stat431 a3 q2

Yiming Shen 20891774

16/11/2023

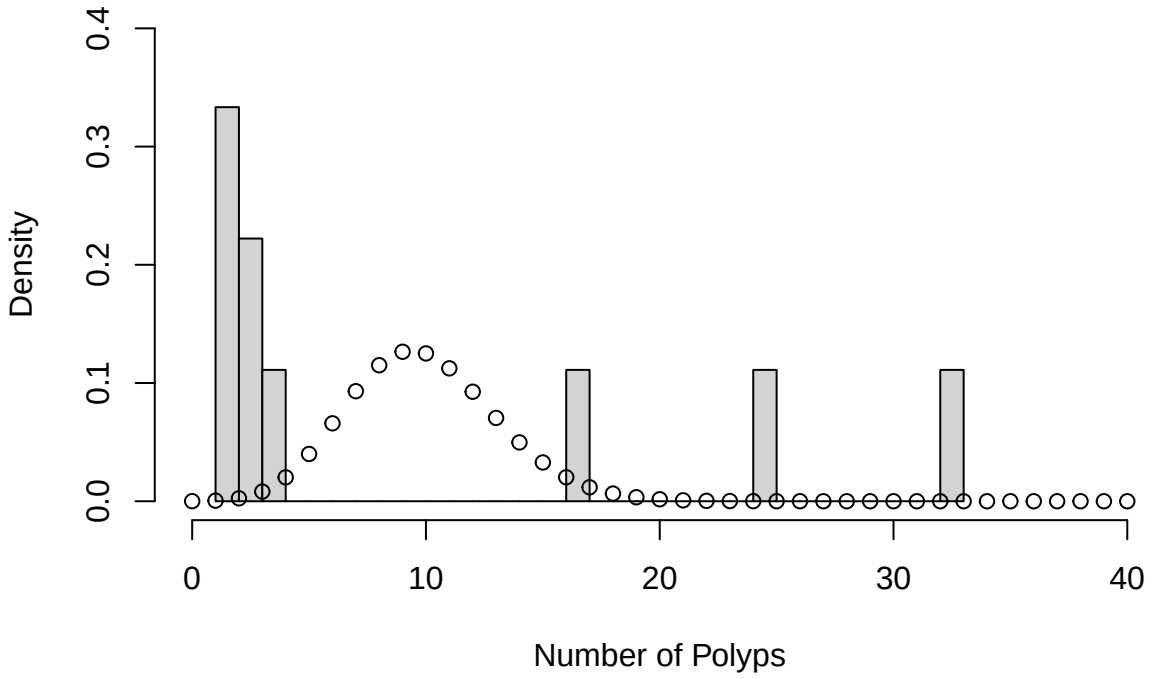
```
library(GLMsData)
data(polyps)
polyps
```

```
##    Number Treatment Age
## 1       1      Drug  22
## 2       1      Drug  23
## 3       2      Drug  16
## 4       3      Drug  23
## 5       3      Drug  23
## 6       4      Drug  42
## 7      17      Drug  22
## 8      25      Drug  17
## 9      33      Drug  23
## 10      7   Placebo  34
## 11     10   Placebo  30
## 12     15   Placebo  50
## 13     28   Placebo  18
## 14     28   Placebo  22
## 15     40   Placebo  27
## 16     44   Placebo  19
## 17     46   Placebo  22
## 18     50   Placebo  34
## 19     61   Placebo  13
## 20     63   Placebo  20
```

(a)

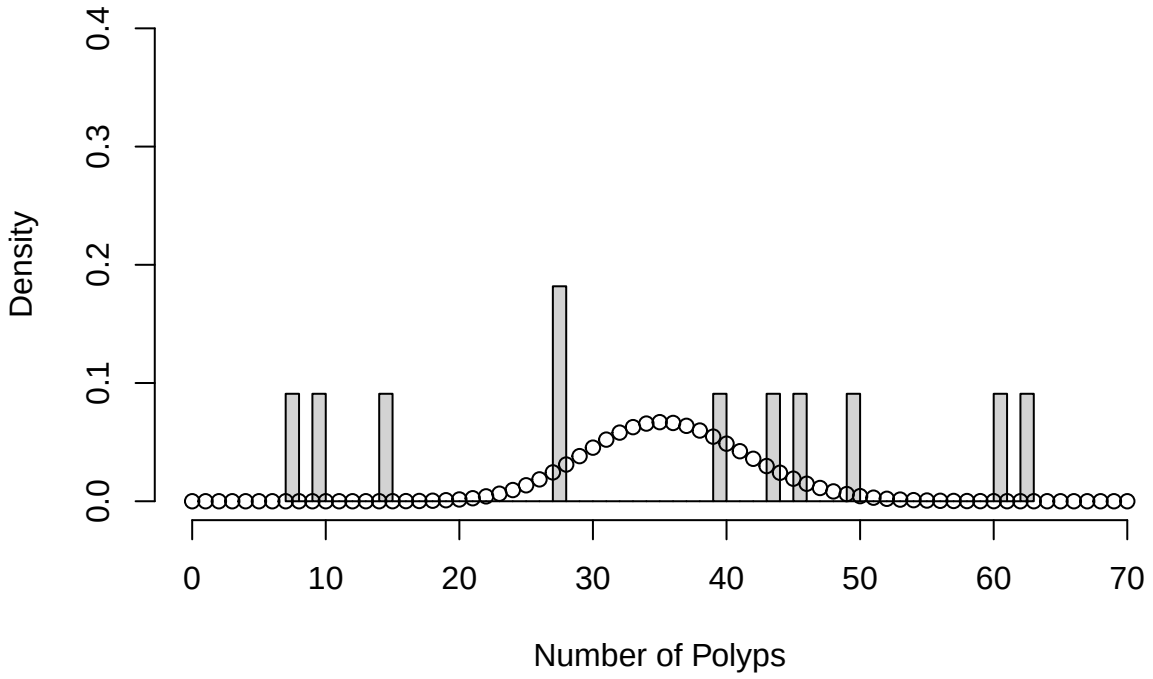
```
# hist for treatment
hist(polyps$Number[polyps$Treatment=="Drug"],
     xlab="Number of Polyps",
     main="The number of polyps in the treatment group",
     breaks=40,
     xlim=c(0,40),
     ylim=c(0,0.4),
     probability = T)

points(x=seq(0,40),y=dpois(x=seq(0,40),lambda = mean(polyps$Number[polyps$Treatment=="Drug"])))
```




```
# hist for placebo
hist(polyps$Number[polyps$Treatment=="Placebo"],
     xlab="Number of Polyps",
     main="The number of polyps in the placebo group",
     breaks=40,
     xlim=c(0,70),
     ylim=c(0,0.4),
     probability = T)

points(x=seq(0,70),y=dpois(x=seq(0,70),lambda = mean(polyps$Number[polyps$Treatment=="Placebo"])))
```



Comments for 'hist for treatment': The observed distribution is not fitted with the poisson distribution. More observations are concentrated around 0 while the poisson distribution has the highest predicated probability around range 5-15

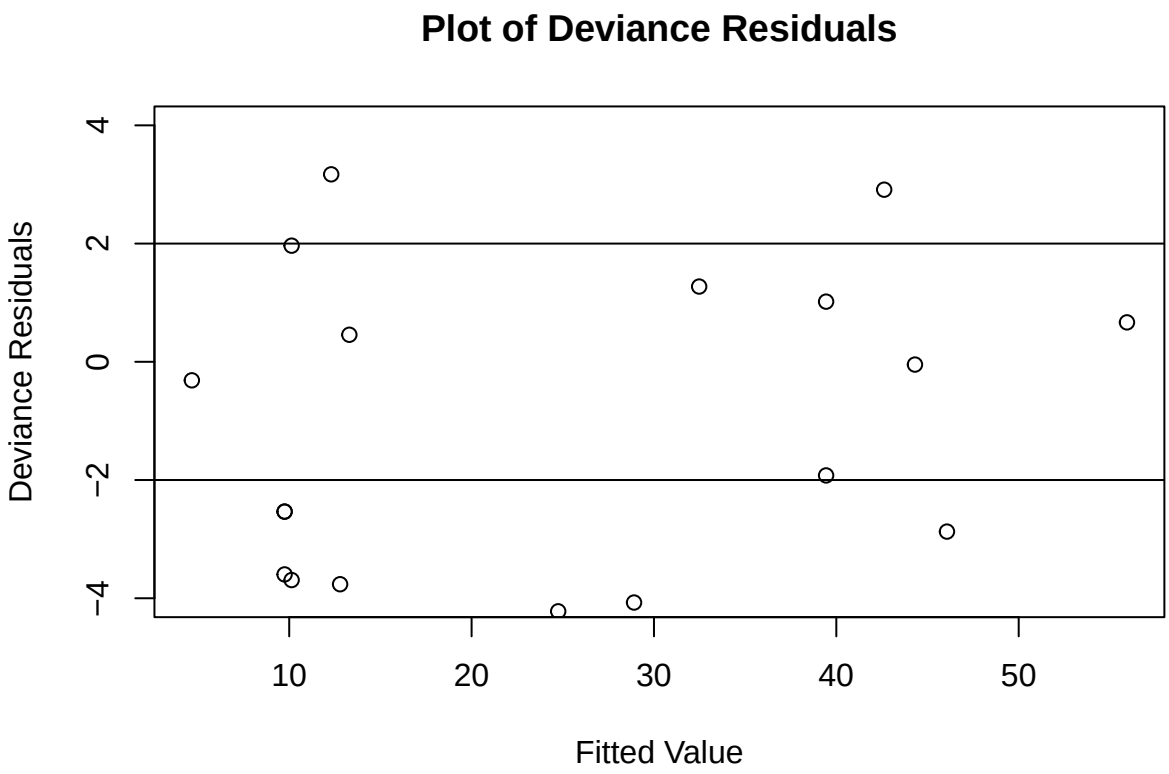
Comments for 'hist for placebo': The observed distribution is relatively fitted with the poisson distribution. Observations are concentrated around the range 25-35 while the poisson distribution has the highest predicated probability around the same range.

(b)

```r
# Fit model
mod <- glm(Number ~ Treatment + Age, family = poisson, data = polyps)
summary(mod)
```

```
##
## Call:
## glm(formula = Number ~ Treatment + Age, family = poisson, data = polyps)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2212  -3.0536  -0.1802   1.4459   5.8301
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       3.169941   0.168210   18.84  < 2e-16 ***
## TreatmentPlacebo  1.359083   0.117643   11.55  < 2e-16 ***
## Age              -0.038830   0.005955   -6.52 7.02e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 378.66  on 19  degrees of freedom
## Residual deviance: 179.54  on 17  degrees of freedom
## AIC: 273.88
##
## Number of Fisher Scoring iterations: 5
```

```r
# Plot the Deviance residuals
rd <- residuals.glm(mod,"deviance")
fv <- mod$fitted.values
plot(fv,rd,xlab="Fitted Value", ylim=c(-4,4),
     ylab="Deviance Residuals",main="Plot of Deviance Residuals")
abline(h=-2)
abline(h=2)
```



Evidence of overdispersion: Since D=179.54 and (n-p)=17, D/(n-p) » 1, so there is evidence of overdispersion. Besides, we observe that there are a lot of points are distributed outside (-1.96,+1.96) based on the plot of deviance residuals, which means that the variance is very large and proves overdispersion as well.

(c)

Firstly, we estimate the dispersion parameter.

```
# by using ad-hoc method
D <- 179.54
df <- 17
estimate_dispersion <- D/df
estimate_dispersion
```

```
## [1] 10.56118
```

Does accounting for the overdispersion change the treatmtent-outcome conclusion? We conduct the hypothesis: H0: beta1=0 vs. HA: beta1!=0 under ad-hoc method.

```
estimate_beta1 <- 1.359083
se <- 0.117643
se_adj <- sqrt(estimate_dispersion) * se
t_adj <- estimate_beta1/se_adj
t_adj
```

```
## [1] 3.55487
```

```
p_value_adj <- 2 * (1-pnorm(abs(t_adj)))
p_value_adj
```

```
## [1] 0.0003781658
```

Comments: Since p-value(adj) still < 0.05, accounting for the overdispersion does not change our conclusion that there is a significant treatment-outcome association.

How about age-outcome conclusion? We conduct the hypothesis: H0: beta2=0 vs. HA: beta2!=0 under ad-hoc method.

```
estimate_beta2 <- -0.038830
se2 <- 0.005955
se2_adj <- sqrt(estimate_dispersion) * se2
t_adj2 <- estimate_beta2/se2_adj
t_adj2
```

```
## [1] -2.006455
```

```
p_value_adj2 <- 2 * (1-pnorm(abs(t_adj2)))
p_value_adj2
```

```
## [1] 0.0448077
```

Comments: Since p-value(adj) still < 0.05, accounting for the overdispersion does not change our conclusion that there is a significant age-outcome association.

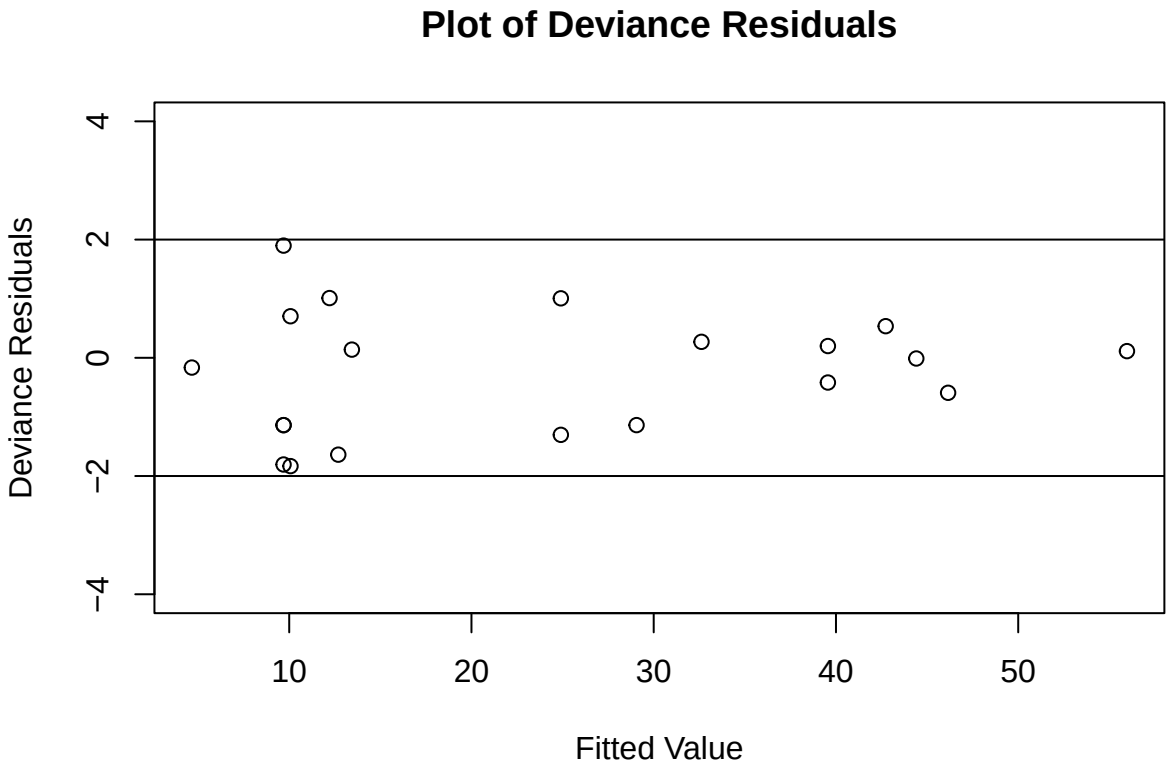
(d)

```
library(MASS)

# we fit negative binomial
mod2 <- glm.nb(Number ~ Treatment + Age, data = polyps, link = log,
               init.theta = 1, trace = F)
summary(mod2)
```

```
##
## Call:
## glm.nb(formula = Number ~ Treatment + Age, data = polyps, trace = F,
##     init.theta = 1.719491001, link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.83270  -1.13898  -0.08851   0.33637   1.89785
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        3.15791    0.55753   5.664 1.48e-08 ***
## TreatmentPlacebo   1.36812    0.36903   3.707 0.000209 ***
## Age               -0.03856    0.02095  -1.840 0.065751 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.7195) family taken to be 1)
##
##     Null deviance: 36.734  on 19  degrees of freedom
## Residual deviance: 22.002  on 17  degrees of freedom
## AIC: 164.88
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.719
##           Std. Err.:  0.607
##
##  2 x log-likelihood:  -156.880
```

```
# plot deviance residual
rd2 <- residuals.glm(mod2,"deviance")
fv2 <- mod2$fitted.values
plot(fv2,rd2,xlab="Fitted Value", ylim=c(-4,4),
     ylab="Deviance Residuals",main="Plot of Deviance Residuals")
abline(h=-2)
abline(h=2)
```



Comments: Based on plot of deviance residual, we found that most points distributed within range (-1.96,+1.96), which means that the model is proper. So I am satisfied with this model.

(e)

Based on (b) and (d), our final model is negative binomial glm.

For beta0 (intercept)

```
est_beta0 <- 3.15791
se_beta0 <- 0.55753
exp_est_beta0 <- exp(est_beta0)
exp_est_beta0
```

```
## [1] 23.52138
```

```
# 95% CI
c <- 1.96
L <- est_beta0 - c * se_beta0
U <- est_beta0 + c * se_beta0
exp(L)
```

```
## [1] 7.88649
```

```
exp(U)
```

```
## [1] 70.15231
```

Interpretation: When using drug for treatment and at age 0, the expected number of polyps observed (rate/risk) is 23.52138 in a unit time. The 95% CI is [7.88649, 70.15231].

For beta1

```
est_beta1 <- 1.36812
se_beta1 <- 0.36903
exp_est_beta1 <- exp(est_beta1)
exp_est_beta1
```

```
## [1] 3.927959
```

```
# 95% CI
L <- est_beta1 - c * se_beta1
U <- est_beta1 + c * se_beta1
exp(L)
```

```
## [1] 1.905646
```

```
exp(U)
```

```
## [1] 8.096394
```

Interpretation: The expected number of polyps observed when using placebo (risk) will be 3.927959 times of the expected number of polyps observed when using drug (risk), while holding other variables unchanged. The 95% CI is [1.905646, 8.096394].

For beta2

```
est_beta2 <- -0.03856
se_beta2 <- 0.02095
exp_est_beta2 <- exp(est_beta2)
exp_est_beta2
```

```
## [1] 0.962174
```

```
# 95% CI
L <- est_beta2 - c * se_beta2
```

```
U <- est_beta2 + c * se_beta2
exp(L)
```

```
## [1] 0.9234654
```

```
exp(U)
```

```
## [1] 1.002505
```

Interpretation: When age increase 1 and holding other variables unchanged, the expected number of polyps observed in next year (risk at age+1) will be 0.962174 times of the expected number of polyps observed in this year (risk at age). The 95% CI is [0.9234654, 1.002505].