ABF\_Week5\_HW

AI MBA 이태환

2024-02-01

Odds:오즈

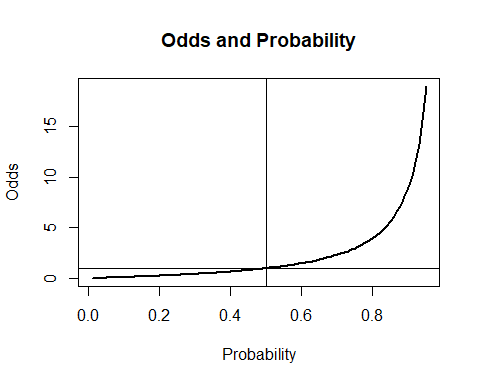
오즈의 특징

•음수의 값을 가질 수 없음

•확률 0 이면 오즈도 0, 확률이 0.5 이면 오즈는 1, 확률이 1 이면 오즈는 ∞

•오즈와 확률의 관계식의 문제 -> 오즈값의 불균형성

p=seq(0.01,0.99,0.01) #0.01에서 0.99까지 0.01 단위로 숫자 생성  
odds=p/(1-p)  
plot(p[1:95],odds[1:95], type="l", lwd=2, main="Odds and Probability", xlab="Probability",   
ylab="Odds")  
abline(h=1); abline(v=0.5)



Log Odds:로그오즈

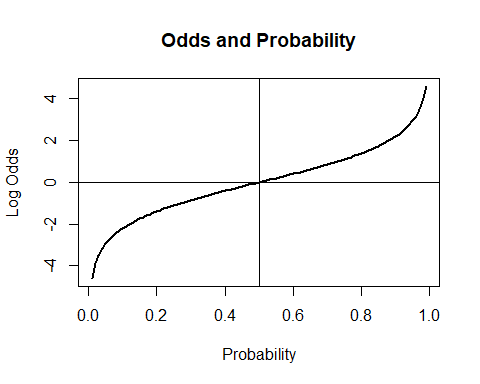
로그 오즈의 특징

•–∞ ~ + ∞ 값을 가짐

•확률 0 이면 로그 오즈는 –∞ , 확률이 0.5 이면 로그 오즈는 0, 확률이 1 이면 +∞

•로그 오즈는 부드러운 균형성을 가짐

p=seq(0.01,0.99,0.01) #0.01에서 0.99까지 0.01 단위로 숫자 생성  
ln\_odds=log(p /(1-p))  
plot(p , ln\_odds, type="l", lwd=2, main="Odds and Probability", xlab="Probability",ylab="Log Odds")  
abline(h=0); abline(v= 0.5)

 예제

wt=c(40,143,10,101)  
gender=factor(c(rep("girl",2),rep("boy",2))) #gender [1] girl girl boy boy // Levels: boy girl  
y=as.factor(rep(c("1buy","0N\_buy"),2)) #앞에 숫자를 붙인 이유는 not\_buy레벨을 0으로 만들기 위함  
levels(gender) #알파벳 순으로 레벨 순서 결정

## [1] "boy" "girl"

levels(y)

## [1] "0N\_buy" "1buy"

model1 <- glm(y ~ gender, weights=wt, family=binomial())  
summary(model1)

##   
## Call:  
## glm(formula = y ~ gender, family = binomial(), weights = wt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.3125 0.3315 -6.976 3.04e-12 \*\*\*  
## gendergirl 1.0386 0.3767 2.757 0.00583 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 268.12 on 3 degrees of freedom  
## Residual deviance: 259.40 on 2 degrees of freedom  
## AIC: 263.4  
##   
## Number of Fisher Scoring iterations: 5

exp(coef(model1)) # Odds Ratio: exponentiated coefficients

## (Intercept) gendergirl   
## 0.0990099 2.8251748

신뢰구간

exp(confint(model1)) # 95% CI for Odds Ratio

## 프로파일링이 완료되길 기다리는 중입니다...

## 2.5 % 97.5 %  
## (Intercept) 0.04843014 0.1801265  
## gendergirl 1.39911044 6.2142389

exp(confint(model1, level = 0.99)) # 99% CI for Odds Ratio

## 프로파일링이 완료되길 기다리는 중입니다...

## 0.5 % 99.5 %  
## (Intercept) 0.03748465 0.2135141  
## gendergirl 1.13495047 8.1657195

예측확률 추정

predict(model1, list(gender=c("girl","boy")), type="response")

## 1 2   
## 0.21857923 0.09009009

wt1=c(3,5,4,8,7)  
wt2=10 - wt1  
  
wt=as.numeric(rbind(wt1,wt2)) #표를 토대로 가중치를 만들기  
# rbind(wt1,wt2)  
# [,1] [,2] [,3] [,4] [,5]  
# wt1 3 5 4 8 7  
# wt2 7 5 6 2 3  
  
# wt  
# [1] 3 7 5 5 4 6 8 2 7 3  
x=rep(1:5,rep(2,5))  
# rep(2,5)  
# [1] 2 2 2 2 2  
  
# rep(1:5,rep(2,5)) 1~5를 2번씩 반복  
# [1] 1 1 2 2 3 3 4 4 5 5  
  
y=factor(rep(c("1buy","0N\_buy"),5))  
# [1] 1buy 0N\_buy 1buy 0N\_buy 1buy 0N\_buy 1buy 0N\_buy 1buy 0N\_buy  
# Levels: 0N\_buy 1buy  
model2 <- glm(y ~ x, weights=wt, family=binomial())  
summary(model2)

##   
## Call:  
## glm(formula = y ~ x, family = binomial(), weights = wt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.2324 0.7086 -1.739 0.0820 .  
## x 0.4701 0.2195 2.141 0.0323 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 68.994 on 9 degrees of freedom  
## Residual deviance: 63.978 on 8 degrees of freedom  
## AIC: 67.978  
##   
## Number of Fisher Scoring iterations: 4

wt=stack(as.data.frame(cbind(wt1,wt2)))$values  
x=rep(1:5,2)  
y=factor(c(rep("1buy",5),rep("0N\_buy",5)))  
  
model2 <- glm(y ~ x, weights=wt, family=binomial())  
summary(model2)

##   
## Call:  
## glm(formula = y ~ x, family = binomial(), weights = wt)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.2324 0.7086 -1.739 0.0820 .  
## x 0.4701 0.2195 2.141 0.0323 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 68.994 on 9 degrees of freedom  
## Residual deviance: 63.978 on 8 degrees of freedom  
## AIC: 67.978  
##   
## Number of Fisher Scoring iterations: 4

exp(coef(model2)) # Odds Ratio: exponentiated coefficients

## (Intercept) x   
## 0.2915834 1.6000891

exp(confint(model2)\*2) # 95% CI for Odds Ratio of 2 level difference

## 프로파일링이 완료되길 기다리는 중입니다...

## 2.5 % 97.5 %  
## (Intercept) 0.004485575 1.25501  
## x 1.120901928 6.40521

predict(model2, list(x=3), type="response")

## 1   
## 0.5443205

#모델 선택 backward 방법

# da1=read.table("clipboard", h=T) #finance.txt의 데이터부분을 복사(ctrl + c)한 후 이 코드를 실행한다.  
da1=read.csv("finance.csv", h=T)  
da1$y=as.factor(da1$y)  
model4 <- glm(y ~ x1+x2+x3+x4, family=binomial(), data=da1)  
  
scope = formula(~ x1+x2+x3+x4)  
null = glm(y ~ 1, family=binomial(), data=da1)  
step(model4, direction="backward")

## Start: AIC=37.66  
## y ~ x1 + x2 + x3 + x4  
##   
## Df Deviance AIC  
## - x2 1 27.745 35.745  
## - x4 1 28.420 36.420  
## - x1 1 29.538 37.538  
## <none> 27.660 37.660  
## - x3 1 42.122 50.122  
##   
## Step: AIC=35.74  
## y ~ x1 + x3 + x4  
##   
## Df Deviance AIC  
## - x4 1 28.573 34.573  
## <none> 27.745 35.745  
## - x1 1 33.353 39.353  
## - x3 1 43.130 49.130  
##   
## Step: AIC=34.57  
## y ~ x1 + x3  
##   
## Df Deviance AIC  
## <none> 28.573 34.573  
## - x1 1 35.700 39.700  
## - x3 1 43.200 47.200

##   
## Call: glm(formula = y ~ x1 + x3, family = binomial(), data = da1)  
##   
## Coefficients:  
## (Intercept) x1 x3   
## -5.807 6.416 2.983   
##   
## Degrees of Freedom: 45 Total (i.e. Null); 43 Residual  
## Null Deviance: 63.42   
## Residual Deviance: 28.57 AIC: 34.57

#모델 선택 forward 방법

step(null, scope, direction="forward")

## Start: AIC=65.42  
## y ~ 1  
##   
## Df Deviance AIC  
## + x3 1 35.700 39.700  
## + x1 1 43.200 47.200  
## + x2 1 43.695 47.695  
## <none> 63.421 65.421  
## + x4 1 63.372 67.372  
##   
## Step: AIC=39.7  
## y ~ x3  
##   
## Df Deviance AIC  
## + x1 1 28.573 34.573  
## + x2 1 31.028 37.028  
## + x4 1 33.353 39.353  
## <none> 35.700 39.700  
##   
## Step: AIC=34.57  
## y ~ x3 + x1  
##   
## Df Deviance AIC  
## <none> 28.573 34.573  
## + x4 1 27.745 35.745  
## + x2 1 28.420 36.420

##   
## Call: glm(formula = y ~ x3 + x1, family = binomial(), data = da1)  
##   
## Coefficients:  
## (Intercept) x3 x1   
## -5.807 2.983 6.416   
##   
## Degrees of Freedom: 45 Total (i.e. Null); 43 Residual  
## Null Deviance: 63.42   
## Residual Deviance: 28.57 AIC: 34.57

#모델 선택 Stepwise(both) 방법

step(null, scope, direction="both")

## Start: AIC=65.42  
## y ~ 1  
##   
## Df Deviance AIC  
## + x3 1 35.700 39.700  
## + x1 1 43.200 47.200  
## + x2 1 43.695 47.695  
## <none> 63.421 65.421  
## + x4 1 63.372 67.372  
##   
## Step: AIC=39.7  
## y ~ x3  
##   
## Df Deviance AIC  
## + x1 1 28.573 34.573  
## + x2 1 31.028 37.028  
## + x4 1 33.353 39.353  
## <none> 35.700 39.700  
## - x3 1 63.421 65.421  
##   
## Step: AIC=34.57  
## y ~ x3 + x1  
##   
## Df Deviance AIC  
## <none> 28.573 34.573  
## + x4 1 27.745 35.745  
## + x2 1 28.420 36.420  
## - x1 1 35.700 39.700  
## - x3 1 43.200 47.200

##   
## Call: glm(formula = y ~ x3 + x1, family = binomial(), data = da1)  
##   
## Coefficients:  
## (Intercept) x3 x1   
## -5.807 2.983 6.416   
##   
## Degrees of Freedom: 45 Total (i.e. Null); 43 Residual  
## Null Deviance: 63.42   
## Residual Deviance: 28.57 AIC: 34.57

model5 <- glm(y ~ x1+x3, family=binomial(), data=da1)  
anova(model4, model5, test="Chisq") # model5 is our final model!

## Analysis of Deviance Table  
##   
## Model 1: y ~ x1 + x2 + x3 + x4  
## Model 2: y ~ x1 + x3  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 41 27.660   
## 2 43 28.573 -2 -0.91228 0.6337

이 분산 분석 표(Analysis of Deviance Table)는 로지스틱 회귀분석 또는 일반화 선형 모델(GLM)에서 두 모델의 적합도를 비교하는 데 사용됩니다. 여기서, Model 1은 변수 x1, x2, x3, x4를 포함하고, Model 2는 변수 x1, x3만을 포함합니다. 분석의 목적은 두 모델 간의 차이가 통계적으로 유의한지를 평가하는 것입니다.

분산 분석 표의 주요 항목을 해석해 보겠습니다:

Resid. Df: 잔차의 자유도(residual degrees of freedom)입니다. 모델이 데이터에 적합할수록 이 값은 작아집니다. Model 1의 경우 41, Model 2의 경우 43입니다.

Resid. Dev: 잔차의 편차(residual deviance)입니다. 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표로, 낮을수록 모델의 적합도가 좋다고 해석됩니다. Model 1의 경우 27.660, Model 2의 경우 28.573입니다.

Df: 자유도 차이로, Model 1과 Model 2 간의 변수 개수 차이를 의미합니다. 여기서는 2입니다.

Deviance: 두 모델 간의 편차 차이로, Model 2의 잔차 편차에서 Model 1의 잔차 편차를 뺀 값입니다. 여기서는 0.91228입니다.

Pr(>Chi): 카이제곱 검정을 통해 계산된 p-value로, 두 모델 간의 차이가 통계적으로 유의한지를 나타냅니다. 여기서는 0.6337입니다. p-value가 0.05보다 크므로, 통계적으로 유의미한 수준에서 Model 1과 Model 2 간에는 유의미한 차이가 없다고 결론지을 수 있습니다. 이는 x2와 x4 변수를 제거한 Model 2가 Model 1에 비해 유의미하게 성능이 떨어지지 않음을 의미합니다. 즉, 더 단순한 Model 2도 데이터를 설명하는 데 있어 Model 1만큼 효과적이라고 볼 수 있습니다.

결론적으로, Model 1과 Model 2는 통계적으로 유의미한 차이가 없으므로, 변수의 수를 줄인 Model 2를 선호할 수 있습니다. 이는 Occam’s razor(오캄의 면도날) 원칙에 따라, 불필요한 복잡성을 피하고 가능한 가장 단순한 모델을 선택하는 것이 좋다는 원칙과 일치합니다.

# ROC (Receiver Operating Characteristic) curve

install.packages("ROCR")

## 'C:/Users/luisf/OneDrive/바탕 화면/WorkSpace/R\_LIB'의 위치에 패키지(들)을 설치합니다.  
## (왜냐하면 'lib'가 지정되지 않았기 때문입니다)

## 패키지 'ROCR'를 성공적으로 압축해제하였고 MD5 sums 이 확인되었습니다  
##   
## 다운로드된 바이너리 패키지들은 다음의 위치에 있습니다  
## C:\Users\luisf\AppData\Local\Temp\RtmpGORTjp\downloaded\_packages

library(ROCR)

## Warning: 패키지 'ROCR'는 R 버전 4.3.2에서 작성되었습니다

score = predict(model5, type = 'response', da1)  
pred = prediction(score, da1$y)  
perf = performance(pred, "tpr", "fpr")  
plot(perf, colorize = TRUE, main = "ROC Curve")  
text(0.2, 0.6, adj = 0, labels = "AUROC(the Area Under an ROC Curve")  
text(0.2, 0.45, adj = 0, labels = "- ROC Curve 아래면적을 가르킴")

