536HW5

Coco_Luo

2022-11-13

problem 1

1.

```
# create the 2x2 table
crohn1 = c(2037, 1757, 958, 218)
crohn1.array = array(crohn1, c(2,2))
crohn1.array
```

```
##      [,1] [,2]
## [1,] 2037  958
## [2,] 1757  218
```

```
# fit the log-linear model of independence
indep.loglin1 = loglin(crohn1.array, margin = list(1, 2), fit = T, param = T)
```

```
## 2 iterations: deviation 0
```

```
# fit the saturated log-linear model
saturated.loglin1 = loglin(crohn1.array, margin = list(c(1,2)), fit = T, param = T)
```

```
## 2 iterations: deviation 0
```

```
# X^2 statistics
p.pearson1 = 1 - pchisq(indep.loglin1$pearson - saturated.loglin1$pearson,
                        indep.loglin1$df - saturated.loglin1$df)
# likelihood ratio test statistic G^2
p.lrt1 = 1 - pchisq(indep.loglin1$lrt - saturated.loglin1$lrt,
                    indep.loglin1$df - saturated.loglin1$df)
```

When we did the likelihood ratio test and get our p values based on χ^2 and G^2 test statistics, we got 0 as a result for both. We reject the null hypothesis of independence and concluded that the log-linear model of independence does not fit the data well.

2

From the previous problem, we obtained p values smaller than 0.05, so we concluded that the log-linear model of independence does not fit the data well, instead we choose to use the saturated log-linear model.

```
mydata1 = matrix(c(rep(c(1,1),2037),rep(c(1,2),958),
                   rep(c(2,1),1757),rep(c(2,2),218)),
                 ncol = 2, byrow = TRUE)
mylogit1 =loglin(crohn1.array, margin = list(c(1, 2)), fit = T, param = T)
```

```
## 2 iterations: deviation 0
```

```
mylogit1$param
```

```
## $`(Intercept)`
## [1] 6.834985
##
## $`1`
## [1]  0.4070558 -0.4070558
##
## $`2`
## [1]  0.7103134 -0.7103134
##
## $`1.2`
##             [,1]       [,2]
## [1,] -0.3331206  0.3331206
## [2,]  0.3331206 -0.3331206
```

From the above we can obtain the estimates for μ terms: $\mu = 6.835$, $\mu_{2(1)} = 0.710$, $\mu_{2(2)} = -0.710$, $\mu_{1(1)} = 0.407$, $\mu_{1(2)} = -0.406$, $\mu_{12(22)} = -0.333$, $\mu_{12(12)} = 0.333$, $\mu_{12(11)} = -0.333$, $\mu_{12(21)} = 0.333$.

```
mydata1 = matrix(c(rep(c(1,1),2037),rep(c(1,2),958),
                   rep(c(2,1),1757),rep(c(2,2),218)),
                 ncol = 2, byrow = TRUE)
mylogit1 = glm(factor(mydata1[, 1]) ~ factor(mydata1[, 2]),
               family = binomial(link = logit))
#mylogit1
```

Fitting the logistic regression, we obtained a model as below:

$$log\frac{P(X_1 = 2|X_2)}{P(X_1 = 1|X_2)} = \beta_0 + \beta_1 X_2 = -0.1479 - 1.3325X_2$$

This equation tells that for every unit change in X_2, the log-odds of X_1 given X_2 is expected to change by -1.3325. When X_2 is zero, the log-odds of X_1 equals to -0.1479. In addition, the saturated linear model can be introducted as a logistic regression model with explanatory "SNP2" and outcome "Disease."

$$log\frac{P(X_1 = 2|X_2 = i)}{P(X_1 = 1|X_2 = i)} = log(m_{2i}) - log(m_{1i})$$
$$= (\mu + \mu_{1(2)} + \mu_{12(2i)} + \mu_{2(i)}) - (\mu + \mu_{1(1)} + \mu_{12(1i)} + \mu_{2(i)})$$
$$= (\mu_{1(2)} - \mu_{1(1)}) + (\mu_{12(2i)} - \mu_{12(1i)})$$

We can see that "SNP2" and logit of "Disease" interact with each other and from what we obtained for the μ terms, we can then compute

$$log\frac{P(X_1 = 2|X_2 = 1)}{P(X_1 = 1|X_2 = 1)} = (\mu_{1(2)} - \mu_{1(1)}) + (\mu_{12(21)} - \mu_{12(11)}) = -0.407 - 0.407 + 0.333 + 0.333 = -0.148$$

$$log\frac{P(X_1 = 2|X_2 = 2)}{P(X_1 = 1|X_2 = 2)} = (\mu_{1(2)} - \mu_{1(1)}) + (\mu_{12(22)} - \mu_{12(12)}) = -0.407 - 0.407 - 0.333 - 0.333 = -1.481$$

The former equation tells us if the allele b is absent at SNP2, the odds of having the disease against not having the disease is $e^{-0.148} = 0.862$.

The second equation tells us if the allele b is present at SNP2, the odds of having the disease against not having the disease is $e^{-1.481} = 0.227$.

3.

To see whether there is an association between occurrence of Crohn's disease and obsence or presence of the minor allele b at location "SNP2" versus the alternative of no association using the cross product $\alpha = \frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \frac{\text{P(not having the disease | SNP2 = BB)}}{\text{P(not having the disease | SNP2 = Bb or bb)}}$, we want to test:

$$H_0 : \alpha = 1 \quad vs. \quad H_a : \alpha \neq 1$$

```
fisher.test(crohn1.array, simulate.p.value = T, B = 1e5)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  crohn1.array
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2237137 0.3103244
## sample estimates:
## odds ratio
##  0.2638874
```

Our odds ratio is around 0.2639 and our p value is less than 0.05. We reject the null and concluded that there is an association between occurrence of Crohn's disease and absence or presence of the minor allele b at location "SNP2".

The Fisher's Exact test allows us to further test whether the association is negative or positive. Now, we would like to test:

$$H_0 : \alpha = 1 \quad vs. \quad H_a : \alpha < 1$$

As we want to observe if the odds of minor allele b at location "SNP2" in the no disease (i.e. p_{11}/p_{12}) group is smaller than the has disease (i.e. p_{21}/p_{22}) group:

```
fisher.test(crohn1.array, simulate.p.value = T, alternative = 'greater', B = 1e5)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  crohn1.array
## p-value = 1
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.2296763       Inf
## sample estimates:
## odds ratio
##  0.2638874
```

```
fisher.test(crohn1.array, simulate.p.value = T, alternative = 'less', B = 1e5)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  crohn1.array
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is less than 1
## 95 percent confidence interval:
##  0.0000000 0.3026314
## sample estimates:
```

```
## odds ratio
##  0.2638874
```

When the alternative is set to "greater", the exact p value is 1 suggests that we fail to reject the nul. When the alternative is set to "less", the exact p value is nearly 0 suggests that we have reject the null. Therefore, we have enough evidence to conclude that the occurrence of Crohn's disease decrease the absence or presence of the minor allele b at location "SNP2".

problem 2

1.

```
# create the 2x2 table
crohn2 = c(2037, 1757, 631, 18, 327, 200)
crohn2.array = array(crohn2, c(2,3))
crohn2.array
```

```
##      [,1] [,2] [,3]
## [1,] 2037  631  327
## [2,] 1757   18  200
```

```
# fit the log-linear model of independence
indep.loglin2 = loglin(crohn2.array, margin = list(1, 2), fit = T, param = T)
```

```
## 2 iterations: deviation 4.547474e-13
```

```
# fit the saturated log-linear model
saturated.loglin2 = loglin(crohn2.array, margin = list(c(1, 2)), fit = T, param = T)
```

```
## 2 iterations: deviation 2.842171e-14
```

```
# X^2 statistics
p.pearson2 = 1 - pchisq(indep.loglin2$pearson - saturated.loglin2$pearson,
                        indep.loglin2$df - saturated.loglin2$df)

# likelihood ratio test G^2 statistics
p.lrt2 = 1 - pchisq(indep.loglin2$lrt - saturated.loglin2$lrt,
                    indep.loglin2$df - saturated.loglin2$df)
p.pearson2
```

```
## [1] 0
```

```
p.lrt2
```

```
## [1] 0
```

Our null hypothesis is that the occurrence of Crohn's disease provides no information about the absence or presence of the minor allele b at location "SNP2". When we did the likelihood ratio test and get our p values based on χ^2 and G^2 test statistics, we got 0 as a result for both. In addition, table 2 has quite large counts in each cell and does not have any structural zeros so we expect the p values to be accurate. Thus we reject the null hypothesis of independence and concluded that there is a relationship between the occurrence of Crohn's disease and the absence or presence of the minor allele b at location "SNP2"

2.

```
library(exactLoglinTest)
set.seed(536)
mydata2 = data.frame(y = c(2037, 631, 327, 1757, 18, 200),
```

```
                    disease = rep(1:2, each = 3), SNP2 = rep(1:3, times = 2))

# importance sampling
mcx = mcexact(y ~ factor(disease) + factor(SNP2), data = mydata2)
mcx
```

```
##               deviance  Pearson
## observed.stat 575.9075 439.4396
## pvalue          0.0000   0.0000
## mcse            0.0000   0.0000
```

We will use the exactLoglinTest to examine the exact p value for this task. First of all, using importance sampling, we obtained nearly the same value for G^2 and χ^2 as when we used 'loglin' function, and the estimated exact p value is 0 for both statistics.

```
# Markov chain Monte Carlo algorithm
mcx2 = mcexact(y ~ factor(disease) + factor(SNP2), data = mydata2,
               method = "cab", p = 0.5, nosim = 10^4, batchsize = 100)
mcx2
```

```
##               deviance  Pearson
## observed.stat 575.9075 439.4396
## pvalue          0.0000   0.0000
## mcse            0.0000   0.0000
```

Next, we use MCMC to perform the same calculation. We can see that both algorithm gives us an exact p value of zero, which agrees with the asymptotic p-values we calculated in 2.1. This suggest dependence between the genotype and occurrence of Crohn's disease, so we choose the saturated log-linear model.

3.

```
# hat u's
saturated.loglin2$param
```

```
## $`(Intercept)`
## [1] 5.919425
##
## $`1`
## [1]  0.6994079 -0.6994079
##
## $`2`
## [1]  1.6258730 -1.2505865 -0.3752865
##
## $`1.2`
##            [,1]      [,2]       [,3]
## [1,] -0.6254727  1.079059 -0.4535865
## [2,]  0.6254727 -1.079059  0.4535865
```

Usingt the u terms of the log-linear model of independence, we would like to examine the odds of not having disease versus having disease, we get:

$$log\frac{P(D=No)}{P(D=Yes)} = \widehat{u}_{1(1)} - \widehat{u}_{1(2)} = 0.208 - (-0.208) = 0.416$$

The ratio of people not having disease and those have the disease in the sample is 1.52. The equation above tells us that the odds of not having disease versus having disease is $e^{0.416} = 1.52$. Similarly, the odds of two categories of X_2 under the model of independence.

$$log\frac{P(X_2 = i_1)}{P(X_2 = i_2)} = \mu_{2(i1)} - \mu_{2(i2)}$$

One example is $log\frac{SNP2=Bb}{SNP2=bb} = \mu_{2(2)} - \mu_{2(3)} = 0.208$ which tells us that the odds of a person with Bb versus bb is $e^{0.208} = 1.23$, which is equal to the ratio of the number of people has genotype Bb and the number of people has genotype bb in the sample: $e^{0.208} = 649/527 = 1.23$.

However, the independence model does not fit the data in Table 2 well, which means that there is an relationship between having the disease or not and the single-locus genotype SNP2 of a person. Or we say that the odds of somebody having the disease depends on their genotype. Using the u terms corresponding with the saturated log linear model, we got the odds of having disease versus not having disease for each possible genotype at SNP2 as below:

$$\begin{aligned}
log\frac{P(X_1 = i_1|X_2 = j)}{P(X_1 = i_2|X_2 = j)} &= (u + u_{1(i_1)} + u_{2(j)} + u_{12(i_1j)}) - (u + u_{1(i_2)} + u_{2(j)} + u_{12(i_2j)}) \\
&= (u_{1(i_1)} - u_{1(i_2)}) + (u_{12(i_1j)} - u_{12(i_2j)}) \\
log\frac{\text{P(disease = No | SNP2 = BB)}}{\text{P(disease = Yes | SNP2 = BB)}} &= (\widehat{u_{1(1)}} - \widehat{u_{1(2)}}) + (\widehat{u_{12(11)}} - \widehat{u_{12(21)}}) \\
&= (0.6994 - (-0.6994)) + (-0.6255 - 0.6255) = 0.1478 \\
log\frac{\text{P(disease = No | SNP2 = Bb)}}{\text{P(disease = Yes | SNP2 = Bb)}} &= (\widehat{u_{1(1)}} - \widehat{u_{1(2)}}) + (\widehat{u_{12(12)}} - \widehat{u_{12(22)}}) \\
&= (0.6994 - (-0.6994)) + (1.079 - (-1.079)) = 3.557 \\
log\frac{\text{P(disease = No | SNP2 = bb)}}{\text{P(disease = Yes | SNP2 = bb)}} &= (\widehat{u_{1(1)}} - \widehat{u_{1(2)}}) + (\widehat{u_{12(13)}} - \widehat{u_{12(23)}}) \\
&= (0.6994 - (-0.6994)) + (-0.4536 - 0.4536) = 0.4916
\end{aligned}$$