p8131_hw3_ps3194

Pangsibo Shen

2/15/2021

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Question 1 Esophageal Cancer and Daily Alcohol Consumption/Ages

```
age = c(25,35,45,55,65,75,25,35,45,55,65,75)
alc_level = c(0,0,0,0,0,0,1,1,1,1,1,1) #0 stands for daily alcohol 0-79g and 1 stands for daily alcohol
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)
```

```
# create a dataframe for the question
df1 = tibble(age,alc_level,case,control)
# fit a prospective model using logit link
logit_fit_1 = glm(cbind(case,control)~age+alc_level, family = binomial(link = "logit"), data = df1)
```

```
summary(logit_fit_1)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ age + alc_level, family = binomial(link = "logit"),
##     data = df1)
##
## 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 ***
## alc_level    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
```

```
#anti log for interpretation of beta
exp(logit_fit_1$coefficients)
```

```
## (Intercept)         age   alc_level
## 0.006581788 1.063514159 5.929853538
```

For the interpretation of $\beta's$, we need to first exponentiate all the coefficients and the relative odds of disease among non-exposure is 0.0066; the odds ratio of getting esophageal cancer between people whose daily alcohol consumption greater than 80g and people whose daily alcohol consumption less than 80g is 5.9299 while holding the age as constant;the odds ratio of getting esophageal cancer between one year older people with younger people is 1.063514159 while holding the daily alcohol intake status as constant.

Question 2 Orobanche Seeds Study

```
df2 = tibble(
  yi = c(10,23,23,26,17,8,10,8,23,0,5,53,55,32,46,10,3,22,15,32,3), #number of germinating
  mi = c(39,62,81,51,39,16,30,28,45,4,6,74,72,51,79,13,12,41,30,51,7),#number of seeds
  root_extract = c(c(rep(0,10)),c(rep(1,11))),
  type = c(c(rep(0,5)),c(rep(1,5)),c(rep(0,6)),c(rep(1,5)))
) %>%
  relocate(type,root_extract,yi,mi) %>%
  mutate(type = as.factor(type), # 0 stands for O. aegyptiaca 75 and 1 stands for O. aegyptiaca 73
         root_extract = as.factor(root_extract), # 0 stands for Bean root extraction and 1 stands for c
         yi_0 = mi - yi) # number of seeds not germinating
```

```
# fit a prospective model using logit link
logit_fit_2 = glm(cbind(yi,yi_0)~type+root_extract, family = binomial(link = "logit"), data = df2)
summary(logit_fit_2)
```

a)

```
##
## Call:
## glm(formula = cbind(yi, yi_0) ~ type + root_extract, family = binomial(link = "logit"),
```

```
##     data = df2)
##
## 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)    -0.4300     0.1137  -3.781 0.000156 ***
## type1          -0.2705     0.1547  -1.748 0.080435 .
## root_extract1   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
```

```
#anti log for interpretation of beta
exp(logit_fit_2$coefficients)
```

```
##   (Intercept)         type1 root_extract1
##     0.6504882     0.7630352     2.9001133
```

For the intercept, we first apply anti log to get valid number for interpretations. The odds for O. aegyptiaca 75 grown on Bean root extract media is 0.6504882; the odds ratio for growing in Bean root extract media between O. aegyptiaca 75 and O. aegyptiaca 73 is 0.7630352; the odds ratio for O. aegyptiaca 75 growing in between Bean root extract media and Cucumber root extract media is 2.9001133. Types of seed seems to have weak effect on the germination rate for the seeds since its p-value is 0.080435 which is greater than 0.05 yet less than 0.1.

```
#calculate G_0 and phi
G_0 = sum(residuals(logit_fit_2,type='pearson')^2) # pearson chisq
G_0
```

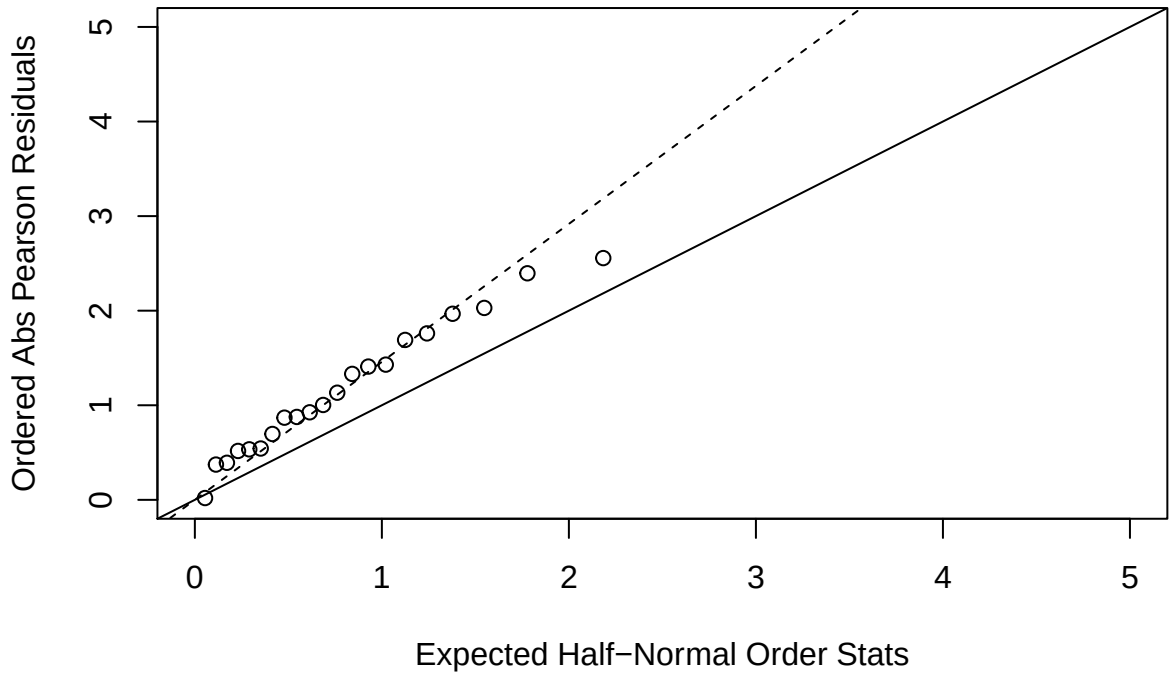
b)

```
## [1] 38.31062
```

```
phi = G_0/(21-3)
phi
```

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

```r
# test over-dispersion (half normal plot)
res = residuals(logit_fit_2,type='pearson')
plot(qnorm((21+1:21+0.5)/(2*21+1.125)),sort(abs(res)),xlab='Expected Half-Normal Order Stats',
ylab='Ordered Abs Pearson Residuals', ylim=c(0,5),xlim=c(0,5))
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)
```



From the Half Normal Plot, there is linear deviation from the reference line which indicates constant over-dispersion. The dispersion parameter $\hat{\phi}$ is approximated to be 2.1283678

```r
# refit model with constant over-dispersion
summary(logit_fit_2,dispersion=phi)
```

```
##
## Call:
## glm(formula = cbind(yi, yi_0) ~ type + root_extract, family = binomial(link = "logit"),
##     data = df2)
##
## 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)    -0.4300     0.1659  -2.592  0.00955 **
## type1          -0.2705     0.2257  -1.198  0.23081
```

```
## root_extract1   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 I updated my model with dispersion parameter, all the beta coefficients stay the same. Hence the interpretation for the coefficients stay the same as 2 a). However, the std.error for each coefficients became larger after I updated my model with dispersion parameter and the p-value for type iS 0.23 which is greater than 0.05. Hence we could conclude that the types of seed does not affect the seed germination rate.

c) First, within each batch, the germination of each seeds might not be strict Bernoulli process. In other words, the germination of one seed might impact its neighboring seeds. Second, within each batch, the germination rate for each seed is likely to be different.