```
rm(list=ls())
##### Chapter 3 로지스틱 회귀


### 3.1 서론
# 로지스틱 모형 : 반응 변수가 범주형인 경우 적용되는 회귀분석 모형
# 이 방법은 새로운 설명변수 or 예측변수의 값이 주어질 때
# 반응변수의 각 범주 또는 집단에 속할 확률을 알려주며(예측모형)
# 추정확률의 기준치에 따라 분류의 목적으로 사용될 수 있다. (분류모형)
# 이때 모형의 적합을 통해 추정된 확률을 사후확률이라고 한다.
# 이번 장에서는 반응 변수의 범주가 이진형인 경우만을 다루기로 한다.

# 참고
# 선형 회귀모형을 적용했을 경우 종속변수의 특성이 무시당하는 경우가 생길 수 있는데, 이를 해결하기 위해
# 예측확률을 종속변수의 값 증감에 따라 증감하고, 0과 1사이 값으로 가지는 확률로 변환해서 사용한다.
# 로지스틱 회귀모델은 odds라는 상대적인 비율 개념을 이용해 선형회귀 모델을 변형시킨 모델이다.


### 3.2 로지스틱 회귀
# 이진 반응변수 Y에 대해 다중 로지스틱 회귀모형의 일반적인 형태는 다음과 같다.
log(π(x)/(1−π(x))) = α+β1x1 + ... + βkxk

π(x) = P(Y=1|x = (x1, x2, ..., xk)) = Pr(사건발생) = 성공확률

     = exp(α+β1x1 + ... + βkxk) / (1− exp(α+β1x1 + ... + βkxk)) = 1 / (1+exp−(α+β1x1 + ... + βkxk))

     = F(α+β1x1 + ... + βkxk)

## 로지스틱 회귀모형은 오즈(odds)의 관점에서 해석될 수 있는 장점을 지닌다.
# 예) exp(β1)의 의미: 나머지 변수(x2, x3, ..., xk)가 주어졌을 때, x1이 한 단위 증가할 때마다
# 성공의 오즈 (Y=1가 될 확률) 가 몇 배 증가하는지를 나타내는 값이다.

# 표준 로지스틱 분포의 c.d.f를 F(x)라고 할 때, F(x)로 성공의 확률을 설명한다는 의미이다.
# 프로빗 모형(로지스틱 모형과 비슷한 모형)은 위에 식에서 F(.) 대신 Φ(.)로 성공의 확률을 모형화한 것이다.

Φ^−1(π(x)) = α+β1x1 +...+βkxk

π(x) = Φ(α+β1x1 +...+βkxk)

# 그래프의 형태는 설명변수가 한 개(x1)인 경우 해당 회귀계수 β1의 부호에 따라 S자 모양(β1 > 0) 또는 역 S자 모양(β1 <0)을 가진다.
# 그래프(예측변수가 1개 일 경우와 2개일 경우)

# 로지스틱 회귀가 분류의 목적으로 사용될 경우, π(x)가 기준값보다 크면 Y=1, 작으면 Y=0 집단으로 분류
# 분류의 기준값 결정은 사전정보 또는 손실함수를 사용하거나, 정분류율, 민감도, 특이도를 고려하는 등의 다양한 방법 사용 가능

## 일반화선형모형(GLM)에서의 이탈도(deviance)
# NULL Deviance : 절편모형의 완전모형으로부터의 이탈도, = 2{LL(포화모형) - LL(영모형)}, df=n-1
# Residual Deviance : 제안모형의 완전모형으로 부터의 이탈도, = 2{LL(포화모형) - LL(제안모형)}, df = n-(p+1)
# LL = 로그가능도
# 포화모형(완전모형, Full Model) = 추정해야 할 모수의 수가 데이터의 수와 같은 모형
# 영모형(Null Model) = 절편항만 가지는 모형으로 추정할 모수가 1개인 모형
# 제안모형(Proposed Model) = (P개의 모수 + 절편항)을 포함하는 모형, 추정할 모수가 (P+1)개인 모형
```

NULL Deviance와 Residual Deviance는 값이 작을수록 해당 모형이 잘 적합함을 알 수 있다.
이탈도에 근거한 검정은 두 종류의 Deviance이 해당 모형이 참일 때, 근사적으로 카이제곱에 따른다는 사실에 기초에 근거한다.
이때 자유도 = "(포화모형의 모수의 수)-(해당 모형의 모수의 수)"
영모형과 제안모형 간의 비교(검정)는 근사적으로 자유도가 {(n-1)-(n-p-1)}=p인 카이제곱분포를 따른다는 사실에 기초한다.

다중회귀에서의 변수선택법
전진선택법 : 가장 작은 모형에서 설명변수를 추가해 나가는 방법
후진 제거법 : 가장 큰 모형에서 설명변수를 제거해 나가는 방법
단계별 선택법 : 가장 작은 모형에서 설명변수를 추가하거나 제거해 나가는 방법

예제 1
반응 변수의 범주가 2개로 로지스틱 회귀를 계산하기 위해 iris 자료의 일부분만 이용하기로 한다.

데이터 선언
아래의 결과에서 Species는 Factor형 변수로 setosa는 1, versicolor는 2로 인식되고 있음을 알 수 있고,
로지스틱 회귀를 적용할 때, 큰 숫자인 versicolor일 오즈를 모형화 하므로 해석에 유의할 필요가 있다.
data(iris) # 데이터 불러오기
a <- subset(iris, Species == "setosa" | Species == "versicolor") # 종이 setosa거나 versicolor인 경우의 데이터만 저장
a$Species <- factor(a$Species) # 종의 종류가 3개에서 2개로 바뀐것을 다시 저장해준다
str(a) # 5개의 변수(Sepal.len, Sepal.Wid, Petal.len, Petal.Wid, Species)와 100개 데이터

glm()함수를 이용하여 로지스틱 회귀모형을 적합한다. 이때 family= binomial 옵션을 사용한다.
b <- glm(Species~Sepal.Length, data=a, family=binomial) # Sepal.Length를 통해 종을 구분
summary(b) # summary를 통해 결과를 확인할 수 있다.

회귀계수 검정에서 Sepal.length의 p값이 거의 0이므로 Sepal.length는 매우 유의한 변수가 된다.
Sepal.Length의 Estimate(추정계수) = 5.14
Sepal.length가 한 단위 증가할때 versicolor일 오즈가 170배 증가함을 알 수 있다.
exp(5.140) # 170.7158
exp(5.140336) # 170.7731
coef(b) # 적합값(coeficient)의 Estimate, β_0, β_1, 출력값 : -27.831 5.140
exp(coef(b)["Sepal.Length"]) # Y=2일 오즈값 ($\exp(\beta_1)$), 출력값 : 170.7158

Null deviance : 절편만 포함하는 모형($H_0 : \beta = 0$)의 완전모형(포화모형)으로부터의 이탈도(deviance)
p값이($\chi^2(n-1) < Nll\ deviance$, $1-pchisq(Nll\ deviance,\ n-1)$) 0.005302078이므로 귀무가설을 기각하며,
통계적으로 유의하여 적합 결여를 나타낸다.
1 - pchisq(138.629, 99) # 0.005302078

Residual deviance : Sepal.length가 추가된 적합모형의 이탈도
자유도 1 기준에 이탈도가 74.4(138.629(Null deviance) -> 64.211(Residual deviance) 정도의 큰 감소량을 보인다.
p값이($\chi^2(n-(p+1)) < Residual\ deviance$, $1-pchisq(Residual\ deviance,\ n-(p+1))$) 0.9966935이므로 귀무가설에 기각되지 않으며,
적합 값이 잘 적합 되고 있다고 할 수 있다.
1 - pchisq(64.211, 98) # 0.9966935

anova() 함수는 모형의 적합(변수가 추가되는) 단계별 이탈도의 감소량과 유의성 검정 결과를 제시해준다.
anova(b , test="Chisq")

confint() 함수를 이용하여 회귀계수 β와 오즈의 증가량 exp(β)에 대한 신뢰구간을 구할 수 있다.
confint(b, parm="Sepal.Length") # 회귀계수에 대한 신뢰구간, (3.421613 7.415508)
exp(confint(b, parm="Sepal.Length")) # 오즈의 증가량에 대한 신뢰구간, (30.61878 1661.55385)

fitted() 함수를 통해 적합 결과를 확인할 수 있다.
fitted(b)[c(1:5, 96:100)] # 1~5, 96~100번 데이터의 Y=2(versicolor)일 확률의 추정값, 높은 값일수록 versicolor일 확률이 높다고 해석 가능

predict() 함수를 이용해 새로운 자료에 대한 예측을 수행한다. type="response"로 지정한다.
predict(b, newdata=a [c(1, 50, 51, 100),], type="response")

cdplot() 함수는 로지스틱 회귀의 탐색적 분석에 유용하며, Sepal.Length(연속형 변수)의 변화에 따른 범주형 변수의 조건부분포를 보여준다.
# 그래프를 보아, sepal Length가 커짐에 따라 versicolor의 확률이 증가함을 보여준다.
cdplot(Species~Sepal.Length, ylevels= 2:1, data=a)

적합된 로지스틱 회귀모형의 그래프
plot(a$Sepal.Length, a$Species, xlab="Sepal.Length") # X, Y축 설정
x=seq(min(a$Sepal.Length), max(a$Sepal.Length), 0.1) # X 축 범위 설정
lines(x, 1+(1/(1+(1/exp(-27.831+5.140*x)))), type="l", col="red") # 그래프 선 그리기

예제 2
데이터 선언, 32종류의 자동차에 대해 11개의 변수를 측정한 자료
attach(mtcars) # 데이터 선언
str(mtcars) # 데이터 요약, 11개의 변수로 32개의 데이터가 있다.

glm() 함수를 사용하여 로지스틱 회귀모형 적합을 하며, 예측변수가 많아져도 적용 시 family 옵션을 binomial로 선언
# 이항변수 vs(0: flat engine, 1: straight engine)를 반응변수로
# mpg(miles:0 / gallon:0)와 am(변속기, 0:자동, 1:수동)을 예측변수로 하는 로지스틱 회귀 적합
glm.vs <- glm(vs~mpg+am, data=mtcars, family=binomial)
summary(glm.vs) # summary를 통하여 적합 결과를 확인한다.

다중 로지스틱에서 추정된 회귀계수 $\hat{\beta}$에 대한 해석
am이 주어질 때 mpg의 값이 한 단위 증가할 때마다, vs가 1일 오즈가 1.975배 즉 98%가 증가한다.
즉 am이 주어질 때, miles일 때 보다 gallon일 때 vs가 straight engine일 오즈가 98% 증가한다.
exp(glm.vs$coefficients[2]) # 1.975696
mpg이 주어질 때 am의 값이 한 단위 증가할 때마다, vs가 1일 오즈가 0.049배 즉 95%가 감소한다.
즉 mpg가 주어질 때, 변속기가 자동일 때보다 수동일 때 vs=1의 오즈가 95% 감소
exp(glm.vs$coefficients[3]) # 0.04942624

Null deviance(절편만 포함하는 모형의 완전모형으로부터의 이탈도)
p값이($\chi^2(n-1) < Nll\ deviance$, $1-pchisq(Nll\ deviance,\ n-1)$) 0.06273542이므로 귀무가설을 기각할 수 있으며,
통계이므로 유의하므로 적합 결여를 나타낸다.
1 - pchisq(43.860, 31) # 0.06273542

Residual deviance(mpg, am이 추가된 이탈도)
p값이($\chi^2(n-(p+1)) < Residual\ deviance$, $1-pchisq(Residual\ deviance,\ n-(p+1))$) 0.8717172이므로 귀무가설에 기각되지 않으며
적합 값이 잘 적합 되고 있다고 할 수 있다.
1 - pchisq(20.646, 29) # 0.8717172

step()함수의 direction 옵션을 사용하여 예측변수가 여러 개인 모형의 적합 시 변수 선택법을 설정할 수 있다.
# direction= "both(단계별 선택법)", "backward(후진제거법)", "forward(전진선택법)" / 디폴트값 : "backward"
step.vs <- step(glm.vs, direction="backward") # glm.vs : 로지스틱 회귀모형 적합한 정보
summary(glm.vs) # 단순 glm()함수로만 적합한 결과
summary(step.vs) # 위에 따로 옵션을 주지 않은 경우와 디폴드 값이 같으므로(="backward") 결과가 같다.

glm()함수는 다양한 유용한 결과를 제공, ls(), str() 함수를 통해 확인할 수 있다.
head(str(glm.vs),5) : 너무 길어서 주석처리
ls(glm.vs)

anova() 함수는 모형의 적합(변수가 추가되는) 단계별 이탈도의 감소량과 유의성 검정 결과를 제시해준다.
# Deviance Resid. : 영모양에서 예측변수가 모형에 추가됨에 따라 발생하는 이탈도의 감소량
# mpg 추가 : 18.327 감소
# 위에에서 am 추가 : 위에에서 4.887 감소
# P값 : 이탈도의 감소량에 대한 적합값으로 작으면 통계적으로 유의하다고 할수 있다.
anova(glm.vs, test="Chisq")

영모양에서 mpg가 모형에 추가됨에 따라 발생하는 이탈도의 감소량의 P값 ($\chi^2(1) < Residual\ deviance$, $1-pchisq(Residual\ deviance,\ 1)$)이
1.860515e-05이므로 통계적으로 유의하다고 할 수 있다.
1-pchisq(18.327, 1) # 이탈도 감소량(18.327), 자유도(1), 결과 : 1.860515e-05

mpg모형에서 am모형이 추가됨에 따라 발생하는 이탈도의 감소량의 P값 ($\chi^2(1) < Residual\ deviance$, $1-pchisq(Residual\ deviance,\ 1)$)이
0.02705967이므로 통계적으로 유의하다고 할 수 있다.
1-pchisq(4.887, 1) # 이탈도 감소량(4.887), 자유도(1), 결과 : 0.02705967

참고
로지스틱 회귀모형은 일반화 선형모형의 특별한 경우로 로짓모형이라고도 한다.
반응변수의 범주가 3개 이상인 경우에는 범주의 유형(명목형 또는 순서형)에 따라 다양한 다범주 로짓모형을 적합할 수 있다.