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.

This is a retrospective study i.e. case-control study. We therefore model (Z_1, n_1) and (Z_0, n_0) with age, diseased status as predictors. Model age as a continuous variable taking values 25, 35, 45, 55, 65, and 75:

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

Fit a logit model:

```
# Fit GLM
# Exposed v.s. Unexposed
fit.logit = glm(cbind(exposed, unexposed) ~ diseased + age, family = binomial(link = 'logit'), data = da
summary(fit.logit)
```

```
##
## Call:
## glm(formula = cbind(exposed, unexposed) ~ diseased + age, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0316  -0.9954   0.3196   0.9712   1.2647
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.7727684  0.3163332  -5.604 2.09e-08 ***
## diseased     1.7381306  0.1874862   9.271  < 2e-16 ***
## age         -0.0008152  0.0065648  -0.124    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 121.331  on 11  degrees of freedom
## Residual deviance:  24.883  on  9  degrees of freedom
## AIC: 73.471
##
## Number of Fisher Scoring iterations: 4
```

The model gives us $\alpha_0 = -1.7727683803$, and $\alpha_1 = 1.7381305722$, $\alpha_2 = -0.0008151921$. So the logit model is $log(\frac{\rho}{1-\rho}) = \alpha_0 + \alpha_1 D + \alpha_2 Age$, $D = (0, 1)$. Note that the age variable has a very large p-value so it is not significant for response prediction in this case (this happens because we treated as continuous variable).

Then the model of alcohol consumption with respect to age and disease is given by:

$$P(E = Exposure|D = Disease, X = Age) = \frac{e^{-1.7727683803+1.7381305722D-0.0008151921x}}{1 + e^{-1.7727683803+1.7381305722D-0.0008151921x}}$$

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

```r
# odds of E given no disease (control)
exp(coef(summary(fit.logit))[1])
```

```
## [1] 0.1698621
```

```r
# odds of E between case and control group
exp(coef(summary(fit.logit))[2])
```

```
## [1] 5.686703
```

```r
# odds of E given age
exp(coef(summary(fit.logit))[3])
```

```
## [1] 0.9991851
```

The model means that the odds of exposure to daily alcohol consumption of 80+g is 0.1698621 given the person does not have esophageal cancer.

And the odds of exposure to daily alcohol consumption of 80+g for people with esophageal cancer is 5.686703 times the odds of people without esophageal cancer.

Also, for a one year increase in age, the odds of exposure to daily alcohol consumption of 80+g for people with esophageal cancer is 0.9991851 times the odds of exposure of people without esophageal cancer.

(b) Comparing odds ratio between age groups

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

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

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

summary(M0)
```

```
##
## Call:
## glm(formula = cbind(exposed, unexposed) ~ age_group, family = binomial(link = "logit"),
##     data = data1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6945  -1.7579   0.8174   2.2907   4.8698
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3609     0.3308  -7.137 9.56e-13 ***
## age_group2    0.6322     0.3856   1.639 0.101127
## age_group3    1.2809     0.3664   3.496 0.000472 ***
## age_group4    1.4417     0.3601   4.003 6.25e-05 ***
## age_group5    1.1515     0.3802   3.029 0.002454 **
## age_group6    0.3067     0.5788   0.530 0.596176
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 121.331  on 11  degrees of freedom
## Residual deviance:  90.563  on  6  degrees of freedom
## AIC: 145.15
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(M1)
```

```
##
## Call:
## glm(formula = cbind(exposed, unexposed) ~ diseased + age_group,
##     family = binomial(link = "logit"), data = data1)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  1.49346  -0.06957   0.14775   0.43128  -1.30444   1.10080  -0.22828   0.02195
##        9        10        11        12
## -0.10162  -0.38727   1.33001  -1.92431
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -2.3878     0.3319  -7.194 6.28e-13 ***
## diseased      1.6699     0.1896   8.807  < 2e-16 ***
## age_group2    0.5414     0.3885   1.394   0.1635    
## age_group3    0.8486     0.3759   2.258   0.0240 *  
## age_group4    0.8299     0.3739   2.220   0.0264 *  
## age_group5    0.4428     0.3993   1.109   0.2675    
## age_group6   -0.4002     0.6042  -0.662   0.5078    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 121.331  on 11  degrees of freedom
## Residual deviance:  11.041  on  5  degrees of freedom
## AIC: 67.63
##
## Number of Fisher Scoring iterations: 4
```

Check if they are nested:

```
M0$coefficients
```

```
## (Intercept)  age_group2  age_group3  age_group4  age_group5  age_group6 
##  -2.3608540   0.6321527   1.2809338   1.4416689   1.1514903   0.3067303
```

```
M1$coefficients
```

```
## (Intercept)    diseased  age_group2  age_group3  age_group4  age_group5 
##  -2.3878270   1.6698900   0.5414210   0.8485817   0.8299038   0.4427683 
##  age_group6 
##  -0.4001646
```

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 $j = 1, 2, 3, 4, 5, 6$

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

```
## [1] 6
```

```
# D_0 - D_1 ~ Chisquare(df=p2)
diff = dev0 - dev1;diff
```

```
## [1] 79.52203
```

```
pchisq(diff, p2, lower.tail = FALSE)
```

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

The difference between deviance of M_0 and M_1 is 79.52203. The number of predictors of M_1 is 6. Therefore we get a very small p-value 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

```
data2 = tibble(species = c("o_a_75", "o_a_75", "o_a_75", "o_a_75", "o_a_75",
                           "o_a_75", "o_a_75", "o_a_75", "o_a_75", "o_a_75", "o_a_75",
                           "o_a_73", "o_a_73", "o_a_73", "o_a_73", "o_a_73", "o_a_73",
                           "o_a_73", "o_a_73", "o_a_73", "o_a_73"),
               rootMedia = c("b", "b", "b", "b", "b",
                             "c", "c", "c", "c", "c", "c",
                             "b", "b", "b", "b", "b",
                             "c", "c", "c", "c", "c"),
               germ = c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3),
               total = c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51,
data2
```

```
## # A tibble: 21 x 4
##    species rootMedia  germ total
##    <chr>   <chr>     <dbl> <dbl>
##  1 o_a_75  b            10    39
##  2 o_a_75  b            23    62
##  3 o_a_75  b            23    81
##  4 o_a_75  b            26    51
##  5 o_a_75  b            17    39
##  6 o_a_75  c             5     6
##  7 o_a_75  c            53    74
##  8 o_a_75  c            55    72
##  9 o_a_75  c            32    51
## 10 o_a_75  c            46    79
## # ... with 11 more rows
```

Fit the model

```
# Build Model
fit.logit2 = glm(cbind(germ, total-germ) ~ species + rootMedia, family = binomial(link = 'logit'),
          data = data2)
summary(fit.logit2)
```

```
##
## Call:
## glm(formula = cbind(germ, total - germ) ~ species + rootMedia,
##     family = binomial(link = "logit"), data = data2)
```

```
##
## 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 ***
## specieso_a_75   0.2705     0.1547   1.748   0.0804 .
## rootMediac      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 model gives us $\alpha = -0.7005$, and $\beta_1 = 0.2705$, $\beta_2 = 1.0647$. So the model is

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

where X_1 indicates the species of the Orobanche seeds (1 = O. aegyptiaca 75, 0 = O. aegyptiaca 73), X_2 indicates the root extract media (1 = cucumber, 0 = bean). The risk ratio (RR) of disease corresponding to unit change in different covariates is:

```
exp(coef(summary(fit.logit2))[1])
```

```
## [1] 0.4963454
```

```
# for species
exp(coef(summary(fit.logit2))[2])
```

```
## [1] 1.310555
```

```
# for root extract media RR
exp(coef(summary(fit.logit2))[3])
```

```
## [1] 2.900113
```

```
# for root extract = c, species = 75
exp(coef(summary(fit.logit2))[2] + coef(summary(fit.logit2))[3])
```

```
## [1] 3.800759
```

The model means that for a O. aegyptiaca 73 seed in bean root extract media, it has 0.4963454 germination rate.

And $e^{\beta_1} = 1.310555$, meaning that using bean as the root extract media, a O. aegyptiaca 75 seed is expected to have a gernimation rate of 1.31 times the germination rate of a O. aegyptiaca 73 seed.

$e^{\beta_2} = 2.900113$ means that for O. aegyptiaca 73 seed, cucumbers root media makes the seed to have a gernimation rate of 2.900113 times the germination rate of the seed with beans root media.

And for a O. aegyptiaca 75 seed using cucumbers root media, it has 3.800759 times the germination rate of a O. aegyptiaca 73 seed in bean root extract media.

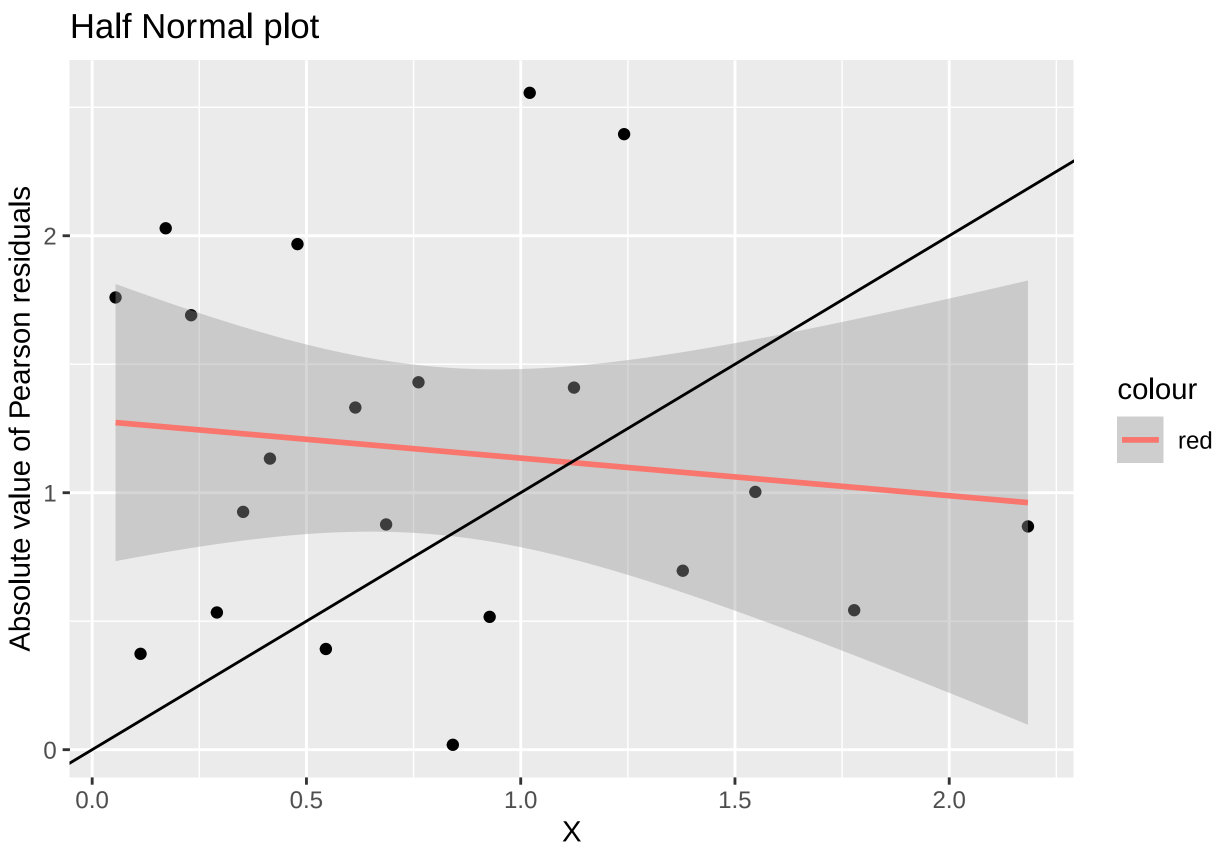
(b) Is there over dispersion? If so, what is the estimate of dispersion parameter? Update your model and reinterpret the result.

check for overdispersion using half normal plot:

```
r = abs(residuals(fit.logit2, type = 'pearson'))
# n = 21
x = 1:21
# x axis is inverse normal
x = qnorm((21 + x + 0.5)/(21*2 + 1.125))

plot_num = tibble(x = x, y = r)
plot_num %>% ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", aes(color = "red")) +
  geom_abline(slope = 1) +
  labs(
    title = "Half Normal plot",
    x = "X",
    y = "Absolute value of Pearson residuals")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The distribution of points is clearly off the reference line of slope = 1, this indicates constant overdispersion.

The estimate of the dispersion parameter ϕ is:

```
G.0 = sum(residuals(fit.logit2, type = 'pearson')^2)
# degree of freedom is 21 - 3
phi = G.0/(21 - 3);phi
```

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

The estimated dispersion ϕ is 2.128368, so there is indeed overdispersion.

Update the model and reinterpret the result:

```
summary(fit.logit2, dispersion = phi)
```

```
##
## Call:
## glm(formula = cbind(germ, total - germ) ~ species + rootMedia,
##     family = binomial(link = "logit"), data = data2)
##
## 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 **
## specieso_a_75   0.2705     0.2257   1.198  0.23081
## rootMediac      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 adjusting for the dispersion parameter, the coefficients of the model did not change, but the sample error has increased for each parameter.

For the odds of germination for a O. aegyptiaca 73 seed in bean root extract media, the SE increased from 0.1507 to 0.2199, and the odds of germination for a O. aegyptiaca 75 seed using bean as the root extract media, the SE increased from 0.1547 to 0.2257. For odds of germination of O. aegyptiaca 73 seed in cucumbers root media, the SE increased from 0.1442 to 0.2104.

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

Over-dispersion indicates that the germination rate does not follow our hypothetical binomial distribution. There might be intra-class correlation in each seed set. For example, some seeds germinated first and occupy the resource of the media so others are less likely to germinate later. The germination rate could be correlated to source of supply, so there is hierarchichal sampling effect between some set of seeds. For example, O. aegyptiaca 75 seeds collected from one specific supplier are lessly like to germinate than others.