HW3

Anna Ma

2/24/2022

```
library(tidyverse)
```

Problem 1

a)

1. Enter data

```
cancer_alch_df =
  tibble(age = rep(c(25, 35, 45, 55, 65, 75),2),
         exposure = rep(c(0, 1), each = 6),
         # exposure = 0: consumption 0-79g, exposure = 1: consumption >= 80g
         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))
```

2. Model

```
prosp_mod = glm(cbind(case, control) ~ age + exposure, data = cancer_alch_df, family = binomial(link =

prosp_mod %>% broom::tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)  -5.02     0.418      -12.0  3.10e-33
## 2 age           0.0616   0.00729      8.45 3.01e-17
## 3 exposure      1.78     0.187        9.51 1.83e-21
```

The fitted prospective model for the data is

$$P(D = Disease|E = Exposure, X = Age) = \frac{e^{-5.02+0.0616X+1.78E}}{1+e^{-5.02+0.0616X+1.78E}}$$

From the logit model, we can see that the log odds ratio of esophageal cancer between exposed (people with 0-79 g daily alcohol consumption) and unexposed group (people with 80+ g daily alcohol consumption) is 1.0635142 for 1 unit change in age, given the same alcohol consumption

The log odds ratio of esophageal cancer between the exposed and unexposed group is 5.9298535 for 1 unit change in daily alcohol consumption, given the same age.

Therefore, we can conclude that age and daily alcohol consumption of 80+ g are both positively associated with esophageal cancer.

b)

1. Fit both model

Let M0 be the smaller model where ψ_j,the odds ratio relating alcohol consumption and disease in the jth age group is 1; and M1 be the larger model where $\psi_j = \psi$

```
cancer_alch_df = cancer_alch_df %>%
  mutate(age_group = as.factor(c("1", "2", "3", "4", "5", "6", "1", "2", "3", "4", "5", "6")))

M0 = glm(cbind(case, control) ~ age_group, data = cancer_alch_df, family = binomial(link = 'logit'))

M1 = glm(cbind(case, control) ~ age_group+exposure, family = binomial(link = 'logit'),data = cancer_alc

M0 %>% broom::tidy()
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic    p.value
##   <chr>          <dbl>     <dbl>     <dbl>      <dbl>
## 1 (Intercept)    -4.74      1.00     -4.72 0.00000231
## 2 age_group2      1.70      1.06      1.60 0.110
## 3 age_group3      3.46      1.02      3.39 0.000688
## 4 age_group4      3.96      1.01      3.91 0.0000924
## 5 age_group5      4.09      1.02      4.02 0.0000590
## 6 age_group6      3.88      1.06      3.67 0.000246
```

```
M1 %>% broom::tidy()
```

```
## # A tibble: 7 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -5.05     1.01      -5.01 5.53e- 7
## 2 age_group2      1.54     1.07       1.45 1.48e- 1
## 3 age_group3      3.20     1.02       3.13 1.77e- 3
## 4 age_group4      3.71     1.02       3.65 2.66e- 4
## 5 age_group5      3.97     1.02       3.88 1.06e- 4
## 6 age_group6      3.96     1.07       3.72 1.99e- 4
## 7 exposure        1.67     0.190      8.81 1.28e-18
```

2. Check if the models are nested

```
M0$coefficients
```

```
## (Intercept)  age_group2  age_group3  age_group4  age_group5  age_group6
##   -4.744932    1.695133    3.455580    3.963678    4.088826    3.875894
```

```
M1$coefficients
```

```
## (Intercept)  age_group2  age_group3  age_group4  age_group5  age_group6
##   -5.054348    1.542294    3.198762    3.713490    3.966882    3.962190
##    exposure
##    1.669890
```

$M_0 = \beta_0 + \beta_1 * age_2 + \beta_2 * age_3 + \beta_3 * age_4 + \beta_4 * age_5 + \beta_5 * age_6$

M0 has 6 parameters and 5 predictors: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ & age_group 2-5

$M_1 = \beta_0 + \beta_1 * age_2 + \beta_2 * age_3 + \beta_3 * age_4 + \beta_4 * age_5 + \beta_5 * age_6 + \beta_6 * alcohol$

M1 has 7 parameters and 6 predictors: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_5$ & age_group 2-5 and alcohol consumption

Therefore, we can see that M_0 is nested in M_1, and we should perform deviance analysis.

Let $H_0 : \beta_j = 0$, $H_a : \beta_j \neq 0$

```
# Deviance
dev0 = M0$deviance #D1
dev1 = M1$deviance #D2
diff = dev0 - dev1 # difference between the two deviance
pchisq(diff, 6, lower.tail = FALSE) #p2 = the number of predictors in the larger model = 6
```

```
## [1] 4.484692e-15
```

The difference of deviance between M_0 and M_1 is 79.52203, which gives a small p value of $4.4846916 \times 10^{-15}$. Therefore, we should abject the null hypothesis and conclude that M_1 fit the data better.

Problem 2

a)

1. Enter Data

```
germ_df = tibble(
  seed = c(rep("o75", 11), rep("o73", 10)),
  root = c(rep("b", 5), rep("c", 6), rep("b", 5), rep("c", 5)),
  germ = c(c(10, 23, 23, 26, 17), c(5, 53, 55, 32, 46, 10), c(8, 10, 8, 23, 0), c(3, 22, 15, 32, 3)),
  total = c(c(39, 62, 81, 51, 39), c(6, 74, 72, 51, 79, 13), c(16, 30, 28, 45, 4), c(12, 41, 30, 51, 7)
```

2. Fit the logit regression model

```
logit_mod = glm(cbind(germ, total-germ) ~ seed + root, data = germ_df, family = binomial(link = 'logit'

logit_mod %>% broom::tidy()
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   -0.700     0.151     -4.65 3.36e- 6
## 2 seedo75        0.270     0.155      1.75 8.04e- 2
## 3 rootc          1.06      0.144      7.38 1.55e-13
```

Let G = Germinated, X1 = Seed, X2 = Root The fitted model is

$$P(G|X_1 = x_1, X_2 = x_2) = \frac{e^{-0.7005+0.2705x_1+1.0647x_2}}{1 + e^{-0.7005+0.2705x_1+1.0647x_2}}$$

when $X_1 = 1$, the seed is O. aegyptiaca 75, and when $X_1 = 0$, the seed is O.aegyptiaca 73. When $X_2 = 1$, the root is cumcumber, and when $X_2 = 0$, the root is bean.

```
# intercept: germination rate for O73 seed in a bean root
exp(logit_mod$coefficients[1])
```

```
## (Intercept)
##   0.4963454
```

```
# b1: odds ratio of germination rate of O75 compare to O73, holding root constant yo be bean
exp(logit_mod$coefficients[2])
```

```
##  seedo75
## 1.310555
```

```
#b2: the odds ratio of germinating on a cucumber root
exp(logit_mod$coefficients[3])
```

```
##    rootc
## 2.900113
```

From the model, we can see that the O.aegyptiaca 73 seed in bean root media has a 0.4963454 germination rate.

Holding the root extract media constant to be bean, the odds ratio of germinating of O.aegyptiaca 75 seed is 1.3105554 time of an O.aegyptiaca 73 seed

Holding the seed type constant to be O.aegyptiaca 73, the odds ratio of germinating on a cucumber root is 2.9001133 time the odds ratio of a bean root.

From those results, we can conclude that the root extract medium is highly associated with germinating rate.

b) dispersion

1. Goodness of fit

```
logit_mod$deviance
```

```
## [1] 39.68589
```

```
qchisq(0.95,18)
```

```
## [1] 28.8693
```

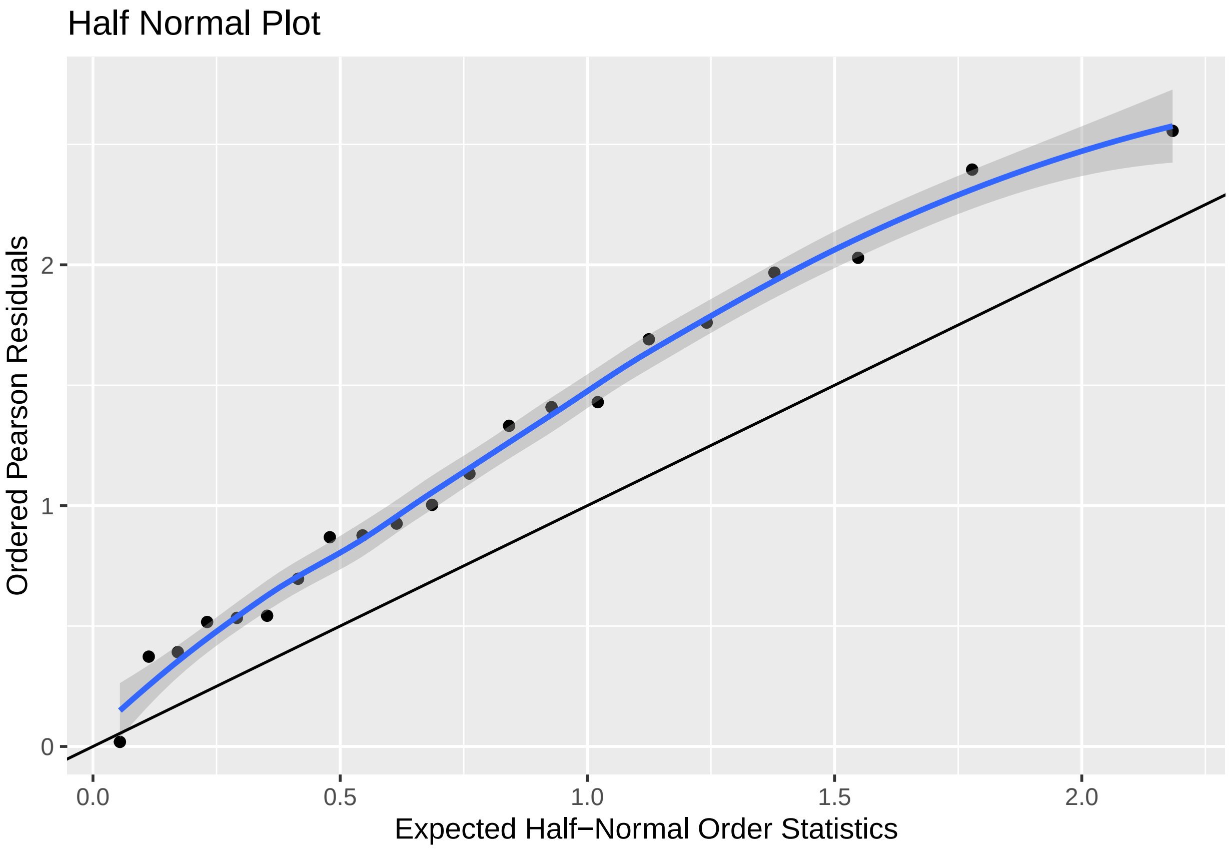
The deviance of the model is 39.6858896, which is larger than the chi-square critical value of 28.8692994 with 18 degrees of freedom. Therefore, the model is not a good fit.

2. Check for over dispersion with half normal Plot

```
e = residuals(logit_mod, type = "pearson")
n_i = 1:21
res = tibble(
  x = qnorm((21 + n_i + 0.5)/(2 * 21 + 1.125)),
  y = sort(abs(e))
)

res %>% ggplot(aes(x = x,y = y)) +
  geom_point() +
  geom_smooth() +
  geom_abline(slope = 1) +
  labs(title = "Half Normal Plot",
       x = "Expected Half-Normal Order Statistics",
       y = "Ordered Pearson Residuals")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From the plot, we can see that there is a linear deviation from the reference line. Therefore, we can conclude that there is over dispersion.

3.Dispersion Parameter

```
# pearson statistics
G = sum(residuals(logit_mod, type = "pearson")^2)
phi = G/(21 - 3)
```

The dispersion parameter $\phi = 2.1283678$

4. Update Model

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

```
##
## Call:
## glm(formula = cbind(germ, total - germ) ~ seed + root, family = binomial(link = "logit"),
##     data = germ_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)  -0.7005     0.2199  -3.186  0.00144 **
## seedo75       0.2705     0.2257   1.198  0.23081
## rootc         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
```

After updating the model, the estimate of the model did not change. However, the standard error was increased, which caused the p-value to change. In the updated model, we can clearly see that the p-value for seeds, 0.231 is greater than 0.05, indicating that the species of the seed might be insignificant to the germination rate, holding the root extract media constant.

c) A plausible cause of the over dispersion is intra-class correlation. For example, germination in one spot can influence its neighbors, causing the germination to be dependent and violating the assumption that the trials are independent.