P8131 HW3

Shihui Zhu, sz3029

Problem 1

(a) Fit a prospective model to the data to study the relation consumption, age, and disease. Interpret the result.

Model age as a continuous variable taking values 25, 35, 45, 55, 65, and 75

```
## # A tibble: 6 x 5
##     age case_0_79 case_80 control_0_79 control_80
##   <dbl>     <dbl>   <dbl>        <dbl>      <dbl>
## 1    25         0       1          106          9
## 2    35         5       4          164         26
## 3    45        21      25          138         29
## 4    55        34      42          139         27
## 5    65        36      19           88         18
## 6    75         8       5           31          0
```

n_1 = total case, n_0 total control, z_1 case_80, z_0 control_80

Fit a logit model:

```
# Fit GLM
fit.logit = glm(cbind(case_80, control_80) ~ age, family = binomial(link = 'logit'),
          data = data1)
fit.logit$coefficients
```

```
## (Intercept)         age
## -3.27242625  0.06214693
```

The model gives us $\alpha^* = -3.27242625$, and $\beta = 0.06214693$. So the model is

$$P(D = 1|X, S = 1) = \frac{e^{-3.27242625+0.06214693x}}{1 + e^{-3.27242625+0.06214693x}}$$

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

```
exp(coef(summary(fit.logit))[2])
```

```
## [1] 1.064119
```

The model means that for a one year increase in age, we expect to see 6.41% increase in the odds of having more than 80g daily alcohol consumption among group with esophageal cancer, comparing with non-esophageal cancer group.

(b) Comparing odds ratio between age groups

Two Model: $M_0 : \psi_j = 1$ for all j, and $M_1 : \psi_j = \psi$:

```
# Add group j index 1 - 6
data1["age_group"] = c("1", "2", "3", "4", "5", "6")

# Build Model 0, only the intercept is used
M0 = glm(cbind(case_80, control_80) ~ 1, family = binomial(link = 'logit'),
          data = data1)
# Build Model 1
M1 = glm(cbind(case_80, control_80) ~ age_group, family = binomial(link = 'logit'),
          data = data1)

summary(M0)
```

```
##
## Call:
## glm(formula = cbind(case_80, control_80) ~ 1, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -2.5270  -3.9186  -0.0785   2.3388   0.5506   2.7544
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.127      0.140  -0.907    0.364
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 35.107  on 5  degrees of freedom
## Residual deviance: 35.107  on 5  degrees of freedom
## AIC: 55.292
##
## Number of Fisher Scoring iterations: 4
```

```
summary(M1)
```

```
##
## Call:
## glm(formula = cbind(case_80, control_80) ~ age_group, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.1972     1.0541  -2.084   0.0371 *
## age_group2     0.3254     1.1830   0.275   0.7833
## age_group3     2.0488     1.0888   1.882   0.0599 .
## age_group4     2.6391     1.0826   2.438   0.0148 *
```

```
## age_group5      2.2513     1.1042   2.039   0.0415 *
## age_group6     26.7337 57729.9201   0.000   0.9996
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.5107e+01  on 5  degrees of freedom
## Residual deviance: 2.2078e-10  on 0  degrees of freedom
## AIC: 30.185
##
## Number of Fisher Scoring iterations: 22
```

Check if they are nested:

```
M0$coefficients
```

```
## (Intercept)
##  -0.1269997
```

```
M1$coefficients
```

```
## (Intercept)  age_group2  age_group3  age_group4  age_group5  age_group6
##  -2.1972246   0.3254224   2.0488046   2.6390573   2.2512918  26.7337242
```

M_0 is nested in M_1 because it only contains the intercept.

Use Deviance Analysis to compare the two model:

$H_0 : \beta_j = 0$, $H_1 : \beta_j \neq 0$, for $j = 1, 2, 3, 4, 5$

```
# Deviance
dev0 = M0$deviance
dev1 = M1$deviance
p2 = M1$df.null - M1$df.residual;p2
```

```
## [1] 5
```

```
# D_0 - D_1 ~ Chisquare(df=p2)
pchisq(dev0-dev1, p2, lower.tail = FALSE)
```

```
## [1] 1.432565e-06
```

The deviance of M_0 is 35.10683, and the deviance of M_1 is approximately 0. The number of predictors of M_1 is 5. Therefore we get a p-value of 1.432565e-06 < 0.05 and we reject the null hypothesis. M_1 better fits the data.

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

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

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