lec15_1

Jiansong Xu

March 6, 2019

1

```
workers <- read.csv("healthcare_worker.csv")
head(workers)
```

```
##          Occup.group Hepatitis Size
## 1     Exposure prone         5 2205
## 2      Fluid contact        17 6207
## 3          Lab staff         3  533
## 4    Patient contact         2 1238
## 5 No patient contact         3  471
```

```
levels(workers$Occup.group)
```

```
## [1] "Exposure prone"     "Fluid contact"      "Lab staff"
## [4] "No patient contact" "Patient contact"
```

a.

```
library(tidyverse)
workers %>% mutate(Presence = Hepatitis, Absence = Size - Hepatitis) %>% select(-Hepatitis, -Size) %>%
gather(key = status, value = Number, Absence, Presence) -> workers
xtabs(Number ~ Occup.group + status, data = workers) -> wrk.table
wrk.table
```

```
##                    status
## Occup.group          Absence Presence
##   Exposure prone        2200        5
##   Fluid contact         6190       17
##   Lab staff              530        3
##   No patient contact     468        3
##   Patient contact       1236        2
```

b.

Let I,J denote Occup.group and status

H0: Pi_ij = Pi_i+ + Pi_+j for each i,j;

Ha: Pi_ij != Pi_i+ + Pi_+j for some i,j

```
library(package = vcd)
```

```
## Loading required package: grid
```

```
assocstats(x = wrk.table)##LR test and Pearson
```

```
##                    X^2 df P(> X^2)
## Likelihood Ratio 3.7350  4  0.44305
## Pearson          4.5043  4  0.34204
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.021
## Cramer's V        : 0.021
```

We cannot reject H0 in either of the tests. So we can conclude type of occupational group and status of hepatitis are independent.

c.

```
ind.test <- chisq.test(x=wrk.table, correct=FALSE)
```

```
## Warning in chisq.test(x = wrk.table, correct = FALSE): Chi-squared
## approximation may be incorrect
```

```
prop.table(wrk.table, margin=1)
```

```
##                     status
## Occup.group              Absence    Presence
##   Exposure prone     0.997732426 0.002267574
##   Fluid contact      0.997261157 0.002738843
##   Lab staff          0.994371482 0.005628518
##   No patient contact 0.993630573 0.006369427
##   Patient contact    0.998384491 0.001615509
```

```
round(ind.test$stdres, digits=1)
```

```
##                     status
## Occup.group          Absence Presence
##   Exposure prone         0.5     -0.5
##   Fluid contact          0.2     -0.2
##   Lab staff             -1.3      1.3
##   No patient contact    -1.5      1.5
##   Patient contact        0.8     -0.8
```

Workers who have no patient contacts tends to have higher conditional probability of carrying hepatitis, and those who have patient contact have the lowest conditional probability among all groups(This was not my expection). None of the residuals appear to be unusual(absolute value greater than 2 or 3), but 'No patient contact' group has relatively large residuals.

d.

H0: beta_1 = beta_2 = beta_3 = beta_4 = 0;

Ha: not all beta_s are 0

```
wrk2 <- read.csv("healthcare_worker.csv")
wrk2 <- mutate(wrk2, Occup.group = as.factor(Occup.group))
lgfit <- glm(Hepatitis/Size ~ Occup.group, weights = Size, family = binomial(link = "logit"), data = wrk2)
library(car)
Anova(lgfit, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Hepatitis/Size
##             LR Chisq Df Pr(>Chisq)
## Occup.group    3.735  4     0.4431
```

We cannot reject H0. So we can conclude all occupational groups have the same probability of hepatitis.

e. Because assuming 2 variables in a contingency table are independent is basiclly same as assuming one has 0 effect on another, putting this into logistic regression we should expect to see the one as explanatory variable is not signifacant in model fit.