범주형자료분석방법론 HW2

이름: 김연주

학번: 2021250461

1) Table 4.3 데이터 로드

```
> #범주형자료분석방법론 HW2
>
> crabs=read.table("http://www.stat.ufl.edu/~aa/cat/data/Crabs.dat", header
= TRUE)
> head(crabs)
  crab sat y weight width color spine
1    1   8 1   3.05  28.3     2     3
2    2   0 0   1.55  22.5     3     3
3    3   9 1   2.30  26.0     1     1
4    4   0 0   2.10  24.8     3     3
5    5   4 1   2.60  26.0     3     3
6    6   0 0   2.10  23.8     2     3
```

2) Poisson regression model – log link

```
> #poisson regression model
> #log link
> model1=glm(sat~width, data=crabs, family=poisson(link=log))
> summary(model1)

Call:
glm(formula = sat ~ width, family = poisson(link = log), data = crabs)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.30476    0.54224  -6.095  1.1e-09 ***
width        0.16405    0.01997   8.216  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: 927.18

Number of Fisher Scoring iterations: 6
```

Model1에 따르면 $\log\hat{\mu}(x) = -3.30476 + 0.16405x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 e^0.16405 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 e^0.16405배 증가한다는 것이다. 또한 회귀계수의 z검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> #LRT
> 632.79-567.88
[1] 64.91
> qchisq(0.95,df=1)
[1] 3.841459
```

모델의 null deviance와 residual deviance의 차이를 통해 LR test를 하면 검정통계량인 64.91이 임계값 3.84보다 훨씬 크기 때문에 p-value는 0.05보다 작다. 그러므로 width는 유의한 변수라는 것을 알 수 있다.

3) poisson regression model – identity link

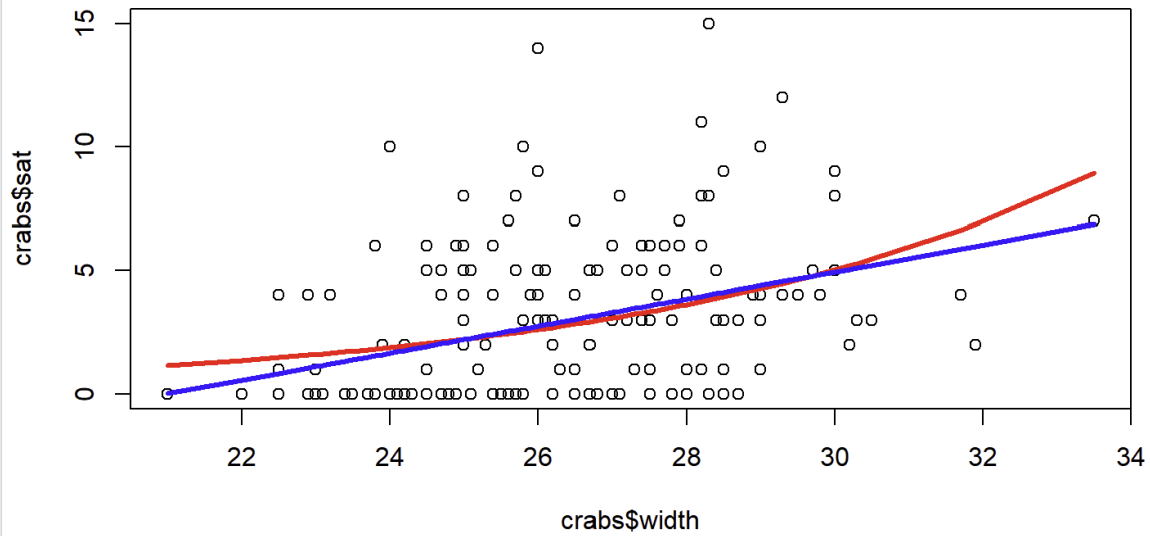
```
> #identity link
> model2=glm(sat~width, data=crabs, family=poisson(link=identity), start=coe
f(model1))
> summary(model2)

Call:
glm(formula = sat ~ width, family = poisson(link = identity),
    data = crabs, start = coef(model1))

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.52547    0.67767  -17.01   <2e-16 ***
width         0.54925    0.02968   18.50   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 557.71  on 171  degrees of freedom
AIC: 917.01

Number of Fisher Scoring iterations: 22
```

Model2에 따르면 $\hat{\mu}(x) = -11.52547 + 0.54925x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 0.54925 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 0.54925배 증가한다는 것이다. 또한 회귀계수의 z검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> #LRT
> 632.79-557.71
[1] 75.08
```

모델의 null deviance와 residual deviance의 차이를 통해 LR test를 하면 검정통계량인 75.08이 임계값 3.84보다 훨씬 크기 때문에 p-value는 0.05보다 작다. 그러므로 width는 유의한 변수라는 것을 알 수 있다.

```
> #비교
> ind=order(crabs$width)
> plot(crabs$width,crabs$sat)
> lines(x=crabs$width[ind],y=fitted(model1)[ind],col='red',lwd=3)
> lines(x=crabs$width[ind],y=fitted(model2)[ind],col='blue',lwd=3)
>
>
```



4) negative binomial regression model – log link

```
> #negative binomial regression
> #log link
> library(MASS)
> model3=glm.nb(sat~width, data=crabs, link=log)
> summary(model3)

Call:
glm.nb(formula = sat ~ width, data = crabs, link = log, init.theta = 0.90456
808)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.05251    1.17143  -3.459 0.000541 ***
width        0.19207    0.04406   4.360  1.3e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.9046) family taken to be 1)

    Null deviance: 213.05  on 172  degrees of freedom
Residual deviance: 195.81  on 171  degrees of freedom
AIC: 757.29

Number of Fisher Scoring iterations: 1


              Theta:  0.905
          Std. Err.:  0.161

 2 x log-likelihood:  -751.291
```

Model3에 따르면 $\log\hat{\mu}(x) = -4.05251 + 0.19207x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 e^0.19207 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 e^0.19207배 증가한다는 것이다. 또한 회귀계수의 z검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> #LRT
> 213.05-195.81
[1] 17.24
> qchisq(0.95,df=171)
[1] 202.5126
>
```

모델의 null deviance와 residual deviance의 차이를 통해 LR test를 하면 검정통계량인 17.24가 임계값 3.84보다 크기 때문에 p-value는 0.05보다 작다. 그러므로 width는 유의한 변수라는 것을 알 수 있다. 또한, residual deviance인 195.81은 자유도가 171인 카이제곱분포의 임계값보다 작으므로 모델이 제이터를 잘 적합함을 알 수 있다.

5) Negative binomial regression model – identity link

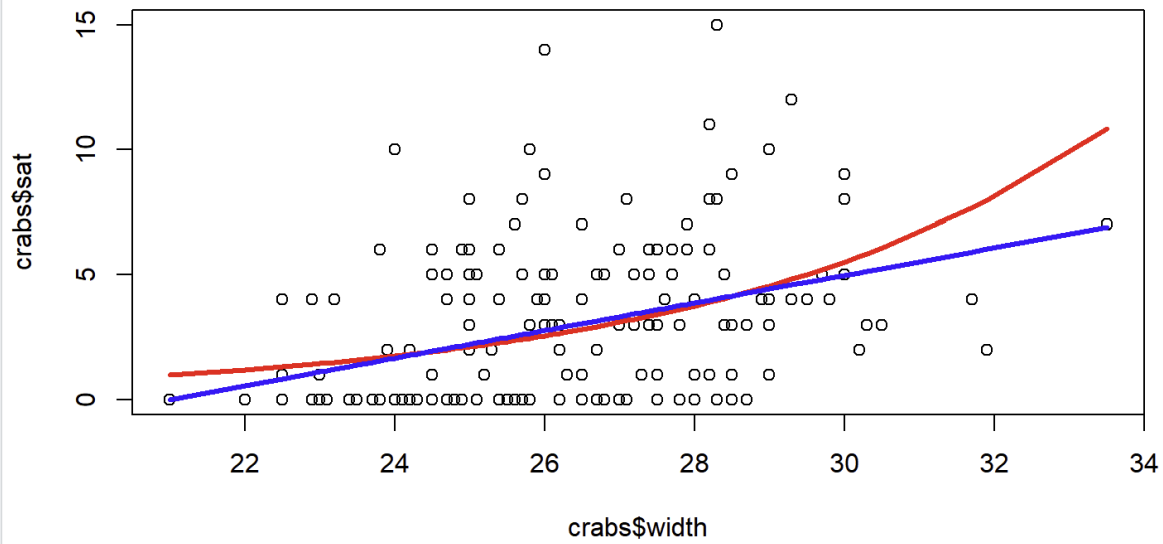
```
> #identity link
> model4=glm.nb(sat~width, data=crabs, link=identity, start=coef(model1))
> summary(model4)

Call:
glm.nb(formula = sat ~ width, data = crabs, start = coef(model1),
    link = identity, init.theta = 0.9316967133)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.63354    1.07112  -10.86   <2e-16 ***
width         0.55398    0.05101   10.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.9317) family taken to be 1)

    Null deviance: 216.51  on 172  degrees of freedom
Residual deviance: 195.52  on 171  degrees of freedom
AIC: 753.93

Number of Fisher Scoring iterations: 1


              Theta:  0.932
          Std. Err.:  0.168

 2 x log-likelihood:  -747.928
```

Model4에 따르면 $\hat{\mu}(x) = -11.63354 + 0.55398x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 0.55398 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 0.55398배 증가한다는 것이다. 또한 회귀계수의 z검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> #LRT
> 216.51-195.52
[1] 20.99
>
```

모델의 null deviance와 residual deviance의 차이를 통해 LR test를 하면 검정통계량인 20.99가 임계값 3.84보다 훨씬 크기 때문에 p-value는 0.05보다 작다. 그러므로 width는 유의한 변수라는 것을 알 수 있다. 또한, residual deviance인 195.52가 자유로 171인 카이제곱분포의 임계값보다 작으므로 이 모델도 데이터를 잘 적합한다.

```
> #비교
> plot(crabs$width,crabs$sat)
> lines(x=crabs$width[ind],y=fitted(model3)[ind],col='red',lwd=3)
> lines(x=crabs$width[ind],y=fitted(model4)[ind],col='blue',lwd=3)
>
```



6) Poisson regression model with quasi-likelihood – log link

```
> #Poisson regression with quasi-likelihood
> #log link
> model5=glm(sat~width, data = crabs, family=quasi(link = "log", variance =
"mu"))
> summary(model5)

Call:
glm(formula = sat ~ width, family = quasi(link = "log", variance = "mu"),
    data = crabs)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.30476    0.96729  -3.417 0.000793 ***
width        0.16405    0.03562   4.606 7.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 3.182205)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 567.88  on 171  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

Model5에 따르면 $\log\hat{\mu}(x) = -3.30476 + 0.16405x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 e^0.16405 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 e^0.16405배 증가한다는 것이다. 또한 회귀계수의 t검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> chisq=sum(resid(model1,type='pearson')^2)
> chisq
[1] 544.157
> phi=sqrt(chisq/171)
> phi
[1] 1.783874
> phi*0.01997
[1] 0.03562395
```

Quasi likelihood 에서 width의 standard error는 기존 width 변수의 standard error에 φ를 곱하여 나온 결과임을 알 수 있다.

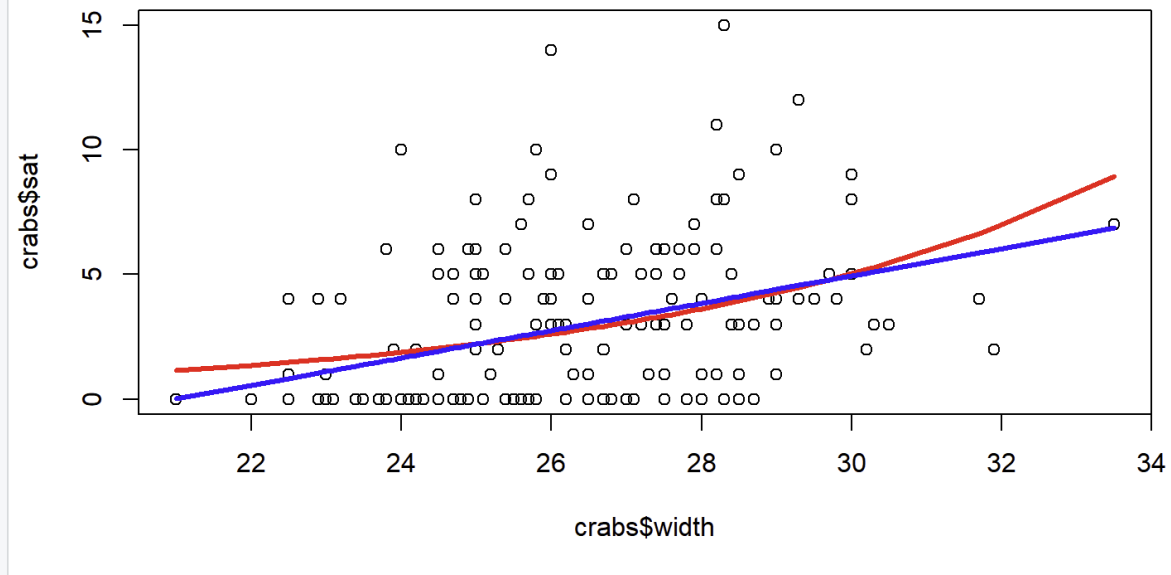
7) Poisson regression model with quasi-likelihood – identity link

```
> #identity link
> model6=glm(sat~width, data = crabs, family=quasi(link = "identity", varian
ce= "mu"),start = coef(model1))
> summary(model6)

Call:
glm(formula = sat ~ width, family = quasi(link = "identity",
    variance = "mu"), data = crabs, start = coef(model1))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.52547    1.20694  -9.549   <2e-16 ***
width         0.54925    0.05286  10.390   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi family taken to be 3.17204)

    Null deviance: 632.79  on 172  degrees of freedom
Residual deviance: 557.71  on 171  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 22
```

Model6에 따르면 $\hat{\mu}(x) = -11.52547 + 0.54925x$ 이다. 이는 width의 한 단위 증가는 μ에 대해 0.54925 만큼의 영향을 미친다는 것을 의미한다. 즉, width가 한 단위 증가하면 satellite의 수는 0.54925배 증가한다는 것이다. 또한 회귀계수의 t검정 결과가 유의하므로 width는 satellite수를 예측하기에 유의한 변수이다.

```
> chisq2=sum(resid(model2,type='pearson')^2)
> chisq2
[1] 542.4253
> phi2=sqrt(chisq2/171)
> phi2
[1] 1.781033
> phi2*0.02968
[1] 0.05286106
```

```
> #비교
> plot(crabs$width,crabs$sat)
> lines(x=crabs$width[ind],y=fitted(model5)[ind],col='red',lwd=3)
> lines(x=crabs$width[ind],y=fitted(model6)[ind],col='blue',lwd=3)
> 
```



분석 결과 앞선 poisson regression model과 residual deviance가 각각 동일했다. 그러므로 poisson regression model과 같은 그림이 완성되었다.