https://doi.org/10.21186/IPR.2020.5.1.001

생존분석 모형을 활용한 산업재해 데이터의 분석

백재욱
한국방송통신대학교

Analysis of Industrial Accidents Data with Survival Model

Jaiwook Baik
Korea National Open University

요 약 본 연구에서는 정부정책이 효과가 있었는지 파악하기 위하여 과거 10년간의 산업재해 데이터를 살펴보았다. 이들 데이터로부터 중요한 두 개 또는 세 개의 변수간의 관계를 EDA 방법으로 살펴보았다. 근로자수(사업장규모)와 생존확률 간의 관계를 살펴본 결과 근로자수가 많을수록 시간이 지남에 따라 생존확률이 더욱 더 떨어짐(산업재해가 더 많이 일어남)을 알 수 있다. Cox의 비례위험모형을 적용해본 결과 사업장에서 발생한 총산업재해수가 많을수록 해당 사업장에서 산업재해가 발생할 위험성(hazard)이 높아지고, 근로자수가 적을수록 산업재해가 발생할 위험성이 높으며, 업종별로는 농업, 어업 및 임업이 건설업에 비해 산업재해를 당할 위험성이 더 크다. 공단, 민간 및 고용노동부의 역할은 고용노동부만 효과가 있고, 나머지 두 조직은 효과가 없는 것으로 나온다. recurrent event data를 Cox의 비례위험모델로 분석해본 결과 비슷한 결과가 나온다.

주제어 : 산업재해 데이터, 산업재해방지정책, 생존모형, Cox의 비례위험모형, 재발사건 데이터

Abstract The purpose of this study is to analyze the industrial accidents data with survival model. EDA approach is used to explore the relationship between two variables and among three variables for the past 10 years of industrial accidents data. Survival models are also tried. Survival curve drops more rapidly for the business with fewer employees as time goes by. Industrial accidents occur more often as the total number of industrial accidents gets larger and as the number of employees gets smaller. Agriculture, fishing and forestry have a higher level of industrial accidents than construction while service industry and 'transportation · storage and telecommunication' have a fewer number of industrial accidents than construction. Korea Safety and Health Agency's and Ministry of Employment and Labor's involvement were not effective but Civilian's was. Recurrent event data analysis reveals all most the same result as for non-recurrent data analysis.

Key Words : Industrial Accidents Data, Industrial Accident Prevention Policy, Survival Model, Cox's Proportional Hazard Model, Recurrent Event Data

* 이 논문은 2018년 한국방송통신대학교 학술연구비지원을 받아 작성된 것임.
Received 31 December 2019, Revised 08 January 2020
Accepted 15 January 2020
Corresponding Author: Jaiwook Baik
(Korea National Open University)
Email: jbaik@knou.ac.kr
ISSN: 2466-1139

1. 서론

현대사회는 도시화 및 기술의 발달로 인하여 안락하면서도 복잡한 사회가 구축되고 있다. 이런 복잡한 사회는 위험 또한 치명적일 수 있으며, 이런 고위험은 우리 삶의 다양한 분야까지 스며들어 있다. 이런 위험 중 특히 주목해야 할 것은 산업체 내외의 사건 및 사고이며, 정부에서는 이런 사건 및 사고를 파악하고, 이를 미리 예방하기 위하여 여러 산업재해조사를 실시하고 있다. 이런 산업재해조사는 산업재해의 산업별, 규모별, 지역별, 발생시기별, 원인별 분포와 재해 근로자의 성별, 연령별, 근속기간별 등 특성을 파악하여 산업재해예방 정책을 수립하는 기초자료로 제공되고 있다.

우리나라는 1970년대부터 성장 위주의 경제성장 전략으로 인하여 근로현장에서 산업재해가 발생하는 것을 어쩔 수 없다고 인식했었다([1]). 더욱이 사회장과 기업에서의 산업재해가 근로자 개인의 부주의 때문에 생각다고 하여, 산업재해 발생의 본질을 왜곡시켰으며, 산업재해의 문제를 해결하는 게 더욱 어려웠다([2]).

최근에는 산업재해 보상내용과 보상범위 등 법제도의 정비와 관련된 내용들([3, 4]), 자동차 관련업, 병원, 제조업, 건설업 등에서 특정 직업군의 재해특성을 분석한 연구([5-9]), 일부 특정 사업장의 특성과 재해발생 간의 연관성에 대한 연구([10]), 특정집단의 집단 및 개인적 차원의 재해원인과 관련 예방대책([11, 12]) 등 다각적인 측면에 대한 연구가 진행되었다.

근래에 어떤 연구자들은 2002~2008년까지 산업재해 발생현황을 이용하여, 국내 전체산업에 걸쳐 재해유형별로 산업재해 발생 위험도를 비교・분석하고, 재해발생시간에 대한 수리적 모델인 순환모델에서 모수를 추정하였다([1]).

어떤 연구자는 C4.5 알고리즘을 이용하여 의사결정나무를 만들었고([13]), C4.5 알고리즘을 사용하기 위한 툴로 통계소프트웨어 SAS의 Enterprise Miner를 사용했다. 이외에 CHIAD 알고리즘을 활용하여 정량적 위험성 평가기법이 부재한 제조업에서 산업재해에 대한 정량적 평가가 가능한 특성분석 (feature analysis)을 실시하였다([14, 15]).

데이터 마이닝 기법은 대량의 과거 데이터로부터 미래를 예측하기 위해 활용된다. 기존의 데이터 마이닝 기법은 이동통신사의 이탈 고객 예측, 취업고객 분석 및 예측, 의학적 진단 예측 등에 활용되었다([16]). 하지만 근래에는 데이터 마이닝 기법을 활용하여 강원도 내의 제조업 재해자 총 10,536명의 자료를 바탕으로 사망 및 부상자 수를 예측하기 위하여 의사결정나무 알고리즘인 CHAID, C4.5, CART, 로지스틱회귀(Logistic Regression) 및 신경망(Neural Network) 등의 다양한 기법을 적용했다([17]).

한편, 산업안전보건법에서는 산업재해와 관련한 기본적인 규정을 제시하고 있는데, 이는 이러한 기초적인 규정과 제도를 지키는 것이 산업재해를 예방하는 가장 근본적인 방법이기 때문이다. 이에 어떤 연구자는 t검정과 로지스틱 회귀모델(logistic regression)을 이용하여 산업안전보건법의 이행여부와 산업재해 발생간의 관계를 고찰하였다([18]).

정부에서는 산업재해를 줄여서 보다 안전한 대한민국을 만들기 위해 여러 가지 정책 활동을 시행했다([19, 20]). 그 결과 고용노동부에서는 특정 재해에 대해 시정조치, 고소 등 적절한 활동을 취하고, 산업재해공단의 경우 사업장 단위별로 클린사업, 기술지원 등 여러 가지 예방활동을 펼쳐나가고 있다. 하지만 이런 조치가 제대로 효과를 발휘했는지 보기 위해서 기존에는 빈도분석의 방법으로 검증을 했는데, 최근에는 과거 2006~2015년까지의 산업재해 데이터를 대상으로 case control study, logistic regression, Poisson regression의 방법으로 산업재해에 영향을 미치는 risk factor가 무엇인지 살펴보았다([21]). 본 연구에서는 똑같은 데이터를 가지고 또 다른 분석을 실시하고자 한다. 구체적으로 2절에서는 이들 산업재해 데이터에서 두 변수 간의 관계를 탐색적 자료분석 방법으로 살펴본다. 다음으로 3절에서는 신뢰성 분야에서 많이 활용되는 신뢰성분석 모델을 활용하여 산업재해에 영향을 미치는 risk factor의 영향력을 살펴본다. 마지막으로 4절에서는 지금까지 살펴본 것을 요약하며 추후에 해야 할 일을 기술한다.

2. 산업재해 데이터에 대한 탐색적 자료분석

산업재해와 관련된 데이터는 산업안전보건연구원에서 여러 가지의 형태로 수집하고 분석하고 있다. 본 연

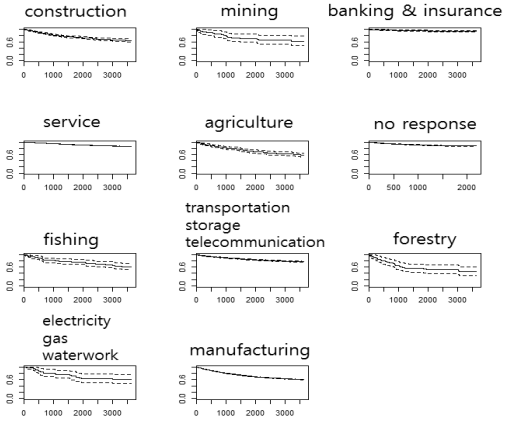
구에서는 이들 여러 데이터를 병합한 하나의 파일을 중심으로 산업재해를 분석하고자 한다. 이 파일에는 2006년도에 사업을 개시한 78,743개의 사업장에서 발생한 산업재해정보가 있다. 구체적으로 사업장별로 사업장관리번호, 사업개시번호, 사업개시일(date_start), 사업종료일(date_end) 및 산업재해 횟수와 각 산업재해일에 대한 정보가 있으며, 2006년부터 2015년까지 각 년도마다

(*) 사업장상태(정상, 소멸), 업종(제조업, 건설업, 서비스업 등), 근로자수 및 재해자수(=사고사망자수+질병사망자수+사고부상자수+질병이환자수)에 대한 정보가 있으며.

(**) 산업안전보건공단(Korea Safety and Health Agency, KSHA), 민간(Civilian) 및 고용노동부(Ministry of Employment and Labour, MEL)가 각 사업장의 안전 및 보건의 확보를 위해 각 사업장에 개입한 횟수에 대한 정보가 있다.

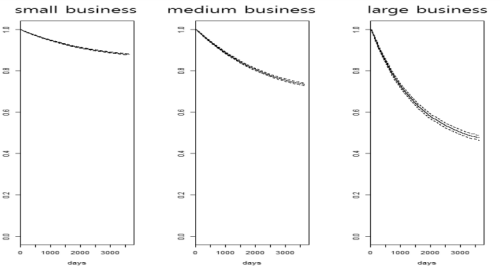
2.1 두 개의 중요 변수들 간의 관계

우선 '첫 번째 산업재해가 일어나기까지 걸린 시간'과 업종과의 관계를 살펴보면 [Fig. 1]과 같다. 이로부터 광업, 농업, 어업, 임업, '전기・가스 및 상수도업' 등에서 시간이 흐를수록 재해가 일어날 가능성은 높아진다. 한편 '금융 및 보험업'과 '기타의 사업(서비스업)'은 다른 업종에 비해 시간이 흘러도 재해의 가능성이 그다지 높지 않다.



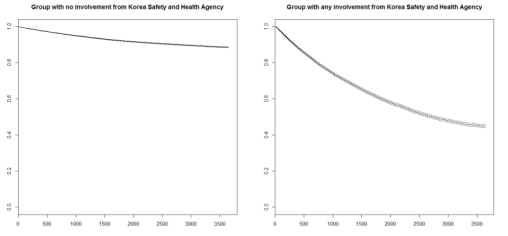
[Fig. 1] Survival probabilities for diverse business categories

다음으로 근로자수(사업장규모)별로 '첫 번째 산업재해가 발생할 때까지의 시간'에 대한 survival probability의 추이가 달라지는지 살펴보면 [Fig. 2]와 같다. 여기에서 근로자수는 편의상 3인 이하이면 소기업, 3인보다 많고 10인 이하이면 중기업, 그리고 10명보다 많으면 대기업이라고 지칭한다(각각의 사업장 수는 51,360개, 20,025개, 7,281개임).



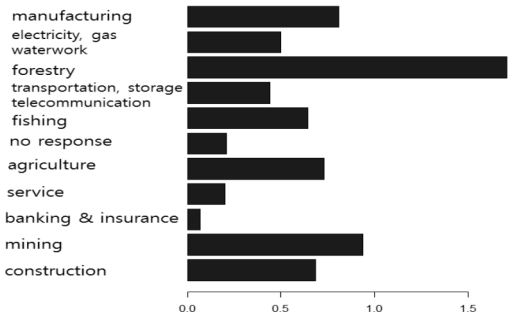
[Fig. 2] Survival probabilities for small, medium and large businesses

한편, 사업장들은 산업안전보건공단, 민간 및 고용노동부로부터 여러 가지 지도를 받을 수 있다. 따라서 각각으로부터 지도받은 group과 지도받지 않은 group의 survival probability의 추이를 파악하고 싶다. 이들 중 산업안전보건공단으로부터 지도받지 않은 group과 한 번이라도 지도받은 group의 survival probability의 추이를 살펴보면 [Fig. 3]과 같다. 이로부터 산업안전보건공단으로부터 한 번이라도 지도를 받은 group이 한 번도 지도를 받지 않은 group에 비해 시간이 흐르면서 산업재해가 훨씬 더 많이 일어난다는 것을 알 수 있다. 산업안전보건공단에서는 산업재해가 많이 일어나는 사업장에 대해 여러 가지 활동을 하는 것으로 판단된다.



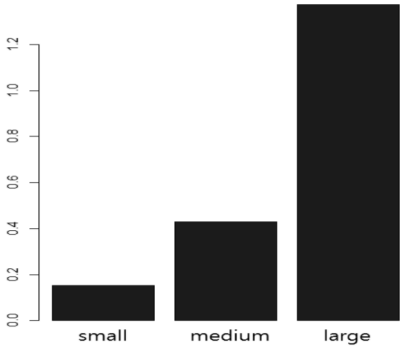
[Fig. 3] Survival probabilities for those who have not had any involvement from Korea Safety and Health Agency and for those who have had

다음으로 산업재해수와 업종간의 관계를 살펴보면 [Fig. 4]와 같다. 이로부터 임업이 다른 업종에 비해 평균재해수가 많으며, 그 다음으로 광업, 제조업, 농업, 건설업인 것을 알 수 있다. 산업재해가 가장 적은 업종은 금융 및 보험업이며, 이어서 '기타의 산업(서비스업)'임을 알 수 있다.



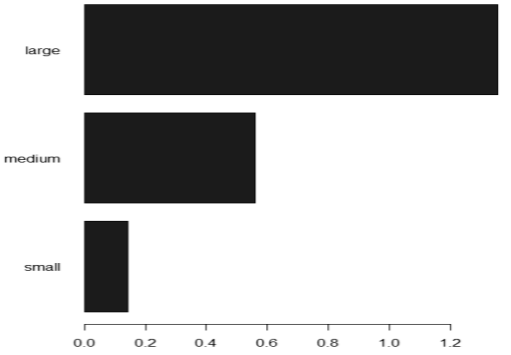
[Fig. 4] Average number of industrial accidents for each business category

다음으로 근로자수(사업장규모)별로 산업재해가 얼마나 많이 일어나는지 알고 싶을 수 있다. 근로자수는 3인 이하의 소기업, 3인보다 많고 10인 이하인 중기업, 그리고 10명보다 많은 대기업으로 편의상 나누어, 각 group에서 산업재해가 얼마나 많이 일어나는지 살펴본 결과 [Fig. 5]와 같다. 이로부터 사업장규모가 커질수록 산업재해가 더 많아짐을 알 수 있다. 사업장규모는 근로자수를 말하는데, 근로자수가 많을수록 산업재해가 더 많이 일어날 수 있음은 당연한 것이다.



[Fig. 5] Average number of industrial accidents for small, medium and large business

다음으로 근로자수와 산업안전보건공단에서 지도받은 횟수와의 관계를 살펴보면 [Fig. 6]과 같다. 이로부터 사업장규모가 클수록 산업안전보건공단으로부터 지도를 더욱 많이 받음을 알 수 있다.

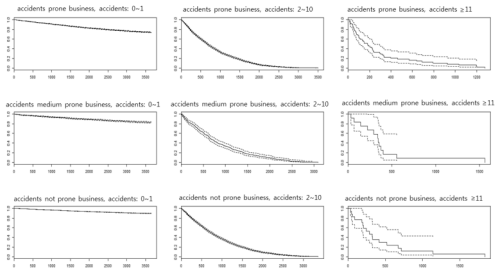


[Fig, 6] Number of involvements from 'Korea Safety and Health Agency' for small, medium and large business

2.2 세 개의 중요 변수들 간의 관계

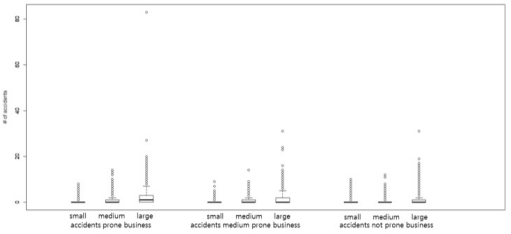
여러 변수들 중 '첫 번째 산업재해가 일어나기까지 걸린 시간'이 다른 변수들의 여러 조건에서 어떤 양상을 보이는지 알고 싶을 수 있다. 첫 번째 예로 업종별로 재해가 가장 많은 업종(임업, 광업, 제조업, 농업), 재해가 보통인 업종(건설업, 어업, '운수・창고 및 통신업', '전기・가스 및 상수도업'), 재해가 가장 적은 업종{기타의 사업(서비스업), 무응답, '금융 및 보험업'}으로 나누고, 그 다음으로 산업재해수를 적음(1회 이하), 보통(1회 보다 많지만 10회 이하), 많음(10회 보다 많음)으로 나누어 총 9개의 범주로 세분화 하여, 각 cell에 속하는 사업장의 '첫 번째 산업재해가 발생할 때까지의 시간'을 survival probability로 나타내고 싶을 수 있다.

[Fig. 7]은 각 cell에 속한 사업장의 survival probability를 나타낸다. 이로부터 재해가 많은 (accident proness) 업종일수록 그리고 재해수 (accident number)가 많을수록 해당 사업장의 survival probability는 시간에 따라 급격히 떨어짐을 알 수 있다. 이는 이들 cell에 속한 사업장의 경우 시간이 흐를수록 산업재해가 더욱 많이 나타날 수 있음을 의미한다.



[Fig. 7] Survival probabilities for each category of accident proness and accident number

다음으로 여러 변수들 중 산업재해횟수와 다른 변수들 간의 관계를 알고 싶을 수 있다. 구체적으로 사업장을 소, 중, 대기업별로 나누고, 그 다음으로 재해 많은 업종, 재해 보통 업종, 재해 적은 업종으로 나누어 총 9개의 범주로 나누고, 각 cell에 속한 사업장에서 산업재해가 얼마나 많이 일어나는지 알고 싶을 수 있다. [Fig. 8]은 각 cell에 속한 사업장의 산업재해수를 나타낸다.



[Fig. 8] Boxplot of the number of industrial accidents for each cetegory of business size and accident proness

이로부터 당연하지만 재해가 많은 업종이 그렇지 않은 업종에 비해 재해가 더욱 많이 일어남을 알 수 있다. 재해가 많이 나는 업종에서는 기업의 규모가 클수록 더욱 더 많은 재해가 일어난다.

3. Cox의 proportional hazard model

어느 사업장에서나 산업재해는 일어날 수 있다. 이 절에서는 사업장에서 첫 번째 산업재해가 일어나기까지 걸리는 시간에 대한 모델링을 어떻게 할 것인지 살펴보고자 한다. 그런데 본 연구를 위해 살펴보는 파일에서는 사업장에서 첫 번째 산업재해가 일어나기까지 걸리는 시간이 명확히 주어져 있지 않다. 더구나 생존분석 데이터가 모두 그러하듯이, 일정 시점까지 관측했는데 그 때까지 산업재해가 일어나지 않은 경우도 많다. 따라서 다음과 같은 방법으로 첫번째 산업재해가 일어나기까지 걸리는 시간 또는 마지막 관측시간을 구하고 해당 시간이 고장시간인지 중도중단시간인지 표시한다. 즉,

t_1=파일에서 COL1의 값(사업개시 후 첫 번째 산업재해가 일어나기까지의 기간)

t_2=사업개시(date_start)부터 사업종료(date_end)까지의 기간

t_3=사업개시(date_start)부터 정상영업(survival2006)까지의 기간이라고 할 때 $t_1 < \min(t_2, t_2)$이면 수명 $t = t_1$이면서 status=1이고 (즉, 고장시간임) $t_1 \geq \min(t_2, t_2)$이면 수명 $t = \min(t_2, t_3)$이면서 status=0(즉, 중도중단시간임)으로 나타낸다.

따라서 본 연구에 주어진 산업재해 데이터의 경우 여러 공변수까지 고려한 데이터베이스는 <Table 1>에서와 같이 구축할 수 있다.

〈Table 1〉 Example of database for Cox's PH analysis

id	t	status	x1	x2	business category	# KSHA involvements	# Civilian involvements	# MEL involvements
1	2481	0	0	0.3	service	NaN	NaN	NaN
2	3408	0	0	4.6	service	0	4	0
3	2770	0	0	0.5	service	NaN	NaN	NaN
4	3497	0	0	2.4	service	NaN	NaN	NaN
5	1743	0	0	0.5	service	NaN	NaN	NaN
6	2921	0	0	1.6	service	0	1	0
7	2190	0	0	0.7	service	NaN	NaN	NaN
8	3503	0	0	4.2	service	0	3	0
9	2172	0	0	1.0	service	NaN	NaN	NaN
10	3651	0	0	8.9	service	0	3	0
11	2190	0	0	0.7	service	NaN	NaN	NaN
12	3650	0	0	1.6	manufa-cturing	NaN	NaN	NaN
13	3561	0	0	2.3	service	0	3	0
14	3651	0	0	2.4	service	0	3	0
15	3378	0	0	2.2	service	0	2	0
16	1053	1	1	3.5	manufa-cturing	0	1	0
17	3642	0	0	1.2	service	NaN	NaN	NaN
18	3561	0	0	0.7	service	NaN	NaN	NaN
19	1979	0	0	0.5	manufa-cturing	NaN	NaN	NaN
20	2740	0	0	1.6	service	0	1	0

Cox의 비례위험모델{proportional hazard(PH) model}은 수명에 대한 실험에서 특정 사건이 일어날 때까지 걸리는 시간에 대해 모델링할 때 많이 적용하는 모델이다. 산업재해가 처음 발생할 위험은 해당 사업장이 처해 있는 업종, 근로자수, 그리고 해당 사업장에서 발생한 총 산업재해수 등에 의해 영향을 받을 수 있다. 본 연구의 경우 Cox의 비례위험모델은 다음과 같이 쓸 수 있다.

$$h(t,x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

여기에서 $h(t,x)$는 공변수 x_1, x_2, ,…, x_p를 고려했을 때의 위험함수(hazard function)이며 $h_0(t)$는 $x_1 = x_2 = ... = x_p = 0$일 때의 기저위험함수(baseline hazard function)이다. 비례위험모델의 특징은 기저위험함수는 다른 공변수에 의존하지 않고 단지 시간 t에만 영향을 받으며, 공변수는 위험함수에 비례적으로 작용한다는 것이다. 현재의 산업재해 데이터에서는 변수들을 다음과 같이 놓을 수 있다.

t: 사업장에서 첫 번째 산업재해가 발생할 때까지 걸리는 시간
status: 1 if 산업재해 발생
: 0 if 산업재해 미발생
x_1: 사업장에서 발생한 총산업재해수
x_2: 사업장 근로자수
x_3: 업종
x_{31}: 광업
x_{32}: 금융 및 보험업
x_{33}: 기타의 사업(서비스업)
x_{34}: 농업
x_{35}: 무응답
x_{36}: 어업
x_{37}: 운수・참고 및 통신업
x_{38}: 임업
x_{39}: 전기・가스 및 상수도업
x_{310}: 제조업
x_4: 산업안전보건공단이 해당 사업장에 지도한 횟수
x_5: 민간이 해당 사업장에 지도한 횟수
x_6: 고용노동부가 해당 사업장에 지도한 횟수

Cox의 PH model에서 모수 β_i들은 다음의 partial likelihood를 최대로 하는 값들이다.

$$L(\beta) = \prod_{i=1}^{n} [\frac{\exp(\beta^T X_i)}{\sum_{t_j \geq t_i} \exp(\beta^T X_j)}]^{\delta_i}$$

여기에서 δ_i는 해당 시점에서 사건이 관측되었으면 1이고, 중도중단되었으면 0의 값을 받는다. partial likelihood라고 하는 이유는 앞의 likelihood는 사건이 관측된 시점에서의 hazard만을 고려하기 때문이다. 앞에서와 같은 likelihood를 가지고 있는 경우 통상적인 최적화 방법대로 patial likelihood에 로그를 취해 최적해(optimal solution)를 찾는다.

Cox의 비례위험모델을 본 연구에 주어진 산업재해 데이터에 적용해본 결과 <Table 2>에서와 같은 표를 얻었다.

〈Table 2〉 Results from Cox's PH analysis

```
Call:
coxph(formula = Surv(var, status) ~ num.jaehae + no_employees +
    business_category + KSHA + Civilian + MEL, method = "breslow")

  n= 47604, number of events= 11812
   (31062 observations deleted due to missingness)
                                                  coef  exp(coef)
 se(coef)     z      Pr(>|z|)
num.jaehae                                    0.2931684  1.3406685
0.0036779  79.712    < 2e-16 ***
no_employees                                 -0.0053485  0.9946658
0.0003956 -13.518    < 2e-16 ***
business_category: mining                     0.3435616  1.4099603
0.4341857   0.791   0.428782
business_category: banking &                  0.0097931  1.0098412
0.5952362   0.016   0.986873
business_category: service                   -0.8831368  0.4134839
0.1499239  -5.891  3.85e-09 ***
business_category: agriculture                0.7776069  2.1762581
0.2050065   3.793   0.000149 ***
business_category: no response                0.2176906  1.2432023
0.1597177   1.363   0.172892
business_category: fishing                    0.7550843  2.1277910
```

```
0.3656492    2.065   0.038918 *
business_category: transportation &        -0.2940770   0.7452191
0.1620473   -1.815   0.069561 .
business_category: forestry                 1.1962685   3.3077511
0.3066680    3.901   9.59e-05 ***
business_category: electricity &            0.4478034   1.5648710
0.3612509    1.240   0.215127
business_category: manufacturing           -0.1024774   0.9025986
0.1497325   -0.684   0.493721
KSHA                                0.2348306  1.2646945  0.0066810
35.149   < 2e-16 ***
Civilian                            0.1338040  1.1431687  0.0083828
15.962   < 2e-16 ***
MEL                                -0.0304860  0.9699740  0.0144277
-2.113   0.034599 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                                    exp(coef)
exp(-coef) lower .95 upper .95
num.jaehae                                          1.3407      0.7459
  1.3310      1.3504
no_employees                                        0.9947      1.0054
  0.9939      0.9954
business_category: mining                           1.4100      0.7092
  0.6020      3.3021
business_category: banking &                        1.0098      0.9903
  0.3145      3.2428
business_category: service                          0.4135      2.4185
  0.3082      0.5547
business_category: agriculture                      2.1763      0.4595
  1.4562      3.2525
business_category: no response                      1.2432      0.8044
  0.9091      1.7002
business_category: fishing                          2.1278      0.4700
  1.0392      4.3568
business_category: transportation &                 0.7452      1.3419
  0.5424      1.0238
business_category: forestry                         3.3078      0.3023
  1.8134      6.0335
business_category: electricity &                    1.5649      0.6390
  0.7709      3.1767
business_category: manufacturing                    0.9026      1.1079
  0.6730      1.2105
KSHA                                                1.2647      0.7907
  1.2482      1.2814
Civilian                                            1.1432      0.8748
  1.1245      1.1621
MEL                                                 0.9700      1.0310
  0.9429      0.9978

Concordance= 0.846  (se = 0.003 )
Rsquare= 0.243   (max possible= 0.995 )
Likelihood ratio test= 13224  on 15 df,   p=0
Wald test            = 22982  on 15 df,   p=0
Score (logrank) test = 69807  on 15 df,   p=0
```

이로부터 적합모델은 다음과 같이 쓸 수 있다.

$$h(t,x) = h_0(t)\exp(0.29x_1 - 0.005x_2 - 0.883x_{33} + 0.778x_{34} + 0.755x_{36} - 0.294x_{37} + 0.235x_4 + 0.134x_5 - 0.03x_6)$$

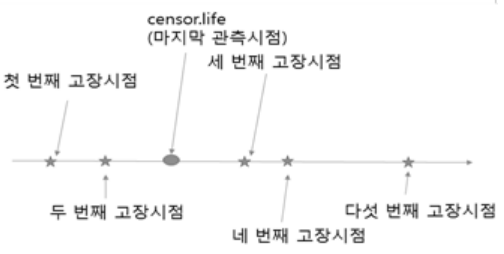
x_1의 계수가 양(+)의 값을 가지므로 사업장에서 발생한 총산업재해수가 많을수록 해당 사업장에서 첫 번째 산업재해를 당할 위험성(hazard)은 커짐을 알 수 있다. x_1의 경우 hazard ratio는 $\exp(0.32)$=1.38로 산업재해수가 1번 증가하면 산업재해가 발생할 위험률은 38% 증가함을 알 수 있다(95% 신뢰구간은 1.36부터 1.39까지임). 다음으로 x_2의 계수는 음(-)의 값을 가지므로 앞에서 양(+)의 값을 갖는 경우와 달리 근로자수가 적을수록 해당 사업장이 산업재해를 당할 위험성이 커지고 있다. 다음으로 업종별로는 농업, 어업 및 임업이 건설업에 비해 산업재해를 당할 위험성이 더 크며, '기타의 사업(서비스업)'과 '운수・창고 및 통신업'이 건설업에 비해 산업재해를 당할 위험성이 더 작다. 마지막으로 x_4와 x_5의 경우 회귀계수가 양(+)의 값을 가지므로 모델대로 해석하면 공단 및 민간으로부터 지도가 많으면 많을수록 해당 사업장은 산업재해를 당할 위험성이 더욱 커진다는 것을 알 수 있다. 하지만 이런 해석은 현실적으로 문제가 있을 것이다. 대신, 공단 및 민간은 산업재해를 당하는 사업장 위주로 지도를 많이 나가는 것으로 보아야 할 것이다. 한편, x_6의 경우 회귀계수가 음(-)의 값을 가지므로 고용노동부의 지도가 많으면 많을수록 산업재해를 당할 위험성이 줄어듬을 알 수 있다.

4. Recurrent event data analysis

사업장의 입장에서는 사업기간 동안에 재해가 한 번 이상 발생할 수 있다. 사실 전체 78,743개의 사업장 중에서 2006년부터 2015년까지 10년간 64,408개의 사업장에서는 산업재해가 한 번도 일어나지 않았지만 8,750개의 사업장에서는 산업재해가 한 번 일어났고, 2,892개의 사업장에서는 산업재해가 2번 일어났으며, 1,202개의 사업장에서는 산업재해가 3번 일어났다. 따라서 첫 번째 산

업재해가 일어날 때까지의 시간뿐만 아니라 순차적으로 일어나는 시간까지 고려할 수 있는 recurrent event data 분석방법이 필요하다. 본 연구에서는 사업장에서 산업재해가 5회까지 일어나는 시간만을 추적한다. 참고로 10년간 산업재해가 5회까지 일어난 사업장은 78,156개의 사업장으로 전체 78,743개 사업장의 99.35%를 차지한다.

이런 recurrent event data의 경우에도 Cox의 비례위험모델(proportional hazard model)을 적용할 수 있다. Cox의 비례위험모델을 적용하기 위해서는 우선 원래의 데이터를 사업장별로 산업재해 순서대로 정렬해야 한다. 예를 들어 3,456번째 사업장은 [Fig. 9]에서와 같이 사업개시 후 1,957일째에 첫 번째 산업재해가 일어났고, 사업개시 후 2,089일째에 두 번째 산업재해가 일어났고, 사업개시 후 2,189일이 될 때까지 관측이 된 경우 recurrent event data로는 <Table 3>에서와 같이 표현한다.



[Fig. 9] Cases where industrial accidents happened twice after start-up

〈Table 3〉 Data array of cases where industrial accidents happened twice after start-up

id	order	event	start	stop
3456	1	1	0	1957
3456	2	1	1957	2089
3456	3	0	2089	2189

본 연구에서 살펴보는 산업재해 데이터의 경우 recurrent event data 분석을 하기 위해 만든 데이터베이스는 <Table 4>에서와 같다. 여기에서 id는 사업장을 가리키며, interval은 산업재해 순서, start는 해당 순서의 시작, stop은 해당 순서의 끝을 나타낸다. 예를 들어 16번째 사업장의 경우 사업개시 후 1,053일이 지난 시점에서 산업재해가 한 번 일어났고, 그 이후 3,347일까지 관측되었으나 산업재해는 더 이상 일어나지 않았다. 16번째 사업장의 경우 공단과 고용노동부로부터 지도를 받은 적 없으나 민간으로부터는 지도를 1번 받았다. 16번째 사업장은 사업기간 중 산업재해를 1번 당했으며, 이 사업장의 업종은 제조업이다.

〈Table 4〉 Example of recurrent event data

id	event	intti me	inter val	start	stop	# KSHA involve ments	# Civilian involve ments	# MEL involve ments	total # indus- trial accid- ents	# of workers	busi ness cate gory
1	0	2481	1	0	2481	NaN	NaN	NaN	0	0.2727273	service
2	0	3408	1	0	3408	0	4	0	0	4.1818182	service
3	0	2770	1	0	2770	NaN	NaN	NaN	0	0.4545455	service
4	0	3497	1	0	3497	NaN	NaN	NaN	0	2.1818182	service
5	0	1743	1	0	1743	NaN	NaN	NaN	0	0.4545455	service
6	0	2921	1	0	2921	0	1	0	0	1.4545455	service
7	0	2190	1	0	2190	NaN	NaN	NaN	0	0.6363636	service
8	0	3503	1	0	3503	0	3	0	0	3.8181818	service
9	0	2172	1	0	2172	NaN	NaN	NaN	0	0.9090909	service
10	0	3651	1	0	3651	0	3	0	0	8.0909091	service
11	0	2190	1	0	2190	NaN	NaN	NaN	0	0.6363636	service
12	0	3650	1	0	3650	NaN	NaN	NaN	0	1.4545455	manufa cturing
13	0	3561	1	0	3561	0	3	0	0	2.0909091	service
14	0	3651	1	0	3651	0	3	0	0	2.1818182	service
15	0	3378	1	0	3378	0	2	0	0	2	service
16	1	1053	1	0	1053	0	1	0	1	3.1818182	manufa cturing
16	0	2294	2	1053	3347	0	1	0	1	3.1818182	manufa cturing
17	0	3642	1	0	3642	NaN	NaN	NaN	0	1.0909091	service
18	0	3561	1	0	3561	NaN	NaN	NaN	0	0.6363636	service
19	0	1979	1	0	1979	NaN	NaN	NaN	0	0.4545455	manufa cturing

recurrent event data라고 하더라도 사용하는 생존분석 방법은 Cox의 비례위험모델이므로 위험함수는 3절에서 제시한 모형 $h(t,x)=h_0(t)\exp(\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)$와 똑같다. <Table 4>의 recurrent event data에 대해 Cox의 비례위험모델을 적용해본 결과 <Table 5>에서와 같이 나오며, 따라서 적합 모델은 다음과 같다.

$$h(t,x)=h_0(t)\exp(0.12x_1+0.001x_2+0.683x_{31}-1.04x_{33}+0.391x_{34}+0.666x_{36}+0.979x_{38}-1.027x_{39}-0.398x_{310}+0.246x_4+0.088x_5+0.064x_6)$$

〈Table 5〉 Results from recurrent event data analysis

```
Call:
coxph(formula = Surv(start, stop, event) ~ num.jae + no_emp +
    business_cat + KSHA + Civilian + MEL, data = data.recur, method = "breslow")

  n= 68212, number of events= 21111
   (34139 observations deleted due to missingness)

                                        coef exp(coef)  se(coef)      z Pr(>|z|)
num.jae                            0.1243619 1.1324256 0.0015705 79.186  < 2e-16 ***
no_emp                             0.0012259 1.0012267 0.0001837  6.675 2.48e-11 ***
business_cat: mining               0.6827178 1.9792496 0.2362998  2.889 0.003862 **
business_cat: banking &           -0.9353386 0.3924530 0.5822515 -1.606 0.108182
business_cat: service             -1.0421860 0.3526829 0.0775847 -13.433  < 2e-16 ***
business_cat: agriculture          0.3905235 1.4777542 0.1274927  3.063 0.002191 **
business_cat: no response          0.0195409 1.0197330 0.0880612  0.222 0.824391
business_cat: fishing              0.6664663 1.9473439 0.2222970  2.998 0.002717 **
business_cat: transportation &    -0.0818879 0.9213752 0.0878333 -0.932 0.351176
business_cat: forestry             0.9789744 2.6617250 0.1741082  5.623 1.88e-08 ***
business_cat: electricity &       -1.0266541 0.3582035 0.3106852 -3.304 0.000952 ***
business_cat: manufacturing       -0.3977024 0.6718619 0.0774014 -5.138 2.77e-07 ***
KSHA                               0.2456366 1.2784349 0.0045648 53.811  < 2e-16 ***
Civilian                           0.0880828 1.0920785 0.0062914 14.000  < 2e-16 ***
MEL                                0.0641474 1.0662496 0.0083757  7.659 1.88e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                                   exp(coef) exp(-coef) lower .95 upper .95
num.jae                               1.1324     0.8831    1.1289    1.1359
no_emp                                1.0012     0.9988    1.0009    1.0016
business_cat: mining                  1.9792     0.5052    1.2456    3.1451
business_cat: banking &               0.3925     2.5481    0.1254    1.2286
business_cat: service                 0.3527     2.8354    0.3029    0.4106
business_cat: agriculture             1.4778     0.6767    1.1510    1.8972
business_cat: no response             1.0197     0.9806    0.8581    1.2118
business_cat: fishing                 1.9473     0.5135    1.2596    3.0107
business_cat: transportation &        0.9214     1.0853    0.7757    1.0945
business_cat: forestry                2.6617     0.3757    1.8922    3.7442
business_cat: electricity &           0.3582     2.7917    0.1948    0.6585
business_cat: manufacturing           0.6719     1.4884    0.5773    0.7819
KSHA                                  1.2784     0.7822    1.2670    1.2899
Civilian                              1.0921     0.9157    1.0787    1.1056
MEL                                   1.0662     0.9379    1.0489    1.0839

Concordance= 0.824  (se = 0.002 )
Rsquare= 0.292   (max possible= 0.999 )
Likelihood ratio test= 23538  on 15 df,   p=0
Wald test            = 30571  on 15 df,   p=0
Score (logrank) test = 119509  on 15 df,   p=0
```

<Table 5>로부터 x_1의 계수가 양(+)의 값을 가지므로 사업장에서 발생한 총산업재해수가 많을수록 해당 사업장에서 산업재해를 당할 위험성은 커짐을 알 수 있다. x_1의 경우 hazard ratio는 $\exp(0.124)$=1.132로 산업재해수가 1번 증가하면 산업재해가 발생할 위험률은 13.2% 증가함을 알 수 있다(95% 신뢰구간은 1.129부터 1.136까지임).

다음으로 x_2의 계수 또한 양(+)의 값을 가지므로 사업장에서 근로자수가 많을수록 산업재해를 당할 위험성은 커진다는 것을 알 수 있다. 하지만 hazard ratio가 $\exp(0.0012259)$=1.0012이고, 95% 신뢰구간이 (1.0009, 1.0016)로 거의 1에 가까우므로 현실적으로는 근로자수는 산업재해가 큰 관련이 없는 것으로 보인다.

다음으로 업종 중에서는 광업, 농업, 어업 및 임업이 건설업에 비해 산업재해를 당할 위험이 더 크며, ‘기타의 사업(서비스업)’, ‘전기・가스 및 상수도업’ 및 제조업이 건설업에 비해 산업재해를 당할 위험성이 더 작다.

마지막으로 x_4, x_5, x_6의 경우 회귀계수가 모두 양(+)의 값을 가지므로 변수대로 해석하면 공단, 민간 및 고용노동부로부터 지도가 많을수록 해당 사업장은 산업

재해를 당할 위험성이 더욱 커진다는 것을 뜻한다. 하지만 이런 해석은 현실적으로 문제가 있으므로 앞의 다른 분석에서와 마찬가지로 산업재해를 당할 위험성이 많은 곳을 공단, 민간 및 고용노동부가 지도를 많이 나가는 것으로 보는 것이 더 타당할 것이다.

5. 결론 및 추후 연구

본 연구에서는 정부정책이 효과가 있었는지 파악하기 위하여 과거 10년간의 산업재해 데이터를 살펴보았다. 이들 데이터로부터 중요한 두 개 또는 세 개의 변수간의 관계를 탐색적 자료분석방법으로 살펴보았다. 두 개의 변수들간의 관계를 예로 들면, 근로자수(사업장규모)와 생존확률 간의 관계를 살펴본 결과 근로자수가 많을수록 시간이 지남에 따라 생존확률이 더욱 더 떨어짐(산업재해가 더 많이 일어남)을 알 수 있다. 세 변수들간의 관계를 예로 들면, 사업장규모별 그리고 재해 다수 업종별 산업재해횟수를 boxplot으로 그려본 결과, 대기업일수록 그리고 재해가 많이 나는 업종일수록 산업재해가 더욱 많이 일어난다는 것을 알 수 있다.

다음으로 산업재해 데이터에 대해 Cox의 비례위험모형을 적용해본 결과 사업장에서 발생한 총산업재해수가 많을수록 해당 사업장에서 산업재해가 발생할 위험성(hazard)이 높아지고, 근로자수가 적을수록 산업재해가 발생할 위험성이 높으며, 업종별로는 농업, 어업 및 임업이 건설업에 비해 산업재해를 당할 위험성이 더 크며, '기타의 사업(서비스업)'과 '운수・창고 및 통신업'이 건설업에 비해 산업재해를 당할 위험성이 더 작다. 공단, 민간 및 고용노동부의 역할은 고용노동부만 효과가 있고, 나머지 두 조직은 효과가 없는 것으로 나온다.

마지막으로 recurrent event data를 Cox의 비례위험모델로 분석해본 결과 사업장에서 발생한 총산업재해수가 많을수록 그리고 근로자수가 많을수록 해당 사업장에서 연속적으로 산업재해가 발생할 위험성(hazard)이 높아지고, 업종별로는 광업, 농업, 어업 및 임업이 건설업에 비해 산업재해를 당할 위험성이 더 크며, '기타의 사업(서비스업)', '전기・가스 및 상수도업' 및 제조업이 건설업에 비해 산업재해를 당할 위험성이 더 작다. 산업안전보건공단, 민간 및 고용노동부는 산업재해가 보다 많이 일어나는 사업장에 더욱 더 많은 지도를 하는 것으로 판단된다.

지금까지 산업재해 데이터베이스에 들어간 정보는 재해가 일어날 때까지 걸린 시간, 재해자가 속한 사업장 정보(업종, 근로자수 등) 및 '공단, 민간, 고용노동부의 총지도횟수' 등이다. 이외에 재해자에 대한 정보, 사고와 관련된 상황 정보, '공단, 민간, 고용노동부의 정책활동 내역 및 실시일', 사업장 지리정보 등 산업재해와 관련된 정보 등을 산업재해 데이터베이스로 구축할 수 있다면 보다 다양한 분석을 할 수 있을 것이다.

다음으로 지금까지의 분석에서는 산업안전보건공단, 민간 및 고용노동부의 지도가 효과가 있었는지 판단할 때 과거 10년간 누적지도횟수를 가지고 판단하였다. 하지만 총 관측기간을 각 기관의 지도 전과 후로 나누어 산업재해의 발생여부 또는 빈도 등을 통계적인 방법으로 비교한다든지 또는 의학통계에서 repeated measure에 대한 효과 분석방법을 이용하여 각 기관의 지도가 효과가 있었는지 판단할 수도 있을 것이다.

다음으로 지금까지는 사업장에 대한 분석을 실시했는데, 새로운 산업재해 데이터베이스에 사람 위주로 산업재해와 관련된 정보를 데이터베이스화 할 수 있다면 지금까지 사업장을 대상으로 실행해온 분석을 사람을 대상으로 실행하여 여러 가지 유용한 정보들을 얻을 수 있을 것이다. 아울러 각 사업장 또는 각 사람의 지리정보까지 산업재해 데이터베이스에 넣을 수 있다면 공간분석까지도 가능할 것이다.

마지막으로 본 프로젝트는 주로 산업예방정책이 효과가 있었는지 판단할 수 있는 모델의 구축에 중점을 두었는데, 이런 모델을 더욱 발전시켜 tree 구조의 모델 또는 support vector machine을 만들어 산업재해가 가장 잘 일어날 것 같은 사업체 또는 사람을 산업재해가 발생하기 전에 미리 발견(예측)하여 그 사업체 또는 사람에게 feedback을 사전에 줄 수 있는 시스템을 구축할 수 있다면 더욱 좋을 것이다.

References

[1] Kim, H. Y. and Heo, T. Y. (2010). "An Analysis of relative injury risk by industry and estimation of a circular distribution model for industrial injury". Seoul City Research, 11, 127-138.

[2] Ju, J. H. (1997). “An analysis of factors and structure affecting industrial accidents in Korea”. Kyungsung University Ph.D. Dissertation.
[3] Kim, H. S. (2008). “A Study on the relations between industrial accident insurance and the automobile insurance”. Labour Law, 26, 303-325.
[4] Park, J. S. (2006). “Occupational accidents due to colleague’s abusive act and right to reimbursement of industrial accidents compensation insurance”. Labour Law, 22, 363-386.
[5] Kim, S. K (1998). “A status of the report for industrial injuries and illnesses at an automobile related plant”. Annals of Occupational and Environmental Medicine, 10, 562-570.
[6] Lee, C. J., Jun, Y. U., Choi, Y. H. and Jo, A. (2002). “Causes and preventive measures for low back pain industrial accidents suffered by automobile assembly workers”. Korea Ergonomics Society Conference Proceeding, 119-123.
[7] Lee, J. C., Shin, S. W. and Lee C. S. (2007). “Accident analysis of middle-aged & advanced-aged construction workers”. Korea Architecture Association Conference Proceeding(Structural System), 27, 797-800.
[8] Lee, K. S. and Jung, B. Y. (2008). “An analysis of industrial accidents in small-scale fiber business”. Korea Ergonomics Society Spring Conference Proceeding, 252-255.
[9] Park, H. J. (2007). “Research on industrial disaster in hospital”. Korea Ergonomics Society Fall Conference Proceeding, 492-495.
[10] Lee, G. S. et al (2006). “Relationship between Injury Occurrence and Workplace Organization in Small-sized Manufacturing Factories”. Korea Industrial Medicine Society, 18, 73-86.
[11] Lee, S. W., Kim, K. S. and Kim, T. W. (2008). “The status and characteristics of industrial accidents for migrant workers in Korea compared with native workers”. Korea Industrial Medicine Society, 20, 351-361.
[12] Kim, H. H. et al (2009). “An analysis of characteristics of musculoskeletal disorders risk factors”. Korea Ergonomics Society, 28, 17-25.
[13] Leem, Y. M, Kwag, J. K. and Hwang, Y. S. (2005). “A feature analysis of industrial accidents using C4.5 algorithm”. Korea Safety Society, 20, 130-137.
[14] Leem, Y. M. and Hwang, Y. S. (2006). “Data Analysis of Industrial Accidents in Manufacturing Industries Using CHIAD Algorithm”. Korea Safety Management Society/Korea Cyber Terrorism Information Transfer Society Spring Conference Proceeding, 45-50.
[15] Song, J. M. and Yoon, S. U. (2004). “A study on split selection algorithms in decision tree”. Yonsei University Master’s Thesis.
[16] Lee, K. N. and Lee, H. C. (2003). “A Study on the combined decision tree(C4.5) and neural network algorithm for classification of mobile telecommunication customer”. Korea Intelligence Information System Society, 9, 139-155.
[17] Leem, Y. M. and Ryu, C. H. (2006). “A comparison of data mining techniques for predicting model of industrial accidents”. Korea Industrial Management System Society Conference Proceeding, 107-113.
[18] Jung, W. I. and Jun, Y. I. (2014). “Working conditions and industrial accidents in accordance with safety and health environment in the workplace”. Korea Crisis Management Society, 10, 323-344.
[19] Korea Labor Institute. (2018). Report on Labor Force survey at establishments.
[20] Baek, E. M. and Jung, H. S. (2019). “A study of factors impacting work-related health problems in different work-hour groups”, Journal of Korean Society Occupational Environmental Hygiene, 29, 383-393.
[21] Kim, Y. S., Jo, J. N. and Baik, J. W. (2016). “Comparative study of working conditions of Korea and Europe”, Journal of the Korean Data & Information science Society, 27, 1-21.

백 재 욱(Baik, Jaiwook)

・중앙대학교 응용통계학과 학사
・미국 Virginia Polytechnic Institute and State University 통계학박사
・2019년 12월 현재 : 한국방송통신대학교 정보통계학 교수
・관심분야 : 통계학, 생산관리
・E-Mail : jbaik@knou.ac.kr