hw3

Shan Jiang

1. Problem 1

```r
## 1.Active packages
library(ggplot2)
library(tidyverse)
library(aod)
library(auditor)
```

```r
### 2.import in the esophageal data
cancer_df = readxl::read_xlsx("./cancer.xlsx") %>%
      janitor::clean_names()

cancer_df
```

```
## # A tibble: 12 x 4
##    alcol   age disease undisea
##    <dbl> <dbl>   <dbl>   <dbl>
##  1     1    25       1       9
##  2     1    35       4      26
##  3     1    45      25      29
##  4     1    55      42      27
##  5     1    65      19      18
##  6     1    75       5       0
##  7     0    25       0     106
##  8     0    35       5     164
##  9     0    45      21     138
## 10     0    55      34     139
## 11     0    65      36      88
## 12     0    75       8      31
```

```r
### 3. fit a prospective model
# fit logit model
y = cbind(cancer_df$disease, cancer_df$undisea)

logit.prosp = glm(y ~  alcol + age, family = binomial(link = 'logit'), data = cancer_df)

summary(logit.prosp) ## we cannot interpret the intercept.
```

```
##
## Call:
## glm(formula = y ~ alcol + age, family = binomial(link = "logit"),
##     data = cancer_df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.59974  -1.72957   0.06822   1.19015   1.50808
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.023449   0.418224 -12.011   <2e-16 ***
## alcol        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
```

Model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \cdot Alcol + \beta_2 \cdot Age$$

where Alcol $=$ 1 means alcolhol consumption $>$ 80g;

Interpretation:

- The $\beta_1 = 1.78$ means that the log odds ratio of esophageal cancer is 1.78 between alcohol consumption below 79g group and alcholhol consumption $>$ 80g for people ranging from 25 to 75 years old.
- The β_2 means that the log odds ratio of esophageal cancer is 0.06 corresponding to one unit change of age.
- The β_0 cannot be interpreted.

2. Problem 2

```
## import data
gertest_df = readxl::read_xlsx("./germin.xlsx") %>%
      janitor::clean_names() %>%
      mutate(seeds = as.factor(seeds),
             nutri = as.factor(nutri) )

head(gertest_df)
```

```
## # A tibble: 6 x 4
##   seeds nutri    germi total
##   <fct> <fct>    <dbl> <dbl>
## 1 A     bean        10    39
## 2 A     bean        23    62
## 3 A     bean        23    81
## 4 A     bean        26    51
## 5 A     bean        17    39
## 6 A     cucumber     5     6
```

Notation: Seed.A = O. aegyptiaca 75; Seed.B = O. aegyptiaca 73;

```
## 3. fit a prospective model
# fit logit model
y = cbind(gertest_df$germi, gertest_df$total - gertest_df$germi)

logit.prosp = glm(y ~ seeds + nutri, family = binomial(link = 'logit'), data = gertest_df)
```

```r
summary(logit.prosp)
```

```
## 
## Call:
## glm(formula = y ~ seeds + nutri, family = binomial(link = "logit"), 
##     data = gertest_df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.3919  -0.9949  -0.3744   0.9831   2.4766  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -0.4300     0.1137  -3.781 0.000156 ***
## seedsB         -0.2705     0.1547  -1.748 0.080435 .  
## nutricucumber   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
```

Model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 \cdot seed_B + \beta_2 \cdot nut_{cuc}$$

Interpretation:

- The β_1 means that the log odds ratio of germination is -0.2705 between seed O. aegyptiaca 75 veesus O. aegyptiaca 73.
- The test of β_2 shows that the p-value is $0.08 > 0.05$, which exceeds the threshold for rejection at the sig. level of 0.05, meaning that we cannot reject the null hypothesis that the parameter is not different from 0, which implies there may be no significant impact of seed type on log odds ratio of germination at the sig. level of 0.05.
- However, based on a sig. level of 0.1, we can say The β_2 implies that the log odds ratio of germination is 1.06 between cucumber and Bean nutrition extract media.
- The β_0 means that the log odds of germination when the seed is O. aegyptiaca 75 and nutrition extract media is Bean.

2.2 Overdispersion

```r
# goodness of fit
pval = 1 - pchisq(logit.prosp$deviance, 21-3)
pval # bad fit
```

```
## [1] 0.00230277
```

Because of p-value is under 0.05, meaning that we can reject the null hypothesis at the sig. level of $\alpha = 0.05$, and accept the alternative hypothesis that this model is not a good fit.

In this situation, we need to find out is there any dispersion in this dataset that biased our simulation.

```
# calc dispersion parameter
G.stat = sum(residuals(logit.prosp,type = 'pearson')^2) # pearson chisq

G.stat
```

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

```
phi = G.stat/(21 - 3)
phi
```

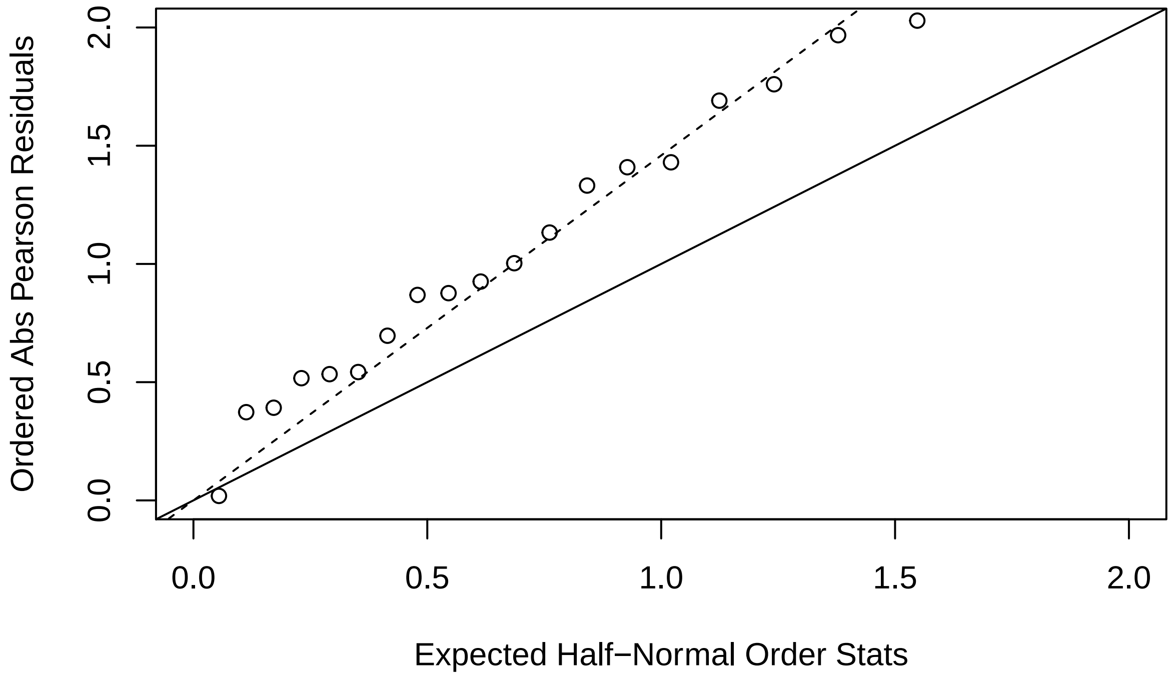
```
## [1] 2.128368
```

```
tilde.phi = logit.prosp$deviance/logit.prosp$df.residual
tilde.phi # similar to the one estimated from pearson chisq
```

```
## [1] 2.204772
```

The dispersion parameter result is 2.128368.

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



The half-normal plot suggests that the absolute value of residuals obtained from original model suffered a linear deviation from the refrence line, indicating a constant over-dispersion, so we need to adjust our model for a better fit.

```
# fit model with constant over-dispersion
summary(logit.prosp, dispersion = phi)
```

```
##
## Call:
## glm(formula = y ~ seeds + nutri, family = binomial(link = "logit"),
##     data = gertest_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -0.4300     0.1659  -2.592  0.00955 **
## seedsB         -0.2705     0.2257  -1.198  0.23081
## nutricucumber   1.0647     0.2104   5.061 4.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 2.128368)
##
##     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
```

Based on the quasi-likelihood estimation, we fit the model again. In this case, the coefficient β_0, β_1, and β_2 remains the same as the original one, while `new` β_2 cannot pass the significance test as it was in the original model, suggesting that after ruling out the possibility of dispersion, variable `seed` has yet no significant influence on the germination indeed, we may reconsider this variable in the future modelling.

Meanwhile, all Std. Error turns bigger as the effect of a parameter ϕ in dispersion.

2.3 Causes for over-dispersion

The cause of over-dispersion comes from Intra-Group correlation exists, for example, these plants may share same gene traits or features, then the independent assumption is violated, so the binomial distribution is not followed exactly.