HW3 yq2378

Qi Yumeng

2024-02-27

Q1

(a) Fit a prospective modeland interpret the result.

```r
df = tibble(
  DAC = c(rep("0-79g",6),rep("80+g",6)),
  age = rep(seq(25,75,by =10), 2),
  case = c(0, 5,21,34,36,8,1,4,25,42,19,5),
  control = c(106,164,138,139,88,31,9,26,29,27,18,0))

# retrospective studym diseased fixed
M1 =glm(cbind(case, control)~DAC + age, data = df, family=binomial(link='logit'))
summary(M1)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ DAC + age, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.023449   0.418224 -12.011   <2e-16 ***
## DAC80+g      1.780000   0.187086   9.514   <2e-16 ***
## age          0.061579   0.007291   8.446   <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: 211.608  on 11  degrees of freedom
## Residual deviance:  31.932  on  9  degrees of freedom
## AIC: 78.259
##
## Number of Fisher Scoring iterations: 4
```

- Intercept: The intercept represents the log odds of being a case (having the disease) for a reference group, which in this case is the group with daily alcohol consumption of 0-79g and the youngest age category (since age is continuous, technically it's the log odds for an individual of age 0, which doesn't make sense in this context, so we consider it as the baseline category). The intercept is highly significant with a p-value less than 2e-16, indicating a very strong effect.

- DAC80+g: This coefficient represents the log odds ratio for the group with daily alcohol consumption of 80+ grams compared to the reference group (0-79 grams). The positive coefficient (1.780) suggests that higher alcohol consumption is associated with higher odds of having the disease, and it is statistically significant with a p-value less than 2e-16.

- Age: This coefficient represents the change in the log odds of being a case for each one-unit increase in age. Since age was measured in years and modeled as a continuous variable, a one-unit increase corresponds to one year. The positive coefficient (0.061579) suggests that the odds of having the disease increase with age, and this effect is also statistically significant (p-value < 2e-16).
- The residual deviance shows how well the response variable is predicted by the model on the actual predictors. Compared to the null deviance (211.608), the residual deviance reduces greatly to 31.932. This suggests that the model fits the data pretty well.

(b) nested model

Model M1 (Null Model): This model suggests that there is a constant effect of alcohol consumption on disease across all age groups (no interaction term).

Model M2 (Alternative Model): This model assumes that the effect of alcohol consumption on the disease is not the same across all age groups.

```
#M0 <- glm(cbind(case, control) ~ DAC + age, data = df, family = binomial(link = 'logit'))
M2 <- glm(cbind(case, control) ~ DAC * age, data = df, family = binomial(link = 'logit'))
anova(M1, M2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(case, control) ~ DAC + age
## Model 2: cbind(case, control) ~ DAC * age
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1         9     31.932
## 2         8     31.929  1 0.0022516   0.9622
```

It is a nested model. And the from the result we know the p-value is 0.9622. Under significance value of 0.05, we fail to reject the null hypothesis (M1), indicating that alcohol consumption does not have a significant effect on the disease outcome across all age groups.

Q2

(a)

```
df = tibble(
  seed = c(rep("OA 75", 11), rep("OA 73",10)),
  root = c(rep("Bean",5), rep("Cucumber",6),rep("Bean",5), rep("Cucumber",5)),
  germinate = c(10,23,23,26,17,5,53,55,32,46,10,8,10,8,23,0,3,22,15,32,3),
  ttl = c(39,62,81,51,39,6,74,72,51,79,13,16,30,28,45,4,12,41,30,51,7)
)

M1 = glm(cbind(germinate, ttl-germinate)~ seed + root, data = df, family=binomial(link='logit'))
summary(M1)
```

```
##
## Call:
## glm(formula = cbind(germinate, ttl - germinate) ~ seed + root,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
```

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.7005     0.1507  -4.648 3.36e-06 ***
## seedOA 75      0.2705     0.1547   1.748   0.0804 .
## rootCucumber   1.0647     0.1442   7.383 1.55e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

- Intercept: The intercept represents the log odds of germinating for a reference group, which in this case is the group with O.aegyptiaca 73 and bean as the seed. The intercept is highly significant with a p-value 3.36e-06, indicating a very strong effect.
- seed OA 75 : This coefficient represents the log odds ratio for the group with O.aegyptiaca 75 compared to the reference group O.aegyptiaca 73. The positive coefficient (0.2705) suggests that O.aegyptiaca 75 is associated with higher odds of germinating, but it is not statistically significant with a p-value of 0.0804.
- root Cucumber: This coefficient represents the log odds ratio for the group with cucumber root compared to the reference group bean root. The positive coefficient (1.0647) suggests that cucumber root is associated with higher odds of germinating, and it is statistically significant with a p-value of 1.55e-13.
- The residual deviance shows how well the response variable is predicted by the model on the actual predictors. Compared to the null deviance (98.719), the residual deviance reduces greatly to 39.686. This suggests that the model fits the data pretty well.

(b) Over dispersion

```
pval=1-pchisq(M1$deviance,21-2)
pval # bad fit, reject the fitting
```

```
## [1] 0.003597422
```

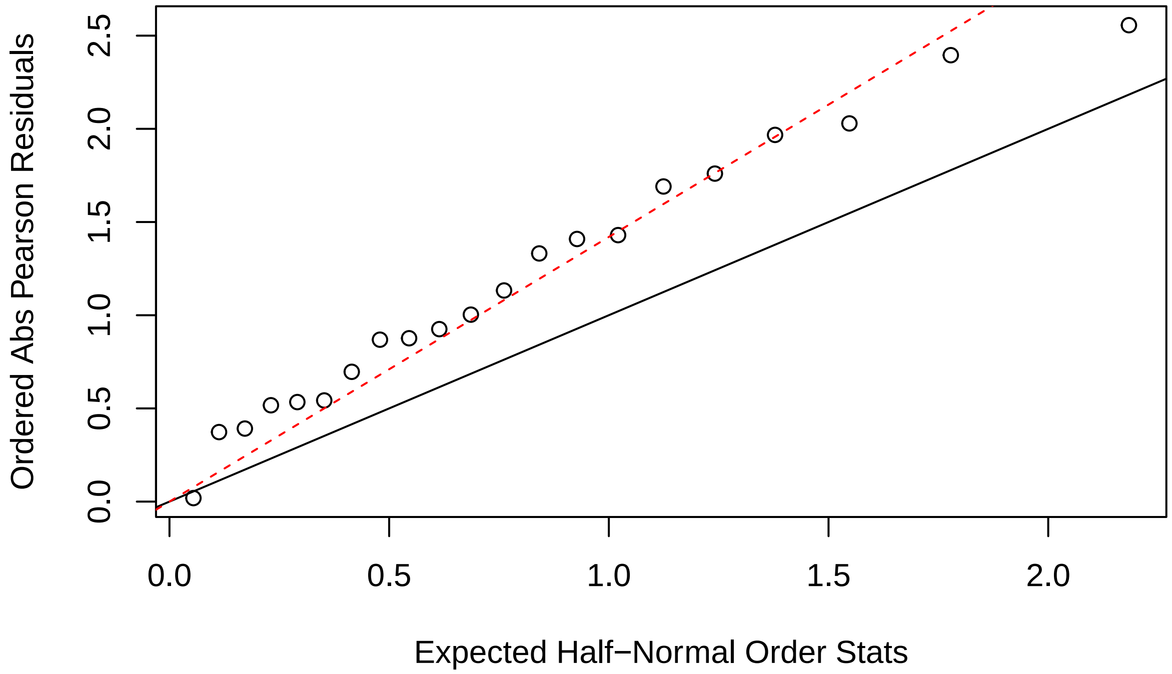
```
# calc dispersion param
G.stat=sum(residuals(M1,type='pearson')^2) # pearson chisq
G.stat
```

```
## [1] 38.31062
```

```
phi=G.stat/(21-2)
phi
```

```
## [1] 2.016348
```

```
# test over-dispersion (half normal plot)
res=residuals(M1,type='pearson')
plot(qnorm((21+1:21+0.5)/(2*21+1.125)),sort(abs(res)),xlab='Expected Half-Normal Order Stats',ylab='Orde
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2, col = 'red')
```



```
summary(M1,dispersion=phi)
```

```
##
## Call:
## glm(formula = cbind(germinate, ttl - germinate) ~ seed + root,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.7005     0.2140  -3.273  0.00106 **
## seedOA 75      0.2705     0.2197   1.231  0.21828
## rootCucumber   1.0647     0.2048   5.199    2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 2.016348)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

Yes, from the expected half-normal order stat plot, there's clearly over dispersion. The estimate of dispersion parameter is 2.0163484. After updating the model.

- Intercept: The intercept represents the log odds of germinating for a reference group, which in this case is the group with O.aegyptiaca 73 and bean as the seed. The intercept is highly significant with a p-value 0.00106, indicating a very strong effect.
- seed OA 75 : This coefficient represents the log odds ratio for the group with O.aegyptiaca 75 compared to the reference group O.aegyptiaca 73. The positive coefficient (0.2705) suggests that O.aegyptiaca 75 is associated with higher odds of germinating, but it is not statistically significant with a p-value of 0.21828. Notice that the sd of the estimate is larger then before.

- root Cucumber: This coefficient represents the log odds ratio for the group with cucumber root compared to the reference group bean root. The positive coefficient (1.0647) suggests that cucumber root is associated with higher odds of germinating, and it is statistically significant with a p-value of 2e-07. Notice that the sd of the estimate is larger then before.
- The residual deviance shows how well the response variable is predicted by the model on the actual predictors. Compared to the null deviance (98.719), the residual deviance reduces greatly to 39.686. This suggests that the model fits the data pretty well.

(c)

- Unaccounted Variability: The variability in germination rates may be due to factors not included in the model. For instance, there might be micro-environmental differences within the experimental setup that affect germination but are not captured by the variables 'seed' and 'root'.
- Clustered Data: If the seeds are not independent of one another — for example, seeds from the same pod or batch may be more similar to each other than to seeds from different batches — this could lead to overdispersion. The standard binomial model assumes each trial is independent, and clustering violates this assumption.
- Incorrect Model Structure: If the true relationship between the predictors and the response is not linear on the logit scale, or if important interaction terms or non-linear effects have been omitted, the model might not fit well, leading to overdispersion.