code for HW3

Hana Akbarnejad

2/24/2020

All code chunks and output used for HW3

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.4
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

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

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
knitr::opts_chunk$set(
    echo = TRUE,
    warning = FALSE,
    fig.width = 8,
  fig.height = 6,
  out.width = "90%"
)
options(
  ggplot2.continuous.colour = "viridis",
  ggplot2.continuous.fill = "viridis"
)
scale_colour_discrete = scale_colour_viridis_d
scale_fill_discrete = scale_fill_viridis_d
theme_set(theme_minimal() + theme(legend.position = "bottom"))
```

```
# help myself understand the data (retrospective data):

##############################################################################
#
#                       case(cancer+)         control(cancer-)
#
# #exosed (alcohol80+)      96                        109             n1=205
# #unexposed (alc 80-)      104                       666             n0=770
#
#                           m1=200                    m0=775
#
#
##############################################################################

cancer_data = tibble(
  age = rep(c(25, 35, 45, 55, 65, 75), 2),
  alcohol_consump = c(rep("0-79 g", 6), rep("80+ g", 6)),
  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)
)

cancer_data
```

```
## # A tibble: 12 x 4
##      age alcohol_consump  case control
##    <dbl> <chr>           <dbl>   <dbl>
##  1    25 0-79 g              0     106
##  2    35 0-79 g              5     164
##  3    45 0-79 g             21     138
##  4    55 0-79 g             34     139
##  5    65 0-79 g             36      88
##  6    75 0-79 g              8      31
##  7    25 80+ g               1       9
##  8    35 80+ g               4      26
##  9    45 80+ g              25      29
## 10    55 80+ g              42      27
## 11    65 80+ g              19      18
## 12    75 80+ g               5       0
```

```
m1_df = cancer_data %>%
  group_by(alcohol_consump) %>%
  select(case) %>%
  summarize(sum_case = sum(case))
```

```
## Adding missing grouping variables: `alcohol_consump`
```

```
m1 = sum(m1_df$sum_case)

m0_df = m1 = cancer_data %>%
  group_by(alcohol_consump) %>%
  select(control) %>%
  summarize(sum_cntrl = sum(control))
```

```
## Adding missing grouping variables: `alcohol_consump`
```

```
m0 = sum(m0_df$sum_cntrl)

# using table to calculate n1 and n0
n1 = 205
n0 = 770
# now I know what my table looks like!

#
response = cbind(cancer_data$case, cancer_data$control)
response
```

```
##       [,1] [,2]
##  [1,]    0  106
##  [2,]    5  164
##  [3,]   21  138
##  [4,]   34  139
##  [5,]   36   88
##  [6,]    8   31
##  [7,]    1    9
##  [8,]    4   26
##  [9,]   25   29
## [10,]   42   27
## [11,]   19   18
## [12,]    5    0
```

```
# fitting a prospective model
logit_prosp = glm(response ~ cancer_data$alcohol_consump + cancer_data$age, family=binomial(link='logit
summary(logit_prosp)
```

```
##
## Call:
## glm(formula = response ~ cancer_data$alcohol_consump + cancer_data$age,
##     family = binomial(link = "logit"))
##
## 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 ***
## cancer_data$alcohol_consump80+ g  1.780000   0.187086   9.514   <2e-16 ***
## cancer_data$age                   0.061579   0.007291   8.446   <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
```

```
# seed: 0 >> O. aegyptiaca 75, seed: 1 >> O. aegyptiaca 73
# root: 0 >> Bean, root: 1 >> cucumber

# making data df
germ_data = tibble (
  seed = c(rep(0, 11), rep(1, 10)),
  root = c(rep(0, 5), rep(1, 6), rep(0, 5), rep(1, 5)),
  y = c(c(10, 23, 23, 26, 17), c(5, 53, 55, 32, 46, 10), c(8, 10, 8, 23, 0), c(3, 22, 15, 32, 3)),
  m = c(c(39, 62, 81, 51, 39), c(6, 74, 72, 51, 79, 13), c(16, 30, 28, 45, 4), c(12, 41, 30, 51, 7))
)

germ_data
```

```
## # A tibble: 21 x 4
##     seed  root     y     m
##    <dbl> <dbl> <dbl> <dbl>
##  1     0     0    10    39
##  2     0     0    23    62
##  3     0     0    23    81
##  4     0     0    26    51
##  5     0     0    17    39
##  6     0     1     5     6
##  7     0     1    53    74
##  8     0     1    55    72
##  9     0     1    32    51
## 10     0     1    46    79
## # ... with 11 more rows
```

```
germ_rate = cbind(germ_data$y, germ_data$m - germ_data$y)
germ_rate
```

```
##       [,1] [,2]
##  [1,]   10   29
##  [2,]   23   39
##  [3,]   23   58
##  [4,]   26   25
##  [5,]   17   22
##  [6,]    5    1
##  [7,]   53   21
##  [8,]   55   17
##  [9,]   32   19
## [10,]   46   33
## [11,]   10    3
## [12,]    8    8
## [13,]   10   20
## [14,]    8   20
## [15,]   23   22
## [16,]    0    4
## [17,]    3    9
## [18,]   22   19
```

```
## [19,]   15   15
## [20,]   32   19
## [21,]    3    4
```

```
germ_logit = glm(germ_rate ~ seed + root, family = binomial(link = 'logit'), data = germ_data)
summary(germ_logit)
```

```
##
## Call:
## glm(formula = germ_rate ~ seed + root, family = binomial(link = "logit"),
##     data = germ_data)
##
## 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 ***
## seed         -0.2705     0.1547  -1.748 0.080435 .
## root          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
```

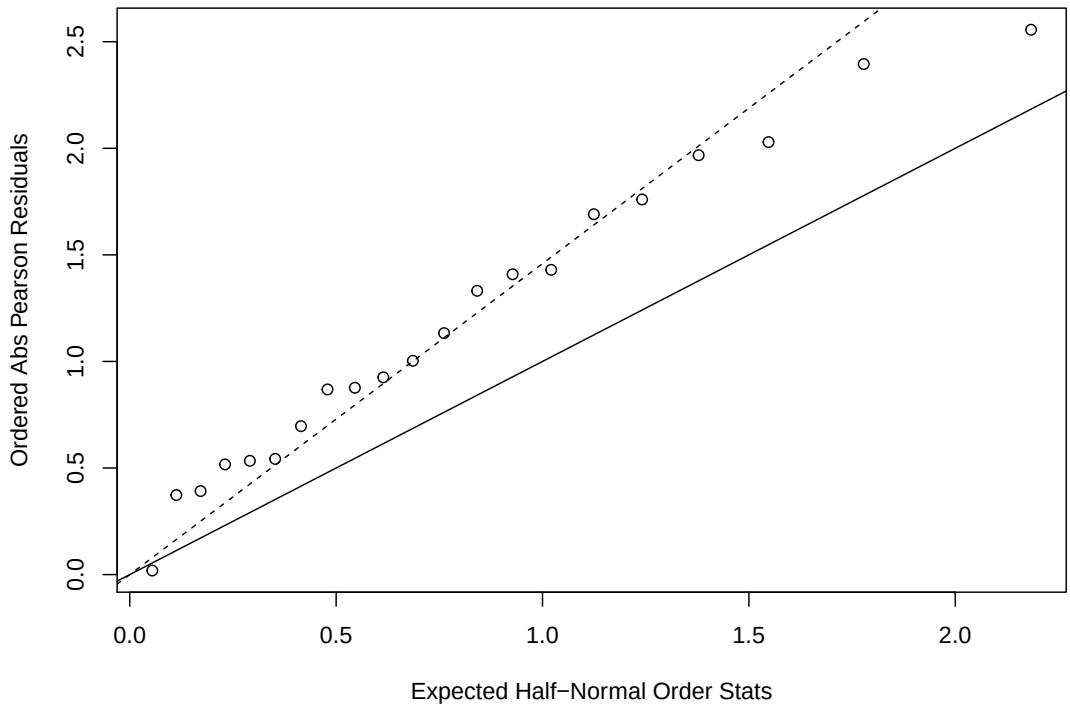
```
# n = 21 (total number)
# p = 3 (seed and root)
# n - p = 18

G_0 = sum(residuals(germ_logit,type='pearson')^2) # pearson chisq statistics (G0)
phi_hat_G = G_0 / (21-3) # over dispersion parameter estimated using G_0
phi_hat_G
```

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

```
# phi_hat_D = germ_logit$deviance/(21-3)
# tilde.phi=germ_logit$deviance/germ_logit$df.residual his!
```

```
res=residuals(germ_logit,type='pearson')
plot(qnorm((21+1:21+0.5)/(2*21+1.125)),sort(abs(res)),xlab='Expected Half-Normal Order Stats',ylab='Ord
abline(a=0,b=1)
abline(a=0,b=sqrt(phi_hat_G),lty=2)
```




```
summary(germ_logit, dispersion = phi_hat_G)
```

```
##
## Call:
## glm(formula = germ_rate ~ seed + root, family = binomial(link = "logit"),
##     data = germ_data)
##
## 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 **
## seed         -0.2705     0.2257  -1.198  0.23081
## root          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
```