에 미국 여성의 HIV 바이러스(human immunodeficiency virus) 감염를 조사하기 위해 시작되었다. 뉴욕의 두 개 지역, 워싱턴 DC, 로스 앤젤레스, 샌프란시스코, 시카고 이렇게 여섯 지역에

Table 4.1. Competing risk regression results for five different models

		원인 1 (HAAR)			원인 2 (AIDS)		
		Estimator	s.e.	p-value	Estimator	s.e.	p-value
(1) Cause-specific 위험 모형	IDU	−0.395	0.088	7.2e−06	0.537	0.113	1.9e−06
	CD4	−0.230	0.019	2.0e−16	−0.199	0.026	2.3e−14
	age	0.010	0.005	0.036	0.013	0.007	0.064
(2) Fine과 Gray 모형	IDU	−0.522	0.089	5.3e−09	0.711	0.112	1.9e−10
	CD4	−0.072	0.015	1.6e−06	−0.084	0.029	4.3e−03
	age	0.004	0.005	0.376	0.008	0.006	0.246
(3) 이항 회귀 모형 (model = "additive")	IDU	−0.061	0.011	1.1e−07	0.054	0.010	2.1e−07
	CD4	−0.011	0.001	8.3e−15	−0.004	0.002	0.011
	age	0.001	0.001	0.122	0.001	5.3e−04	0.207
(4) 절대 위험 회귀 모형	IDU	−0.370	0.068	6.5e−08	0.505	0.103	1.0e−06
	CD4	−0.069	0.013	7.2e−08	−0.238	0.030	5.2e−15
	age	0.006	0.003	0.054	2.8e−04	6.4e−04	0.965
(5) 비례 오즈 모형	IDU	−0.697	0.123	1.8e−08	0.835	0.135	1.5e−09
	CD4	−0.154	0.021	2.4e−11	−0.104	0.034	0.001
	age	0.008	0.006	0.155	0.009	0.008	0.259

서 표본을 수집한다. 이 데이터는 WIHS에 등록된 HIV 양성 판정을 받았지만 AIDS가 발병하지 않은 1995년 6월에 생존했었던 1,164명의 여성에 관한 것으로 2006년 9월까지 추적한 데이터이다. 사망의 원인을 두 가지로 나눌 수 있는데, 원인 1은 강력한 항레트로바이러스 치료 highly active antiretroviral therapy(HAART)이고, 원인 2는 AIDS 발병이다. 그 밖에 중도 절단된 경우가 있다. 공변량으로는 마약 주사 사용(injection of drug use) 여부 IDU, 1995년 6월 시점에서 각 여성의 나이 ageatfda, 세포표면항원무리 4의 개수(CD4 count nadir) CD4가 있다.

Table 4.1에서 이 논문에서 설명한 다섯 가지 방법의 분석 결과를 확인할 수 있다. (1) Cause-specific 위험 모형은 식 (2.1), (2) Fine과 Gray 모형은 식 (2.4), (3) 이항 회귀 모형은 식 (2.6), (4) 절대 위험 회귀 모형은 식 (2.9), (5) 비례 오즈 모형은 식 (2.11)에 따라 추정하였다. 이항 회귀 모형에서는 공변량의 효과가 시간에 따라 일정한 모형을 고려했다. 모든 분석은 R을 이용하였으며 두 가지 패키지가 이용 가능한 Fine과 Gray 모형에는 R 패키지 cmprsk을 이용하였다. 이항 회귀 모형은 R 패키지 timereg에서 model = "additive"로 설정하였다. 다섯 가지 방법의 모형이 모두 다르기 때문에 각각의 모형에 맞는 해석을 하는 것이 중요하다. 원인 1 HAAR에 대한 회귀 모형을 생각해보자. Cause-specific 위험 모형에서는 IDU, CD4, 나이 모두 사망에 미치는 영향이 유의하다고 해석할 수 있고, 다른 네 가지 모형에서는 IDU, CD4만 사망에 미치는 영향이 유의한 것으로 보인다. 절대 위험 모형의 경우, 원인 1 HAAR의 회귀 분석 결과 $\exp(\widehat{\beta}_{\mathrm{IDU}}) = 0.69$, $\exp(\widehat{\beta}_{\mathrm{CD4}}) = 0.93$, $\exp(\widehat{\beta}_{\mathrm{age}}) = 1.01$이다. 마약 주사 경험이 있다면, 원인 1로 인한 누적 사망률이 31% 감소, CD4 한 단위가 증가할 때 누적 사망률 7% 감소한다. 나이는 원인 1의 누적 사망률에 큰 영향을 미치지 못한다. 비례 오즈 모형 (2.11)은 이 데이터에 적합하게 다음과 같은 식으로 쓸 수 있다.

$$\mathrm{Odds}\{F_1(t;Z)\} = \frac{F_1(t;Z)}{1-F_1(t;Z)} = \Lambda_{10}(t)e^{\beta_1 \mathrm{IDU}} e^{\beta_2 \mathrm{CD4}} e^{\beta_3 \mathrm{age}}.$$

원인 1 HAAR의 회귀 분석 결과 $\exp(\widehat{\beta}_{\mathrm{IDU}}) = 0.50$, $\exp(\widehat{\beta}_{\mathrm{CD4}}) = 0.86$, $\exp(\widehat{\beta}_{age}) = 1.01$이다. 원인 1 HAAR로 인한 사망률의 오즈는 HAAR로 인한 누적 사망률과 누적 생존률의 비율이다. 마약 주사 경험이 있다면 오즈, 즉 원인 1로 인한 사망률과 생존률의 비율이 50% 감소, CD4 한 단위가 증가

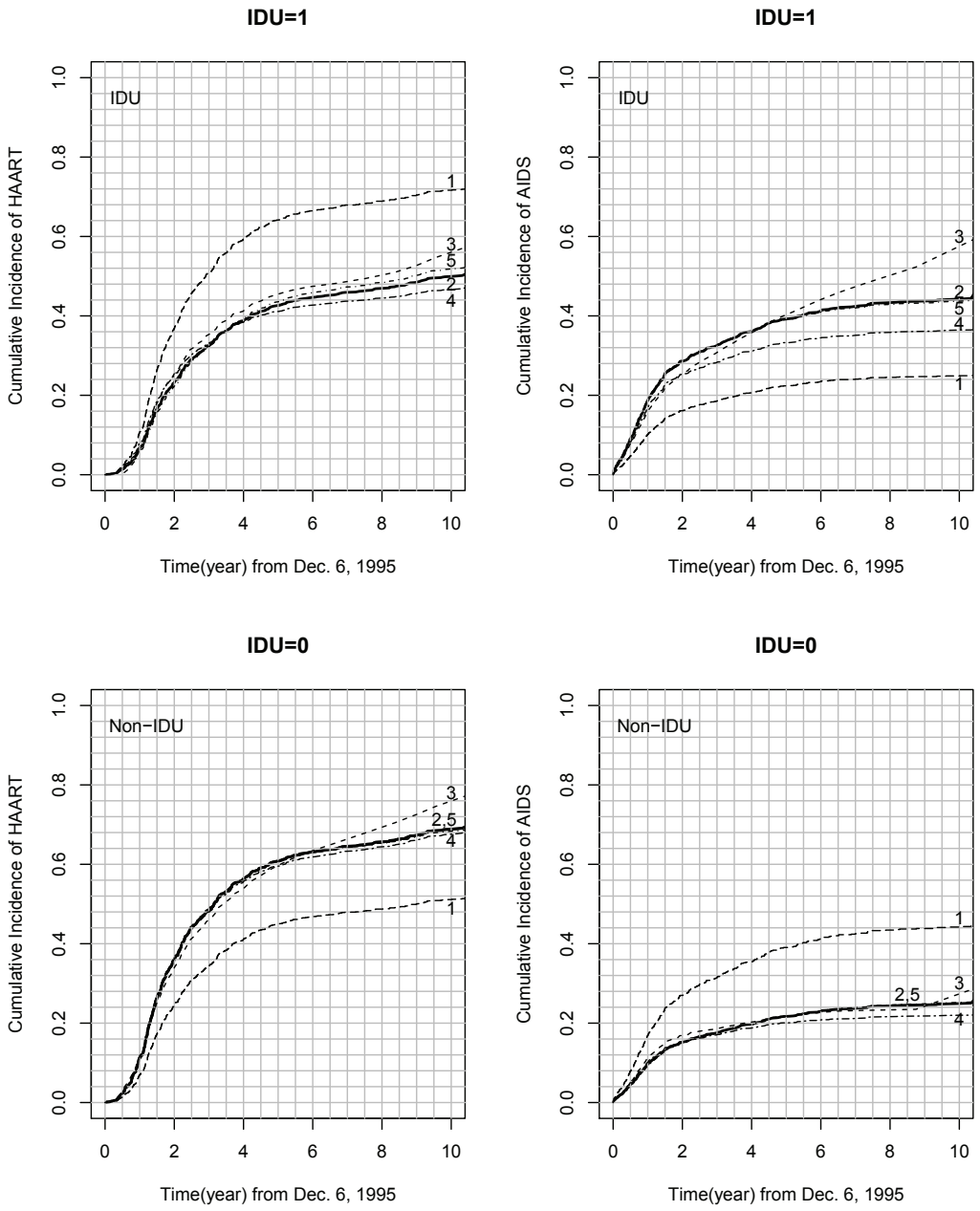

Figure 4.1. Predicted cumulative incidence of the person who is 33 year-old with 349 CD4 nadir count. Left figure is corresponding to the cumulative incidence death from HAAR(highly active antiretroviral therapy) before occurrence of AIDS. Right figure is corresponding to the cumulative incidence from AIDS before having HAAR. The thick lines represent Fine and Gray model. 1: cause-specific hazard model 2: Fine and Gray model, 3: binomial additive model, 4:absolute risk regression model, 5: proportional odds model.

할 때 오즈가 14% 감소한다. 나이는 원인 1의 오즈에 큰 영향을 미치지 못한다. 원인 2 AIDS의 경우를 살펴보자. 모든 모형에서 IDU와 CD4가 유의한 변수이다. 나이는 유의한 변수가 아닌 것으로 드러났다. 절대 위험 모형에서 원인 2 AIDS의 회귀 분석 결과 $\exp(\widehat{\beta}_{\text{IDU}}) = 1.66$, $\exp(\widehat{\beta}_{\text{CD4}}) = 0.79$, $\exp(\widehat{\beta}_{\text{age}}) = 1.00$이다. 마약 경험이 있다면 그렇지 않은 사람에 비해 원인 2 ADIS로 인한 누적 사망률이 66% 증가, CD4가 한 단위 증가할수록 누적 사망률 21% 감소하고 나이는 원인 2의 누적 사망률에 유의한 영향을 미치지 못한다. 비례 오즈 모형에서는 $\exp(\widehat{\beta}_{\text{IDU}}) = 2.30$, $\exp(\widehat{\beta}_{\text{CD4}}) = 0.90$, $\exp(\widehat{\beta}_{\text{age}}) = 1.01$이다. 마약 주사 경험이 있다는 그렇지 않을 때보다 오즈 비 즉, 원인 2로 인한 오즈, 누적 사망률과 누적 생존률률의 비율이 130% 증가, CD4 한 단위 증가시 오즈가 10% 감소하고 나이는 오즈의 변화에 영향을 미치지 않는다.

이 데이터에서 여성들의 나이의 중간값은 33세이고, CD4 nadir의 중간값은 349이다. Figure 4.1은 CD4 nadir가 349이고, 33세인 여성에 대해 마약 주사 경험이 있는 경우(IDU)와 그렇지 않은 경우(Non-IDU)로 나누어서 각각의 방법에 해당하는 누적 발생률 함수의 예측치를 표시한 그림이다. 그림에서의 번호는 모형에 해당하는 번호이다. 예측면에서 효율적인 모형으로 알려진 (2) Fine과 Gray 모형을 굵은 실선으로 표시하고, (1) cause-specific 모형에는 R 그래픽 옵션 lty = 5, (3) 이항 회귀 모형에는 lty = 2, (4) 절대 위험 모형에는 lty = 6, (5) 비례 오즈 모형에는 lty = 4로 표시하였다. (2) Fine과 Gray 모형과 (5) 비례 오즈 모형이 거의 같은 패턴을 보이는 것을 확인할 수 있다. (3) additive 이항 모형은 10년 후에 다른 모형보다 비교적 높은 사망률을 보이는 것이 특이한 점이다. 절대 위험 모형은 10년 후에 다른 모형보다 낮은 사망률을 보이는 패턴을 보이고 있다. 비례 오즈 모형은 subdistribution에 바탕을 두지만, Fine과 Gray 모형과 거의 같은 패턴을 보이면서 회귀 계수에 대한 해석이 좀 더 수월하다는 점에서 Fine과 Gray 모형을 대체할 수 있는 모형으로 생각된다.

5. 결론

Cause-specific 모형과 Fine과 Gray (1999) 모형은 경쟁 위험이 존재하는 생존 자료에 주로 이용되는 모형이다. 이후에 많은 모형이 제시되었지만, 알고리즘으로 구현하기 어렵거나 추정방법이 신뢰할만 하지 않다는 이유로 쓰이지 못한 모형이 많다. Fine과 Gray 모형의 단점은 subdistribution 위험에 근거한 위험률 모형의 회귀 계수의 해석이 어렵다는 점이다. 회귀 계수의 해석이 용이한 모형으로 Gerds 등 (2012)에서 절대 위험 모형을 제시했지만 누적 발생률 확률이 1을 초과할 수 있는 단점이 있다. Scheike 등 (2008)에 유연한 모형을 만들 수 있는 이항 회귀 모형이 제시되었고, 이항 모형으로 Fine과 Gray 모형, 비례 오즈 모형을 만들 수 있다는 장점이 있다. Eriksson 등 (2015)의 비례 오즈 모형이 이항 모형의 비례 오즈 모형보다 일치성과 효율성 측면에서 더 낫다는 것을 보였다. 이 모형 역시 subdistribution 위험에 근거한 모형이지만, 위험률이 아닌 오즈를 모형화한다는 점에서 해석하기는 조금 더 수월해진다. 데이터 분석을 통해 Fine과 Gray 모형이 Eriksson 등 (2015)의 비례오즈 모형과 누적 발생률이 아주 비슷한 패턴으로 나타난다는 것을 볼 수 있으며, 비슷한 예측치를 제공한다면 회귀 계수의 해석이 가능한 비례 오즈 모형을 쓰는 것이 해석 측면에서 유리할 수 있다. 각 모형이 장단점을 가지고 있으므로, 분석 목적에 맞게 잘 이용하는 것이 중요하다.

References

Beyersmann, J., Dettenkofer, M., Bertz, H., and Schumacher, M. (2007). A competing risks analysis of bloodstream infection after stem-cell transplantation using subdistribution hazards and cause-specific hazards, *Statistics in Medicine*, **26**, 5360–5369.

Chen, K., Jin, Z., and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data, *Biometrika*, **89**, 659–668.

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

Eriksson, F., Li, J., Scheike, T., and Zhang, M. J. (2015). The proportional odds cumulative incidence model for competing risks, *Biometrics*, **71**, 687–695.

Fine, J. P. (2001). Regression modeling of competing crude failure probabilities, *Biostatistics*, **2**, 85–97.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk, *Journal of the American Statistical Association*, **94**, 496–509.

Gail, M. H. (2005). *Relative Hazard*, Encyclopedia of Biostatistics, 7.

Gerds, T. A., Scheike, T. H., and Andersen, P. K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction, *Statistics in Medicine*, **31**, 3921–3930.

Gray, R. J. (1988). A class of K-sample tests for comparing the cumulative incidence of a competing risk, *The Annals of Statistics*, **16**, 1141–1154.

Holt, J. D. (1978). Competing risk analyses with special reference to matched pair experiments, *Biometrika*, **65**, 159–165.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

Lau, B., Cole, S. R., and Gange, S. J. (2009). Competing risk regression models for epidemiologic data, *American Journal of Epidemiology*, kwp107.

Lin, D. Y. (1997). Non-parametric inference for cumulative incidence functions in competing risks studies, *Statistics in Medicine*, **16**, 901–910.

Prentice, R. L., Kalbfleisch, J. D., Peterson Jr, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks, *Biometrics*, 541–554.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association*, **90**, 122–129.

Scheike, T. H. and Zhang, M. J. (2011). Analyzing competing risk data using the R timereg package, *Journal of Statistical Software*, **38**.

Scheike, T. H., Zhang, M. J., and Gerds, T. A. (2008). Predicting cumulative incidence probability by direct binomial regression, *Biometrika*, **95**, 205–220.

경쟁 위험 회귀 모형의 이해와 추정 방법

김미정[a,1]

[a]이화여자대학교 통계학과

(2016년 7월 18일 접수, 2016년 9월 3일 수정, 2016년 10월 7일 채택)

요 약

경쟁위험에 대한 연구 중 주로 쓰이는 방법은 Cause-specific 위험 모형과 subdistribution을 이용한 비례 위험 모형 방법이다. 그 이후에도 많은 모형이 제시되었지만, 추정 방법 면에서 설명력이 부족하거나 알고리즘으로 구현하기 어려운 단점을 가지고 있어서 잘 활용되고 있지 않다. 이 논문에서는 Cause-specific 위험 모형, subdistribution을 이용한 비례 위험 모형과 비교적 최근에 제시된 이항 회귀 모형(direct binomial model), 절대 위험 회귀 모형(absolute risk regression model), Eriksson 등 (2015)의 비례 오즈 모형(proportional odds model)을 소개하고 추정 방법을 간단히 설명하고자 한다. 각 모형에 대하여 SAS와 R을 이용한 활용 방법을 제시하고, 두 가지 경쟁위험이 존재하는 데이터를 R을 이용하여 분석하였다.

주요용어: 비례 위험 모형, 비례 오즈 모형, 경쟁 위험, 누적 발생률 함수

[1](03760) 서울특별시 서대문구 이화여대길 52, 이화여자대학교 통계학과. E-mail: m.kim@ewha.ac.kr