P31 HW3

Yihan Feng

2/15/2021

```
library(tidyverse)
```

Question 1

```
alcohol = read_csv("./alcohol.csv") %>%
  pivot_longer(
    case_079:control_80,
    names_to = "treatment",
    values_to = "count"
  ) %>%
  separate(treatment, c("treatment", "consumption"), "_") %>%
  pivot_wider(
    names_from = treatment,
    values_from = count
  )
```

a. Fit a prospective model to the data to study the relation between alcohol consumption, age, and disease (model age as a continuous variable taking values 25, 35, 45, 55, 65, and 75). Interpret the result.

```
case_control = cbind(alcohol$case, alcohol$control)

alcohol.prosp = glm(data = alcohol,
                case_control ~ age + consumption,
                family = binomial(link = 'logit'))
summary(alcohol.prosp)
```

```
##
## Call:
## glm(formula = case_control ~ age + consumption, family = binomial(link = "logit"),
##     data = alcohol)
##
## 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 ***
## age            0.061579   0.007291   8.446   <2e-16 ***
## consumption80  1.780000   0.187086   9.514   <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
```

- The model doesn't fit well. The deviance of the model is 31.9315101, while chi-squared critical value is 16.9189776. Therefore, we reject the null hypothesis, and conclude that the model doesn't fit well.
- The odds ratio of disease between exposed and non-exposed group is 1.0635142 for one unit change in age. And the 95% confidence interval is (1.0484252, 1.0788202).
- The odds ratio of disease between exposed and non-exposed group is 5.9298535 for one unit change in alcohol consumption.And the 95% confidence interval is (4.1095668, 8.556416).

b.

M_0: smaller model. $log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 * age_2 + \beta_2 * age_3 + ... + \beta_5 * age_6$ β_6 (coefficient for alcohol consumption)$= 0$, which implies that the disease is not related to alcohol consumption.

M_1: larger model. $log(\pi_i/(1-\pi_i)) = \beta_0 + \beta_1 * age_2 + \beta_2 * age_3 + ... + \beta_5 * age_6 + \beta_6 * alcohol$

```
alcohol.small = glm(
  data = alcohol.b,
  case_control ~ age,
  family = binomial(link = "logit")
)
summary(alcohol.small)
```

```
##
## Call:
## glm(formula = case_control ~ age, family = binomial(link = "logit"),
##     data = alcohol.b)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.477  -1.299   0.368   2.481   5.028
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.745      1.004  -4.725 2.31e-06 ***
## age35-44       1.695      1.061   1.598 0.110006
## age45-54       3.456      1.018   3.394 0.000688 ***
## age55-64       3.964      1.014   3.910 9.24e-05 ***
## age65-74       4.089      1.018   4.017 5.90e-05 ***
## age75+         3.876      1.057   3.666 0.000246 ***
## ---
## 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:  90.563  on  6  degrees of freedom
## AIC: 142.89
##
## Number of Fisher Scoring iterations: 6
```

```
alcohol.large = glm(
  data = alcohol.b,
  case_control ~ age + consumption,
  family = binomial(link = "logit")
)
summary(alcohol.large)
```

```
##
## Call:
## glm(formula = case_control ~ age + consumption, family = binomial(link = "logit"),
##     data = alcohol.b)
```

```
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -1.16129   0.96641   0.04747  -0.05538  -0.11628   0.13652  -0.35391   0.45905
##        9        10        11        12
##  0.96513  -1.59342  -0.67850   2.11053
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.0543     1.0094  -5.007 5.53e-07 ***
## age35-44        1.5423     1.0659   1.447 0.147916
## age45-54        3.1988     1.0232   3.126 0.001770 **
## age55-64        3.7135     1.0185   3.646 0.000266 ***
## age65-74        3.9669     1.0231   3.877 0.000106 ***
## age75+          3.9622     1.0650   3.720 0.000199 ***
## consumption80   1.6699     0.1896   8.807  < 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:  11.041  on  5  degrees of freedom
## AIC: 65.369
##
## Number of Fisher Scoring iterations: 5
```

```
alcohol.dev = deviance(alcohol.small) - deviance(alcohol.large)
alcohol.chi = qchisq(0.95, (1))
```

The difference of deviance between M_0 and M_1 is 79.5220267, which is greater than the chi-squared distribution with degree freedom of 1 (7 parameters - 6 parameters) 3.8414588. Therefore, it is able to reject the null hypothesis/M_0.

Question 2

a. Fit a logistic regression model to study the relation between germination rates and different types of seed and root extract. Interpret the result.

```
germ.model = glm(data = germ.df,
                 cbind(germinating, nseed - germinating) ~ seed + root,
                 family = binomial(link = 'logit'))
summary(germ.model)
```

```
##
## Call:
## glm(formula = cbind(germinating, nseed - germinating) ~ 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     0.1507  -4.648 3.36e-06 ***
## seedcucumber   1.0647     0.1442   7.383 1.55e-13 ***
## rootoa_75      0.2705     0.1547   1.748   0.0804 .
## ---
## 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
```

- The model doesn't fit well. The deviance of the model is 39.6858896, while chi-squared critical value with degree of freedom 9 is 28.8692994. Therefore, we reject the null hypothesis, and conclude that the model doesn't fit well.
- The odds ratio of germination is 2.9001133 between bean and cucumber. And the 95% confidence interval is (2.1860477, 3.847426).
- The odds ratio of germination is 1.3105554 between O. aegyptiaca 75 and O. aegyptiaca 73. And the 95% confidence interval is (0.9677646, 1.7747659).

b. Is there over dispersion? If so, what is the estimate of dispersion parameter? Update your model and reinterpret the result.

```
G.stat=sum(residuals(germ.model,type = 'pearson')^2)

germ_row = nrow(germ.df)
phi = G.stat/(germ_row - 3)
```

There is over dispersion in this model. The estimate of dispersion parameter ϕ is 2.128.

```
summary(germ.model, dispersion = phi)
```

updated model:

```
##
## Call:
## glm(formula = cbind(germinating, nseed - germinating) ~ 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 **
## seedcucumber   1.0647     0.2104   5.061 4.18e-07 ***
## rootoa_75      0.2705     0.2257   1.198  0.23081
## ---
## 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
```

- The model doesn't fit well. The deviance of the model is 39.6858896, while chi-squared critical value with degree of freedom 9 is 28.8692994. Therefore, we reject the null hypothesis, and conclude that the model doesn't fit well.
- The odds ratio of germination is 2.9001133 between bean and cucumber. And the 95% confidence interval is (1.9201189, 4.3802792).
- The odds ratio of germination is 1.3105554 between O. aegyptiaca 75 and O. aegyptiaca 73. And the 95% confidence interval is (0.8420544, 2.0397204).

c. What is a plausible cause of the over dispersion?

The plausible cause of the over dispersion can be the violation of Bernoulli trial assumptions, which are: 1. underlying independent trials.
2. each trial should have same probability of success.

In this problem, it might be the potential correlation that the trials are not independent. For example, the germination has impact on its neighbor. Also, the germination rate could be different, because there might be unobserved material in the soil.