Women and children first. Is that all to survive?

2020020344 통계학과 정해성

· 서 론

1912년 발생한 타이타닉 침몰 사고는 2224명의 탑승 인원 중 1514명이 사망하고 생존자가 711명에 불과한 역대 최악의 한 해상 사고로 꼽힌다. 이 사고를 바탕으로 제작한 영화 타이타닉에서는 여자와 아이들을 먼저 구출하는 장면이 나올 정도로 실제 타이타닉호 침몰 당시 선장과 승무원이 여성과 어린이를 먼저 구명정에 태운 사실은 널리 알려져 있다. 이번 프로젝트를 통해 타이타닉호 탑승객 2201명에 대한 데이터를 분석하여 실제 여자와 아이들의 생존율이 통계적으로 유의하게 높은지 알아보고자 한다. 또한, 경제력을 간접적으로 보여주는 econo.class 변수를 이용해 경제력이 생존 여부에도 영향을 미치는지, 만약 영향을 미친다면 성별과 비교해 생존 여부에 어떠한 방식으로 영향을 미치는지 알아보고자 한다.

· EDA

먼저 survived와 econo.class, age, sex 각각의 변수들에 대하여 교차분석표를 사용해 생존율을 살펴본 뒤, 독립성 검정을 시행하고자 한다. 이때 검정 통계량은 피어슨 카이제곱 통계량을 사용했다. 또한, econo.class 변수를 순서형 변수로 여기고 맨텔-헨젤 통계량(Linear Trend alternative to Independence; $M^2 = (n-1)r^2$)도 사용해 분석해보았다.

(1) survived

survived		
0	1	total
1490	711	2201

전체 2201명의 탑승객 중 1490명은 사망, 711명은 생존했다.

(2) econo.class와 survived

	survived		
econo.class	0	1	total
1	122(0.38)	203(0.62)	325
2	167(0.59)	118(0.41)	285
3	528(0.75)	178(0.25)	706
4	673(0.76)	212(0.24)	885

count	age	
econo.class	0	1
1	6	319
2	24	261
3	79	627
4	0	885

$\chi^2 = 190.4$ (df=3), p-value < 2.2e-16

* 첫 번째는 빈도, 괄호 안은 econo.class 별 survived 비율

위 표는 econo.class와 survived 간의 빈도표를 나타낸다. 여기서 주목할 점은 econo.class=4에는 어린 아이가 없다는 점이다. 이는 추후 로그선형모형 및 로지스틱모형 적합 시 유의해야 할 부분이다.

독립성 검정 결과, 피어슨 카이제곱 통계량 값이 190.4(df=3)을 나타냈다. 즉, p-value는 유의수준 0.05보다 매우 작은 값이며 'H_0 : econo.class와 survived는 서로 독립이다'를 기각한다. 따라서, econo.class에 따라 survived 여부에 연관성이 있다.

한편, econo.class는 좌석클래스를 나타내는 변수이므로 1>2>3>4의 순서를 갖는 순서형 변수로 여기고 멘텔-헨젤 통계량을 이용해 독립성 검정을 수행해볼 수 있다. econo.class의 score를 (1,2,3,4)로 부여했을 때의 통계량 $M^2=(n-1)r^2=(2201-1)(-0.2713954)^2=162.042$가 되며, $\chi^2_{1,0.05}=3.841$과 비교했을 때 매우 큰 값이므로 'H_0 : econo.class와 survived는 서로 독립이다'를 기각한다. 즉, nonzero correlation이 존재한다는 또한 $r^2=-0.2713954$로 약한 음의 선형 관계가 존재한다고 해석할 수 있다.

(3) age와 survived

	survived		
age	0	1	total
0	52(0.48)	57(0.52)	109
1	1438(0.69)	654(0.31)	2092

$\chi^2=20.005$ (df=1), p-value = 7.725e-06

* 첫 번째는 빈도, 괄호 안은 age 별 survived 비율

위 표는 age와 survived 간의 빈도표를 나타낸다. 독립성 검정 결과, 피어슨 카이제곱 통계량 값이 20.005(df=1)을 나타냈다. 즉, p-value는 유의수준 0.05보다 매우 작은 값이며 'H_0 : age와 survived는 서로 독립이다'를 기각한다. 따라서, age와 survived 간 연관성이 있다.

(4) sex와 survived

	survived		
sex	0	1	total
0	126(0.27)	344(0.73)	470
1	1364(0.79)	367(0.21)	1731

$\chi^2=454.5$ (df=1), p-value < 2.2e-16

* 첫 번째는 빈도, 괄호 안은 sex 별 survived 비율

위 표는 sex와 survived 간의 빈도표를 나타낸다. 독립성 검정 결과, 피어슨 카이제곱 통계량 값이 454.5(df=1)을 나타냈다. 즉, p-value는 유의수준 0.05보다 매우 작은 값이며 'H_0 : sex와 survived는 서로 독립이다'를 기각한다. 따라서, sex와 survived 사이에 연관성이 있다.

이처럼 생존 여부는 econo.class, age, 그리고 sex와 모두 관련이 있다. 특히 교차분석표를 통해 여성이 남성보다, 어린이가 어른보다, 그리고 좌석클래스가 높을수록 생존확률이 높아 보인다. 이러한 EDA의 결과를 토대로 첫째, 로그 선형 모형을 통해 성별과 나이, 좌석클래스 그리고 생존 여부 사이의 연관성과 교호작용을 보다 깊게 살펴보고 둘째, 성별과 나이에 경제적인 요인을 포함하여 생존 여부를 설명하는 로지스틱 모형을 선택한 뒤 변수간 연관성을 살펴보고자 한다.

· 3원 분할표에 대한 로그 선형 모형

(1) 로그 선형 모형에서 적합도의 측도로 피어슨 카이제곱 통계량과 LR 통계량, 그리고 AIC를 기준으로 정리한 표는 다음과 같다. 여기서 표현의 편의를 위해 econo.class: E, age: A, sex: G, survived: S로 표기하였다.

<모형 적합 결과>

로그선형모형	$Pearson's\text{-}\chi^2$	$G^2(M)(df)$	$G^2(M_2\|M_1)(df)$	$\chi^2_{0.05}(df)$
(AGS)	0	0		
(AG, AS, GS)	18.80	16.32(1)		3.84
(AG, AS)	463.89	437.12(2)	420.8(1)	3.84
(AG, GS)	26.18	22.21(2)	5.89(1)	
(AS, GS)	40.90	25.94(2)	9.62(1)	
(A, GS)	50.23	45.50(3)	29.18(2)	5.99
(G, AS)	475.74	460.40(3)	444.08(2)	
(S, AG)	477.28	456.68(3)	440.36(2)	
(A, G, S)	505.09	479.96(4)	463.64(3)	7.82

위 표에는 적합도 검정 통계량인 피어슨의 카이제곱 통계량과 우도비 검정 통계량 $G^2(M)$를 제시했다. 또한, 비교하고자 하는 모형을 M_1 current model을 M_2라고 할 때, 두 모형의 비교를 위한 $G^2(M_2|M_1)$ 통계량 값이 나와 있다. saturated model인 (AGS)와 homogeneous association model인 (AG, AS, GS)의 검정 결과를 비롯한 이후의 모든 축소모형의 검정 결과가 임계치인 $\chi^2_{0.05}(df)$보다 크다. 즉 축소된 모형이 모두 적합하지 않다. 그러나 해석의 편의성을 위해 포화모형을 제외한 모형 중 2요인 교호작용만을 고려한 모형을 선택하였다.

주된 관심은 교호작용인 age와 sex, age와 survived, 그리고 sex와 survived간의 해석이다. 그러므로 본 과제에서는 주 효과에 대한 해석은 생략하고 관심 변수 모수에 대한 해석을 통해 연관성을 살펴보고자 한다. 다음 표는 각 교호작용에 대한 모수 추정값을 보여준다.

	sex	
age	0	1
0	0.1819	−0.1819
1	−0.1819	0.1819

	survived	
age	0	1
0	−0.1392	0.1392
1	0.1392	−0.1392

	survived	
sex	0	1
0	−0.5735	0.5735
1	0.5735	−0.5735

2요인 교호작용들의 계수 추정값들을 살펴보면, $\lambda_{11}^{AG} = 0.1819$로, 어린아이의 여성 비율이 높으며 $\lambda_{11}^{AS} = -0.1392$, $\lambda_{11}^{GS} = -0.5735$로, 각각 나이가 어린 집단, 성별이 여성일 때 사망하지 않을 경향이 있다.

그러므로 실제 데이터 역시 알려진 바와 같이 여성이고 어린아이일수록 생존하는 경향이 존재한다는 것을 뒷받침하고 있다.

· 4원 분할표에 대한 로그 선형 모형

주어진 titanic.data를 4차원의 분할표로 나타냈을 때의 문제점은 0인 칸이 많으며 이러한 칸의 비율이 8/32=0.25 즉, 25%에 달한다. 따라서 로그 선형 모형의 최대우도 추정량과 같은 카이제곱 검정 통계량 값에 영향을 미치며, 포화 모형을 비롯한 3요인과 일부 2요인 교호작용일 포함한 모형의 경우 모수 추정값이 생성되지 않았다. 하지만, survived와 나머지 변수 간의 관계를 알아보는 것이 분석의 목표인 점과 해석의 간결성을 위해 2요인 교호작용항인 (ES, AS, GS)만을 포함한 로그 선형 모형을 적합하고 결과를 해석하고자 한다. 2요인 교호작용들에 대한 모수는 다음과 같다.

	survived	
econo. class	0	1
1	−0.5147	0.5147
2	−0.0864	0.0864
3	0.2836	−0.2836
4	0.3175	−0.3175

	survived	
age	0	1
1	−0.2199	0.2199
2	0.2199	−0.2199

	survived	
sex	0	1
1	−0.5793	0.5793
2	0.5793	−0.5793

앞서 살펴본 바와 같이 age와 sex의 수준에 따라 survived의 경향이 달라지는 것을 확인할 수 있다. 또한, econo.class와 survived 간 교호작용에 대한 계수 추정값을 살펴보면, $\lambda_{11}^{ES} = 0.5147$, $\lambda_{21}^{ES} = 0.0864$로 econo.class가 1과 2일 때는 생존자일 경향이 높으며, 반대로 $\lambda_{31}^{ES} = -0.2836$, $\lambda_{41}^{ES} = -0.3175$로 econo.class가 3과 4일 때는 사망자일 경향을 보인다.

따라서, 성별과 나이 외에 경제적 요인 역시 생존 여부에 영향을 미치는 중요한 요인임을 확인했다.

· 로지스틱 회귀 모형

앞서 성별과 나이, 그리고 경제 지위가 생존 여부에 영향을 준다는 것을 확인했다. 특히 여성과 아이일수록 생존율이 높다는 사실은 영화를 통해 확인할 수 있듯 널리 알려진 사실이다. 그러나, '같은 여성일지라도 경제 지위에 따라 생존 여부가 다르지 않을까'라는 의문이 들 수 있다. 이를 로지스틱 회귀 모형을 통해 알아보고자 한다. 그러므로 age, sex, econo.class를 독립변수로 갖고 종속변수는 survived인 로지스틱 회귀 모형을 고려하기로 한다. 이때 EDA에서 살펴보았듯 econo.class=4에는 age=0의 도수가 존재하지 않으므로 age와 econo.class의 교호작용항도 제거한 모형들만 고려한다.

Model	$\mathrm{Pr}edictors^{*}$	Deviance G^2	df	AIC	model compared	Deviance Difference
1	(A*G+G*E)	37.263	4	103.88		
2a	(A+G*E)	45.899	5	110.52	(2a)−(1)	8.637(df=1)
2b	(E+A*G)	94.548	7	155.17	(2b)−(1)	57.286(df=3)
3	(A+G+E)	112.57	8	171.19	(3)−(2a) (3)−(2b)	66.667(df=3) 18.018(df=1)

* A: age, G: sex, E: econo.class

AIC 기준에서 2번째로 낮지만, 이번 목표는 성별과 좌석클래스 간 유의한 교호작용을 확인하는 것이 목표이므로 모형 2a을 최종적으로 선택한다. 모형 2a에 대한 회귀계수 추정 결과는 다음과 같다.

	Estimated	Std.Error	z value	Pr(>|z|)
intercept	4.6115	0.5567	8.283	<0.001 ***
age1	−1.0537	0.2304	−4.573	<0.001 ***
sex1	−4.2331	0.5310	−7.972	<0.001 ***
econo.class2	−1.6806	0.5878	−2.859	0.00425 **
econo.class3	−3.8854	0.5287	−7.350	<0.001 ***
econo.class4	−1.6608	0.8003	−2.075	0.03797 *
sex1:econo.class2	0.4483	0.6460	0.694	0.48772
sex1:econo.class3	2.8625	0.5633	5.082	<0.001 ***
sex1:econo.class4	1.0862	0.8197	1.325	0.18516

결과를 살펴보면, 유의수준 0.05하에서 각 econo.class가 2, 4와 sex와의 교호작용항을 제외한 나머지 회귀계수가 모두 유의하다. 최종적으로 추정된 회귀식은 다음과 같다.

$$\widehat{\text{logit}(S)} = 4.61 - 1.05A_1 - 4.23G_1 - 1.68E_2 - 3.89E_3 - 1.66E_4 + 0.45G_1{*}E_2 + 2.86G_1{*}E_3 + 1.09G_1{*}E_4$$

1. sex와 econo.class가 일정할 때, age 여부에 따라 어른(age=1) 대비 아이(age=0)의 생존할 확률의 추정 오즈비는 exp(1.05)=2.86으로, 아이가 생존할 추정 오즈는 어른보다 2.86배 높다.

2. age가 일정할 때, 남성(sex=1) 대비 여성(sex=0)의 생존할 확률에 대한 추정 오즈비는 $\exp(4.23 - 0.45E_2 - 2.86E_3 - 1.09E_4)$다. 따라서 econo.class가 1,2,3,4일 때 여성이 남성보다 생존할 확률에 대한 추정 오즈비는 각각 68.72, 43.82, 3.94, 23.10배 높다. 즉, 여성이더라도 어느 econo.class에 속했는지 여부에 따라 남성에 비해 생존할 추정 오즈비가 크게 달라진다.

3. 좌석클래스 별 추정된 회귀식은 다음과 같다.

$$E_1 : \widehat{\text{logit}(S)} = 4.61 - 1.05A_1 - 4.23G_1 \qquad E_2 : \widehat{\text{logit}(S)} = 2.93 - 1.05A_1 - 3.78G_1$$
$$E_3 : \widehat{\text{logit}(S)} = 0.72 - 1.05A_1 - 1.37G_1 \qquad E_4 : \widehat{\text{logit}(S)} = 2.95 - 1.05A_1 - 3.14G_1$$

그러므로 age가 일정할 때, 1st class의 생존할 확률의 추정 오즈비는 2nd class 대비 여성의 경우 5.37배, 남성의 경우 3.42배 높았다. 이처럼 age가 일정할 때, 남성과 여성의 각 econo.class 대비 생존할 추정 오즈비는 다음과 같다.

<여성: 각 econo.class 대비 survived=1 추정 오즈비>

	1st	2nd	3rd	crew
1st		5.37	48.91	5.26
2nd	0.19		9.12	0.98
3rd	0.02	0.11		0.11
crew	0.19	1.02	9.09	

<남성: 각 econo.class 대비 survived=1 추정 오즈비>

	1st	2nd	3rd	crew
1st		3.42	2.80	1.79
2nd	0.29		0.82	0.52
3rd	0.36	1.22		0.63
crew	0.56	1.92	1.59	

이를 통해 성별에 따라 각 좌석 클래스 대비 추정 오즈비가 달라지지만, 1st class는 성별에 무관하게 나머지 2nd, 3rd class 그리고 crew 대비 생존할 확률이 높은 것을 확인했다. 또한, 남성의 2nd class 대비 3rd class의 생존할 확률의 추정 오즈비를 제외한 나머지 모든 3rd class의 추정 오즈비가 다른 좌석클래스 대비 낮다. 즉, 좌석클래스 역시 생존율에 큰 영향을 미치고 있다. 다시 말해, 비싼 좌석에 앉은 승객이 생존할 확률이 더 높으며, 저렴한 좌석에 탑승한 승객은 생존할 확률이 낮다.

· 결론

본 과제에서 타이타닉 탑승객 데이터를 바탕으로 성별과 나이, 그리고 탑승객의 좌석클래스를 바탕으로 생존 여부에 어떤 영향을 주는지 파악하기 위해 교차분석표와 로그선형모형을 통해 연관성을 살펴본 후, 로지스틱 회귀모형으로 각 변수가 생존 여부에 어떻게 영향을 주는지 살펴보았다.

이 과정에서 서론에 제시한 궁금증인 좌석클래스 즉, 탑승객의 경제 수준을 간접적으로 나타낸다고 할 수 있는 변수가 생존 여부에 어떠한 영향을 미치는지 확인했다. 여성과 아이의 생존율은 이미 알려진 바와 같이 높은 것을 확인할 수 있었다. 하지만 타이타닉호 탑승객의 생존 여부에는 성별과 나이뿐만이 아닌, 좌석의 클래스도 영향을 미쳤다. 특히 높은 등급 혹은 비싼 좌석에 탑승할수록 성별에 무관하게 생존할 확률이 높으며, 낮은 등급 혹은 저렴한 좌석일수록 생존할 확률이 낮다는 것은 주목할만한 결과이다.