흡연기간과 직업에 따른 폐암 발생 여부에 대한 로지스틱 회귀분석

이화여자대학교
인공지능대학 데이터사이언스학과
2392032 정지영

목차

이화여자대학교
EWHA WOMANS UNIVERSITY

- 분석 계기
- 데이터 및 변수 설명
- R을 이용한 로지스틱 회귀분석과정
- 결론 및 고찰
- 참고문헌
- 느낀점
- Q & A
- 부록

분석 계기

이화여자대학교 EWHA WOMANS UNIVERSITY

폐암: 폐에 발생한 악성 종양

▪ 성별 10대암 조발생률: 2021

 # 조발생률

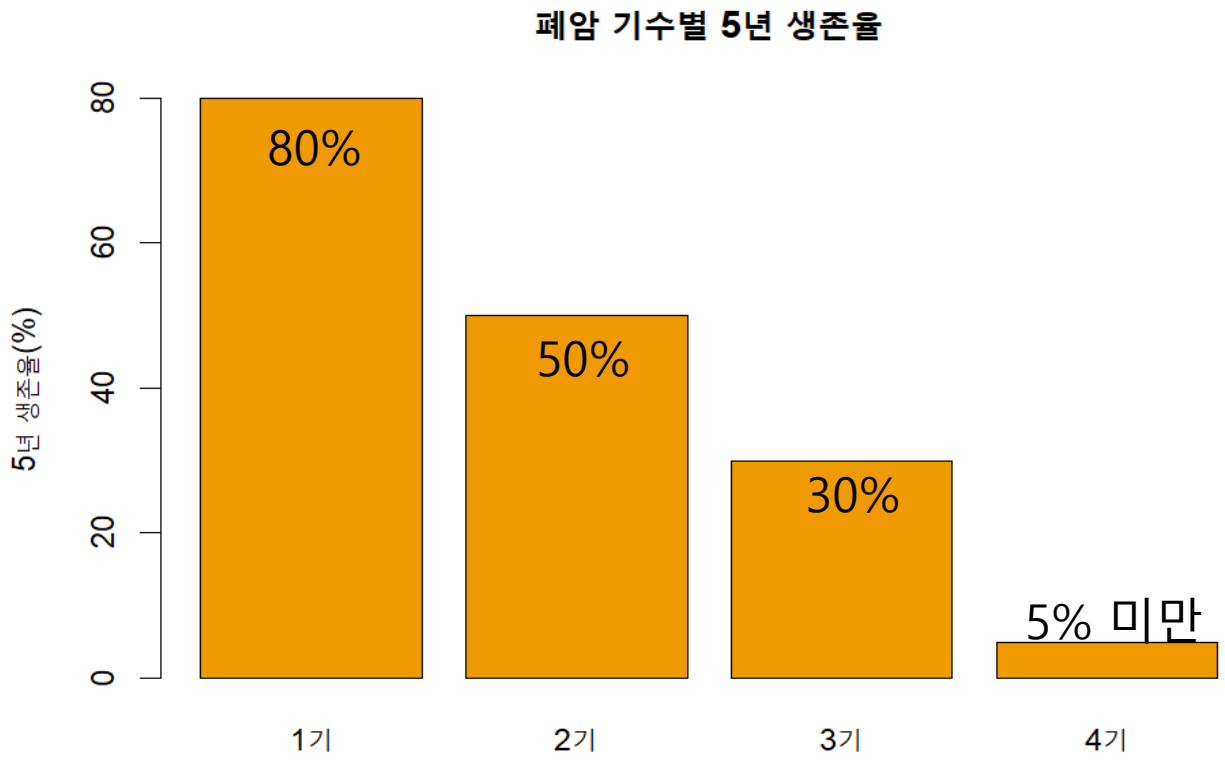
: 해당 관찰기간 동안 특정 인구집단에서 새로이 발생한 암환자수로, 해당 인구집단에서 암발생 정도를 절대적으로 평가할 때 사용

국가암정보센터 > 홈 > 통계로 보는 암 > 발생률 > 암종별 발생 현황

분석 계기

이화여자대학교 EWHA WOMANS UNIVERSITY

폐암의 예후



[4기 폐암 평균 여명]

10년 전: 6개월 ~ 1년

최근: 3년 이상 생존 흔함

고려대학교의료원 > KUMM Vol.21 Summer 2023 > 재발 위험 높은 조기 폐암, 수술 통해 완치율 높인다

분석 계기

이화여자대학교
EWHA WOMANS UNIVERSITY

폐암의 원인

직업적 요인

유전적 요인

흡연

환경적 요인

방사능

충북대학교병원 충북지역암센터 > 암정보 > 암종별정보 > 폐암 > 원인

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

코드북

S7	과거 흡연_기간(년)	( )년	
		(.) 결측	만 19세 미만 또는 이전조사 미참여자 또는 성인 가구원 이전조사 미응답 과거에는 흡연했으나 현재 피우지 않는다고 응답하지 않은 경우(S3≠3)
CD1_LCA	만성질환 유무_폐암	(1) 예	
		(2) 아니오	
		(.) 결측	이전조사 미참여자 또는 이전조사 사망가구원
ECO9	직업분류	[CODE-직업분류] 참고	
		(.) 결측	"경제활동 상태" 조사대상이 아닌 경우 및 비경제활동인구(AGE<15 or ECO1=6)

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

산업분류

코드	설명
1	농업, 임업 및 어업
2	광업
3	제조업
4	전기, 가스, 증기 및 공기조절 공급업
5	수도, 하수 및 폐기물 처리, 원료 재생업
6	건설업
7	도매 및 소매업
8	운수 및 창고업
9	숙박 및 음식점업
10	정보통신업
11	금융 및 보험업
12	부동산업
13	전문, 과학 및 기술 서비스업
14	사업시설 관리, 사업 지원 및 임대 서비스업
15	공공행정, 국방 및 사회보장 행정
16	교육 서비스업
17	보건업 및 사회복지 서비스업
18	예술, 스포츠 및 여가관련 서비스업
19	협회 및 단체, 수리 및 기타 개인 서비스업
20	가구 내 고용활동 및 달리 분류되지 않은 자가소비 생산활동
21	국제 및 외국기관

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

직업분류

코드	설명
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

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

데이터 불러오기

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

변수 선택

```
11 # dplyr 패키지의 함수를 이용해 전체 데이터 중 이용할 변수만 선택
12 library(dplyr)
13 test_data <- a_ind %>% select(S7, EC09, CD1_LCA)
```

S7: 과거흡연기간_년

EC09: 직업 분류

CD1_LCA: 만성질환 유무_폐암

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

데이터 특성 관찰 & 결측치 확인

```
15  # 데이터 특성 관찰
16  summary(test_data)
17  table(test_data$CD1_LCA)   # 범주형 자료이므로 table 관찰
```

```
> summary(test_data)
       S7              ECO9          CD1_LCA
 Min.   : 0.00   Min.   : 1.00   Min.   :1.000
 1st Qu.:11.25   1st Qu.:39.00   1st Qu.:2.000
 Median :22.00   Median :61.00   Median :2.000
 Mean   :23.81   Mean   :57.48   Mean   :1.997
 3rd Qu.:35.00   3rd Qu.:84.00   3rd Qu.:2.000
 Max.   :66.00   Max.   :99.00   Max.   :2.000
 NA's   :14169   NA's   :8917    NA's   :2753
```

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

데이터 가공

```
# 데이터 가공
## 결측치 제거(S7, CD1_LCA)
test_data <- test_data %>% filter(!is.na(S7))
test_data <- test_data %>% filter(!is.na(CD1_LCA))
## 종속변수 로지스틱 모형 대입을 위해 범위 보정
test_data$CD1_LCA <- test_data$CD1_LCA - 1
## 직업분류 중
## 건설 및 채굴관련 기능직 및 건설 및 광업 관련 단순 노무직(1)과
## 아닌 직업들(0)로 구분해 새로운 변수 생성
test_data <- test_data %>%
  mutate(job = ifelse(ECO9 %in% c(78, 91), 1, 0))
```

* 결측치 제거
- S7과 CD1_LCA에서 결측치가 존재하는 행을 모두 삭제해준다

* CD1_LCA 값 보정
- 해당 분석에서의 종속변수로, 로지스틱 회귀모형의 결과가 되기 위해서는 {0,1} 중에서 값을 가져야 한다
- 0: 폐암 있음
- 1: 폐암 없음

* 직업 분류 이용해 새로운 변수 생성
- 석면에 많이 노출되는 직업과 아닌 직업으로 나누어 관찰하고 싶음
- 1: 석면이 많이 노출되는 직업
- 0: 석면에 많이 노출되지 않는 직업

데이터 및 변수 설명

이화여자대학교 EWHA WOMANS UNIVERSITY

최종 데이터

a_ind	16587 obs. of 294 variables

test_data	2416 obs. of 4 variables

	S7	ECO9	CD1_LCA	job
1	5	NaN	1	0
2	19	74	1	0
3	15	NaN	1	0
4	25	NaN	1	0
5	33	87	1	0
6	20	91	1	1
7	32	75	1	0
8	44	NaN	1	0
9	15	NaN	1	0

변수명	해석
S7	과거흡연기간_년
ECO9	직업 분류
CD1_LCA	만성질환 유무_폐암 – 0: 폐암 있음 – 1: 폐암 없음
job	새로운 직업분류 – 0: 석면에 많이 노출되지 않는 직업 – 1: 석면에 많이 노출되는 직업

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

데이터 확인 및 시각화

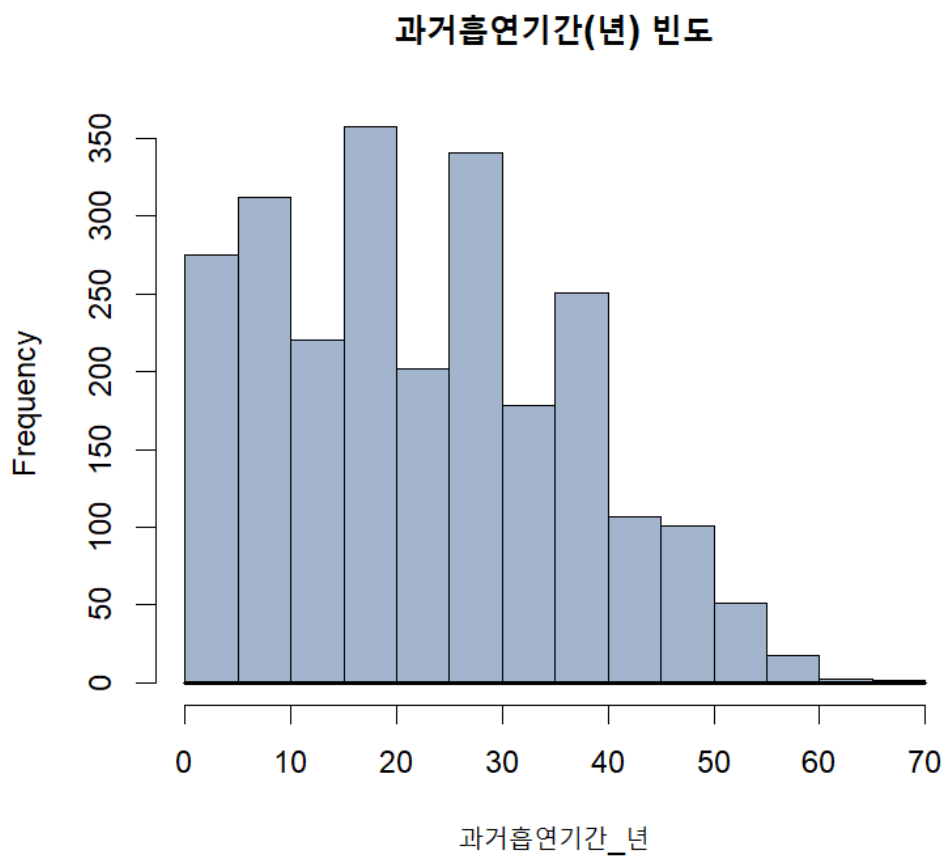
```
39  # 데이터 확인
40  summary(test_data)
41  table(test_data$job)
42  table(test_data$CD1_LCA)
```

```
> summary(test_data)
       S7              ECO9          CD1_LCA           job
 Min.   : 0.00   Min.   : 1.00   Min.   :0.0000   Min.   :0.00000
 1st Qu.:11.00   1st Qu.:41.00   1st Qu.:1.0000   1st Qu.:0.00000
 Median :22.00   Median :61.00   Median :1.0000   Median :0.00000
 Mean   :23.81   Mean   :61.33   Mean   :0.9917   Mean   :0.03022
 3rd Qu.:35.00   3rd Qu.:87.00   3rd Qu.:1.0000   3rd Qu.:0.00000
 Max.   :66.00   Max.   :99.00   Max.   :1.0000   Max.   :1.00000
                 NA's   :795
```

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

```
43 hist(test_data$S7, main = "과거흡연기간(년) 빈도", xlab = "과거흡연기간_년", col = "lightsteelblue3")
44 x <- test_data$S7
45 curve (dnorm(x, mean=mean(test_data$S7), sd=sd(test_data$S7)), lwd=2, add=TRUE)
```



H0: S7은 정규분포를 이룬다

H1: S7은 정규분포를 이루지 않는다

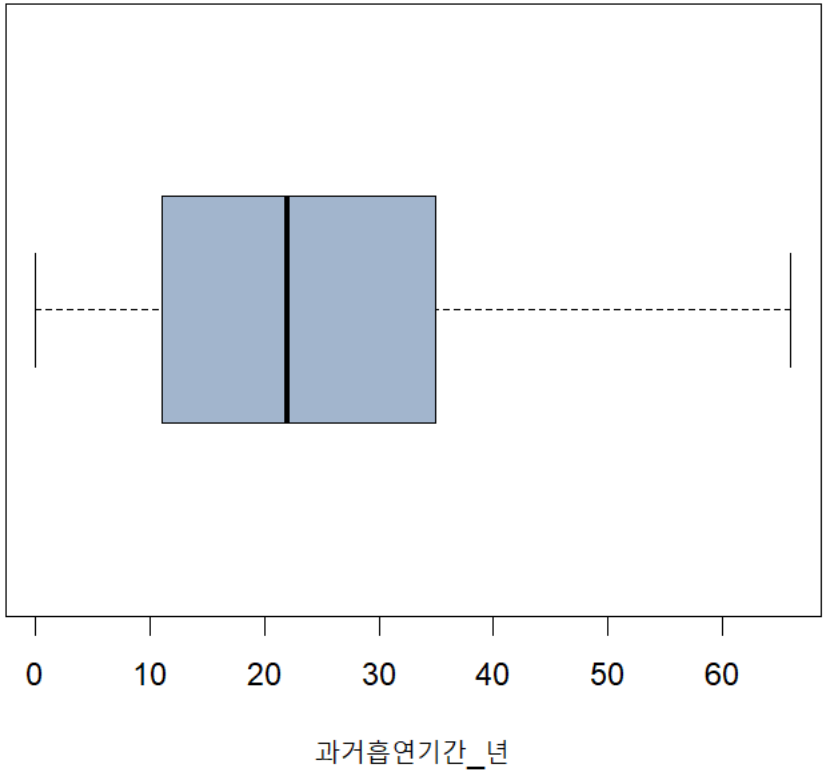
```
> shapiro.test(test_data$S7)

        Shapiro-Wilk normality test

data:  test_data$S7
W = 0.97317, p-value < 2.2e-16
```

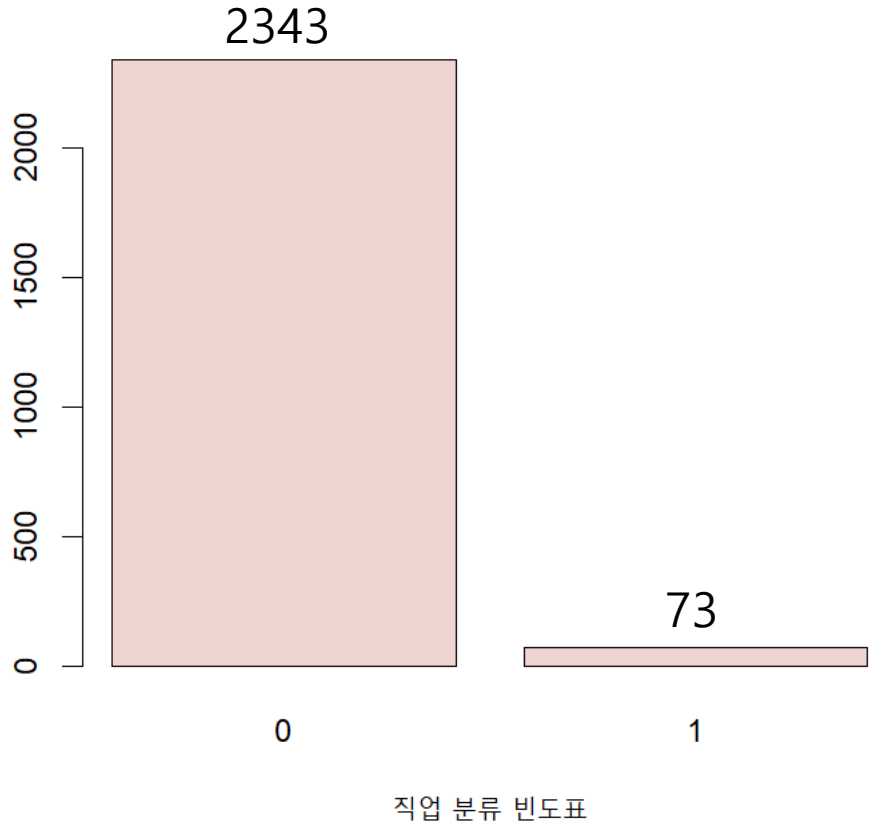
H0 기각

데이터 및 변수 설명



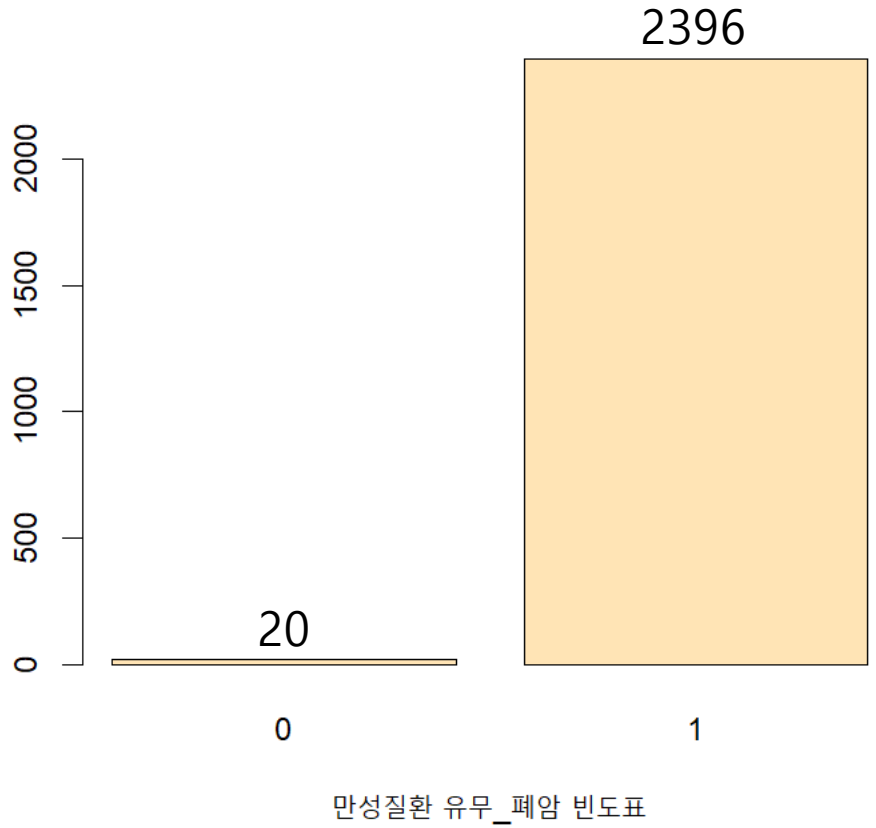
데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY



1: 석면이 많이 노출되는 직업
0: 석면에 많이 노출되지 않는 직업

데이터 및 변수 설명

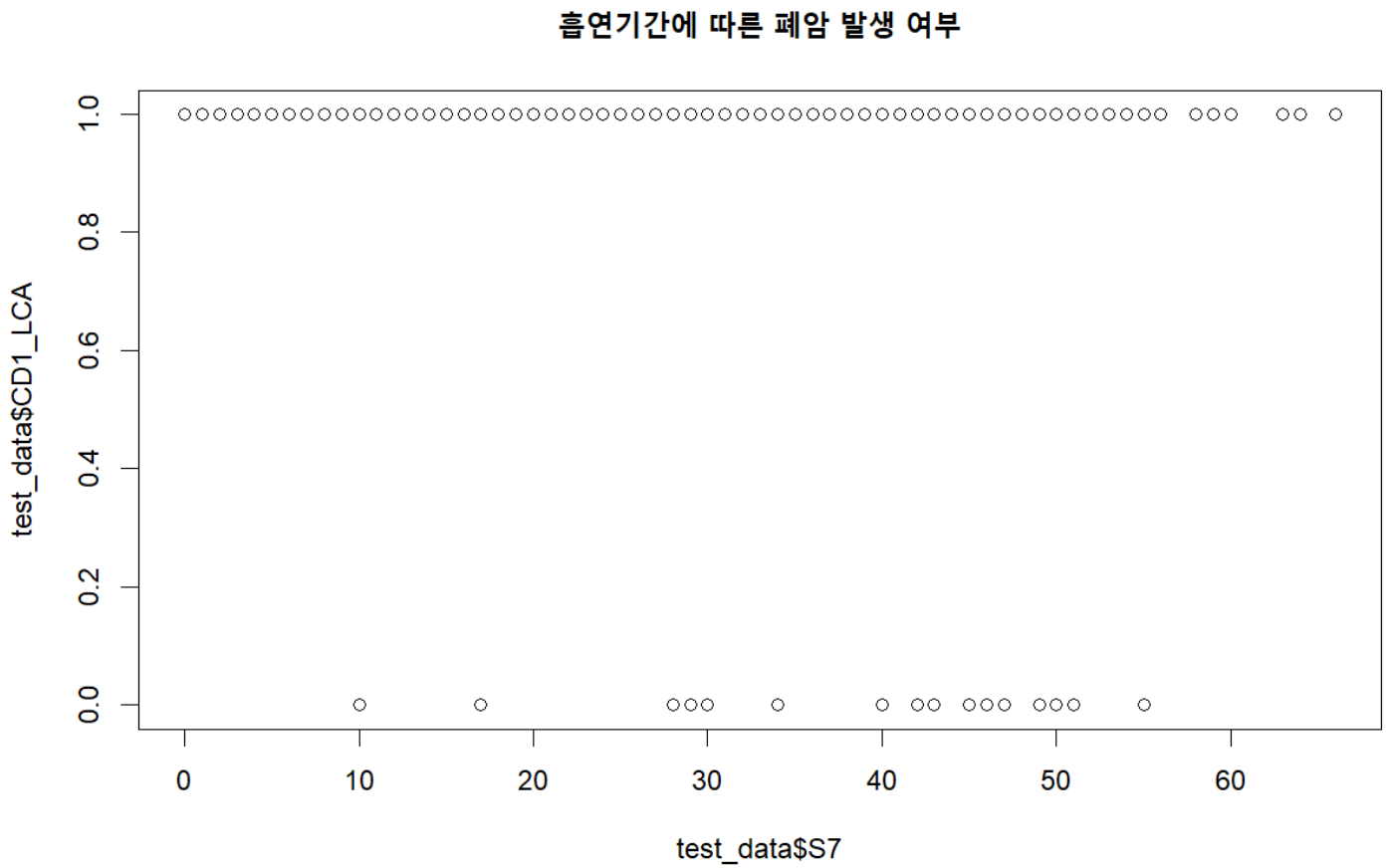


0: 폐암 있음
1: 폐암 없음

데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

S7에 대한 CD1_LCA 분포 산점도



데이터 및 변수 설명

이화여자대학교
EWHA WOMANS UNIVERSITY

job에 대한 CD1_LCA 빈도표

```
> result
```

	폐암 있음	폐암 없음
비노출직업	20	2323
노출직업	0	73

R을 이용한 로지스틱 회귀분석과정

가설

S7은 CD1_LCA에 영향을 미칠것이다.

Job은 CD1_LCA에 영향을 미치지 않을 것이다.

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

S7, job 에 대한 로지스틱 회귀모형 적합, 왈드 검정

```
# 범주형 변수 지정
test_data$CD1_LCA <- as.factor(test_data$CD1_LCA)

# 로지스틱 회귀모형 적합
## 과거흡연기간_년, 산업분류, 직업분류
fit_0 <- glm(CD1_LCA ~ S7 + job, data=test_data, family=binomial("logit"))
summary(fit_0)
```

```
> summary(fit_0)

Call:
glm(formula = CD1_LCA ~ S7 + job, family = binomial("logit"),
    data = test_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.07633    0.69205  10.225  < 2e-16 ***
S7           -0.07422    0.01706  -4.350 1.36e-05 ***
job          14.55372 1209.38476   0.012     0.99
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 231.60  on 2415  degrees of freedom
Residual deviance: 209.05  on 2413  degrees of freedom
AIC: 215.05

Number of Fisher Scoring iterations: 18
```

H0: $\beta_{S7} = \beta_{job} = 0$

H1: not H0

S7: H0 기각
Job: H0 기각 X

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

S7 에 대한 로지스틱 회귀모형 적합, 왈드 검정

```
52  # 과거흡연기간_년 만을 이용한 이분형 로지스틱 단순회귀모형 다시 fit 하기
53  fit <- glm(CD1_LCA ~ S7, data=test_data, family=binomial("logit"))
54  summary(fit)
```

```
> summary(fit)

Call:
glm(formula = CD1_LCA ~ S7, family = binomial("logit"), data = test_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.13161    0.69578  10.250  < 2e-16 ***
S7          -0.07498    0.01716  -4.368 1.25e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 231.60  on 2415  degrees of freedom
Residual deviance: 210.08  on 2414  degrees of freedom
AIC: 214.08

Number of Fisher Scoring iterations: 8
```

H0: $\beta_{S7} = 0$

H1: not H0

S7: H0 기각

$$\ln\left(\frac{px}{1-px}\right) = 7.13161 - 0.07498 * S7$$

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

우도비 검정

```
56  # 우도비 검정_LR test
57  anova(fit_0,test="Chisq")
```

```
> # 우도비 검정_LR test
> anova(fit_0,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: CD1_LCA

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                  2415     231.60
S7    1  21.5197      2414     210.08 3.502e-06 ***
job   1   1.0305      2413     209.05      0.31
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H0: $\beta_{S7} = \beta_{job} = 0$

H1: not H0

S7: H0 기각
Job: H0 기각 X

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

회귀계수의 신뢰구간 구하기

```
59  # 회귀계수의 95% 신뢰구간 구하기
60  confint(fit_0)
61  confint(fit)
```

```
> confint(fit_0)
프로파일링이 완료되길 기다리는 중입니다...
                  2.5 %      97.5 %
(Intercept)   5.8444477   8.57240477
S7           -0.1091385  -0.04187395
job         -155.6753562          NA
There were 22 warnings (use warnings() to see them)
```

H0: $\beta_{S7} = \beta_{job} = 0$

H1: not H0

```
> confint(fit)
프로파일링이 완료되길 기다리는 중입니다...
                 2.5 %      97.5 %
(Intercept)  5.8927258  8.63506082
S7          -0.1100955 -0.04243979
```

H0: $\beta_{S7} = 0$

H1: not H0

S7: H0 기각
Job: H0 기각 X

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

오즈비, 오즈비의 신뢰구간

변수명	해석
S7	과거흡연기간_년
EC09	직업 분류
CD1_LCA	만성질환 유무_폐암 - 0: 폐암 있음 - 1: 폐암 없음
job	새로운 직업분류 - 0: 석면에 많이 노출되지 않는 직업 - 1: 석면에 많이 노출되는 직업

```
63  # 오즈비 보기
64  odds_ratios <- exp(fit$coefficients)
65  odds_ratios
```

```
> odds_ratios
 (Intercept)          S7
1250.8952221   0.9277648
```

흡연기간이 1년 늘어날 때 폐암에 안 걸릴 확률이 0.9278배 된다.

```
67  # 오즈비 95% 신뢰구간
68  exp(confint(fit))
```

```
> exp(confint(fit))
프로파일링이 완료되길 기다리는 중입니다...
                  2.5 %       97.5 %
(Intercept) 362.3917495 5625.4758566
S7            0.8957486    0.9584482
```

H0: $\beta_{S7} = 0$

H1: not H0

H0 기각

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

Hosmer-Lemeshow 검정

```
77  ## Hosmer-Lemeshow 검정
78  library(ResourceSelection)
79  hoslem.test(test_data$CD1_LCA, fitted(fit))
80  ### 요인(factor)을 숫자로 변환
81  test_data$CD1_LCA <- as.numeric(as.character(test_data$CD1_LCA))
82  ### 검정 다시 실행
83  hoslem.test(test_data$CD1_LCA, fitted(fit))
```

```
> hoslem.test(test_data$CD1_LCA, fitted(fit))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  test_data$CD1_LCA, fitted(fit)
X-squared = 2416, df = 8, p-value < 2.2e-16

Warning message:
In Ops.factor(1, y) : 요인(factors)에 대하여 의미있는 '-'가 아닙니다.
```

```
> hoslem.test(test_data$CD1_LCA, fitted(fit))

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  test_data$CD1_LCA, fitted(fit)
X-squared = 10.909, df = 8, p-value = 0.2069
```

H0: 회귀모형이 적합하다

H1: 회귀모형이 적합하지 않다

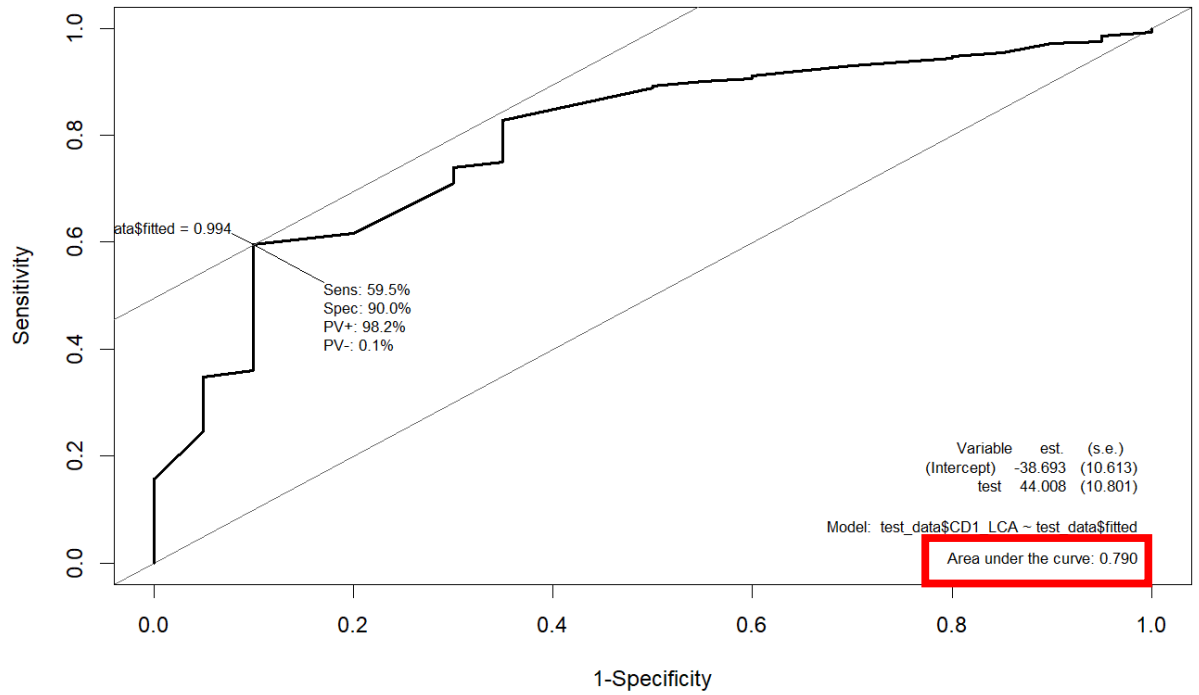
H0 기각 X

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

ROC 곡선

```
85  ## ROC 곡선_적합도 검정
86  library(Epi)
87  test_data$fitted <- fitted.values(fit)
88  ROC(test_data$fitted, test_data$CD1_LCA, grid = FALSE)
```



곡선 아래의 면적이 0.79로
모형의 적합도가 좋은 것으로 판단된다

R을 이용한 로지스틱 회귀분석과정

이화여자대학교 EWHA WOMANS UNIVERSITY

R^2

```
73  ## R 통계량
74  library(rcompanion)
75  nagelkerke(fit)
```

```
> nagelkerke(fit)
$Models

Model: "glm, CD1_LCA ~ S7, binomial(\"logit\"), test_data"
Null:  "glm, CD1_LCA ~ 1, binomial(\"logit\"), test_data"

$Pseudo.R.squared.for.model.vs.null
                             Pseudo.R.squared
McFadden                           0.09291780
Cox and Snell (ML)                 0.00886762
Nagelkerke (Cragg and Uhler)       0.09700990

$Likelihood.ratio.test
 Df.diff LogLik.diff Chisq    p.value
      -1      -10.76 21.52 3.5021e-06

$Number.of.observations

Model: 2416
Null:  2416

$Messages
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"

$Warnings
[1] "None"
```

R^2 이 0.0089 ~ 0.097로
높은 설명력을 보이지 않는다

결론 및 고찰

이화여자대학교
EWHA WOMANS UNIVERSITY

로지스틱 회귀분석 결과,

석면에 많이 노출되는 직업을 가졌나 여부는 폐암 유무를 잘 설명하지 못하는 변수였다.

반면에 과거흡연기간은 폐암유무를 잘 설명하는 변수였으며 과거흡연기간이 1년 증가함에 따라 폐암이 아닐 확률이 0.9278배 되었다. 즉, 과거흡연기간이 길수록 폐암에 걸릴 확률이 높아졌다.

결론 및 고찰

이화여자대학교
EWHA WOMANS UNIVERSITY

과거흡연기간에 대한 폐암여부 모형의 적합성을 분석해 보았을 때,

R 통계량에 의하면 적합하지 않은 모형이라는 결론이 나왔다.

Hosmer-Lemeshow 검정과 ROC 곡선을 이용한 적합도 검정에 따르면 모형의 적합도가 좋다는 결론이 나왔다.

이처럼 상반되는 결과의 해석은 추가적인 데이터 관찰 및 분석, 그리고 의학 분야 전문가들의 의견을 고려해 결론을 내려야 할 것이다.

참고문헌

- 국가암정보센터,「주요암 사망분율」, 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 완치율 높인다」,『KUMM vol.21 Summer 2023』, https://www.kumc.or.kr/seasonPress/KUMM_vol21/kumm12.jsp, 2024.04.10.
- 차재형,『R과 함께하는 의학통계』, 자유아카데미, 2023.

느낀점

Q & A

감사합니다

부록

이화여자대학교
EWHA WOMANS UNIVERSITY

```r
# 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
library(dplyr)
test_data <- a_ind %>% select(S7, EC09, CD1_LCA)

# 데이터 특성 관찰
summary(test_data)
table(test_data$CD1_LCA)   # 범주형 자료이므로 table 관찰
## 결측치 확인
table(is.na(test_data$S7))
table(is.na(test_data$EC09))
table(is.na(test_data$CD1_LCA))

# 데이터 가공
## 결측치 제거(S7, CD1_LCA)
test_data <- test_data %>% filter(!is.na(S7))
test_data <- test_data %>% filter(!is.na(CD1_LCA))
## 종속변수 로지스틱 모형 대입을 위해 범위 보정
test_data$CD1_LCA <- test_data$CD1_LCA - 1
## 직업분류 중
## 건설 및 채굴관련 기능직 및 건설 및 광업 관련 단순 노무직(1)과
## 아닌 직업들(0)로 구분해 새로운 변수 생성
test_data <- test_data %>%
  mutate(job = ifelse(EC09 %in% c(78, 91), 1, 0))
## 자료형 확인
class(test_data$S7)
class(test_data$job)
class(test_data$CD1_LCA)

# 데이터 확인
summary(test_data)

# 데이터 시각화
hist(test_data$S7, main = "과거흡연기간(년) 빈도", xlab = "과거흡연기간_년", col = "lightsteelblue3")
x <- test_data$S7
curve (dnorm(x, mean=mean(test_data$S7), sd=sd(test_data$S7)), lwd=2, add=TRUE)
shapiro.test(test_data$S7)                    # 정규분포 따르지 않아 곡선 나타나지 않음
boxplot(test_data$S7, horizontal = T, xlab = "과거흡연기간_년", col = "lightsteelblue3")
barplot(table(test_data$job), xlab = "직업 분류 빈도표", col = "mistyrose2", width = 0.5)
barplot(table(test_data$CD1_LCA), xlab = "만성질환 유무_폐암 빈도표", col = "moccasin")
plot(x=test_data$S7, y=test_data$CD1_LCA, main = "흡연기간에 따른 폐암 발생 여부")
result <- table(test_data$job, test_data$CD1_LCA)
rownames(result) <- c("비노출직업", "노출직업")
colnames(result) <- c("폐암 있음", "폐암 없음")
result
table(test_data$job)
table(test_data$CD1_LCA)
```

부록

이화여자대학교
EWHA WOMANS UNIVERSITY

```r
# 범주형 변수 지정
test_data$CD1_LCA <- as.factor(test_data$CD1_LCA)

# 로지스틱 회귀모형 적합
## 과거흡연기간_년, 산업분류, 직업분류
fit_0 <- glm(CD1_LCA ~ S7 + job, data=test_data, family=binomial("logit"))
summary(fit_0)

# 과거흡연기간_년 만을 이용한 이분형 로지스틱 단순회귀모형 다시 fit 하기
fit <- glm(CD1_LCA ~ S7, data=test_data, family=binomial("logit"))
summary(fit)

# 우도비 검정_LR test
anova(fit_0,test="Chisq")

# 회귀계수의 95% 신뢰구간 구하기
confint(fit_0)
confint(fit)

# 오즈비 보기
odds_ratios <- exp(fit$coefficients)
odds_ratios

# 오즈비 95% 신뢰구간
exp(confint(fit))

# 오즈비 그래프

# 모형의 적합성
## R 통계량
library(rcompanion)
nagelkerke(fit)

## Hosmer-Lemeshow 검정
library(ResourceSelection)
hoslem.test(test_data$CD1_LCA, fitted(fit))
### 요인(factor)을 숫자로 변환
test_data$CD1_LCA <- as.numeric(as.character(test_data$CD1_LCA))
### 검정 다시 실행
hoslem.test(test_data$CD1_LCA, fitted(fit))

## ROC 곡선_적합도 검정
library(Epi)
test_data$fitted <- fitted.values(fit)
ROC(test_data$fitted, test_data$CD1_LCA, grid = FALSE)
```

```r
lca <- data.frame(level = c("1기", "2기", "3기", "4기"),
                  live = c(80, 50, 30, 5))
barplot(lca$live, names.arg = lca$level, main = "폐암 기수별 5년 생존율",
        col = "orange2", xlab = "기수", ylab = "5년 생존율(%)")
```