로지스틱 회귀 모델(Logistic regression model)

Yoon-Ho Hong

2023-09-03

## 분류

분류 문제는 결과변수가 범주형 변수(category)일 때, 결과변수가 특정 범주에 속할 확률을 구하는 문제로 볼 수 있습니다.

오늘 다를 로지스틱 회귀는 비록 회귀라는 이름을 가지고 있지만, 결과변수가 특정 범주에 속할 확률을 모델링하는 것이기 때문에 분류기로 사용됩니다.

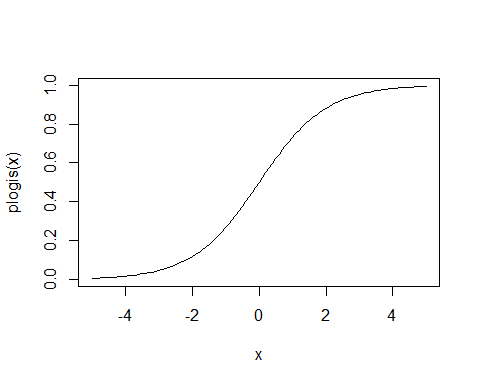
결과변수의 범주가 2개라면(편의상 0과 1이라고 합시다), 예측변수를 이용하여 결과변수의 범주를 예측하는 분류 문제는 P(Y=1|X), 즉 X 변수값이 주어졌을 때, Y=1 확률을 모델링하는 문제입니다(아래에서는 이 조건부 확률 분포를 편의상 P(X)로 표기하겠습니다.)

## 로지스틱 함수

로지스틱 함수를 살펴봅시다.

로지스틱 함수의 출력값은 항상 0에서 1사이입니다. 시그모이드 모양을 가져서 시그모이드 함수라고도 합니다.

curve(plogis(x), -5, 5)



## 오즈(odds)와 로짓(logit)

오즈는 어떤 사건이 일어날 가능성으로 성공확률/실패확률, 즉 P/(1-P)를 말합니다

오즈(odds)를 선형 함수로 모델링한다고 해봅시다.

위 식에서 좌변은 0에서 무한대인데 반해 우변은 모든 범위의 값이 가능합니다. 양변의 범위가 맞지 않죠.

이 문제를 해결하기 위해 좌변에 log 를 취합니다. odds의 로그 변환을 logit 변환이라고 부릅니다.

앞에서 보았던 로지스틱 함수의 형태가 됩니다.

정리하면, logistic regression은 오즈에 로그를 취한 logit값을 결과변수로 하는 선형회귀 모델인 것입니다.

## 타이타닉호의 생존자

타이타닉호의 생존 데이터에서 생존 여부를 예측하는 분류 모델을 만들어봅시다.

carData 패키지를 설치하고, TitanicSurvival 데이터셋의 구조를 봅시다.

library(carData)  
str(TitanicSurvival)

'data.frame': 1309 obs. of 4 variables:  
 $ survived : Factor w/ 2 levels "no","yes": 2 2 1 1 1 2 2 1 2 1 ...  
 $ sex : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...  
 $ age : num 29 0.917 2 30 25 ...  
 $ passengerClass: Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...

각 변수의 요약 통계치를 확인해봅시다.

summary(TitanicSurvival)

survived sex age passengerClass  
 no :809 female:466 Min. : 0.1667 1st:323   
 yes:500 male :843 1st Qu.:21.0000 2nd:277   
 Median :28.0000 3rd:709   
 Mean :29.8811   
 3rd Qu.:39.0000   
 Max. :80.0000   
 NA's :263

결측치(age, NA’s)가 있는 레코드는 일단 제거하기로 합니다.

TitanicSurvival = TitanicSurvival[complete.cases(TitanicSurvival),]

다음으로 logistic regression model을 데이터에 적합시켜봅시다(fit).

glm.fit = glm(survived ~., data = TitanicSurvival, family = "binomial")

모델 적합의 결과를 살펴봅시다.

summary(glm.fit)

Call:  
glm(formula = survived ~ ., family = "binomial", data = TitanicSurvival)  
  
Coefficients:  
 Estimate Std. Error z value Pr(>|z|)   
(Intercept) 3.522074 0.326702 10.781 < 2e-16 \*\*\*  
sexmale -2.497845 0.166037 -15.044 < 2e-16 \*\*\*  
age -0.034393 0.006331 -5.433 5.56e-08 \*\*\*  
passengerClass2nd -1.280570 0.225538 -5.678 1.36e-08 \*\*\*  
passengerClass3rd -2.289661 0.225802 -10.140 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)  
  
 Null deviance: 1414.62 on 1045 degrees of freedom  
Residual deviance: 982.45 on 1041 degrees of freedom  
AIC: 992.45  
  
Number of Fisher Scoring iterations: 4

로지스틱 회귀에서 계수 추정은 최대가능도 추정법(maximum likelihood estimation)을 사용합니다. 이에 대한 자세한 설명은 생략하고, 우리는 모델 적합의 결과를 어떻게 해석할 것인지에 우선 집중하기로 합니다.  
남자와 여자의 생존 확률을 비교해봅시다. 어느 쪽이 얼마나 큽니까?

male의 계수 추정치는 -2.49 입니다. 로지스틱 회귀 모델의 결과 변수는 *odds의 로그 변환, 즉 로짓 값*이라는 것을 기억합시다. 즉, 이 계수 추정값은 오즈비(odds ratio)의 로짓 값입니다.

따라서, 오즈비는 이 계수 추정값의 자연지수 함수값입니다.

exp(coef(glm.fit)[2])

sexmale   
0.08226211

male의 생존 odds는 female의 생존 odds의 0.08배 입니다. 즉, 여성의 생존 오즈는 남성보다 10배 이상 큽니다.

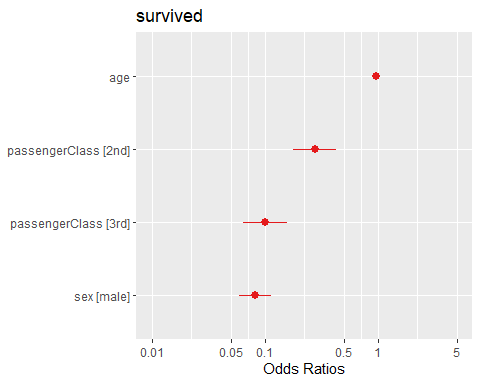
유의성 여부에 대한 통계적 검정 결과가 z 값과 p-value로 제시되어 있습니다. z 값은 계수와 계수의 표준 오차 간의 비율로 계산되는 검정 통계량으로 통계적 유의성에 대한 결정을 내릴 때 사용하는 p-값을 계산하기 위해 사용합니다. (귀무 가설인 계수 추정치가 0이다에 대한 통계적 검정).

1등칸에 비해 2등칸과 3등칸에 탄 승객들의 생존 오즈비는 얼마일까요?  
독립 변수가 범주형 변수일 때, 회귀 모델에서는 이를 dummy variable로 변환합니다.

contrasts(TitanicSurvival$passengerClass)

2nd 3rd  
1st 0 0  
2nd 1 0  
3rd 0 1

library(sjPlot)  
plot\_model(glm.fit, sort.est = T)



로지스틱 회귀 모델의 적합도에 대한 통계적 검정을 위해 일반적으로 가능도 비율 검정법(likelihood ratio test)을 사용합니다.

library(lmtest)

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':  
  
 as.Date, as.Date.numeric

glm.fit.null = glm(survived ~ 1, data = TitanicSurvival, family = "binomial")  
lrtest(glm.fit, glm.fit.null)

Likelihood ratio test  
  
Model 1: survived ~ sex + age + passengerClass  
Model 2: survived ~ 1  
 #Df LogLik Df Chisq Pr(>Chisq)   
1 5 -491.23   
2 1 -707.31 -4 432.17 < 2.2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## 분류 정확도(Classification accurancy)

분류 정확도를 어떻게 평가할까요?

먼저, 위에서 만든 모델을 사용하여 타이타닉호의 생존 결과를 예측해봅시다.

prob = predict(glm.fit, type = "response")   
head(prob)

Allen, Miss. Elisabeth Walton Allison, Master. Hudson Trevor   
 0.9258533 0.7296211   
 Allison, Miss. Helen Loraine Allison, Mr. Hudson Joshua Crei   
 0.9693290 0.4981081   
Allison, Mrs. Hudson J C (Bessi Anderson, Mr. Harry   
 0.9347616 0.3482715

이제, 분류 예측 정확도를 계산해봅시다.

pred = rep("no", length(prob))  
pred[prob>0.5] = "yes"  
mean(pred == TitanicSurvival$survived)

[1] 0.7848948

## 예측 모델의 성능 평가

위의 예제는 training data 에서 성능을 평가한 것이고, 우리가 실제로 관심이 있는 것은 모델 훈련에 사용되지 않은 새로운 데이터셋에서의 성능일겁니다.

아래는 전체 데이터셋을 training set과 test set으로 나누고, training set에서 만든 모델을 test set에 적용해 성능을 평가한 것입니다.

set.seed(1)  
cnt = dim(TitanicSurvival)[1]  
index = sample(cnt, round(cnt/7), replace = F)  
train.data = TitanicSurvival[-index,]  
test.data = TitanicSurvival[index,]

로지스틱 회귀 모델을 학습 데이터에 적합시키고, 검정 데이터에 적용하여 예측값을 얻습니다.

glm.fit = glm(survived ~., data = train.data,   
 family = "binomial")  
prob = predict(glm.fit, test.data, type = "response")

confusion matrix를 출력하고, 정확도를 계산합니다.

pred = rep("no", round(cnt/7))  
pred[prob>0.5] = "yes"  
table(pred, test.data$survived)

pred no yes  
 no 79 15  
 yes 16 39

mean(pred == test.data$survived)

[1] 0.7919463

saveRDS(glm.fit, "data/titanic\_LRmodel.rds")  
saveRDS(test.data, "data/titanic\_testData.rds")