# P8131-hw3-rw2844

Renjie Wei

2/23/2022

## Problem 1

(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.

```
# Fit prospective model
pro_fit.logit = glm(cbind(case, control) ~ age+ alcohol, family = binomial(link = 'logit'), data = data
summary(pro_fit.logit)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ age + alcohol, family = binomial(link = "logit"),
##     data = data1)
##
## 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 ***
## alcoholB     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
```

I use DAC to represent the daily alcohol consumption.

The prospective model gives us $\beta_0 =$-5.023, $\beta_{age} = 0.062$ and $\beta_{DAC \geq 80} = 1.78$ . The probability of esophageal cancer given age and daily alcohol consumption status is:

$$P(D = 1|age, DAC) = \frac{e^{-5.023+0.062\times \text{age}+1.78\times I(\text{DAC}\geq 80)}}{1 + e^{-5.023+0.062\times \text{age}+1.78\times I(\text{DAC}\geq 80)}}$$

The odds ratio of disease corresponding to unit change in different covariates is:

| Coefficients | OR | 95% CI of OR lower limits | 95% CI of OR upper limits |
|---|---|---|---|
| Intercept | 0.007 | 0.003 | 0.015 |
| Age | 1.064 | 1.048 | 1.079 |
| DAC80 | 5.930 | 4.110 | 8.556 |

The model means that hold daily alcohol consumption at the same level, for a year increase in age, we expect to see a 6.35 % increase (or 1.064 times the odds) in the odds of esophageal cancer. On the other hand, if we fixed age, then the odds of esophageal cancer increased by 492.99 % (or 5.93 times the odds) in daily alcohol consumption greater or equal 80 compared with those who consume alcohol less than 80 per day. Since neither of the 95% CI contains 1, we have 95% confidence to conclude that the age and daily alcohol consumption is significantly associated with the risk of esophageal cancer.

(b) Comparing odds ratio between age groups

Two Model: $M_0 : \psi_j = 1$ for all j, and $M_1 : \psi_j = \psi$ (where $\psi$ is an unknown constant):

```
# convert age to age groups
data1["age_group"] = as.factor(data1$age)

# Build alcohol only model means all age groups
M0 = glm(cbind(case, control) ~ age_group, family = binomial(link = 'logit'),
         data = data1)
# Build Model 1
M1 = glm(cbind(case, control) ~ age_group + alcohol, family = binomial(link = 'logit'),
         data = data1)
summary(M0)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ age_group, family = binomial(link = "logit"),
##     data = data1)
##
## 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 ***
## age_group35    1.695      1.061   1.598 0.110006
## age_group45    3.456      1.018   3.394 0.000688 ***
## age_group55    3.964      1.014   3.910 9.24e-05 ***
## age_group65    4.089      1.018   4.017 5.90e-05 ***
## age_group75    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
```

```
summary(M1)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ age_group + alcohol, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -1.16129   0.04747  -0.11628  -0.35391   0.96513  -0.67850   0.96641  -0.05538
##        9        10        11        12
##  0.13652   0.45905  -1.59342   2.11053
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.0543     1.0094  -5.007 5.53e-07 ***
## age_group35   1.5423     1.0659   1.447 0.147916
## age_group45   3.1988     1.0232   3.126 0.001770 **
## age_group55   3.7135     1.0185   3.646 0.000266 ***
## age_group65   3.9669     1.0231   3.877 0.000106 ***
## age_group75   3.9622     1.0650   3.720 0.000199 ***
## alcoholB      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
```

The model we are going to comparing with can be represented by the following model:

$$\log(\frac{P(D=1|age,DAC)}{1-P(D=1|age,DAC)}) = \beta_0 + \beta_1 \cdot I(DAC \geq 80) + \beta_i \cdot I(\text{age} \in \text{age group}_i), (i \in \{1,2,\ldots,5\})$$

Since there are total 6 age groups, we put only 5 dummy variables into the model above. $M_0$ only contains the `age_group`, which means there is no difference in the odds ratio relating to alcohol consumption in each group, that is $\beta_1 = 0$. $M_1$ contains dummy variables for age groups, which means $\beta_1$ in the above model is not 0. Therefore $M_0$ is the nest model of $M_1$.

```
# Deviance Analysis
dev0 = M0$deviance
dev1 = M1$deviance
p2 = M0$df.residual - M1$df.residual
p_val = pchisq(dev0-dev1, p2, lower.tail = FALSE)
```

The null hypothesis is:

$$\mathrm{H}_0 : \beta_1 = 0$$

And the alternative hypothesis is:

$$\mathrm{H}_1 : \beta_1 \neq 0$$

The deviance of $M_0$ is 90.563, the deviance of $M_1$ is 11.041, and the deviance statistics = $deviance(M_0) - deviance(M_1) = 79.522$. And $df(M_0) - df(M_1) = 1$. Therefore we get a p-value of $0 < 0.05$ and we reject the null hypothesis at 0.05 significance level. $M_1$ better fits the data, which means there is a significant association between alcohol consumption and disease.

## Problem 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

I use OA75 as O.ageyptiaca 75 in my dataset. And we need a grouped data grouped by species and root extract.

```
none.disp <- glm(
  cbind(germ,num_seed-germ) ~ type + root, family = binomial("logit")
)
summary(none.disp)
```

```
##
## Call:
## glm(formula = cbind(germ, num_seed - germ) ~ type + root, family = binomial("logit"))
##
## 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 ***
## typeOA75       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
```

The prospective model gives us $\beta_0$ = -0.7, $\beta_{OA75}$ = 0.27 and $\beta_{cucumber}$ = 1.065 . So the probability of germination given species and type of root extract of a seed is:

$$P(germination = 1|\text{species} = \text{OA75}, \text{root extract} = \text{cucumber}) = \frac{e^{-4.30+-0.27\times I(\text{species=OA75})+1.06\times I(\text{root extract=cucumber})}}{1+e^{-4.30+-0.27\times I(\text{species=OA75})+1.06\times I(\text{root extract=cucumber})}}$$

The odds ratio of disease corresponding to unit change in different covariates is:

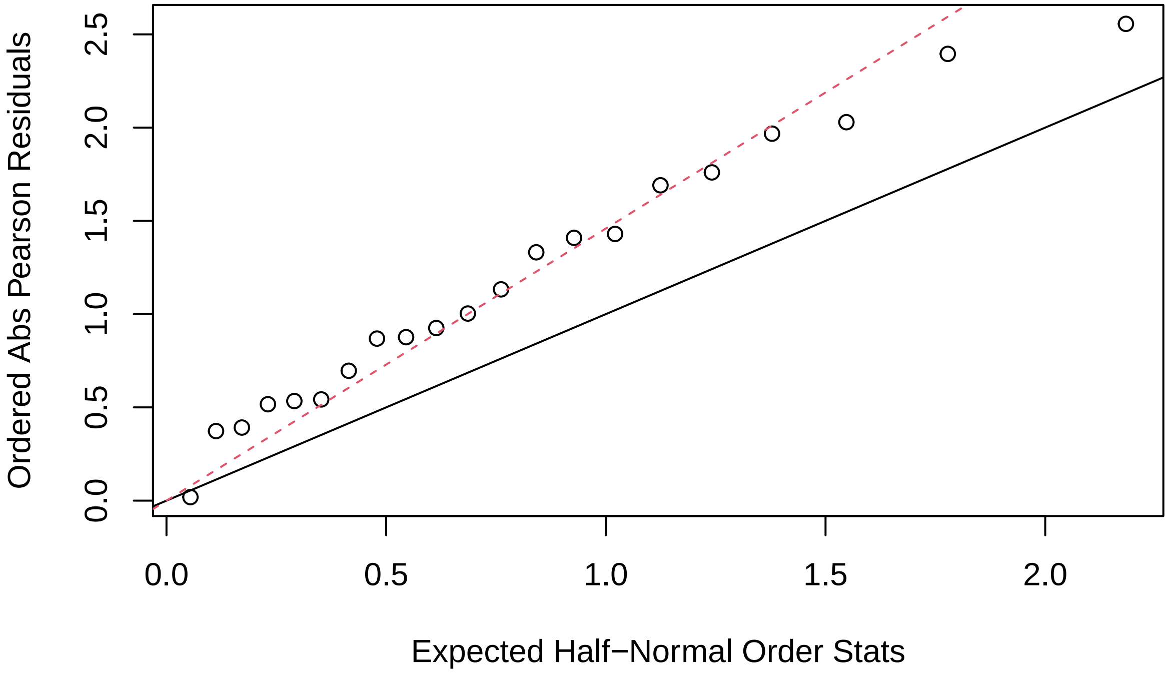
| Coefficients | OR | 95% CI of OR lower limits | 95% CI of OR upper limits |
|---|---|---|---|
| Intercept | 0.496 | 0.369 | 0.667 |
| type=OA75 | 1.311 | 0.968 | 1.775 |
| root extract=cucumber | 2.900 | 2.186 | 3.847 |

The model means that hold species fixed, we expect to see a 190.01 % increase in the odds (or 2.9 times the odds) of germination in cucumber root extract media compared to the bean root extract media. On the other hand, if we fixed root extract media, then the odds of germination decrease by 31.06 % (or 1.31 times the odds) in OA75 species compared to OA73 species. The 95% confidence interval of root extract doesn't contain 1, we have 95% confidence to conclude that the type of root extract is significantly associated with success in germination. However, the 95% CI of species contains 1, we do not have enough confidence to say that the type of species is significantly associated with germination.

(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(none.disp,type='pearson')^2) # pearson chisq
phi=G.stat/(none.disp$df.residual)
```

The true dispersion parameter is represented by $\phi$.From the estimation of dispersion parameter $\hat{\phi} = \frac{G_0}{n-p} = 2.128$, where $G_0$ is the generalized Pearson $\chi^2$ from the original model fitting without over-dispersion. Since $\hat{\phi} > 1$, that means there is an over-dispersion.



And from the Half Normal Plot we can see that the slope of ordered Pearson residuals line is greater than that of the reference line.

```
summary(none.disp,dispersion=phi)
```

```
##
## Call:
## glm(formula = cbind(germ, num_seed - germ) ~ type + root, family = binomial("logit"))
##
## 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 **
## typeOA75       0.2705     0.2257   1.198  0.23081
## rootcucumber   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
```

The interpretation of $\beta$ coefficients in the dispersion model doesn't change, but the standard error of $\beta's$ changed. As the standard error of $\beta's$ increased in the dispersion model, the confidence interval of OR became wider.

(c) What is a plausible cause of the over dispersion?

The over dispersion is caused by the true data not follows the binomial distribution exactly. There may be **intra-class correlations** between each group. In this situation, after a success in germination of a seed, the germinated seed may extract some chemicals that promote the germination of other seeds, and the germination rates of different seeds in different root extract medias may be different, that is **hierarchical sampling**.