2. logistic

Park Ju ho

2022 4 13

설명변수가 1개인 로지스틱

```
library(tidyverse)
data(iris)

#setosa  versicolor
a = subset(iris, Species == "setosa"|Species == "versicolor")
#Species
a= a %>%
  mutate(Species = as.factor(Species))

str(a)
```

```
## 'data.frame':    100 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
#             = logistic
b = glm(Species~Sepal.Length, data = a, family=binomial)
summary(b)
```

```
##
## Call:
## glm(formula = Species ~ Sepal.Length, family = binomial, data = a)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05501  -0.47395  -0.02829   0.39788   2.32915
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -27.831      5.434  -5.122 3.02e-07 ***
## Sepal.Length    5.140      1.007   5.107 3.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.629  on 99  degrees of freedom
## Residual deviance:  64.211  on 98  degrees of freedom
## AIC: 68.211
##
## Number of Fisher Scoring iterations: 6
```

Coefficients

여기선 exp값이 아닌 단순 베타값이다(exp(5.140)=170정도 이므로 VersiCOlor일 오즈가 170배 증가함)
=> 이 때 이진 분류 일경우 factor의 마지막 값이 성공 즉 파이(x)값이 됨
이 둘의 유의확률이 0.05밑이므로 모두 유의하다고 할 수 있다

Deviance

유의미한 회귀인지 판단해 주는 척도
Null Deviance = 절편만을 모수로 가지는 모형
=>p값(자유도 99의 카이제곱 > 138.629의 확률이 0.005)이 0.005정도로 귀무가설을 기각, 적합 결여를 나타냄
Residual Deviance = 현재 모형(Sepal.Length와 절편을 모수로 가지는 모형)
=>p값(자유도 98의 카이제곱 > 64.211의 확률이 0.997)이 0.997정도로 귀무가설을 채택, 적합이 잘 되었다고 할 수 있음
AIC 정보량에 대한 모델 평가 통계량(값이 작을 수록 좋음)
=> 모형끼리 비교할 때 더 나은 모델을 찾기 위하여 사용
회귀 계수와 오즈의 증가량에 대한 신뢰구간

```r
coef(b)
```

```
##  (Intercept) Sepal.Length
##   -27.831451     5.140336
```

```r
exp(coef(b)["Sepal.Length"])
```

```
## Sepal.Length
##     170.7732
```

```r
confint(b, parm = "Sepal.Length")
```

```
##    2.5 %   97.5 %
## 3.421613 7.415508
```

```r
exp(confint(b,parm = "Sepal.Length"))
```

```
##      2.5 %     97.5 %
##   30.61878 1661.55385
```

적합 결과

0.5보다크면 versicolor

```
fitted(b)[c(1:5,96:100)]
```

```
##          1          2          3          4          5         96         97
## 0.16579367 0.06637193 0.02479825 0.01498061 0.10623680 0.81282396 0.81282396
##         98         99        100
## 0.98268360 0.16579367 0.81282396
```

예측

이진분류 이기에 type="response"

```
predict(b,newdata=a[c(1,50,51,100),],type="response")
```

```
##         1        50        51       100
## 0.1657937 0.1062368 0.9997116 0.8128240
```

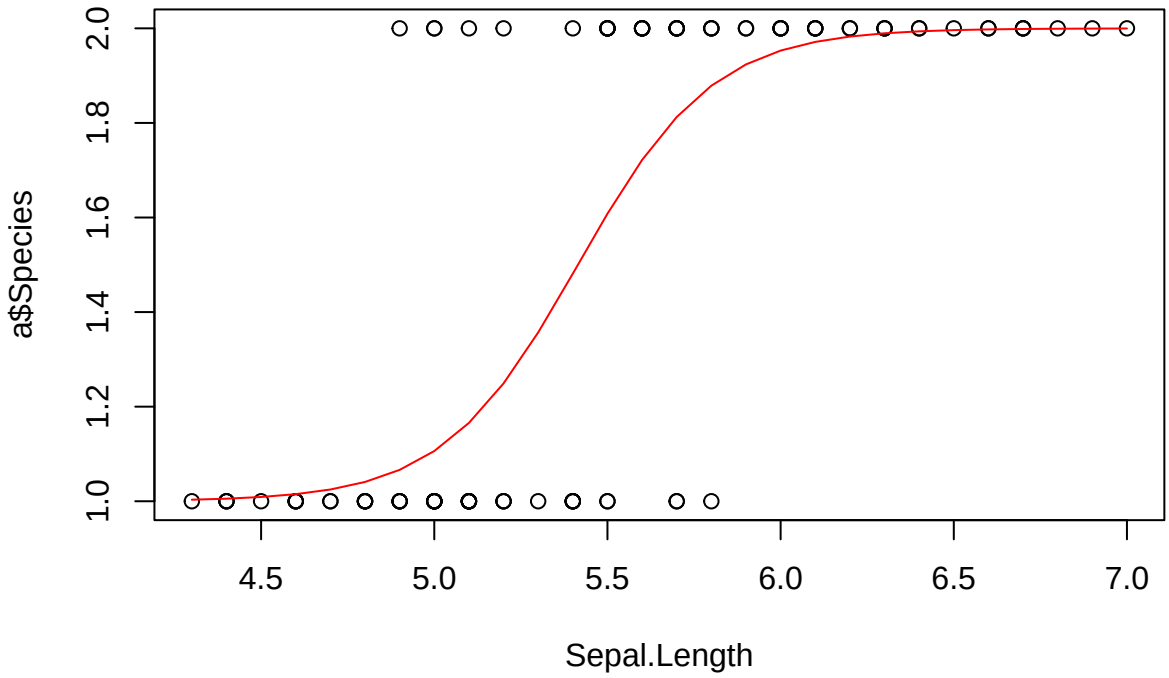
cdplot

연속형 변수 X에 대하여 범주형 변수 Y의 조건부 분포 변화를 보여줌
(Error 발생…)

```
#cdplot(Species~Sepal.Length, data=a)
```

로지스틱 회귀모형 그래프

```
plot(a$Sepal.Length, a$Species, xlab="Sepal.Length")
x=seq(min(a$Sepal.Length), max(a$Sepal.Length), 0.1)
lines(x, 1+(1/(1+(1/exp(-27.831+5.140*x)))), type="l", col="red")
```



다항 로지스틱 예측 변수가 여러개

```
#no need $ when use attach
attach(mtcars)
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
glm.vs = glm(vs~mpg+am, data = mtcars, family = binomial)
summary(glm.vs)
```

```
##
## Call:
## glm(formula = vs ~ mpg + am, family = binomial, data = mtcars)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05888  -0.44544  -0.08765   0.33335   1.68405
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.7051     4.6252  -2.747  0.00602 **
## mpg           0.6809     0.2524   2.698  0.00697 **
## am           -3.0073     1.5995  -1.880  0.06009 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 43.860  on 31  degrees of freedom
## Residual deviance: 20.646  on 29  degrees of freedom
## AIC: 26.646
##
## Number of Fisher Scoring iterations: 6
```

coefficients

해석은 위와 같음
이 때 am의 유의 확률이 0.06으로 유의 수준 0.05보다 크다 => 귀무가설을 기각 할 수 없음…

```
coef(glm.vs)
```

```
## (Intercept)         mpg          am
## -12.7051158   0.6809205  -3.0072739
```

```
exp(coef(b)[c("Sepal.Length","am")])
```

```
## Sepal.Length         <NA>
##     170.7732           NA
```

Deviance

해석 위와 같음

변수 선택

StepWise,Backward,Forward 방식이 있음
Backward = 모든 변수가 추가되어 있는 모델에서 하나씩 제거하면서 유의성 검증
Forward = 변수를 하나씩 추가해 가며 유의성 검증
SteopWise = 둘의 방식을 섞은거
backward 방식

```
step.vs = step(glm.vs,direction="backward")
```

```
## Start:  AIC=26.65
## vs ~ mpg + am
##
##        Df Deviance    AIC
## <none>      20.646 26.646
## - am    1   25.533 29.533
## - mpg   1   42.953 46.953
```

각각 mpg와 am 모두를 사용했을 때,am을 제외하였을 때, mpg를 제외하였을 때의 결과를 보여준다
이 중 아무것도 제거하지 않은 모델의 AIC가 가장 낮기에 AIC를 채택한다
forward 방식

```
step.vs = step(glm.vs,direction="forward")
```

```
## Start:  AIC=26.65
## vs ~ mpg + am
```

AIC가 가장 낮은 모델인 mpg+am 모델을 추천해 주는 것을 확인 할 수 있다
stepwise 방식

```
step.vs = step(glm.vs,direction="both")
```

```
## Start:  AIC=26.65
## vs ~ mpg + am
##
##        Df Deviance    AIC
## <none>      20.646 26.646
## - am    1   25.533 29.533
## - mpg   1   42.953 46.953
```

각각 mpg와 am 모두를 사용했을 때,am을 제외하였을 때, mpg를 제외하였을 때의 결과를 보여준다
이 중 아무것도 제거하지 않은 모델의 AIC가 가장 낮기에 AIC를 채택한다
anova를 활용한 변수 선택
변수를 입력 순서대로 하나씩 입력해 가며 데비언스의 유의성을 판단하는 모델

```
anova(glm.vs,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vs
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    31     43.860
```

```
## mpg   1   18.327        30     25.533 1.861e-05 ***
## am    1    4.887        29     20.646   0.02706 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

mpg를 추가하였을 때 Null Deviance와의 차이가 굉장히 유의하게 나옴
am을 추가하였을 때 유의확률 0.05에선 유의하였으나 mpg보단 덜 한 것을 확인 할 수 있음

```
glm.vs2 = glm(vs~am+mpg, data = mtcars, family = binomial)

anova(glm.vs2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: vs
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    31     43.860
## am    1   0.9071        30     42.953    0.3409
## mpg   1  22.3067        29     20.646 2.324e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

입력순서인 am을 먼저 검정하는 것을 확인 할 수 있음
am을 먼저 입력 시 Deviance의 차이가 유의미하지 않을 것을 확인 할 수 있음