한 지역의 병원에서는 독감예방접종에 대한 홍보문서를 고령층을 중심으로 배부했다. 159명의 환자를 대상으로 실제로 예방접종을 받았는지를 조사했다. 수집된 변수는 아래와 같다.

접종여부 (flushot): 1(받음), 0(받지 않음)

나이 (age)

건강상태에 대한 자각 인덱스(aware): 높을수록 자각이 높음

성별 (gender): 1(남성), 0(여성)

접종여부에 대한 모형

자료는 flushot.csv에 저장되어 있다.

주어진 세 개의 설명변수로 예방접종 확률을 예측하는 모형을 추정하여 추정된 로지스틱 회귀식을 써라.

결과

$$\pi = \frac{\exp(-1.177 + 0.072X_1 - 0.098X_2 + 0.433X_3)}{1 + \exp(-1.177 + 0.072X_1 - 0.098X_2 + 0.433X_3)}$$

콘솔

```
> flueshot=read.csv("flushot.csv")
> model1=glm(flushot~.,data=flueshot,family=binomial)
> summary(model1)

Call:
glm(formula = flushot ~ ., family = binomial, data = flueshot)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4037  -0.5637  -0.3352  -0.1542   2.9394

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
age          0.07279    0.03038   2.396  0.01658 *
aware       -0.09899    0.03348  -2.957  0.00311 **
gender       0.43397    0.52179   0.832  0.40558
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 134.94  on 158  degrees of freedom
Residual deviance: 105.09  on 155  degrees of freedom
AIC: 113.09

Number of Fisher Scoring iterations: 6
```

Exp(b_1),Exp(b_2),Exp(b_3)를 해석하라.

결과
Exp(b_1)= exp(0.072)= 1.074655 Exp(b_2)= exp(-0.098)= 0.9066489 Exp(b_3)= exp(0.433)= 1.541876
콘솔

`> exp(0.072)`
`[1] 1.074655`
`> exp(-0.098)`
`[1] 0.9066489`
`> exp(0.433)`
`[1] 1.541876`

X가 한 단위 증가할 때 Odds가 어떻게 변하는지 해석

해석
Exp(b_1)= exp(0.072)= 1.074655 : 나이 나이가 많을수록 접종받을 확률이 7.46% 증가하고 Exp(b_2)= exp(-0.098)= 0.9066489 :자각 건강상태에 대한 자각이 높을수록 접종받을 확률이 90.6%감소 Exp(b_3)= exp(0.433)= 1.541876: 성별 그리고 남성이 접종할 확률이 54%높다

55세의 건강상태에 대한 자각 인덱스가 60인 남성이 예방접종을 받을 확률은?

파이 햇 = exp(~~)/1-exp(~) 를 계산

계산
$\pi = \frac{\exp(-1.177 + 0.072 * 55 - 0.098 * 60 + 0.433 * 1)}{1 + \exp(-1.177 + 0.072 * 55 - 0.098 * 60 + 0.433 * 1)}$ =0.06966899/1+0.06966899 =0.06513135 결과 : 6.513%

유의하지 않은 설명변수가 있는지 Deviance goodness-of-fit test를 통해 판단하여 최종모형을 추정하라.

과정

1. Wald test를 통해 성별이 유의한지 판단하라

```
> summary(model1)

Call:
glm(formula = flushot ~ ., family = binomial, data = flueshot)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4037  -0.5637  -0.3352  -0.1542   2.9394

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.17716    2.98242  -0.395  0.69307
age          0.07279    0.03038   2.396  0.01658 *
aware       -0.09899    0.03348  -2.957  0.00311 **
gender       0.43397    0.52179   0.832  0.40558
---
```

Gender가 상관관계가 가장 낮으므로 gender를 제외한 것을 model2로 한다.

2. Deviance Goodness-of-fit test를 통해 성별을 모형에서 제거해도 좋을지 판단하라.

```
model2=glm(flushot~.-gender,data=flueshot,family=binomial)
summary(model2)
anova(model1,model2,test = "Chisq")
```

model들을 anova로 돌리면 다음과 같은 결과가 나온다

```
> anova(model1,model2,test = "Chisq")
Analysis of Deviance Table

Model 1: flushot ~ age + aware + gender
Model 2: flushot ~ (age + aware + gender) - gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       155     105.09
2       156     105.80 -1 -0.70221    0.402
> |
```

이때 모델 2의 p값이 0.05보다 크므로 타당하다.

성별을 모형에서 제거해도 좋다

Cutoff를 0.1,0.15,0.2로 두었을 때의 총 error rate과 민감도, 특이도를 계산하라

총 error rate을 최소화 시키는 cutoff는 무엇인가? ROC curve 구하라.

분석

```
jang <- function() {
  x <- seq(0.01,0.5,0.01)



  n <- length(x)

  error_min <- vector(length=n)
  sens <- vector(length=n)
  spec <- vector(length=n)

  for(i in 1:n) {
    tab = xtabs(~flueshot$flushot+(model1$fitted>x[i]))

    res = c(민감도=tab[2,2] / sum(tab[2,]) , 특이도=tab[1,1] / sum(tab[1,]) , ErrorRate=(tab[1,2] +
tab[2,1]) / sum(tab) )

    error_min[i] = (tab[1,2] + tab[2,1]) / sum(tab)   #ErrorRate
    sens[i] = tab[2,2] / sum(tab[2,])
    spec[i] = tab[1,1] / sum(tab[1,])

    print(res)
  }

  print(error_min)
  print(paste("최소의 ErrorRate는 " , min(error_min) , "이다."))
  index = which(error_min<=min(error_min))

  #print(index)

  print(paste("해당하는 민감도=",sens[min(index)],"이다."))
  print(paste("해당하는 특이도=",spec[min(index)],"이다."))
  print(paste("해당하는 에러율=",error_min[min(index)],"이다."))
```

```
   print(paste("해당하는  cutoff=",x[min(index)],"이다."))

   plot(1-spec,sens,type='b')
}
```

결과:

```
[1] "최소의  ErrorRate는   0.132075471698113  이다."
[1] "해당하는  민감도=  0.375  이다."
[1] "해당하는  특이도=  0.955555555555556  이다."
[1] "해당하는  에러율=  0.132075471698113  이다."
[1] "해당하는  cutoff=  0.39  이다."
```