STA303 Assignment 3 Part 2

Haoda Li

Solutions

Question 1

(a) By Fisher's exact test, the p-value is 2.5215×10^{-12}; by 2-sample test for equality of proportions (Binomial sampling), the p-value is 6.704×10^{-12}. From both test results, the p-value is extremely small and we reject the null hypothesis. We have strong evidence that the sex is dependent of a student's preference for playing video games. From the observed data, 80.8 of male students like play video games, while only 46.0 of female students like play video games.Male students tend to like video games more.

(b) The contingency table for the relationship between sex and like for A+ expected grade group is:

```
##          like games
## sex       no yes
##   female 31  11
##   male   26  32
```

The p-value for the contingency test is 0.003861.
The contingency table for the relationship between sex and like for non-A+ expected grade group is:

```
##          like games
## sex        no yes
##   female 103  18
##   male    88  90
```

The p-value for the contingency test is 6.704×10^{-12}.
In both contingency tests, the p-value is very small, and the difference of the two p-values are also very small. Therefore, there is no evidence that the association between sex and student's preference for playing video games changes with the grade expected.

Question 2

(a) **Models being fit**:
Model 2.1 (interaction model)

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1\mathbb{I}_1 + \beta_2\mathbb{I}_2 + \beta_3\mathbb{I}_1\mathbb{I}_2, i = 1, 2, ..., 399$$

Fitted equation is

$$\log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = -0.1574 + 1.7668X_1 - 0.0185X_2 - 0.5231X_1X_2, i = 1, 2, ..., 399$$

Model 2.2 (additive model)

$$\log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1\mathbb{I}_1 + \beta_2\mathbb{I}_2, i = 1, 2, ..., 399$$

Fitted equation is

$$\log(\frac{\pi_i}{1-\pi_i}) = -0.1189 + 1.6111\mathbb{I}_1 - 0.1871\mathbb{I}_2, i = 1, 2, ..., 399$$

The terms are defined as:
$\pi_i = P(\text{"like games"})$ is the probablity that the ith student likes playing video games

$\mathbb{I}_{1i}$ is the indicator that the ith student is male,
$\mathbb{I}_{2i}$ is the indicator that the ith student has expected grade A+.
Test whether we need the interaction term
Wald test:
hypotheses: $H_0 : \beta_3 = 0; H_a : \beta_3 \neq 0$
Test statistic: -0.987
Distribution of the test statistic: $N(0,1)$
p-value: 0.323
Likelihood Ratio Test:
hypotheses: H_0: the additive model (Model 2.2) is better ; H_a: the interaction model (Model 2.1) is better.
Test statistic: $489.37 - 488.41 = 0.96$
Distribution of the test statistic: χ^2_1
p-value: 0.3053
Conclusion: By both test, we get p-value much greater than significance level. We cannot reject the null hypothesis. Therefore, there is some evidence that there is no interaction between the gender and the expected grade and the additive model is adequet.

(b) The practical implication:
When holding the expected grade being the same, the odds of preference for playing video games for a male student are about 5.008 times the odds for a female student. When holding the gender being the same, the odds of preference for playing viedio game for a student with A+ expected grade are about 0.829 times the odds for a student with non-A+ expected grade.
The implication agrees with my answer to Question 1, Specifically, the preference for playing video games is associated with the gender and is independent of the expected grade.

Question 3

count	like	sex	grade
μ_1	no	female	A+
μ_2	no	female	not A+
μ_3	no	male	A+
μ_4	no	male	not A+
μ_5	yes	female	A+
μ_6	yes	female	not A+
μ_7	yes	male	A+
μ_8	yes	male	not A+

(a) **Models being fit**:
Model 2.1 (three-way interaction model)

$$\log(\mu_i) = \beta_0 + \beta_1\mathbb{I}_1 + \beta_2\mathbb{I}_2 + \beta_3\mathbb{I}_3 + \beta_4\mathbb{I}_1\mathbb{I}_2 + \beta_5\mathbb{I}_1\mathbb{I}_3 + \beta_6\mathbb{I}_2\mathbb{I}_3 + \beta_7\mathbb{I}_1\mathbb{I}_2\mathbb{I}_3, i = 1, 2, ..., 8$$

Fitted equation is:

$$\log(\hat{\mu}_i) = 3.4340 - 0.1759\mathbb{I}_1 - 1.0361\mathbb{I}_2 + 1.2007\mathbb{I}_3 + 1.2437\mathbb{I}_1\mathbb{I}_2 + 0.0185\mathbb{I}_1\mathbb{I}_3 - 0.70836\mathbb{I}_2\mathbb{I}_3 + 0.5231\mathbb{I}_1\mathbb{I}_2\mathbb{I}_3, i = 1, 2, ..., 8$$

Model 2.2 (interaction model)

$$\log(\mu_i) = \beta_0 + \beta_1\mathbb{I}_1 + \beta_2\mathbb{I}_2 + \beta_3\mathbb{I}_3 + \beta_4\mathbb{I}_1\mathbb{I}_2 + \beta_5\mathbb{I}_1\mathbb{I}_3 + \beta_6\mathbb{I}_2\mathbb{I}_3, i = 1, 2, ..., 8$$

Fitted equation is:

$$\log(\hat{\mu}_i) = 3.4913 - 0.3061\mathbb{I}_1 - 1.2751\mathbb{I}_2 + 1.1256\mathbb{I}_3 + 1.6111\mathbb{I}_1\mathbb{I}_2 + 0.1871\mathbb{I}_1\mathbb{I}_3 - 0.3547\mathbb{I}_2\mathbb{I}_3, i = 1, 2, ..., 8$$

The terms are defined as

where μ_i is the expected number of students in the ith row of the table above. $\mathbb{I}_{1i}$ is the indicator that the students that belong to the ith row of the table like playing vedio games.

$\mathbb{I}_{2i}$ is the indicator that the students that belong to the ith row of the table are male $\mathbb{I}_{3i}$ is the indicator that the students that belong to the ith row of the table has non-A+ expected grades.

(b)

i. Deviance for Model 3.1 is 4.66×10^{-15}, the deviance is almost 0 because this is the saturated model. Deviance for Model 3.2 is 0.9630. Compare to logistic models in question 2, which have redidual deviance 488.41 and 489.37. The deviances of Possion regression models are much smaller than that of logistic regression models.
   For LRT, The test statistic in Question 3 is $0.963 - 4.66 \times 10^{-15} = 0.963$, The test statistic in Question 2 is 0.96. The test statistics are the same, the p-value is 0.3053 for both logistic models and possion models. We cannot reject the null hypothesis. Therefore, we have some evidence that the model without the interaction term is better.

ii. In both models, the interaction term has test statistic -0.743 (Model 2.2) and 0.743 (Model 2.3). Because normal distribution is symmetric, they give the same p-value, which is 0.3234. We cannot reject the null hypothesis. Therefore, we have some evidence that we should not include the interaction term.

iii. By the test results (p-value=0.3053 in LRT, p-value=0.3234 in Wald test) from both Possion regression models and logistic regression models, we have some evidence that there is no three-way interaction among whether like playing video games, expected grade, and gender. Also, we notice that the Poisson model and the logistic model gives the same test statistics. Therefore, the two models are equivalent.

Appendix

```r
# import and encode data
student <- read.csv('a3data.csv')
like <- NULL
for (i in 1:length(student$Like)){
  like[i] = as.integer(student$Like[i] == 'Somewhat' || student$Like[i] == 'Very much')
}
like <- as.factor(like)
grade <- as.integer(student$Grade == 'A+ ')
sex <- as.factor(student$sex)

# Q1 (a)
count <- c(0, 0, 0, 0)
for (i in 1:length(like)){
  if (like[i] == 1 & sex[i] == "Male"){
    count[1] = count[1] + 1
  } else if (like[i] == 1 & sex[i] == "Female") {
    count[2] = count[2] + 1
  } else if (like[i] == 0 & sex[i] == "Male") {
    count[3] = count[3] + 1
  } else if (like[i] == 0 & sex[i] == "Female"){
    count[4] = count[4] + 1
  }
}
table <- matrix(count, nrow=2, byrow=T)
dimnames(table) <- list(c("like", "not like"), c("Male", "Female"));
names(dimnames(table)) <- c("Like games", "sex")
table
```

```
##           sex
## Like games Male Female
##   like      122    114
##   not like   29    134
```

```r
fisher.test(table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 2.515e-12
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  3.008412 8.248768
## sample estimates:
## odds ratio
##   4.924757
```

```r
prop.test(table, correct=F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  table
```

```
## X-squared = 47.112, df = 1, p-value = 6.704e-12
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2523654 0.4257047
## sample estimates:
##    prop 1    prop 2
## 0.5169492 0.1779141
```

```
# Q1 (b)
yes_m <- matrix(c(31,11,26,32), nrow=2,byrow=TRUE)
dimnames(yes_m) <- list(c('female', 'male'), c('no', 'yes'))
names(dimnames(yes_m)) <- c('sex', 'like games')
yes_m
```

```
##         like games
## sex       no yes
##   female 31  11
##   male   26  32
```

```
no_m <- matrix(c(103,18,88,90), nrow=2,byrow=TRUE)
dimnames(no_m) <- list(c('female', 'male'), c('no', 'yes'))
names(dimnames(no_m)) <- c('sex', 'like games')
no_m
```

```
##         like games
## sex        no yes
##   female 103  18
##   male    88  90
```

```
prop.test(yes_m, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  yes_m
## X-squared = 8.3481, df = 1, p-value = 0.003861
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.1052614 0.4743774
## sample estimates:
##    prop 1    prop 2
## 0.7380952 0.4482759
```

```
prop.test(no_m, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  no_m
## X-squared = 39.757, df = 1, p-value = 2.877e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.2598275 0.4538878
## sample estimates:
##    prop 1    prop 2
```

```
## 0.8512397 0.4943820
```

```r
# Q2
grade <- as.factor(grade)

# Model 2.1
fiti <- glm(like~sex*grade, family=binomial)
summary(fiti)
```

```
##
## Call:
## glm(formula = like ~ sex * grade, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8930  -1.1114   0.6039   1.2449   1.2530
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.1574     0.1452  -1.084    0.278
## sexMale          1.7668     0.2962   5.965 2.45e-09 ***
## grade1          -0.0185     0.3030  -0.061    0.951
## sexMale:grade1  -0.5231     0.5297  -0.987    0.323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 539.70  on 398  degrees of freedom
## Residual deviance: 488.41  on 395  degrees of freedom
## AIC: 496.41
##
## Number of Fisher Scoring iterations: 4
```

```r
# Model 2.2
fita <- glm(like~sex+grade, family=binomial)
summary(fita)
```

```
##
## Call:
## glm(formula = like ~ sex + grade, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8412  -1.1273   0.6369   1.2283   1.3098
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1189     0.1397  -0.851    0.395
## sexMale       1.6111     0.2438   6.610 3.85e-11 ***
## grade1       -0.1871     0.2519  -0.743    0.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 539.70  on 398  degrees of freedom
## Residual deviance: 489.37  on 396  degrees of freedom
## AIC: 495.37
##
## Number of Fisher Scoring iterations: 4
```

```r
# p-value of the LRT
1 - pchisq(1, fita$deviance - fiti$deviance)
```

```
## [1] 0.3053195
```

```r
# Question 3
# import data
count <- c(31, 103, 11, 18, 26, 88, 32, 90)
like <- as.factor(c("no", "no", "no", "no", "yes", "yes", "yes", "yes"))
sex <- as.factor(c("female", "female", "male", "male", "female", "female", "male", "male"))
grade <- as.factor(c("A+", "not A+", "A+", "not A+", "A+", "not A+", "A+", "not A+"))

# Model 3.1
fitpi = glm(count~like*sex*grade, family=poisson)
summary(fitpi)
```

```
##
## Call:
## glm(formula = count ~ like * sex * grade, family = poisson)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    3.4340     0.1796  19.120  < 2e-16 ***
## likeyes                       -0.1759     0.2659  -0.661  0.50835
## sexmale                       -1.0361     0.3509  -2.952  0.00315 **
## gradenot A+                    1.2007     0.2049   5.861 4.59e-09 ***
## likeyes:sexmale                1.2437     0.4392   2.832  0.00463 **
## likeyes:gradenot A+            0.0185     0.3030   0.061  0.95131
## sexmale:gradenot A+           -0.7083     0.4341  -1.632  0.10276
## likeyes:sexmale:gradenot A+    0.5231     0.5297   0.987  0.32341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.9388e+02  on 7  degrees of freedom
## Residual deviance: 4.6629e-15  on 0  degrees of freedom
## AIC: 59.808
##
## Number of Fisher Scoring iterations: 3
```

```r
# Model 3.2
fitpa = glm(count~like*sex+like*grade+sex*grade, family=poisson)
summary(fitpa)
```

```
##
## Call:
```

```
## glm(formula = count ~ like * sex + like * grade + sex * grade,
##     family = poisson)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
## -0.3220   0.1812   0.5849  -0.4170   0.3672  -0.1935  -0.3171   0.1940
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           3.4913     0.1652  21.131  < 2e-16 ***
## likeyes              -0.3061     0.2329  -1.314    0.189
## sexmale              -1.2751     0.2704  -4.715 2.42e-06 ***
## gradenot A+           1.1256     0.1865   6.034 1.60e-09 ***
## likeyes:sexmale       1.6111     0.2438   6.610 3.85e-11 ***
## likeyes:gradenot A+   0.1871     0.2519   0.743    0.458
## sexmale:gradenot A+  -0.3547     0.2523  -1.406    0.160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 193.87673  on 7  degrees of freedom
## Residual deviance:   0.96302  on 1  degrees of freedom
## AIC: 58.771
##
## Number of Fisher Scoring iterations: 4
```