REPORT

수강과목	:	회귀분석(II)
담당교수	:	김충락
학 과	:	통계학과
학 번	:	201611531
이 름	:	정호재
제출일자	:	2019.12.16

Regression Analysis (II)
Project 2.

Due Dec. 16, 2019

You may use any statistical packages like R, minitab, spss, sas, etc.

1. Make your own dataset based on data in Example 8.9 (p. 324). Let $Y<-Y+\epsilon$, where $\epsilon \sim N(0,0.1^2)$. Model and test for (1) the effect of temperature, (2) the effect of pressure, and (3) the interaction effect.

```
> set.seed(201611531)
> temp<-as.factor(c(rep("a1",4),rep("a2",4),rep("a3",4)))
> press<-rep(c(rep("b1",2),rep("b2",2)),3)
> y<-c(6.8,6.6,5.3,6.1,7.5,7.4,7.2,6.5,7.8,9.1,8.8,9.1)
> Y<-y+rnorm(12,0,0.1)
> data<-data.frame(temp,press,Y)
> data
   temp press        Y
1    a1    b1 6.821731
2    a1    b1 6.561201
3    a1    b2 5.216727
4    a1    b2 6.204873
5    a2    b1 7.680650
6    a2    b1 7.312266
7    a2    b2 7.081167
8    a2    b2 6.502727
9    a3    b1 7.826325
10   a3    b1 9.224981
11   a3    b2 8.661723
12   a3    b2 9.145893
```

```
> aov<-aov(Y~temp+press+temp:press)
> summary(aov)
            Df  Sum Sq  Mean Sq  F value   Pr(>F)
temp         2  12.899   6.449   20.887  0.00198 **
press        1   0.569   0.569    1.844  0.22330
temp:press   2   1.032   0.516    1.670  0.26502
Residuals    6   1.853   0.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

R 분석 결과 중 가장 아래에 있는 이원분산분석표를 가지고 해석을 한다.

(1) 온도에 대한 영향이 있는지를 살펴보면, 온도에 대한 P-value는 0.00198으로 유의수준 α 0.05 보다 작다. 따라서 귀무가설 H_0을 기각하고 대립가설 H_1을 채택하여 "온도의 영향은 있다"고 판단할 수 있다.
$(H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ vs $H_1 :$모든 α_i는 0이 아니다.)

(2) 압력에 대한 P-value는 0.22330로 유의수준 α 0.05 보다 크다. 따라서 귀무가설 H_0을 채택하게 되어 "성별에 따른 통계학 성적 차이는 없다"고 판단할 수 있다.
$(H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 :$모든 β_j는 0이 아니다.)

(3) 온도와 압력의 교호작용이 있는지를 살펴보기 위해 temp:press의 P-value를 살펴보면 0.26502으로 유의수준 α 0.05보다 크다. 따라서 귀무가설 H_0을 채택하게 되어 "온도와 압력의 교호작용이 없다."라고 판단할 수 있다.
$(H_0 : (\alpha\beta)_{ij} = 0$ vs $H_1 :$모든 $(\alpha\beta)_{ij}$는 0이 아니다.)

2. Make your own dataset based on data in Table 9.1 (p. 331). Let $X <- X + \epsilon$, where $\epsilon \sim N(0, 0.01^2)$. (1) Fit to the logistic regression model. (2) Obtain 95% approximate C.I. for the median of fitted regression model.

```
> set.seed(201611531)
> x<-c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839)
> m<-c(59,60,62,56,63,59,62,60)
> y<-c(6,13,18,28,52,53,61,60)
> X<-x+rnorm(8,0,0.01)
> t<-y/m
> data<-data.frame(X,t)
> model<-glm(t~X,data=data,family="binomial",weights=m)
> summary(model)

Call:
glm(formula = t ~ X, family = "binomial", data = data, weights = m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6601  -0.4328   0.8701   1.3071   2.2108

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -56.535      4.596  -12.30   <2e-16 ***
X             31.901      2.578   12.37   <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: 284.202  on 7  degrees of freedom
Residual deviance:  18.833  on 6  degrees of freedom
AIC: 49.031

Number of Fisher Scoring iterations: 4
```

(1) 추정치는 $\hat{\beta}_0$=-56.535, $\hat{\beta}_1$=31.901이고 D=18.833이다.

따라서 로지스틱 모형은 $\log(\pi/(1-\pi)) = -56.535 + 31.901x$이다.

```
> med<-(-summary(model)$coefficients[1,1])/summary(model)$coefficients[2,1]
> med
[1] 1.772166
> vcov(model)
            (Intercept)          X
(Intercept)    21.12230  -11.844574
X             -11.84457    6.647353
> p<-(-med)
> var<-(vcov(model)[1,1]-2*p*vcov(model)[1,2]+(p^2)*vcov(model)[2,2])/
summary(model)$coefficients[2,1]^2
> var
[1] 1.73836e-05
> sqrt(var)
[1] 0.004169364
> med+qnorm(0.975)*sqrt(var)
[1] 1.780338
> med-qnorm(0.975)*sqrt(var)
[1] 1.763995
```

(2) $\hat{\pi}=0.5$가 되는 중간값은 $\widehat{\theta_{0.5}}=-\widehat{\beta_0}/\widehat{\beta_1}$=1.772166이다.

신뢰구간을 구하기 위해 $Var(\widehat{\theta_{0.5}})$을 계산해야 되는데 이에 대한 정확한 값은 구할 수 없으므로 다변량 델타법(multivariate delta method)을 이용한다.

즉, $Var[g(\widehat{\beta_0},\widehat{\beta_1})] \simeq (\frac{\partial g}{\partial \widehat{\beta_0}})^2 Var(\widehat{\beta_0})+(\frac{\partial g}{\partial \widehat{\beta_1}})^2 Var(\widehat{\beta_1})+2(\frac{\partial^2 g}{\partial \widehat{\beta_0}\partial \widehat{\beta_1}})Cov(\widehat{\beta_0},\widehat{\beta_1})$ 이므로

$Var(\widehat{\theta_{0.5}}) \simeq (v_{00}-2\hat{\rho}v_{01}+\hat{\rho}^2 v_{11})/\widehat{\beta_1}^2$으로 주어진다.

단, 여기서 $v_{00}=Var(\widehat{\beta_0})=21.12230$, $v_{01}=Cov(\widehat{\beta_0},\widehat{\beta_1})=-11.844574$,

$v_{11}=Var(\widehat{\beta_1})=6.647353$, $\hat{\rho}=\widehat{\beta_0}/\widehat{\beta_1}=-1.772166$

중간값 $\theta_{0.5}$에 대한 $100(1-\alpha)\%$ 근사적 신뢰구간은

$\widehat{\theta_{0.5}}\pm z_{\alpha/2}s.e.(\widehat{\theta_{0.5}})$으로 주어지는데 여기서 $s.e.(\widehat{\theta_{0.5}})=Var(\widehat{\theta_{0.5}})^{1/2}=0.004169364$을 나타낸다.

따라서 신뢰구간은 (1.763995, 1.780338)이다.

3. Make your own dataset based on data in Example 9.4 (p. 341). Let $X <- X + \epsilon$, where $\epsilon \sim N(0, 0.1^2)$. (1) Fit the data to the proportional odds model. (2) After 10 years of serving as a coal miner, what is the risk for being infected by the severe pneumoconiosis?

(1)

```
> install.packages("VGAM")
> set.seed(201611531)
> x<-c(5.8,15,21.5,27.5,33.5,39.5,46.0,51.5)
> X<-x+rnorm(8,0,0.1)
> normal<-c(98,51,34,35,32,23,12,4)
> mild<-c(0,2,6,5,10,7,6,2)
> severe<-c(0,1,3,8,9,8,1,5)
> data<-data.frame(X,normal,mild,severe)
> G<-glm(X~normal+mild+severe)
> summary(G)

Call:
glm(formula = X ~ normal + mild + severe)

Deviance Residuals:
     1       2       3       4       5       6       7       8
 6.142  -7.793  -8.343  -4.378   2.851   2.570   5.782   3.168

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.4971     8.7717   5.643  0.00486 **
normal       -0.5083     0.1201  -4.232  0.01335 *
mild         -0.6201     1.3064  -0.475  0.65979
severe        0.4222     1.1277   0.374  0.72711
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 61.35608)

    Null deviance: 1706.87  on 7  degrees of freedom
Residual deviance:  245.42  on 4  degrees of freedom
AIC: 60.091

Number of Fisher Scoring iterations: 2
```

로지스틱 모형에 적합하기 전에 근무기간 X보다 $\log X$를 사용하는 것이 선형석의 가정에 더 적절하므로 $logit \gamma_{ij} = \theta_j - \beta \log x_{ij},\ j = 1,2\ ;\ i = 1, \cdots, 8$을 사용한다.

```
> fit <- transform(data, let = log(X))
> library(VGAM)
> logit<-vglm(cbind(normal, mild, severe) ~ let, family = cumulative(reverse = FALSE, parallel = TRUE), data = fit)
> summary(logit)

Call:
vglm(formula = cbind(normal, mild, severe) ~ let, family = cumulative(reverse = FALSE,
    parallel = TRUE), data = fit)

Pearson residuals:
                      Min      1Q  Median     3Q    Max
logitlink(P[Y<=1]) -1.150 -0.2116 0.26017 0.4547 0.8629
logitlink(P[Y<=2]) -1.408 -0.3589 0.02701 0.3595 2.1360

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept):1   8.7646     1.2587   6.963 3.32e-12 ***
(Intercept):2   9.7572     1.2809   7.618 2.58e-14 ***
let            -2.3028     0.3654  -6.302 2.93e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])

Residual deviance: 12.7034 on 13 degrees of freedom

Log-likelihood: -28.0425 on 13 degrees of freedom

Number of Fisher scoring iterations: 4

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1'


Exponentiated coefficients:
       let
0.09997925
```

진폐증 감염정도가 3가지로 총 2개의 intercept와 근무기간에 대한 1개의 coefficient가 추정되었다. 비례-오즈모형은 $logit\gamma_{ij} = \theta_j - 2.3028 \times \log x_i$으로 나타난다.

(2) 10년 근무한 광부가 심각한 진폐증에 걸릴 위험은 exp(9.7572-2.3028$\times$ log10)이다.

```
> exp(9.7572-2.3028*log(10))
[1] 86.03955
```

따라서 10년 근무한 광부가 심각한 진폐증에 걸릴 위험은 86명중 1명꼴이다.