<2024 보건의료데이터베이스 기말보고서>

폐암 발생 여부에 대한 로지스틱 회귀분석

이화여자대학교 데이터사이언스학과

2392032 정지영

목차

Ⅰ. 지난 발표 개괄

지난 발표에서 '흡연 기간과 직업에 따른 폐암 발생 여부에 대한 로지스틱 회귀분석'을 주제로 다뤘다. 한국의료패널 원시자료를 이용하여 과거 흡연 기간, 직업 분류에 따른 만성질환 유무_폐암 여부를 관찰하는 분석을 진행했다.

각 변수가 폐암 유무에 영향을 주는지 여부를 알아보기 위해 왈드 검정과 우도비 검정, 회귀계수의 신뢰구간에 대한 분석을 진행하였고, 종속변수에 영향을 미치는 변수를 골라 오즈비를 구해보았다. 모델의 설명력을 확인하기 위해 R 통계량을 확인해 보았고 모델의 적합성을 살펴보기 위해 Hosmer-Lemeshow 검정과 ROC 곡선의 AUC 값 관찰을 수행하였다.

분석 결과, 석면에 많이 노출되는 직업을 가졌나 여부는 폐암 유무를 잘 설명하지 못하는 변수였다. 반면에 과거 흡연 기간은 폐암 유무를 잘 설명하는 변수였다. 과거 흡연 기간이 1년 증가함에 따라 폐암이 아닐 확률이 0.9278배 높아짐을 확인하여 과거 흡연기간이 길수록 폐암에 걸릴 확률이 높아진다는 것을 알 수 있었다. 모델의 적합성 여부 관찰 결과, R 통계량은 적합하지 않다는 결론을, Hosmer-Lemeshow 검정과 ROC 곡선의 AUC 값 관찰 결과는 모델의 적합도가 좋다는 결론이 나왔다.

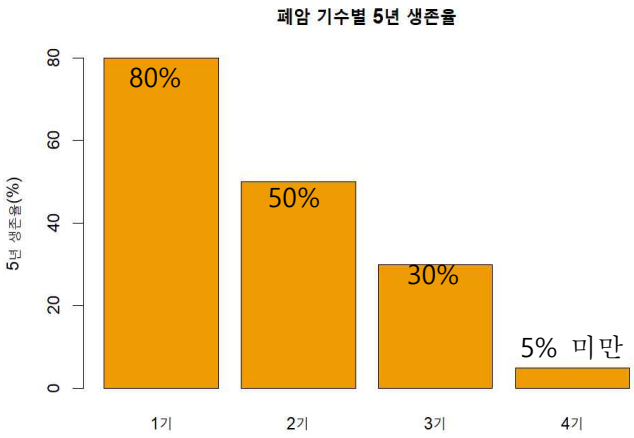
발표에 대한 개선 사항을 몇 가지 제안받았다. 아래 다섯 가지 항목으로 구분해 정리해 보았다.

1. 직업과의 관련성을 알아보는 과정에서, '건설 및 채굴 관련 기능직'과 '건설 및 광업 관련 단순 노무직'만을 위험에 노출된 직업으로 설정한 것은 적절하지 않다.
2. 폐암이 없는 경우를 1, 폐암이 있는 경우를 0으로 설정하여 분석을 진행하면 혼란을 줄 수 있다.
3. 폐암이 있는 환자 집단이 상대적으로 너무 적어 분석 결과에 영향을 준다.
4. 모델의 적합성 검정은 모델의 설명력 검정에 선행하여 수행되어야 한다.
5. R^2과 R 통계량을 혼동하여 사용하면 안 된다.

이 보고서에서는 위 사항들을 포함한 개선 사항들을 추가 적용 및 분석한 결과를 보여주고자 한다.

Ⅱ. 문제 제기 및 가설 설정

폐암이란 폐에 발생하는 악성 종양을 말한다. 위 그림은 국가암정보센터에서 제공한 암종별 발생 현황을 보여주는 그래프이다. 그래프 제목에 있는 '조발생률'이라는 용어는 해당 관찰기간동안 특정 인구집단에서 새로이 발생한 암 환자수라는 뜻이며, 해당 인구집단에서 암 발생 정도를 절대적으로 평가할 때 사용되는 지표이다. 그림에서 볼 수 있듯이, 폐암은 남성과 여성 모두에게서 많이 발생하는 암종이고, 특히 남성에게서는 가장 많이 발생하는 암이 바로 폐암이다.



위 그림은 고려대학교의료원에서 출판한 저널에서 참고한 수치를 바탕으로 R 프로그램을 이용해 그래프로 시각화한 자료이다. 폐암은 기수가 진행됨에 따라 예후가 매우 나빠진다. 비록 과거에 비해 의학과 기술이 많이 발전하여 4기 폐암 환자의 평균 여명이 과거에는 6개월에서 1년 정도로 간주 되었지만, 현재는 3년 이상 생존하는 사례들이 많아졌다고 한다. 그럼에도, 폐암은 여전히 인류에게 있어 극복하기

어려운 질병임은 자명하다.

폐암의 원인은 다양하지만 가장 대표적인 원인은 흡연이다. 충북대학교병원 충북지역암센터에 따르면, 흡연은 폐암의 원인에 70%까지 관련되어 있으며, 흡연자는 비흡연자에 비해 폐암에 걸릴 확률이 20배 높다고 한다. 이 외에도 직업적 요인, 방사성 물질, 환경적 요인, 그리고 유전적 요인 등 다양하다.

이렇듯 인간이 가장 두려워하는 질병인 암 중에서도 가장 흔하며, 위험한 폐암이라는 질병이 흡연이라는 특정 요인이 발병 원인의 대부분을 차지한다는 점이 흥미로워 '폐암 발생 여부에 대한 로지스틱 회귀분석'이라는 분석 주제를 세우게 되었다. 흡연을 포함하여 직업 분류, 출생 연도, 성별이 폐암 발생에 영향을 미치는지 알아보고자 한다.

분석 진행 전, '흡연 여부, 직업 분류, 출생 연도, 성별은 폐암 발생에 영향을 줄 것이다'라는 가설을 설정하였다.

Ⅲ. 분석 과정

우선 데이터 수집 및 전처리 과정을 거쳤다. 한국의료패널의 최신 데이터인 2019-2020년도 연간 데이터를 사용하였다. 그 후 코드북에서 사용할 변수를 선정했다. 아래 표는 분석에 사용할 변수들에 대한 설명을 표로 정리한 것이다.

변수명	한글 변수명	보기 문항 내용
S1	평생 흡연 여부	1: 5갑 미만 2: 5갑 이상 3: 피운 적 없다 (.): 결측
S4	현재 매일_흡연량	( ) 개비 (.): 결측
S7	과거 흡연_기간(년)	( ) 년 (.): 결측
S9	과거 흡연_흡연량	( ) 개비 (.): 결측
SEX	성별	1: 남 2: 여
BIRTH_Y	출생년도	( ) 년
ECO9	직업분류	[CODE-직업분류] 참고 (.): 결측
CD1_LCA	만성질환 유무_폐암	1: 예 2: 아니오 (.): 결측

```
> library(dplyr)
> test_data <- a_ind %>% select(S1, S4, S7, S9, SEX, BIRTH_Y, ECO9, CD1_LCA)
```

dplyr 패키지를 이용해 전체 데이터프레임에서 사용할 변수들만을 뽑아낸, 'test_data'라는 데이터프레임을 새롭게 만들었다. 그 후 summary(), table() 함수를 이용해 데이터 분포의 특성을 확인하고 is.na() 함수를 이용해 각 변수에 대한 결측치 존재 여부도 확인했다. 관찰 결과, 모든 변수에 대해 결측치가 존재함을 확인할 수 있었다.

가장 먼저 CD1_LCA를 제외한 모든 변수를 이용해 폐암 발생 여부를 예측하는 모델을 만들어 보고 싶었다. 하지만, 모든 변수에 대해 결측치를 제거한 후 CD1_LCA의 빈도표를 확인해 본 결과, '폐암 없음'에 해당하는 데이터만 남아있어 빈도표가 적절하게 출력되지 않았다. '폐암 있음'에 대한 데이터가 존재하지 않아 로지스틱 회귀모형에 적용해볼 수 없었다.

```
> table(set1$CD1_LCA)
< table of extent 0 >
```

다음으로, 변수의 수를 줄여 네 개의 변수 S1, SEX, BIRTH_Y, ECO9를 독립변수로 사용하여 분석을 진행해 보았다. 결측치를 제외하고 남은 데이터에 대해 CD1_LCA의 빈도표를 확인해 본 결과, '폐암 있음'이 7291명, '폐암 없음'이 10명 존재했다. 지난 발표 피드백을 반영하여 '폐암 있음'이 1로, '폐암 없음'이 0으로 설정된 빈도표이다.

```
> set2$CD1_LCA <- ifelse(set2$CD1_LCA==2, 0, set2$CD1_LCA)
> table(set2$CD1_LCA)

   0    1
7291   10
```

폐암 여부 데이터가 불균형하게 존재하기 때문에 저번 피드백을 고려해 바로 회귀모형에 적용하지 않고 데이터 불균형을 보정하기 위한 propensity score matching을 진행 후 로지스틱 회귀모형에 적용해 보았다.

```
> set2_match <- matchit(CD1_LCA ~ S1+SEX+BIRTH_Y+ECO9, data = set2, method = "nearest", ratio = 1)
> matched_data2 <- match.data(set2_match)
> model2 <- glm(CD1_LCA ~ S1+SEX+BIRTH_Y+ECO9, data = matched_data2, family = binomial("logit"))
> summary(model2)

Call:
glm(formula = CD1_LCA ~ S1 + SEX + BIRTH_Y + ECO9, family = binomial("logit"),
    data = matched_data2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 15.208066 122.358410   0.124    0.901
S1          -0.972598   1.227712  -0.792    0.428
SEX         -0.102051   1.642777  -0.062    0.950
BIRTH_Y     -0.005411   0.062957  -0.086    0.932
ECO9        -0.031428   0.033856  -0.928    0.353

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.726  on 19  degrees of freedom
Residual deviance: 26.309  on 15  degrees of freedom
AIC: 36.309

Number of Fisher Scoring iterations: 4
```

모델 분석 결과, 모든 변수의 p-value가 유의수준 0.05보다 커 귀무가설 "H0: 종속변수 CD1_LCA에 영향을 주는 변수이다."를 기각하지 못한다. 따라서 S1, SEX, BIRTH_Y, ECO9는 CD1_LCA에 영향을 주지 않는다.

Ⅳ. 결론 및 제언

이번 데이터 분석에서는 지난 발표에 대한 피드백을 참고하여 폐암 유무에 따른 데이터 설정을 조절해 폐암 있음을 1로, 폐암 없음을 0으로 설정하여 분석을 진행하였고, 세분화된 직업 정보의 손실을 줄이고 폐암 원인 물질에 노출될 다양한 가능성을 고려하기 위해 건설환경에 직접적인 노출이 되는 직업을 분류해내기 보다는 원본 직업 분류 데이터를 그대로 사용하였으며, 폐암 유무 데이터의 불균형 문제를 해결하기 위해 propensity score matching을 사용해 데이터 분석을 진행해 보았다.

데이터 분석을 진행해 봤을 때 평생 흡연 여부, 성별, 출생 연도, 직업 분류는 폐암 여부에 영향을 주지 않는 요소들이라는 결과가 나왔다. 폐암 발생 여부에 영향을 미치는 변수를 찾지 못하였기 때문에 적합 완료된 로지스틱 회귀모형에 대한 추가적인 분석은 진행할 수 없었다. 그에 따라 지난 발표 개괄 항목에서 제시된 4, 5번 피드백 사항 또한 적용해보기 어려웠다.

배경지식 조사 과정에서 나온 결과들과는 상반되는 결과가 나와 추가 분석을 진행할 수 없었던 점이 아쉬웠다. 특히 흡연에 대한 변수를 독립변수로 사용하면 폐암 유무 예측에 큰 영향이 있을 것이라 생각했지만 그렇지 않았다. 일반적인 상식과 동떨어진 결과가 나온 이유 중 하나로 선정한 변수들이 충분한 정보를 담고 있지 않았다는 점이 있을 것 같다. '평생 흡연 여부'에 흡연 강도나 빈도는 포함되지 않기 때문이다. 또한 원본 데이터에 폐암이 있는 환자의 데이터 수가 너무 적었기 때문이라는 추측이 간다. 결측치를 다 제거한 후 폐암이 있는 환자의 수 10명을 가지고 진행한 분석에 무리가 있었던 것 같다. propensity score matching을 이용해 데이터 불균형을 조절해 주었더니, 지난 분석과 달리 오히려 흡연 여부가 폐암 유무에 영향을 주지 않는다는 결과가 나온 것으로 보았을 때, 폐암이 있는 환자 데이터들이 특이한 케이스에 속해있을 가능성도 생각해 볼 필요가 있을 것 같다. 앞으로 추가적인 기회가 있다면 한국 의료 패널의 과거 데이터를 함께 사용해 폐암 환자의 수를 많이 확보한 후 분석을 적용해 보고 싶다. 이를 통해 보다 정확하고 통찰력 있는 결과 도출이 가능할 것 같다.

Ⅴ. 부록

* 한국의료패널 코드북의 [CODE-직업분류]

직업분류	
코드	설명
11	공공 기관 및 기업 고위직
12	행정·경영 지원 및 마케팅 관리직
13	전문 서비스 관리직
14	건설·전기 및 생산 관련 관리직
15	판매 및 고객 서비스 관리직
21	과학 전문가 및 관련직
22	정보 통신 전문가 및 기술직
23	공학 전문가 및 기술직
24	보건·사회복지 및 종교 관련직
25	교육 전문가 및 관련직
26	법률 및 행정 전문직
27	경영·금융전문가 및 관련직
28	문화·예술·스포츠 전문가 및 관련직
31	경영 및 회계 관련 사무직
32	금융 사무직
33	법률 및 감사 사무직
39	상담·안내·통계 및 기타 사무직
41	경찰·소방 및 보안 관련 서비스직
42	돌봄·보건 및 개인 생활 서비스직
43	운송 및 여가 서비스직
44	조리 및 음식 서비스직
51	영업직
52	매장 판매 및 상품 대여직
53	통신 및 방문·노점 판매 관련직
61	농·축산 숙련직
62	임업 숙련직
63	어업 숙련직
71	식품가공 관련 기능직
72	섬유·의복 및 가죽 관련 기능직
73	목재·가구·악기 및 간판 관련 기능직
74	금속 성형 관련 기능직
75	운송 및 기계 관련 기능직
76	전기 및 전자 관련 기능직
77	정보 통신 및 방송장비 관련 기능직
78	건설 및 채굴 관련 기능직
79	기타 기능 관련직
81	식품가공 관련 기계 조작직
82	섬유 및 신발 관련 기계 조작직
83	화학 관련 기계 조작직
84	금속 및 비금속 관련 기계 조작직
85	기계 제조 및 관련 기계 조작직
86	전기 및 전자 관련 기계 조작직
87	운전 및 운송 관련직
88	상하수도 및 재활용 처리 관련 기계 조작직
89	목재·인쇄 및 기타 기계 조작직
91	건설 및 광업 관련 단순 노무직
92	운송 관련 단순 노무직
93	제조 관련 단순 노무직
94	청소 및 경비 관련 단순 노무직
95	가사·음식 및 판매 관련 단순 노무직
99	농림·어업 및 기타 서비스 단순 노무직
01	군인

* 전체 분석 코드

```
#sas 데이터 불러오는 방법
#install.packages("sas7bdat")
#library(sas7bdat)
#x<-read.sas7bdat("파일명", debug=TRUE)
#x

# 한국 의료 패널 데이터 불러오기
library(sas7bdat)
a_ind <- read.sas7bdat("a_ind.sas7bdat")   # 가구원 데이터

# dplyr 패키지의 함수를 이용해 전체 데이터 중 이용할 변수만 선택
# S1: 평생 흡연 여부, S4: 현재 매일_흡연량, S7: 과거흡연기간, S9: 과거 흡연량
# SEX: 성별, BIRTH_Y: 출생년도
# ECO9: 직업분류
# CD1_LCA: 만성질환 유무 - 폐암
library(dplyr)
test_data <- a_ind %>% select(S1, S4, S7, S9, SEX, BIRTH_Y, ECO9, CD1_LCA)

# 데이터 특성 관찰
summary(test_data)
table(test_data$CD1_LCA)   # 범주형 자료이므로 table 관찰. (41:폐암, 13793: 폐암X)
## 결측치 확인
table(is.na(test_data$S1))
table(is.na(test_data$S4))
table(is.na(test_data$S7))
table(is.na(test_data$S9))
table(is.na(test_data$SEX))
table(is.na(test_data$BIRTH_Y))
table(is.na(test_data$ECO9))
table(is.na(test_data$CD1_LCA))

# 데이터 불균형 해소를 위한 PSM 수행
## 패키지 설치
#install.packages("MatchIt")
library(MatchIt)

#set1
set1 <- test_data %>% select(S1, S4, S7, S9, SEX, BIRTH_Y, ECO9, CD1_LCA)
set1 <- set1 %>% filter(!is.na(S1))
set1 <- set1 %>% filter(!is.na(S4))
set1 <- set1 %>% filter(!is.na(S7))
set1 <- set1 %>% filter(!is.na(S9))
set1 <- set1 %>% filter(!is.na(SEX))
set1 <- set1 %>% filter(!is.na(BIRTH_Y))
set1 <- set1 %>% filter(!is.na(ECO9))
set1 <- set1 %>% filter(!is.na(CD1_LCA))
set1$CD1_LCA <- ifelse(set1$CD1_LCA==2, 0, set1$CD1_LCA)
table(set1$CD1_LCA)

#set2
set2 <- test_data %>% select(S1, SEX, BIRTH_Y, ECO9, CD1_LCA)
set2 <- set2 %>% filter(!is.na(S1))
set2 <- set2 %>% filter(!is.na(SEX))
set2 <- set2 %>% filter(!is.na(BIRTH_Y))
set2 <- set2 %>% filter(!is.na(ECO9))
set2 <- set2 %>% filter(!is.na(CD1_LCA))
set2$CD1_LCA <- ifelse(set2$CD1_LCA==2, 0, set2$CD1_LCA)
table(set2$CD1_LCA)
set2_match <- matchit(CD1_LCA ~ S1+SEX+BIRTH_Y+ECO9, data = set2, method = "nearest", ratio = 1)
matched_data2 <- match.data(set2_match)
model2 <- glm(CD1_LCA ~ S1+SEX+BIRTH_Y+ECO9, data = matched_data2, family = binomial("logit"))
summary(model2)
```

* 참고문헌

•국가암정보센터, 「주요암사망분율」, 2023.10.16, https://www.cancer.go.kr/lay1/S1T639C641/contents.do, 2024.04.10.

•국가암정보센터, 「폐암」,2019.08.06, https://www.cancer.go.kr/lay1/program/S1T211C215/cancer/view.do?cancer_seq=5237&menu_seq=5244, 2024.04.10.

•국가암정보센터, 「꼭 알아두면 좋은 암 통계 용어_국가암등록통계」, 2021.12.01, https://post.naver.com/viewer/postView.naver?volumeNo=32859362&vType=VERTICAL, 2024.04.10.

•박동원, 「우리나라 10대 암 "폐암" 비흡연자에게 더 많이 발생한다?」, 『한양대학교병원』, 2022.05.19, https://seoul.hyumc.com/seoul/healthInfo/healthLife.do?action=view&bbsId=healthLife&nttSeq=12258, 2024.04.10.

•장치선, 「재발 위험 높은 조기 폐암,수술 통해 완치율높인다」, 『KUMM vol.21 Summer 2023』, https://www.kumc.or.kr/seasonPress/KUMM_vol21/kumm12.jsp, 2024.04.10.

•차재형, 『R과 함께하는 의학통계』, 자유아카데미, 2023.