신뢰성응용연구
제18권, 제2호. pp.130-142, 2018

경쟁적 위험하에서의 회귀분석*

백 재 욱†

한국방송통신대학교 정보통계학과

Competing Risks Regression Analysis*

Jaiwook Baik†

Department of Information Statistics, Korea National Open University

Purpose: The purpose of this study is to introduce regression method in the presence of competing risks and to show how you can use the method with hypothetical data.
Methods: Survival analysis has been widely used in biostatistics division. But the same method has not been utilized in reliability division. Especially competing risks, where more than a couple of causes of failure occur and the occurrence of one event precludes the occurrence of the other events, are scattered in reliability field. But they are not utilized in the area of reliability or they are analysed in the wrong way. Specifically Kaplan-Meier method is used to calculate the probability of failure in the presence of competing risks, thereby overestimating the real probability of failure. Hence, cumulative incidence function is introduced. In addition, sample competing risks data are analysed using cumulative incidence function along with some graphs. Lastly we compare cumulative incidence functions with regression type analysis briefly.
Results: We used cumulative incidence function to calculate the survival probability or failure probability in the presence of competing risks. We also drew some useful graphs depicting the failure trend over the lifetime.
Conclusion: This research shows that Kaplan-Meier method is not appropriate for the evaluation of survival or failure over the course of lifetime in the presence of competing risks. Cumulative incidence function is shown to be useful in stead. Some graphs using the cumulative incidence functions are also shown to be informative.

Keywords: Competing Risks, Cumulative Incidence Function, Competing Risks Regression, R Statistical Software

1. 서 론

생존 데이터는 통상적으로 어떤 사건이 발생할 때까지 걸리는 시간들로 이루어진다. 이때 어떤 사건이 발생할 때까지 시간은 여러 형태의 함수로 나타낼 수 있다 이들 중 가장 폭넓게 쓰이는 것이 생존함수 또는 생존확률 $S(t)$로서, 이는 개체의 수명이 적어도t시간이 될 확률 즉, $P(T>t)$을 나타낸다. 생존분석에 많이 쓰이는 또 다른 함수로 위험률(hazard rate) $h(t)$는 시간t에서의 사건 발생률(rate of occurrence)을 나타낸다[1-2].

* 이 논문은 2016년 한국방송통신대학교 학술연구비지원을 받아 작성된 것임
† 교신저자 jbaik@knou.ac.kr
2018년 4월 27일 접수; 2018년 5월 31일 수정본 접수; 2018년 6월 1일 게재 확정.

신뢰성분석은 생존분석과 똑같은 의미를 나타내지만 신뢰성분석이 공학 분야에서 제품의 수명에 대해 다루는 것이라면 생존분석은 의학 분야에서 사람의 수명에 대해 다루는 것이라고 할 수 있다. 따라서 생존분석에서 생존함수 $S(t)$와 위험률 $h(t)$은 신뢰성분석에서 신뢰도$R(t)$와 고장률(failure rate) $\lambda(t)$와 같은 개념이다.

신뢰성분석에서 어떤 개체에 대한 수명이 관측되지 않고 중도중단(censoring)되는 경우가 많은데, 이런 중도중단이 어떤 개체에서든 상관없이 독립적으로 일어난다면 앞의 생존함수(신뢰도)와 위험률(고장률)은 각각 Kaplan-Meier 추정치[3]와 Nelson-Aslen 추정치로 추정할 수 있다.

신뢰성분석에서는 개체에 대한 수명시험 결과 연구기간이내에 모든 개체가 사망 또는 고장에 이를 수도 있지만 연구기간 내에 개체가 모두 고장에 이르지 못할 수 있다. 그 이유는 일정한 시간동안만 연구가 진행되어 그 기간 안에 모든 개체가 고장에 이르지 못하기 때문일 수 있고, 실험대상인 개체에 피치 못 할 사정이 생겨 연구기간 중간에 관측을 중단해야 하는 상황에 있을 수 있다. 이런 사유로 생길 수 있는 데이터를 중도중단 데이터(censored data)라고 한다.

의학 분야에서 어떤 사건이 발생할 때에는 그 원인이 하나일 수 있으나 때로는 여러 가지일 수 있으며 이들 여러 원인들 중 어떤 하나의 원인이 문제가 되어 실험의 대상인 개체가 사망이라는 사건에 이르게 된다. 마찬가지로 공학 분야에서도 어떤 아이템이 고장날 때 그 원인이 균열(crack)일 수 있으나 때로는 균열 이외에 파괴(fracture), 마모(wear) 등일 수 있고, 이런 여러 원인들 중 어떤 특정한 원인 하나로 인하여 제품이 고장에 이를 수 있다. 이와 같이 여러 고장원인이 있지만 그들 중 어느 특정한 하나가 원인이 되어 개체가 고장에 이르게 되는 것을 경쟁적 위험(competing risks)이라고 한다.

어떤 개체가 경쟁적 위험하에 있는 경우 하나의 특정 고장원인에 의한 생존확률을 보기 위해 나머지 고장원인들을 무시하고 우측 중도중단(right censored)으로 보고, Kaplan-Meier 방법을 적용하여 생존확률을 구하고, 1에서 생존확률을 뺀 값으로 사망확률(사망확률은 '1-생존확률'로 일정한 시간까지 개체가 사망 또는 고장 날 확률을 말함)을 구하기도 한다[4]. 하지만 여러 고장원인에 의한 고장시간이 서로 독립적이지 않고 고장시간과 중도중단시간 간 서로 독립적이지 않을 때에는 앞에서와 같은 방식으로 생존확률 또는 사망확률을 구하면 그 추정치에 편의(bias)가 생기게 된다[5]. 따라서 경쟁위험하에 있는 경우에는 신뢰도나 사망확률을 추정하기 위한 적절한 방법이 필요하다[6-7]. 이에 본 연구의 제2장에서 경쟁적 위험하에서 사망확률을 추정하는데 Kaplan-Meier 방법 보다는 누적사건함수(Cumulative Incidence Function, *CIF*)를 구하는 방법이 더 적절하다는 것을 설명한다. 이어 제3장에서는 경쟁적 위험하에서 어떤 결과(예를 들어 고장)에 영향을 미칠 수 있는 독립변수의 수가 여럿 있는 경우 회귀분석 모델을 활용하여 공변수의 영향력을 어떻게 평가하는지 알아본다. 다음으로 제4장에서는 통계 소프트웨어인 R을 활용하여 구체적으로 회귀분석 모델에서 모수를 추정하고, 모델 선택의 기준으로 AIC와 BIC를 적용한 결과를 살펴보면서 선택된 모델이 적절한지 진단하는 과정을 소개하며, 모델에서 시간에 따라 변하는(time-varying) 공변수를 어떻게 처리하고, 적절한 모델에 기반한 *CIF* 추정은 어떻게 하는지 알아본다. 마지막으로 제5장에서는 논문을 요약하고 추후 연구방향을 살펴본다.

2. Kaplan-Meier 방법의 문제점과 *CIF*의 소개

경쟁적 위험은 연구의 대상인 어떤 개체가 여러 사망원인(고장원인)에 노출되고, 이들 여러 원인들 중에서 하나의 원인으로 인하여 고장에 이르게 되는 것이다. 여러 원인이 반복적으로 발생할 수도 있는데, 본 논문에서는 여러 원인들 중에서 제일 먼저 발생하는 원인으로 인해 개체가 고장에 이르게 되는 경우에 대해서만 살펴본다.

경쟁적 위험이 있든 또는 없든 생존분석 현장에서는 Kaplan-Meier 방법을 활용하여 사망확률(=1-생존확률)을 평가한다. Kaplan-Meier 방법은 신뢰성 분석에서 비모수적인 방법으로 생존확률인 신뢰도를 구하는 데 자주 사용된다[8]. Kaplan-Meier 방법에 의한 생존확률 $S(t)$는 n_j와 m_j를 t_j라는 고장시점(사실 t_j는 순서통계량인 $t_{(j)}$임)에서의 위험집합(risk set)과 고장 개수라고 하면 다음과 같이 표현할 수 있다.

$$\hat{S}(t_j) = \hat{S}(t_{j-1}) \times P(T > T_j | T \geq t_j)$$

$$= \prod_{i=1}^{j} P(T > t_i | T \geq t_i)$$
$$= \frac{n_1 - m_1}{n_1} \times \frac{n_2 - m_2}{n_2} \times \cdots \times \frac{n_j - m_j}{n_j}$$

하지만 Kaplan-Meier 방법에서는 주관심 대상인 고장원인이외의 다른 고장원인에 의한 고장은 중도중단으로 처리하다가 나중에 주관심 대상인 고장원인에 의해 고장이 발생하면 비로서 고장으로 처리하므로 사망확률을 과대추정하게 되고, 역으로 생존확률(신뢰도)을 과소추정하게 된다[5, 9]. 이에 경쟁적 위험하에서는 다음에 설명하는 누적사건함수를 구하여 사망확률을 구하는 것이 더욱 적절하다.

경쟁적 위험하에서 많이 쓰이는 모델로는 특정원인 위험률함수(cause-specific hazard function)와 누적사건함수가 있다. 이 두 함수는 고장시간과 고장원인인 (T, D)의 결합분포를 완전히 명시한다[10]. 첫 번째 모형에서는 j번째의 고장원인에 대한 특정원인 위험률함수를 다음과 같이 정의한다.

$$\lambda_j(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T_j < t + \Delta t, D = j | T \geq t)}{\Delta t}, \quad j = 1, \cdots, m$$

이 함수는 경쟁적 위험하에서 j번째 고장원인에 의해 개체가 고장이 일어날 위험률(고장률)을 나타낸다. 이 경우 특정원인 누적위험률(cause-specific cumulative hazard)은 다음과 같이 정의된다.

$$\Lambda_j(t) = \int_0^t \lambda_j(u) du$$

이제 전체 위험률(total hazard rate) $\lambda(t)$와 전체 생존함수(overall survival function)는 다음과 같이 정의된다.

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T_j < t + \Delta t | T \geq t)}{\Delta t} = \sum_{j=1}^{m} \lambda_j(t)$$

$$S(t) = P(T > t) = \exp(-\int_0^t \lambda(u) du)$$
$$= \exp(-\sum_{j=1}^{m} \int_0^t \lambda_j(u) du) = \exp(-\sum_{j=1}^{m} \Lambda_j(t))$$

전체 생존함수는 시간 t까지 어떤 고장원인에 의해서도 개체가 고장 나지 않을 확률을 나타낸다. 그러면 고장원인 j의 누적사건함수 $CIF_j(t)$는 다음과 같이 정의된다.

$$CIF_j(t) = P(T \leq t, D = j), \; j = 1, \cdots, m$$

이는 개체가 경쟁적 위험하에서 고장원인j에 의해 시간 t 또는 그 이전에 고장 날 확률을 나타내는 것으로, 다음과 같이 특정원인 위험률과 전체생존함수를 이용한 식으로 표현된다.

$$CIF_j(t) = \int_0^t \lambda_j(u) S(u) du$$

이 함수는 종종 부분분포함수(subdistribution function)라고 불린다. 왜냐하면 시간을 무한대로 보내 누적사건함수를 구해도 $CIF_j(\infty) = P(D = j)$이 1이하가 되어 정상적인 분포함수가 되지 않기 때문이다. 통상적인 Kaplan-Meier 방법에 의해 시간 t 이전에 고장 날 확률, 즉 사망확률은

$$1 - S_j(t) = \int_0^t \lambda_j(u) S_j(u) du$$

와 같이 정의된다. 이는 앞의 누적사건함수 $CIF_j(t)$와 그 모양은 비슷해도 $S(t) \leq S_j(t)$이다. 따라서

$$CIF_j(t) \leq 1 - S_j(t)$$

이므로 경쟁적 위험하에서 Kaplan-Meier 방법에 의해 구하는 사망확률은 누적사건함수에 의해 구하는 사망확률보다 더 크다는 것을 알 수 있다.

다음은 구체적인 신뢰성시험 결과 나온 수명 데이터에 대해 CIF를 구하는 과정이다.

1. 관심대상인 고장원인c로 인한 고장시간을 크기 순서대로 나열하여 t_j라고 하는 경우 각 고장시간 t_j에서의 위험률(고장률)을 구한다.

$$\hat{h}_c(t_j) = \frac{m_{cj}}{n_j}$$

여기에서 m_{cj}는 고장시간 t_j에서 고장원인 c로 인해 고장 난 제품의 개수이며 n_j는 고장시간 t_j에서의 위험집합이다.

2. 제품이 시간 t_{j-1}까지 어떤 고장원인으로도 고장 나지 않을 전반적인 생존확률(overall survival probability) $S(t_{j-1})$을 구한다. 이는 모든 고장원인으로부터 자유로운 생존확률을 일컫는다.
3. 시점 t_j의 바로 이전까지 어떤 고장원인으로도 고장 나지 않았는데, 이제 시점 t_j에서 여러 고장원인 중 고장원인c로 인해 고장이 발생할 고장발생가능성(incidence of failing)을 다음과 같이 구한다.

$$\hat{I}_c(t_j) = \hat{S}(t_{j-1}) \times \hat{h}_c(t_j)$$

4. 원점부터 시점 t_j까지 고장원인c로 인해 고장이 발생할 가능성인 *CIF*를 다음과 같이 구한다.

$$CIF_c(t_j) = \sum_{j'=1}^{j} \hat{I}_c(t_{j'}) = \sum_{j'=1}^{j} \hat{S}(t_{j-1}) \times \hat{h}_c(t_j)$$

참고로 Kaplan-Meier 방법에 의한 사망확률은 $\sum_{j'=1}^{j} \hat{S}_c(t_{j-1}) \times \hat{h}_c(t_j)$로 $\hat{S}_c(t_{j-1}) \geq \hat{S}(t_{j-1})$이므로 Kaplan-Meier 방법에 의한 사망확률이 *CIF* 방법에 의한 사망확률보다 더 크다는 것을 알 수 있다. 하지만 고장원인이 하나만 있는 경우에는Kaplan-Meier 방법에 의해 사망확률을 구하나 *CIF* 방법에 의해 사망확률을 구하나 똑같은 결과를 얻게 된다[11, 12].

3. 경쟁적 위험하에서의 회귀분석

경쟁적 위험하에서의 데이터는 고장시간 T, 고장원인 D 및 공변수 벡터 Z로 구성된다. 여기서 T는 연속적인 양(+)의 값을 취하는 확률변수이고, $D = \{1, 2, ..., m\}$로 m은 고장원인이 m개임을 나타낸다. 경쟁적 위험의 경우 예전에는 다변량분석 기법을 많이 사용했었다. 이 모델에서 각 개체는 각각의 고장원인으로 인한 잠재적인 고장시간(potential failure time)이 있는 것으로 가정하고, 이들 여러 고장원인 중에서 가장 빠른 고장원인에 의한 고장시간만 관측하고 나머지 고장원인에 의한 고장시간은 관측하지 못하고 숨어있는(latent) 것으로 본다[12]. 이 모형에서는 m개의 서로 다른 고장원인에 의한 고장시간 $T_1, \cdots, T_m$을 다음과 같이 결합생존함수

$$S(t_1, \cdots, t_m) = P(T_1 > t_1, \cdots, T_m > t_m)$$

의 형태로 나타낸다. 이 경우 주변위험함수(marginal hazard function)는

$$h_j(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T_j < t + \Delta t | T \geq t)}{\Delta t}, \quad j = 1, \cdots, m$$

와 같으며, 주변생존함수(marginal survival function)는 다음과 같다.

$$S_j(t) = P(T_j > t) = S(0, \cdots, 0, t, 0, \cdots, 0)$$

그러나 추가적인 가정이 없이는 결합생존함수나 주변생존함수가 주어진 데이터로는 식별되지(identifiable) 않기 때문에[1, 10, 13], 이런 잠재고장시간(latent failure time) 접근법은 실용성이 떨어진다. 이에 앞에서 기술한 부분분포함수를 기본으로 하여, 여러 독립변수들의 영향력을 살필 수 있는 회귀모형이 활용된다.

생존분석 데이터의 분석에 가장 많이 활용되는 회귀분석 모형은 Cox의 비례위험(proportional hazards) 모형이다. 하지만 통상적인 Cox의 비례위험 모형은 특정원인 위험률함수가 관심 대상인 경쟁적 위험하의 사건을 중도중단(censored)으로 간주하기 때문에 적절하지 못하다. 더욱이 특정원인 위험률함수는 생존확률의 측면에서 직접적인 해석이 불가능하다.

Cox 회귀모형의 변형된 형태로 데이터 확장(data augmentation)을 이용한 방법이 제안되었다[14]. 경쟁위험이 k개 있는 경우 각 개체에 대한 데이터를 k번 만큼 반복해서 복제하며(각 고장형태에 대해 행이 하나씩 필요함), 따라서 어떤 사건이 발생했는지 파악하기 위해서는 k-1개의 지시변수(indicator variable)가 필요하다. 이 모형을 이용하면 층화Cox 회귀분석을 활용하여 비례위험이 아닌 것도 모델링할 수 있다.

하지만 경쟁적 위험하의 데이터의 경우 *CIF*에 대한 공변수의 영향력을 평가하는 데에는 부분분포함수를 이용한 회귀모형이 적절하다는 연구가 진행되었다[15, 16]. 경쟁적 위험하에서 이런 회귀모형에 대한 여러 가지 접근방법은 많은 사람이 연구를 수행했다[17, 18, 19].

*CIF*의 부분분포 위험률(subdistribution hazard)에 대한 모델이 제안된 연구도 있다[15]. 여기서 부분분

포위험률이란 경쟁적 위험상태 하에 있는 개체가‘아직 살아있다’거나 또는 ‘다른 요인에 의해 죽었다’는 가정 하에 특정한 원인에 의해 고장 날 위험으로 정의된다. 따라서 특정한 원인 r에 대한 부분분포 위험률은 다음과 같이 쓸 수 있다.

$$\lambda_r(t) = \lim_{\Delta t \to 0} \frac{\Pr[t \le T < t+\Delta t, R = r | T \ge t \cup (T \le t \cap R \ne r)]}{\Delta t}$$

$$= -\frac{d}{dt} log(1 - I_r(t))$$

여기서 $I_r(t) = \Pr(T \le t, R = r)$는 원인 $r(r = 1, 2, ..., k)$에 대한 CIF이다.

한편, 공변수 벡터가 X인 개체에 대한 원인 r의 부분분포 위험률 모델로 다음과 같은 반모수적 비례위험(semiparametric proportional hazard) 모델이 제안됐다[15].

$$\lambda_r(t|X) = \lambda_{r0}(t)\exp(\beta_r^T X)$$

여기서 λ_{r0}는 원인 r의 기저(baseline) 부분군 위험률이고, β_r은 공변수들에 대한 회귀계수 벡터이다. 이 모델에서 모수는 통상적으로Cox 모델에서 사용하는 부분우도(partial likelihood)를 활용하면 쉽게 추정할 수 있다.

4. R을 활용한 통계적 추론

R은 http://www.R-project.org에서 내려받을 수 있는 통계 소프트웨어로 오픈소스 소프트웨어인 것이 특징이다. R은 여러 기본 분포 또는 추가적인 패키지를 활용하여 많은 통계분석을 실시할 수 있다. 예를 들어 경쟁적 위험 하에서의 데이터 분석을 위해서는cmprsk라는 추가 패키지를 활용하면 된다[7]. 본 논문에서는 데이터 파일을 R 환경에 읽어 들여, 벡터 또는 행렬 데이터를 분석방향에 맞게 조작할 수 있는 기본적인 지식이 있다고 가정하고 기술한다.

4.1 데이터

우선 <Table 1>은 우리 관심의 대상인 제품의 수명에 영향을 미치는 주요 변수들을 표로 정리한 것으로, 177개 제품의 고장시간(단위: 월)이 여러 요인{작업자의 성별(sex), 재료(material), 기계(machine), 작업방법(method), 온도(temperature)}에 의해 영향을 받을 수 있다는 것을 나타낸다. 우리의 주 관심사는 제품이 깨짐(crack)으로 인해 어떤 현상이 발생하는지 살펴보는 것인데, 제품은 깨짐(crack)이외에 마모(wear)로 인해서도 고장이 발생한다. 따라서 본 연구에서는 공변수들인 작업자의 성별(sex), 재료(material), 기계(machine),

Table 1 Variables in the failure.data

Variable name	Description	Summary Statistics
sex	worker sex	m = Male(100) f = Female(77)
material	two kinds of material	mat1(73) mat2(104)
machine	four kinds of machine	mac1(47) mac2(45) mac3(12) mac4(73)
method	two kinds of method	met1(21) met2(156)
temperature	temperature	4~62℃ 30.47(13.04)
ftime	failure time(unit: month)	0.13~131.77 20.28(30.78)
status	failure status	0 = censored(46) 1 = crack(56) 2 = wear(75)

작업방법(method) 및 온도(temperature)가 깨짐에 미치는 영향력을 경쟁적 위험사건인 마모까지도 고려하여 회귀분석을 이용하여 추정하는 것이다.

데이터가 현재의 작업 디렉토리(woriking directory)에 faildata라는 이름의 csv 형식으로 있는 경우(faildata는 link1에서 읽어 들일 수 있음) 다음과 같은 방식으로 주어진 데이터를 읽어 들인다.

```
> failure.data = read.csv("faildata.csv")
```

데이터가 현재의 작업 디렉토리가 아닌 특정한 위치에 있는 경우 file.choose()라는 함수를 사용하면 사용자가 컴퓨터 내의 특정한 위치를 찾아 그 안에 있는 파일을 선택하여 데이터를 읽어 들일 수 있다. 이제 데이터를 정확하게 읽어 들였으면 첫 일부의 제품에 대한 값들이 얼마나 되는지 다음과 같은 방식으로 확인할 수 있다.

```
> head(failure.data)
```

	sex	*material*	*machine*	*method*	*temperature*	*ftime*	*status*
1	*m*	*mat1*	*mac4*	*met1*	*48*	*0.67*	*2*
2	*f*	*mat2*	*mac2*	*met1*	*23*	*9.50*	*1*
3	*m*	*mat1*	*mac3*	*met1*	*7*	*131.77*	*0*
4	*f*	*mat1*	*mac2*	*met1*	*26*	*24.03*	*2*
5	*f*	*mat1*	*mac2*	*met1*	*36*	*1.47*	*2*
6	*m*	*mat1*	*mac4*	*met1*	*17*	*2.23*	*2*

이로부터 상태(status)라는 중도중단(censoring)과 경쟁적 위험사건까지 감안하여 작업자의 성별(sex), 재료(material), 기계(machine), 작업방법(method), 온도(temperature) 등의 공변수들이 고장시간(ftime)에 미치는 영향을 파악하는 모델을 세우고자 한다. 앞의 데이터로부터 첫 번째 제품은 0.67개월 만에 경쟁적 위험사건인 마모로 인해 고장이 났는데, 이때 작업자의 성별(sex)은 남자(m)이며, 재료(material)는 mat1이고, 사용한 기계(machine)는 mac4이고, 작업방법(method)은 met1이었으며, 주위 온도(temperature)는 48℃이었다.

앞의 공변수들 중에서 온도(temperature)를 제외하고 나머지 공변수들은 모두 범주형 변수들이다. 따라서 이들 범주형 변수들의 값이 범주를 나타내도록 코딩을 해야 한다. 범주형 변수를 코딩하는 방법이 여럿 있지만 가장 간단한 방법은 범주형 변수의 특정한 수준을 기준(baseline)으로 잡고 코딩하는 것이다. J개의 수준을 가진 요인 또는 범주형 변수의 경우 J-1개의 더미(dummy)변수를 만들어, 특정 범주에 속한 경우 해당 변수를 1로 하고 그렇지 않은 경우 0으로 코딩을 한다. 이때 하나의 범주는 기준으로 잡으며, 여기에서 모든 더미변수들의 값은 0으로 두면 모형에서 이들 더미 변수는 없어진다. 예를 들어, 작업자의 성별(sex)은 여자(f)와 남자(m)의 두 수준이 있으므로 하나의 변수만 필요하며, 남자를 기준으로 삼는 경우 남자(m)에게는 0을, 여자(f)에게는 1을 할당한다. 범주형 변수인 기계(machine)의 경우 4개의 수준(mac1, mac2, mac3, mac4)이 있으므로 3개의 더미변수가 필요하다. 일반적으로 J개의 수준을 가진 요인은 J-1개의 더미 변수 즉, 지시(indicator)변수로 만들어야 한다(범주형 변수의 요인(factor)을 지시변수로 만드는 함수 factor2ind()는 link2에서 내려받을 수 있음).

따라서 작업자의 성별(sex)을 지시변수로 바꾸고자 하는 경우 다음과 같은 R code를 입력하면 원하는 결과를 얻을 수 있다.

```
> factor2ind(sex, "m")
```

	sex:f
[1,]	*0*
[2,]	*1*
[3,]	*0*
...	
[177,]	*0*

factor2ind() 함수는 하나의 열과 관측수 만큼의 행(행의 값이 1이면 여자를, 0이면 남자를 나타냄)을 가진 행렬을 생성한다. 따라서 첫 번째와 세 번째 제품은 남자가 만들었으며, 두 번째 제품은 여자가 만들었다. factor2ind() 함수를 4개의 수준을 가진 기계(machine)에 적용하는 경우 다음과 같은 결과를 얻는다.

```
> factor2ind(machine, "mac4")
```

	machine:mac1	*machine:mac2*	*machine:mac3*
[1,]	*0*	*0*	*0*
[2,]	*0*	*1*	*0*
[3,]	*0*	*0*	*1*
...			
[177,]	*0*	*0*	*0*

기계(machine)의 경우 세 개의 열이 있는데, 이는 기계의 수에서 1을 뺀 수와 같다. 앞의 표로부터 첫 번째와 마지막 제품은 기계4(mac4)에 의해 생산되었고, 두 번

째와 세 번째 제품은 기계 2(mac2)와 기계 3(mac3)에 의해 생산되었음을 알 수 있다.

온도(temperature), 작업자 성별(sex), 재료(material), 기계(machine) 및 작업방법(method)을 모두 포함하는 행렬 즉, 설계행렬(design matrix)은 R에서 다음과 같이 cbind() 함수를 사용하여 생성할 수 있다.

```
> x = cbind(temperature, factor2ind(sex, "m"), factor2ind(material, "mat1"), factor2ind(machine, "mac4"), factor2ind(method))
> head(x)
```

	temperature	sex:f	material:mat2	machine:mac1	machine:mac2	machine:mac3	method:met2
[1,]	48	0	0	0	0	0	0
[2,]	23	1	1	0	1	0	0
[3,]	7	0	0	0	0	1	0
[4,]	26	1	0	0	1	0	0
[5,]	36	1	0	0	1	0	0
[6,]	17	0	0	0	0	0	0

4.2 회귀모형

경쟁적 위험하의 데이터에 대한 회귀모형을 적합시키는 함수는 crr()인데, 이는 cmprsk 패키지에 포함되어 있다. 따라서 다음과 같이 R에서 cmprsk 패키지를 먼저 불러와야 한다.

```
> require(cmprsk)
```

crr() 함수에는 수명을 나타내는 고장시간(ftime), 각 고장 시 중도중단의 형태를 나타내는 상태(status), 그리고 공변수 행렬(x)을 넣는다. 상태(status)는 기본적으로 0은 중도중단을, 1은 관심사건에 의한 고장{우리의 경우 깨짐(crack)에 의한 고장}을, 2는 다른 경쟁적 사건에 의한 고장{우리의 경우 마모(wear)에 의한 고장}을 나타낸다.

<Table 1>의 데이터에 대한 회귀분석은 다음과 같이 crr() 함수를 이용하여 실시할 수 있으며, 그 결과는 summary 함수를 활용하면 구할 수 있다.

```
> mod1 = crr(ftime, status, x)
> summary(mod1)
```

Competing Risks Regression

Call:
crr(ftime = ftime, fstatus = status, cov1 = x)

	coef	exp(coef)	se(coef)	z	p-value
temperature	-0.0185	0.982	0.0119	-1.554	0.1200
sex:f	-0.0352	0.965	0.2900	-0.122	0.9000
material:mat2	-0.4723	0.624	0.3054	-1.547	0.1200
machine:mac1	-1.1018	0.332	0.3764	-2.927	0.0034
machine:mac2	-1.0200	0.361	0.3558	-2.867	0.0041
machine:mac3	-0.7314	0.481	0.5766	-1.268	0.2000
method:met2	0.9211	2.512	0.5530	1.666	0.0960

	exp(coef)	exp(-coef)	2.5%	97.5%
temperature	0.982	1.019	0.959	1.005
sex:f	0.965	1.036	0.547	1.704
material:mat2	0.624	1.604	0.343	1.134
machine:mac1	0.332	3.009	0.159	0.695
machine:mac2	0.361	2.773	0.180	0.724
machine:mac3	0.481	2.078	0.155	1.490
method:met2	2.512	0.398	0.850	7.426

Num. cases = 177
Pseudo Log-likelihood = -267
Pseudo likelihood ratio test = 24.4 on 7 df

앞 결과의 첫 번째 부분은 설계행렬 내의 각 항에 대한 회귀계수의 추정치 $\hat{\beta}_j$, 상대적 위험도(relative risk) $\exp(\hat{\beta}_j)$, 추정치의 표준편차, z-값 및 각 변수에 대한 통계적 유의성을 점검하는 p-값을 나타낸다. <Table 1>의 데이터의 경우 성별(sex)은 전혀 유의하지 않으며, 이어서 온도(temperature)와 재료(material) 또한 그다지 유의하지 않지만 방법(method)은 10% 유의수준에서 유의하다. 기계(machine)의 경우 네 번째(mac4) 것이 기준(baseline)이므로 각 기계에 대한 p-값은 네 번째 기계와 비교한 결과를 나타낸다. 예를 들어, 기계 1(mac1)은 기계 4(mac4)에 비해 통계적으로 매우 유의한 차이(p-값이 0.34%)를 나타낸다. 요인의 수준수가2개를 넘는 경우 해당 요인에 대한 전반적인 p-값(overall p-value)을 구해 유의성을 점검하는 Wald 검정을 할 필요가 있는데, R에서는 aod라는 패키지를 불러와 구현할 수 있다.

```
> library(aod)
> wald.test(mod1$var, mod1$coef, Terms=4:6)
```

Wald test:

Chi-squared test:
X2 = 14.0, df = 3, P(> X2) = 0.0029

wald.test() 함수에 들어가는 첫 번째 인수는 회귀계수에 대한 분산-공분산행렬을, 두 번째 인수는 해당 회귀계수를, 마지막 인수는 통계적 유의성 검증을 하고 싶은 회귀계수의 위치를 넣는다{좀 더 자세한 설명을 보려면 R에서 help(wald.test)를 입력하면 된다}. <Table 1>의 데이터에서는 p-값이 0.29%로 통상적인 유의수준 5% 또는 1%보다 훨씬 작으므로 "4대의 기계가 수명에 미치는 영향이 모두 같다"고는 할 수 없음을 알 수 있다. 특히 기계 1(mac1)과 기계 4(mac4) 간에는 수명에 차이가 있다는 것을 알 수 있다.

앞의 crr()을 돌려서 나온 결과 중 두 번째 부분은 각 항에 대한 상대적 위험도 $\exp(\hat{\beta}_j)$와 이에 대한 95% 신뢰구간{95% 이외의 수준에서 신뢰구간을 구하려면 summary() 안에서 conf.int을 활용하면 된다. 좀 더 구체적으로 알기 위해서는 R에서 help(summary.crr)을 입력하면 된다}. 범주형 공변수에 대한 상대적 위험도 또는 부분분포 위험률비율(subdistribution hazard ratio)은 다른 공변수의 값이 모두 동일하다는 가정하에 기준 그룹(baseline group)에 비해 대상 그룹의 부분분포 위험률의 값이 얼마나 큰지 나타내는 비율이다. 따라서 해당 공변수가 연속적인 값을 취한다면 $\exp(\hat{\beta}_j)$는 다른 공변수의 값이 일정하다는 가정하에 해당 공변수의 값이 1단위 증가할 때의 효과를 나타낸다. <Table 1>의 데이터는 온도(temperature) 변수의 경우 exp(-0.0185) = 0.982이므로, 이는 온도가 1℃ 증가할 때의 상대적 위험도가 0.982 증가함을 의미한다. 0.982는 1보다 작으므로 온도가 올라가면 위험도가 작아짐을 알 수 있다. 하지만 $\hat{\beta}_j$의 변동인 표준편차가 0.0119인 것을 고려하면 상대적 위험도는0.982~1.005에 걸쳐 있으므로(1을 포함함) 온도의 증가에 따른 상대적 위험도의 차이는 크지 않은 것으로 판단된다. 한편, 성별(sex)의 경우 남자에 비해 여자의 상대적 위험도는 exp(-00352) = 0.965이다. 그런데 상대적 위험도에 대한 95% 신뢰구간이 1을 포함하므로 남자에 비해 여자가 더 위험하다고는 말할 수는 없다.

4.3 모델의 비교

앞의 crr()을 돌려서 나온 결과 중 마지막 부분은 최대 의사 로그가능도(pseudo log-likelihood) 값과 의사 가능도 비율검정(pseudo likelihood ratio test)값 (global null 모델과 최종 모델 간 목적함수값의 차이)를 나타낸다. 하지만 목적함수가 진짜 가능도(likelihood)가 아니므로(검정통계량이 점근적으로 χ^2분포를 따르지는 않음) 가능도 비율검정에 기반한 모델 간 비교는 참고하는 수준에 머물러야 한다. 모델 간 비교는 다음에 설명하는 AIC 또는 BIC를 기준으로 해야 한다.

주어진 모델에 대한 데이터의 가능도 값은 그 모델이 얼마나 적합한지 잴 수 있는 하나의 척도이다. 그러나 이 가능도 값은 모델에 모수의 수가 많아지면 증가하게 되어 있다. 이런 이유로 가능도 값을 모델선택의 기준으로 삼으면 과도한 적합이 이루어질 수 있다. 이를 피하기 위해서는 추정에 사용되는 모수의 수가 많으면 벌칙을 가하는 기준이 필요하다. 이에 적절한 기준이 AIC(Akaike's Information Criteria)와 BIC (Bayesian Information Criteria)이다. AIC는 다음과 같이 정의된다.

$$\text{AIC} = -2l+2d$$

여기서 l은 주어진 모델에 대한 로그 가능도의 최대값, d는 추정되는 모수의 수를 나타낸다. 따라서 AIC는 추정되는 모수의 수가 많아질수록 커지는 벌점체계를 포함한다. 한편, BIC는 다음과 같이 정의된다.

$$\text{BIC} = -2l+\log(n)d$$

여기서 n은 총관측수이다. BIC는 모수의 수가 많으면 AIC보다 벌점이 더욱 더 크다는 것을 알 수 있다. 실용적인 측면에서 AIC와 BIC는 벌점이 약간 다르게 정의된 것임을 알 수 있다. 하지만 이들 식은 전혀 다른 관점에서 도출되었다. AIC는 추정된 모델과 참 모델간의 상대적 Kullback-Leibler 거리의 기댓값의 추정치이고, BIC는 베이지안의 관점에서 도출된 것으로 각 모델에 대한 선험적 확률이 동일한 경우 두 모델을 비교하는 Bayes factor에 로그를 취한 것에 대한 근사치이다.

AIC와 BIC 모두 어떤 모델에 대해 가설검정의 형태로 유의확률을 제공하지는 않는다. 하지만 AIC와 BIC는 주어진 기준에 의거하여 여러 경쟁적인 모델들에 대해 적합의 우선순위를 매길 수 있다. 즉, AIC와 BIC의 값 자체의 크기는 큰 의미가 없지만 가장 작은 값과의 차이는 의미가 있다. 구체적으로 이들 차이값에 대한 해석은 다음과 같다. 예를 들어, $\Delta BIC_i = BIC_i - \min(BIC)$라고 하는 경우 $\Delta BIC_i \geq 10$이면 i번째 모델은 적합도가 떨어지는 반면 $0 < \Delta BIC_i < 2$이면 i

번째 모델은 적합도가 아주 좋아 이 모형을 이용하여 추정을 해도 무리가 없다[19]. 비슷한 방법으로 AIC를 이용할 수 있다.

<Table 1>의 변수에 대한 수명 데이터에 대해 경쟁적 위험하에서의 회귀모형을 적합시켜 본 결과 일부 변수는 제한적으로 유의하고, 또 다른 일부 변수는 유의하지 않으므로 이들은 모형에서 제거하는 것이 바람직해 보였다. 이와 같이 어떤 변수가 모형에 들어가는 것이 좋고 또 어떤 변수가 모형에서 빠지는 것이 좋은지는 크게 ① 모든 가능한 모형(all possible models) ② 전진적 방법(forward approach) ③ 후진적 방법(backward approach)의 세 가지가 있다. ①의 방법은 변수가 총 5개이므로 가능한 모델의 수가 31(= 2^5-1)개로 너무 많아 현실적으로 어느 모델이 가장 적절한지 판단하기가 쉽지 않다. 따라서 여기에서는 ②의 전진적 방법을 이용하여 변수의 추가에 따른 AIC나 BIC의 변화를 보고 최적 모델을 선정한다.

우리의 데이터에서는 기계(machine)가 통계적으로 가장 유의하고, 이어서 작업방법(method)이 유의하여 모형에 포함시키지만 온도(temperature), 작업자의 성별(sex)과 재료(material)는 크게 유의하지 않으므로 이들 3개의 변수를 모형에 한꺼번에 포함시키는 것은 바람직하지 않다. 따라서 기계(machine)와 작업방법(method)이 모델에 들어가 있는 상태에서 다음에서와 같이 온도(temperature), 작업자의 성별(sex)과 재료(material)를 한 개씩 넣어본다.

```
> mod2 = crr(ftime, status, x[, 4:6]) # machine
> mod3 = crr(ftime, status, x[, c(4:6,7)]) #
machine+method
> mod4 = crr(ftime, status, x[, c(4:6,7,1)]) #
machine+method+temperature
> mod5 = crr(ftime, status, x[, c(4:6,7,2)]) #
machine+method+sex
> mod6 = crr(ftime, status, x[, c(4:6,7,3)]) #
machine+method+material
```

이제 다음의 modsel.crr() 함수를 이용하면(이 함수는 link3에서 내려받을 수 있음) 여러 후보모델 중에서 적절한 모델을 선택할 수 있도록 로그 가능도의 값, 추정모수의 수, BIC의 값, BIC diff(각 모델의 BIC 값과 여러 모델 중 가장 작은 BIC 값 간의 차이)를 구할 수 있다. 다음은 그 결과를 나타낸다.

```
> modsel.crr(mod1, mod2, mod3, mod4, mod5, mod6)
Model selection table

Model 0: Null model
Model 1: crr(ftime = ftime, fstatus = status, cov1 = x)
Model 2: crr(ftime = ftime, fstatus = status, cov1 = x[, 4:6])
Model 3: crr(ftime = ftime, fstatus = status, cov1 = x[, c(4:6, 7)])
Model 4: crr(ftime = ftime, fstatus = status, cov1 = x[, c(4:6, 7, 1)])
Model 5: crr(ftime = ftime, fstatus = status, cov1 = x[, c(4:6, 7, 2)])
Model 6: crr(ftime = ftime, fstatus = status, cov1 = x[, c(4:6, 7, 3)])
  Num.obs  logLik Df.fit    BIC BIC diff
0    177 -278.71      0 557.41   0.0000
1    177 -266.52      7 569.28  11.8679
2    177 -271.53      3 558.59   1.1761
3    177 -270.78      4 562.27   4.8545
4    177 -267.81      5 561.49   4.0797
5    177 -270.73      5 567.33   9.9170
6    177 -267.82      5 561.53   4.1125
```

앞에서 Model 0라고 되어 있는 null model은 요구하지 않아도 포함되는 모델로서 어느 공변수(covariate)도 포함하지 않으며, 따라서 공변수를 일부 포함하고 있는 모델과 비교대상이 되는 모델이다. BIC의 값이 가장 작은 모델은 null model이며, 그 다음은 기계(machine)가 포함된 모델이다. 앞의 AIC와 BIC 관련 활용기준으로부터 기계(machine)만을 포함한 모델 2(mod2)가 적절하다는 것을 알 수 있다. 다음은 모델 2를 적합시킨 경우 나온 결과이다.

```
> summary(mod2)
Competing Risks Regression

Call:
crr(ftime = ftime, fstatus = status, cov1 = x[, 4:6])

                coef exp(coef) se(coef)    z p-value
machine:mac1 -1.113     0.329    0.371 -3.00  0.0027
machine:mac2 -0.979     0.376    0.347 -2.82  0.0048
machine:mac3 -0.789     0.455    0.602 -1.31  0.1900

             exp(coef) exp(-coef)  2.5% 97.5%
machine:mac1     0.329       3.04 0.159 0.680
machine:mac2     0.376       2.66 0.190 0.742
machine:mac3     0.455       2.20 0.140 1.480

Num. cases = 177
Pseudo Log-likelihood = -272
Pseudo likelihood ratio test = 14.3  on 3 df,
```

이로부터 machine 1과 machine 2와 관련된 계수는 통계적으로 0이 아니며, 따라서 기준 그룹인 machine 4에 비해 machine 1과 machine 2의 상태적 위험도는 1/3 정도임을 알 수 있다. 반면, machine 3의 계수는 통계적으로 유의하지 않으며, machine 4에 비해 상대적 위험도가 1/2 정도 되지만 95% 신뢰구간은 (0.14, 1.48)로 1을 포함한다는 것을 알 수 있다.

4.4 모델 점검

crr() 함수의 출력물에는 Schoenfeld 잔차의 행렬이 포함되어 있다. 이는 통상적인 생존분석 모형을 주어진 데이터에 적합시키고 난 후 해당 모형이 얼마나 적합한지 살펴보는데 사용하는 Schoenfeld 잔차와 유사하다. 따라서 2차원 상에 (x, y) = (고장시점, 이 행렬의 j번째 열의 값)의 값을 타점하여 어떤 특정한 패턴이 나타나는지 살펴볼 필요가 있다. 이때 점들에서 어떤 특정한 패턴도 나타나지 말아야 해당 모형이 적합함을 의미한다. 우리 데이터의 경우에는 다음과 같이 하면 행렬에 있는 잔차를 확인하고 고장시간에 따른 잔차를 도시할 수 있다(<Fig. 1> 참조).

```
> head(mod2$res)
```

	[,1]	[,2]	[,3]
[1,]	-0.1394277	-0.1525481	-0.04922084
[2,]	-0.1406975	-0.1539373	-0.04966909
[3,]	-0.1419906	-0.1553521	-0.05012558
[4,]	-0.1433077	-0.1567931	-0.05059054
[5,]	-0.1446494	-0.1582611	-0.05106420
[6,]	-0.1460165	-0.1597569	-0.05154682

<Fig. 1>은 앞에서 선택한 최종 모델인 모델2(mod2)에 있는 mac1, mac2, mac3 각각에 대해 고장시간에서 Schoenfeld 형태의 잔차를 타점한 것이다. 각각의 그림에서는 부분분포 위험률이 비례적인지 평가하기 위해 국부 가중 평활회귀선(locally weighted regression smoother)이 그려져 있다. 이 그림에서는 잔차에 대한 평활회귀선이 수평에 가까우므로 부분분포 위험률들이 비례적이라는 것을 파악할 수 있다.

4.5 공변수의 시간 의존성 점검

"시간에 따라 변하는 고정효과(time-varying fixed effect)"는 인위적인 시간 종속적 공변수 의"서로 다른 시간대에의 효과가 다름"을 모형화 하는데 이용되며, 이는 Cox 비례위험 모형의 비례성이 들어맞지 않는 경우 이를 모형화 하는 방법에 속한다. 똑같은 방식으로 부분포 위험률들이 서로 비례적이 아닌 경우 Fine and Gray 모형에서도 "시간에 따라 변하는 고정효과"를 이용하여 "서로 다른 시간대에의 효과가 다름"을 모형화 할 수 있다.

시간 종속적 공변수는 crr() 함수에서 cov2(공변수들의 행렬)에 tf라는 시간의 함수가 곱해지는 형태로 모델링된다. 예를 들어 온도(temperature) 변수에 대해 2차식의 형태인 $\beta_1$temperature+$\beta_2$temperature t+β_3 temperature t^2를 모델링하고 싶은 경우 crr() 함수에서는 cov2와 tf와 같은 인수를 다음과 같이 명시한다.

```
> mod7 = crr(ftime, status, cov1=x[,c(1,4:6)], cov2 =
cbind(temperature, temperature), tf = function(t) cbind(t, t^2))
```

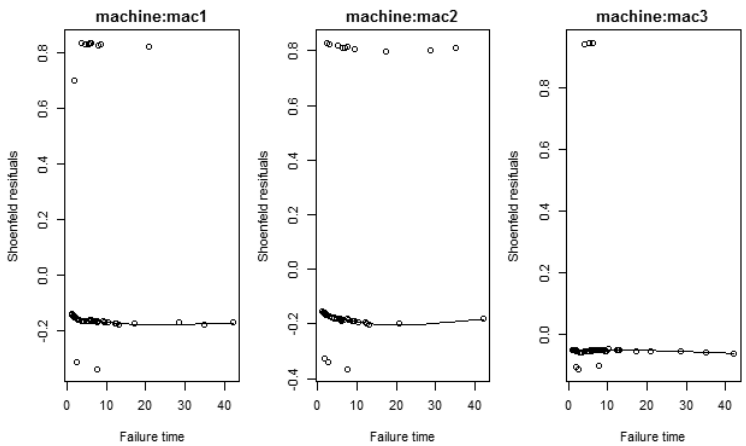

Fig. 1 Plot of Schoenfeld-type residuals against the failure time for each term in the final model

인수 cov1에는 고정 공변수 행렬을 포함하므로 우리의 예에서는 기계(machine)와 온도(temperature)의 공변수와 관련된 항을 포함한다. 그리고 인수 cov2에서는 두 개 열의 행렬{두 열 모두 온도(temperature)의 값을 가짐}이 tf라고 정의되는 2차 함수에 의해 곱해지는 형태가 된다. 그러면 다음과 같은 결과를 얻을 수 있다.

```
> summary(mod7)
Competing Risks Regression

Call:
crr(ftime = ftime, fstatus = status, cov1 = x[, c(1, 4:6)], cov2 =
cbind(temperature, temperature), tf = function(t) cbind(t, t^2))

                  coef exp(coef)  se(coef)      z p-value
temperature   -0.026666     0.974 2.00e-02 -1.335  0.1800
machine:mac1  -1.085768     0.338 3.73e-01 -2.915  0.0036
machine:mac2  -1.010888     0.364 3.48e-01 -2.907  0.0036
machine:mac3  -0.885241     0.413 5.95e-01 -1.488  0.1400
temperature*t  0.000400     1.000 3.17e-03  0.126  0.9000
temperature*tf2 0.000011    1.000 7.09e-05  0.155  0.8800

                  exp(coef) exp(-coef)  2.5%  97.5%
temperature            0.97      41.03 0.936  1.013
machine:mac1          0.338       2.96 0.163  0.701
machine:mac2          0.364       2.75 0.184  0.719
machine:mac3          0.413       2.42 0.129  1.324
temperature*t         1.000       1.00 0.994  1.007
temperature*tf2       1.000       1.00 1.000  1.000

Num. cases = 177
Pseudo Log-likelihood = -269
Pseudo likelihood ratio test = 19.5  on 6 df,
```

앞의 결과로부터 시간 t와 온도(temperature) 간의 1차 또는 2차 항은 통계적으로 유의하지 않다는 것을 알 수 있고, 이는 다시 한 번 부분분포 위험률들이 비례적이라는 가정이 기각되지 않음을 확인할 수 있다

4.6 모델에 기반한 *CIF*의 추정

j번째의 고장원인에 대한 *CIF*의 추정은 다음과 같다.

$$\hat{I}_j(t) = 1 - e^{-\hat{H}_j(t)}$$

여기서 $\hat{H}_j(t)$는 관심의 대상인 j번째의 고장원인에 대한 누적 부분분포 위험률로, Breslow 형태의 추정치를 이용하여 계산되는 추정치이다. 우리의 데이터에서 공변수 기계(machine)의 각 수준에서의 *CIF* 추정치를 구하기 위해서는 기계(machine)의 각 수준에서의 값을 행렬의 형태로 다음과 같이 먼저 정의한다

```
> machine = as.factor(machine)
> x0 = cbind(machine = factor2ind(levels(machine),
"mac4"))
> x0
     levels(machine):mac1 levels(machine):mac2 levels(machine):mac3
[1,]                    1                    0                    0
[2,]                    0                    1                    0
[3,]                    0                    0                    1
[4,]                    0                    0                    0
```

그리고 각 고장시점에서 *CIF*의 추정치는 predict() 함수를 이용하여 다음에서와 같이 구한다

```
> pred = predict(mod2, x0)
> pred
      [,1]       [,2]       [,3]       [,4]       [,5]
[1,]  1.10 0.002962152 0.003384217 0.004093336 0.008984103
[2,]  1.20 0.005942385 0.006787645 0.008206966 0.017968205
[3,]  1.23 0.008940978 0.010210569 0.012341175 0.026952307
[4,]  1.30 0.011958213 0.013653282 0.016496260 0.035936409
[5,]  1.60 0.014994382 0.017116085 0.020672521 0.044920509
..... .....
[53,] 42.17 0.196133976 0.220796778 0.260563415 0.485300353
```

앞에서 첫 번째 열은 고장시점을 나타내고, 나머지 네 개의 열은 각 고장시점에서 x0 행렬의 4개의 행에 있는 공변수들의 값에 해당하는 *CIF*의 추정치이다. 따라서 앞의 결과물의 2, 3, 4, 5열에 있는 값들은 기계(machine)가 mac1, mac2, mac3, mac4인 기계를 사용한 경우 *CIF*의 추정치들을 나타낸다. 마지막으로 앞의 *CIF* 추정치들에 대해서는 다음과 같은 방법으로 <Fig 2>를 그릴 수 있다. 이로부터 기계(machine)는 mac4, mac3, mac2, mac1의 순서로 *CIF*가 높음을 알 수 있다. 이와 같이 회귀분석 모형을 이용하면 공변수의 값에 따라 CIF의 값이 다르게 나타난다는 것을 보여줄 수 있다.

```
> plot(pred, lty=1:4, xlab="Failure times", ylab="CIF")
> legend("bottomright", legend=levels(machine),
lty=1:4, title="machine")
```

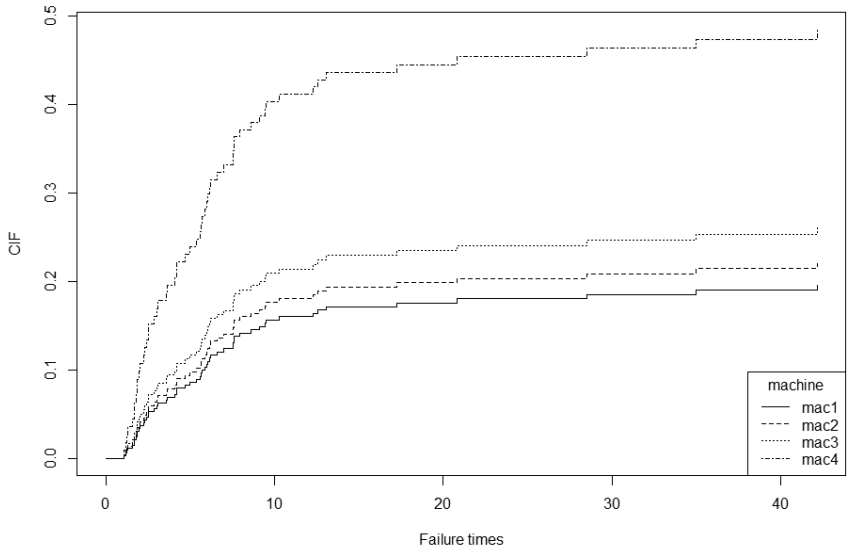

Fig. 2 Cumulative Incidence Curves for different machines at each failure time

5. 결론 및 논의

생존분석 방법론들은 지난 수십 년에 걸쳐 의학 및 약학 분야에 폭넓게 사용된 반면, 신뢰성 분야에서는 많이 활용되지 못했다. 특히, 본 논문의 주제에서와 같이 경쟁적 위험하에 있는 환경에서는 올바른 생존분석 방법이 신뢰성 분야에서 거의 활용되지 못하고 있다. 이에 본 논문에서는 고장의 원인이 세 개 이상이고, 이들 고장원인 중 하나의 고장원인으로 인해 제품이 고장 나는 경쟁적 위험하에서 어떻게 하면 사망확률을 올바로 추정할 수 있는지 살펴보았다. 이런 경쟁적 위험의 경우 기존의 Kaplan-Meier 방법에서는 고장시간과 중도중단시간 간 독립성을 가정하여, 주관심 대상인 고장원인이외의 다른 원인에 의한 고장시간을 중도중단시간으로 처리하다가 이후에 주관심 사인 고장원인에 의해 고장이 발생하면 이전의 모든 중도중단시간을 고장시간으로 처리하므로 사망확률을 과대하게 추정하게 된다. 이에 본 연구에서는 주관심 대상인 고장원인이외의 다른 원인에 의한 고장은 중도중단으로 처리하지만 독립성을 가정하지 않고 사망확률을 구하는 누적사건함수 방법이 적절함을 설명하였다.

다음으로 여러 가지 고장원인에 의한 고장시간에 영향을 미칠 수 있는 공변수의 중요성에 대해 알아보기 위해 경쟁적 위험하에서의 회귀모형을 살펴보았다. 본 연구에서는 여러 고장원인들 중 하나의 고장원인으로 제품 고장이 발생하는 경쟁적 위험하에서 제품의 고장에 영향을 미칠 수 있는 공변수의 영향력을 오프소스 통계 소프트웨어인 R에 있는 library의 하나인 cmprsk를 활용하여 살펴보았다.

구체적으로 제품은 깨짐(crack)이외에 마모(wear)로 인해 고장이 발생할 수 있으며, 이런 고장에 영향을 미칠 수 있는 공변수인 작업자 성별, 재료, 기계, 작업방법 및 온도의 영향력을 회귀모형을 이용하여 살펴보았다. 회귀모형에 대한 설명에서는 우선 적절한 회귀모형을 AIC나 BIC의 값으로 어떻게 찾는지 설명했으며, 선택된 모형이 타당한지 보기 위해 잔차의 특정 패턴 여부를 살펴보았으며, 공변수의 시간종속성 여부를 또한 살펴보고, 마지막으로 최종 모델에 기반하여 CIF를 추정한 결과를 보여주었다.

이 논문에서는 신뢰성에 관심이 있는 사람은 누구나 R을 활용하여 경쟁적 위험하에서의 회귀모형을 통해 통계분석을 실시할 수 있도록 하였다. 사실 경쟁적 위험하에서의 회귀분석이 현장에서 많이 활용되지 못하고 있는 이유 중의 하나는 누구나 쉽게 접할 수 있는 통계 소프트웨어가 없었기 때문일 것이다. 따라서 본 연구에서는 R에 조금이라도 관심이 있는 사람은 R을 활용하여 경쟁적 위험 하에서 나오는 데이터에 대해 쉽게 분석을 할 수 있도록 R code를 포함시켰다.

본 연구에서는 제품의 깨짐(crack) 또는 마모(wear)가 발생하면 해당 제품은 고장이 발생하여 더 이상 기능을 발휘할 수 없다고 가정하고 있다. 하지만 자동차나 일반 설비의 경우 본 논문의 사례에서와 같이 여러 고장의 원인이 있지만 고장이 발생해도 수리가 곧 이루어져 똑같은 고장이 여러 번 발생할 수 있다. 따라서 추후 연구에서는 경쟁적 위험하에 있으며, 각각의 위험요인이 반복되는 사건(recurrent event)이 될 수 있는 일반적인 경우 여러 공변수 중에서 어떤 변수가 통계적으로 유의한지 살펴볼 수 있는 회귀모형에 대해 살펴보기로 한다.

References

[1] Kalbfleisch, J. D. and Prentice, R. L. (2002). “The statistical analysis of failure time data”. John Wiley & Sons, New York.

[2] Klein, J. P. and Moeschberger, M. L. (2003). “Survival analysis. techniques for censored and truncated data”. Springer, New York.

[3] Kaplan, E. and Meier, P. (1958). “Nonparametric estimation from incomplete observations”. Journal of the American Statistical Association, Vol. 53, pp. 457-481.

[4] Porta, N., Gómez, G., and Calle, M. L. (2008). “The Role of survival functions in competing risks”. available at http://www.eio.upc.es/nporta, cited on June 20, 2011.

[5] Putter, H., Fiocco, M., and Geskus, R. B. (2006). “Tutorial in Biostatistics: Competing risks and multi-state events”. Statistics in Medicine, Vol. 26, pp. 2389-2430.

[6] Kim, H. (2007). “Cumulative incidence in competing risks data and competing risks regression analysis”. Clinical Cancer Research, Vol. 13, pp. 559-565.

[7] Scrucca, L., Santucci, A., and Aversa, F. (2007). “Competing risk analysis using R: an easy guide for clinicians”. Bone Marrow Transplantation, Vol. 40, pp. 381-387.

[8] Chastain, T. M., Young, T. M., Geuss, F. M., and León, R. V. (2009). Using reliability tools to characterize wood strand thickness of oriented strand board panels. International Journal of Reliability and Applications, Vol. 10, pp. 89-99.

[9] Gooley, T. A., Leisenring, W., Crowley, J., and Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. Statistics in Medicine, Vol. 18, pp. 695-706.

[10] Lawless, J. (2003). “Statistical models and methods for lifetime data”. John Wiley & Sons, New York.

[11] Baik, J. (2016). “Reliability analysis under the competing risks”. Journal of Applied Reliability, Vol. 16, pp. 56-63.

[12] Fürstová, J. and Valenta, Z. (2011). “Statistical analysis of competing risks: overall survival in a group of chronic myeloid leukemia patients”. EuroMISE s.r.o., 7, en2-en10.

[13] Tsiatis, A. (1975). “A nonidentifiability aspect of the problem of competing risks”. Proceedings of the National Academy of Sciences, USA 72, pp. 20-22.

[14] Lunn, M. and McNeil D. (1995). “Applying Cox regression to competing risks”. Biometrics, Vol. 51, pp. 281-320.

[15] Fine, J. and Gray, R. J. (1999). A proportional hazards model for the subpopulation of a competing risk“. Journal of American Statistical Society, Vol. 94, pp. 496-509.

[16] Klein, J. P. and Anderson, P. K. (2005). “Regression modeling of competing risks”. Biometrics, Vol. 61, pp. 223-229.

[17] Klein, J. P. and Zhang M. J.(2007). “Survival analysis”. Handbook of Statistics, Vol. 27, pp. 281-320.

[18] Logan B. R., Zhang, M. J., and Klein, J. P.(2006). “Regression models for hazard rates versus cumulative incidence probabilities in hematopoietic cell transplantation data”. Biology of Blood Marrow Transplant, Vol. 12, pp. 107-112.

[19] Moeschberger, M. L., Tordoff, K. P., and Kochar, N. (2007). “A Review of Statistical Analyses for Competing Risks”. Handbook of Statistics, Vol. 27, pp. 321-341.

[20] Kass, R. E. and Raftery, A. E. (1995). “Bayes factors”. Journal of American Statistical Association, Vol. 90, pp. 773-795.