범주형자료분석방법론 HW3

이름: 김연주

학과: 통계학과

학번: 2021250461

1) 표 5.6

```
> t5.6=expand.grid(AZT=factor(c("Yes","No"),levels=c("No","Yes")),Race=factor(c("Wh
ite","Black"),levels=c("Black","White")))
> t5.6=data.frame(t5.6,Yes=c(14,32,11,12), No=c(93,81,52,43))
> print(t5.6)
  AZT  Race Yes No
1 Yes White  14 93
2  No White  32 81
3 Yes Black  11 52
4  No Black  12 43
>
> fit1=glm(cbind(Yes,No)~AZT+Race, family=binomial, data=t5.6)
> summary(fit1)

Call:
glm(formula = cbind(Yes, No) ~ AZT + Race, family = binomial,
    data = t5.6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.07357    0.26294  -4.083 4.45e-05 ***
AZTYes      -0.71946    0.27898  -2.579  0.00991 **
RaceWhite    0.05548    0.28861   0.192  0.84755
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.3835  on 1  degrees of freedom
AIC: 24.86

Number of Fisher Scoring iterations: 4
```

logit[π(x)] = -1.07357-0.71946AZT+0.05548Race

모델 요약 결과 AZT 변수의 계수는 유의하지만 Race 변수의 계수는 유의하지 않다.

```
> exp(-0.71946)
[1] 0.4870152
>
> qchisq(df=1,0.95)
[1] 3.841459
```

Residual deviance가 임계값인 3.84보다 작으므로 모델이 saturated model보다 데이터를 더 잘 적합함을 알 수 있다. AZTyes의 계수를 exp처리한 값은 0.487이다. 이는 에이즈 발병의 오즈가

AZT를 사용했을 때 사용하지 않았을 때보다 0.487배라는 것을 의미한다.

```
> fit2=glm(cbind(Yes,No)~AZT*Race,family=binomial, data=t5.6)
> summary(fit2)

Call:
glm(formula = cbind(Yes, No) ~ AZT * Race, family = binomial,
    data = t5.6)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.2763     0.3265  -3.909 9.26e-05 ***
AZTYes            -0.2771     0.4655  -0.595    0.552
RaceWhite          0.3476     0.3875   0.897    0.370
AZTYes:RaceWhite  -0.6878     0.5852  -1.175    0.240
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  8.3499e+00  on 3  degrees of freedom
Residual deviance: -1.5987e-14  on 0  degrees of freedom
AIC: 25.476

Number of Fisher Scoring iterations: 3
```

AZT*Race 변수 추가

```
> fit3=glm(cbind(Yes,No)~AZT, family=binomial, data=t5.6)
> summary(fit3)

Call:
glm(formula = cbind(Yes, No) ~ AZT, family = binomial, data = t5.6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0361     0.1755  -5.904 3.54e-09 ***
AZTYes       -0.7218     0.2787  -2.590  0.00961 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 1.4206  on 2  degrees of freedom
AIC: 22.897

Number of Fisher Scoring iterations: 4
```

AZT 변수만 사용

```
> fit4=glm(cbind(Yes,No)~Race, family=binomial, data=t5.6)
> summary(fit4)

Call:
glm(formula = cbind(Yes, No) ~ Race, family = binomial, data = t5.6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.41838    0.23239  -6.103 1.04e-09 ***
RaceWhite    0.08797    0.28547   0.308    0.758
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.3499  on 3  degrees of freedom
Residual deviance: 8.2544  on 2  degrees of freedom
AIC: 29.731

Number of Fisher Scoring iterations: 4
```

Race 변수만 사용

```
> fit5=glm(cbind(Yes,No)~1, family=binomial, data=t5.6)
> summary(fit5)

Call:
glm(formula = cbind(Yes, No) ~ 1, family = binomial, data = t5.6)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.3606     0.1349  -10.08   <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: 8.3499  on 3  degrees of freedom
Residual deviance: 8.3499  on 3  degrees of freedom
AIC: 27.826

Number of Fisher Scoring iterations: 4
```

Null model

```
> model=c("AZT*Race","AZT+Race","AZT","Race","null")
> df=c(0,1,2,2,3)
> deviance=c(fit1$deviance,fit2$deviance,fit3$deviance,fit4$deviance,fit5$deviance)
> AIC=c(fit1$aic,fit2$aic,fit3$aic,fit4$aic,fit5$aic)
> compare_model=c("_","2-1","3-2","4-3","5-3")
> compare_dev=c("_",c(fit2$deviance-fit1$deviance, fit3$deviance-fit2$deviance, fit4$deviance-fit3
$deviance, fit5$deviance-fit3$deviance))
> compare_df=c(1,1,1,1,1)
> model_diag=data.frame(model,df,deviance,AIC,compare_model,compare_dev,compare_df)
> print(model_diag)
     model df      deviance      AIC compare_model       compare_dev compare_df
1 AZT*Race  0  1.383530e+00 24.85981             _                 _          1
2 AZT+Race  1 -1.598721e-14 25.47628           2-1 -1.38353009248638          1
3      AZT  2  1.420614e+00 22.89689           3-2  1.42061364205088          1
4     Race  2  8.254436e+00 29.73071           4-3  6.83382275208821          1
5     null  3  8.349946e+00 27.82622           5-3  6.92933240995818          1
> #fit3 선택
> residuals(fit3,type='pearson')
         1          2          3          4
-0.4736703  0.5145232  0.6173021 -0.7375014
> residuals(fit3, type = "pearson")/sqrt(1 - lm.influence(fit3)$hat)
         1          2          3          4
-0.7780908  0.8992455  0.7780908 -0.8992455
```

다섯개의 모델을 표로 비교한 결과는 위와 같다. 1,2번 모델을 비교한 결과 compared_dev값이 임계값인 3.84보다 작기에 더 작은 모델인 두번째 모델이 채택된다. 2,3번 모델을 비교하면 같은 이유로 인해 더 작은 모델인 3번 모델이 적합하다. 3,4번 모델을 비교하면 deviance값의 차이가 3.84보다 크기 때문에 3번 모델이 채택된다. 마지막으로 3,5번 모델을 비교하면 같은 이유로 3번 모델이 채택된다. AIC를 비교해 보아도 3번 모델의 AIC가 가장 낮기에 이를 택하는 것이 적합하다.

채택된 모델 (AZT변수만 사용)의 잔차와 표준화잔차를 도출하였다. 그 결과 모델이 잘 적합함을 확인할 수 있었다.

2) 표 6.5, 6.7

```
> #6.5, 6.7
> sp=factor(c("<117","117-126","127-136","137-146","147-156","157-166","167-186",">
186"))
> n=c(156,252,284,271,139,85,99,43)
> obs=c(3,17,12,16,12,8,16,8)
> bp=c(111.5,121.5,131.5,141.5,151.5,161.6,176.5,191.5)
> t6.5=data.frame(sp,bp,n,obs)
> t6.5
       sp    bp   n obs
1    <117 111.5 156   3
2 117-126 121.5 252  17
3 127-136 131.5 284  12
4 137-146 141.5 271  16
5 147-156 151.5 139  12
6 157-166 161.6  85   8
7 167-186 176.5  99  16
8    >186 191.5  43   8
>
```

```
> fit6=glm(obs/n~bp,family=binomial,weights=n,data=t6.5)
> summary(fit6)

Call:
glm(formula = obs/n ~ bp, family = binomial, data = t6.5, weights = n)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.081046   0.724230  -8.397  < 2e-16 ***
bp           0.024330   0.004842   5.024 5.05e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.0226  on 7  degrees of freedom
Residual deviance:  5.9134  on 6  degrees of freedom
AIC: 42.615

Number of Fisher Scoring iterations: 4

> qchisq(0.95,df=6)
[1] 12.59159
>
```

모델 적합 결과 residual deviance가 자유도 6인 카이제곱 분포의 임계값 12.592보다 작으므로 해당 모델이 saturated model보다 데이터를 잘 적합함을 알 수 있다.

```
> #standard residual
> s_res=resid(fit6, type = "pearson")/sqrt(1 - lm.influence(fit6)$hat)
>
> r1=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-1,])
> r2=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-2,])
> r3=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-3,])
> r4=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-4,])
> r5=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-5,])
> r6=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-6,])
> r7=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-7,])
> r8=glm(obs/n~bp,family=binomial,weights=n,data=t6.5[-8,])
>
> G=c(fit6$deviance-r1$deviance,fit6$deviance-r2$deviance,fit6$deviance-r3$devianc
e,fit6$deviance-r4$deviance,fit6$deviance-r5$deviance,fit6$deviance-r6$deviance,fit
6$deviance-r7$deviance,fit6$deviance-r8$deviance)
> G
[1] 1.38682177 5.13817728 0.93478517 0.33606512 0.01599573 0.11325019 0.42199844
[8] 0.03092674
>
> #pearson residual
> p_res=resid(fit6, type = "pearson")
> p_diff=c(p_res[1]^2,p_res[2]^2,p_res[3]^2,p_res[4]^2,p_res[5]^2,p_res[6]^2,p_res
[7]^2,p_res[8]^2)
> p_diff
         1          2          3          4          5          6          7
0.95952629 4.02309509 0.66098125 0.25607027 0.01402959 0.09602632 0.26541626
         8
0.01903237
> #standardized residual (table 6.7)
> s_diff=c(s_res[1]^2,s_res[2]^2,s_res[3]^2,s_res[4]^2,s_res[5]^2,s_res[6]^2,s_res
[7]^2,s_res[8]^2)
> s_diff
         1          2          3          4          5          6          7
1.22311954 5.63944496 0.89286868 0.32714094 0.01613177 0.11038294 0.42763364
         8
0.03075717
```

```
> #df
> df = c((fit6$coefficients[2] - r1$coefficients[2]) / sqrt(vcov(fit6)[2, 2]),(fit6
$coefficients[2] - r2$coefficients[2]) / sqrt(vcov(fit6)[2, 2]),(fit6$coefficients
[2] - r3$coefficients[2]) / sqrt(vcov(fit6)[2, 2]),(fit6$coefficients[2] - r4$coeff
icients[2]) / sqrt(vcov(fit6)[2, 2]),(fit6$coefficients[2] - r5$coefficients[2]) /
sqrt(vcov(fit6)[2, 2]),(fit6$coefficients[2] - r6$coefficients[2]) / sqrt(vcov(fit
6)[2, 2]),(fit6$coefficients[2] - r7$coefficients[2]) / sqrt(vcov(fit6)[2, 2]),(fit
6$coefficients[2] - r8$coefficients[2]) / sqrt(vcov(fit6)[2, 2]))
> df
          bp           bp           bp           bp           bp           bp
 0.484546941 -1.233261336  0.315906335  0.079300520  0.007879018 -0.066291266
          bp           bp
 0.412778680 -0.121788108
>
> #표로 만들어서 비교
> diag2=data.frame(bp,df,p_diff,s_diff,G)
> diag2
     bp           df     p_diff     s_diff          G
1 111.5  0.484546941 0.95952629 1.22311954 1.38682177
2 121.5 -1.233261336 4.02309509 5.63944496 5.13817728
3 131.5  0.315906335 0.66098125 0.89286868 0.93478517
4 141.5  0.079300520 0.25607027 0.32714094 0.33606512
5 151.5  0.007879018 0.01402959 0.01613177 0.01599573
6 161.6 -0.066291266 0.09602632 0.11038294 0.11325019
7 176.5  0.412778680 0.26541626 0.42763364 0.42199844
8 191.5 -0.121788108 0.01903237 0.03075717 0.03092674
```

표 6.7의 결과를 diag2로 표현하였다. 표에서 두번째 행의 값들이 다른 값들보다 훨씬 큰 것을 알 수 있다. 이를 통해 두번째 관측값이 가장 큰 영향력을 가짐을 확인할 수 있다. 이는 두번째 관찰값이 이상치로 작용할 수 있음을 시사한다.

```
> fit7=glm(obs/n~1,family=binomial,weights=n,data=t6.5)
> summary(fit7)

Call:
glm(formula = obs/n ~ 1, family = binomial, data = t6.5, weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5987     0.1081  -24.05   <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: 30.023  on 7  degrees of freedom
Residual deviance: 30.023  on 7  degrees of freedom
AIC: 64.724

Number of Fisher Scoring iterations: 4
```

```
> #6.7
> s_res.ind=resid(fit7, type = "pearson")/sqrt(1 - lm.influence(fit7)$hat)
>
> r1.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-1,])
> r2.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-2,])
> r3.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-3,])
> r4.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-4,])
> r5.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-5,])
> r6.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-6,])
> r7.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-7,])
> r8.ind=glm(obs/n~1,family=binomial,weights=n,data=t6.5[-8,])
>
> G.ind=c(fit7$deviance-r1.ind$deviance,fit7$deviance-r2.ind$deviance,fit7$deviance
-r3.ind$deviance,fit7$deviance-r4.ind$deviance,fit7$deviance-r5.ind$deviance,fit7$d
eviance-r6.ind$deviance,fit7$deviance-r7.ind$deviance,fit7$deviance-r8.ind$devianc
e)
>
> #pearson residual
> p_res.ind=resid(fit7, type = "pearson")
> p_diff.ind=c(p_res.ind[1]^2,p_res.ind[2]^2,p_res.ind[3]^2,p_res.ind[4]^2,p_res.in
d[5]^2,p_res.ind[6]^2,p_res.ind[7]^2,p_res.ind[8]^2)
> #standardized residual (table 6.7)
> s_diff.ind=c(s_res.ind[1]^2,s_res.ind[2]^2,s_res.ind[3]^2,s_res.ind[4]^2,s_res.in
d[5]^2,s_res.ind[6]^2,s_res.ind[7]^2,s_res.ind[8]^2)
>
> diag2.ind=data.frame(p_diff.ind,s_diff.ind,G.ind)
> diag2.ind
   p_diff.ind  s_diff.ind       G.ind
1  6.05140873  6.85619967  9.10672745
2  0.01217915  0.01502886  0.01512068
3  3.20641469  4.07782308  4.53582651
4  0.43624638  0.54798813  0.56812426
5  0.63125065  0.70498497  0.66274271
6  0.81743679  0.87329059  0.79656796
7 13.11561504 14.17126210 10.87035484
8  9.10765714  9.41219000  6.74014336
>
```

독립성 모델. 관측값을 하나씩 제거했을 때 독립성 모델을 살펴보았다.

3) 표 7.1

```
> #7.1
> t7.1=data.frame(LogDose=c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.883
9),n=c(59,60,62,56,63,59,62,60),y=c(6,13,18,28,52,53,61,60))
> fit8=glm(y/n ~ LogDose, weights=n, family=binomial,data=t7.1)
> summary(fit8)

Call:
glm(formula = y/n ~ LogDose, family = binomial, data = t7.1,
    weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
LogDose       34.270      2.912   11.77   <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:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

logit[π(x)] = −60.717 + 34.270LogDose. LogDose의 한 단위 증가는 사건 발생의 오즈를 exp(34.27)배 증가시킨다. 또한 해당 계수는 통계적으로 유의하다.

```
> fit9=glm(y/n ~ LogDose, weights=n,family=binomial(link=probit),data=t7.1)
> summary(fit9)

Call:
glm(formula = y/n ~ LogDose, family = binomial(link = probit),
    data = t7.1, weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935      2.648  -13.19   <2e-16 ***
LogDose       19.728      1.487   13.27   <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: 284.20  on 7  degrees of freedom
Residual deviance:  10.12  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4
```

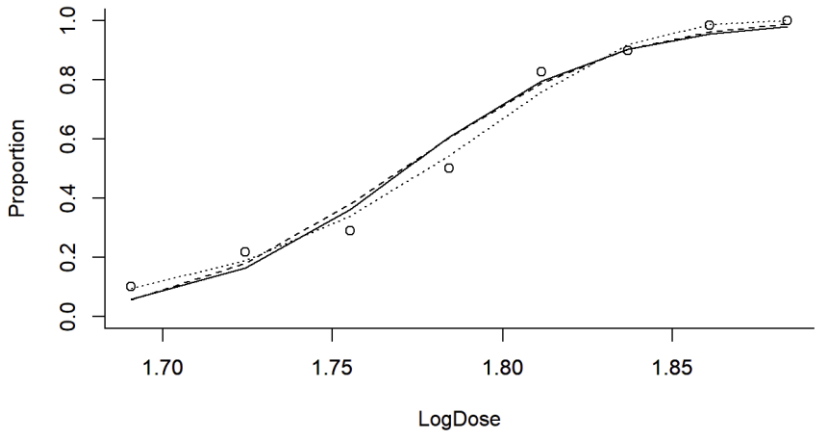
Probit link 사용. $\varphi^{-1}[\pi(x)]$ =-34.935 + 19.728LogDose.

```
> fit10=glm(y/n ~ LogDose, weights=n,family=binomial(link=cloglog),data=t7.1)
> summary(fit10)

Call:
glm(formula = y/n ~ LogDose, family = binomial(link = cloglog),
    data = t7.1, weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -39.572      3.240  -12.21   <2e-16 ***
LogDose       22.041      1.799   12.25   <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: 284.2024  on 7  degrees of freedom
Residual deviance:   3.4464  on 6  degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4
```

cloglog link 사용. Log[-log(1-π(x))] =-39.572 + 22.041LogDose.

해당 모델은 앞선 fit8, fit9와 달리 π(x)가 0으로는 천천히, 1로는 빠른 속도로 가까워진다. 또한 세 모델 중 AIC가 가장 작다.

```
> plot(t7.1$LogDose,t7.1$y/t7.1$n,ylim=c(0,1), xlab="LogDose",ylab="Proportion",bty
="L")
> axis(side=1, at=seq(from=1.5,to=2,by=0.05))
> lines(t7.1$LogDose,predict(fit8, type="response"),lty=1)
> lines(t7.1$LogDose,predict(fit9, type="response"),lty=2)
> lines(t7.1$LogDose,predict(fit10, type="response"),lty=3)
>
```



세 모델을 그래프를 통해 비교하면 실선은 fit8, 점실선은 fit9, 점선은 fit10이다. 비교 결과 세 모델의 예측값이 비슷함을 알 수 있다.