STA303_A3

Part one

1. What was your selected field of interest?

The field I select is to determine if there is an association between longer breast-feeding duration and dental caries in healthy urban children.

2. Write a proper reference for the article, including the author(s), title, journal, year of publication, volume and page indices.

Authors: Peter D. Wong, MBBS, PhD; Catherine S. Birken, MD, MSc; Patricia C. Parkin, MD, MSc; Isvarya Venu, MSc; Yang Chen, MSc; Robert J. Schroth, DMD, PhD; Jonathon L. Maguire, MD, MSc

Title: Total Breast-Feeding Duration and Dental Caries in Healthy Urban Children

Journal:Academic Pediatrics

Year of publication: 2017

Volume and page indices:Volume 17, Issue 3, April 2017, Pages 310-315

3. Which UofT department was the UofT author affiliated with?

Division of Paediatric Medicine, Department of Paediatrics, Faculty of Medicine

4. Provide a link to the article or a soft copy of the article.

https://doi.org/10.1016/j.acap.2016.10.021

5. Which statistical software was used for the data analysis?

Unknown

6. Was the data derived from an observational study or experiment?

Observational study

7. Did the article present summary statistics, tables and/or plots? Explain.

The article presents three tables. First table is population characteristics, including information like age, sex, maternal age, birth weight, maternal ethnicity, self-reported family income, single parent, maternal employment, household smoke exposure, bedtime bottle use, only child, sugar- sweetened beverage consumption, and snacking of sweets, candy, chips, or fried foods.

The second table illustrates the association between the total Breast-Feeding Duration and Dental Caries. It mainly reveals that relative to total breast-feeding dura- tion 0 to 5 months, the odds of caries with total breast- feeding duration 6 to 11 months was 1.17 (95% CI 0.73– 1.88), 12 to 23 months was 1.52 (95% CI 0.97–2.38), and >24 months was 2.75 (95% CI 1.61–4.72, P < .001)

The third table illustrates the probability and corresponding CI of caries with total breast-feeding duration.

8. Did the article present test statistics, their distributions under H0, p-values and/or confidence intervals? Explain.

The article does not present test statistics and distributions under H0. However, it present the p-value and confidence intervals when it presents the odds of caries with different factors. And the article also presents confidence intervals about predicted probability of caries with total breast-feeding duration. Additionally, when the authors use Likelyhood ratio test to compare main effect model and interation model, the P-value > .30 which is sufficiently high to exclude the interation term.

9. To how many decimal places were values reported? Explain.

Two.

10. Identify at least one statistical method used to analyze the data.

The authors use logistic regression to predict the probability of caries with 12 months', 18 months', 24 months', 36 months' total breast-feeding duration. The result is that the predicted probability of caries with total breast-feeding duration of 12 months was 0.07 (95% CI 0.05–0.10), 18 months was 0.08 (95% CI 0.06–0.12), 24 months was 0.11 (95% CI 0.07–0.15), and 36 months was 0.16 (95% CI 0.10–0.25).

And Likelyhood ratio test is used to compare models between main effects model and hypothesized interation model(sex and self-reported family income). The result is interation term is not inclued into the final model.

Part two

Solution

1. Analysis comparing proportions and using contingency tables:

(a) (10 marks) Construct a 2 × 2 table of sex by like. Is there evidence that sex is independent of a student's preference for playing video games? Quote 2 different p-values to support your answer. If there is evidence of association between the variables, explain in practical terms, with illustrative numbers, the nature of the association.

2 x 2 table of sex by like:

```
##         Like           Ruijie Sun 6046
## Sex      yes no
##   male    44  8
##   female  26 12
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  cvd
## X-squared = 3.3314, df = 1, p-value = 0.06797
```

```
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01542393  0.33931057
## sample estimates:
##    prop 1    prop 2
## 0.8461538 0.6842105
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cvd
## p-value = 0.07824
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8195672 8.1070182
## sample estimates:
## odds ratio
##   2.511179
```

Thus, sex is not independent of a student's preference for playing video games since p-values from difference of proportion test and Fisher's Exact Test are both < 0.1. From the output of difference proportion test, we can know the game-preference proportion among male is 0.8461538 which is larger than the game-preference proportion among female 0.6842105. So, compared to female students, male students are more likely to play games. Meanwhile, from the output of Fisher's Exact Test, the odds ratio is 2.511179 which also suppports our previous conclution.

(b) (15 marks) Examine the sex and like relationship separately for each grade type expected. Is there evidence that the association between sex and student's preference for playing video games changes with grade expected? Quote relevant p-values to support your answers.

For grade A type:

```
##         Like           Ruijie Sun 6046
## Sex      yes no
##   male    21  1
##   female   4  5
```

Fisher's Exact Test:

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cvd2
## p-value = 0.003879
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##     1.877541 1284.874786
## sample estimates:
## odds ratio
##   22.44903
```

p-value < 0.1, so there is evidence that sex is not independent of a student's preference for playing video games among grade A students.

For grade nA type:

```
##         Like           Ruijie Sun 6046
## Sex      yes no
```

```
##   male    23  7
##   female  22  7
```

Fisher's Exact Test:

```
##
##  Fisher's Exact Test for Count Data
##
## data:  cvd3
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.2637887 4.1391671
## sample estimates:
## odds ratio
##   1.044688
```

p-value > 0.1, so sex is independent of a student's preference for playing video games among grade nA students.

Conclution: Based on two Fisher's exact test results above, there is evidence that the association between sex and student's preference for playing video games changes with grade expected.

2. Analysis using Logistic Regression:

(a) (20 marks) Write the models being fit; clearly define all terms. Which of the two model should you use? Give the results of two tests that support your choice of logistic regression model. Explain clearly what is being tested for each test.

- **Model 2.1 be the one to include interaction between sex and grade, and**

```
##
## Call:
## glm(formula = like ~ sex * grade, family = binomial, data = data_a3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4864   0.3050   0.7290   0.7433   1.2735
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.2231     0.6708  -0.333  0.73940
## sexmale           3.2677     1.2237   2.670  0.00758 **
## gradenA           1.3683     0.7989   1.713  0.08679 .
## sexmale:gradenA  -3.2232     1.3682  -2.356  0.01848 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 95.347  on 89  degrees of freedom
## Residual deviance: 85.152  on 86  degrees of freedom
## AIC: 93.152
##
## Number of Fisher Scoring iterations: 5
```

The full model: $log(\frac{\pi_i}{1-\pi_i}) = -0.2231 + 3.2677 I_{male,i} + 1.3683 I_{nA,i} - 3.2232 I_{male,i} I_{nA,i}$, where π_i = P("yes"), $I_{male,i} = 1$ if i-th is male otherwise 0 and $I_{nA,i} = 1$ if i-th person's expected grade is nA othwise 0.

- **Model 2.2 be the one without interaction.**

```
##
## Call:
## glm(formula = like ~ sex + grade, family = binomial, data = data_a3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9533   0.5668   0.5861   0.8512   0.8774
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.8288     0.5586   1.484   0.1379
## sexmale       0.9183     0.5291   1.736   0.0826 .
## gradenA      -0.0727     0.5679  -0.128   0.8981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 95.347  on 89  degrees of freedom
## Residual deviance: 92.031  on 87  degrees of freedom
## AIC: 98.031
##
## Number of Fisher Scoring iterations: 4
```

The reduced model: $log(\frac{\pi_i}{1-\pi_i}) = 0.8288 + 0.9183 I_{male,i} - 0.0727 I_{nA,i}$, where π_i = P("yes"), $I_{male,i} = 1$ if i-th is male otherwise 0 and $I_{nA,i} = 1$ if i-th person's expected grade is nA othwise 0.

Compare full model and reduced model:

I use the full model.

Wald tests:

H_0:in full model $\beta_3 = 0$

H_a:in full model $\beta_3 \neq 0$

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.4, df = 2, P(> X2) = 0.024
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.018
```

Since P-value = $0.018 < 0.1$, so there is evidence to reject H_0. So we choose full model.

GOF: H_0: Reduced model H_a: Full model

D0,Da:

$G^2 = D_0 - Da = 92.03091 - 85.15215 = 6.87876 \sim \chi^2_1$

```
## [1] 0.008722606
```

$P(\chi^2_1 > 6.87876) = 0.008722606 < 0.1$ So there is evidence to reject H0. We choose full model.

(b) (10 marks) Give practical implications of the model selected in part (a). What do you conclude? Does it agree with your answer to question 1(b)?

```
## [1] 21.00163
```

```
## [1] 0.8000348
```

```
## [1] 3.286095
```

```
## [1] 3.14307
```

```
## [1] 26.2509
```

```
## [1] 1.045505
```

The odds of game preference for a male-gradeA student is: 21.00163

The odds of game preference for a female-gradeA student is: 0.8000348

The odds of game preference for a male-gradenA student is: 3.286095

The odds of game preference for a female-gradenA student is: 3.14307

Thus, compare the odds of a game preference for a male-gradeA student to a female-gradeA student:

$21.00163/0.8000348 = 26.2509$

compare the odds of a game preference for a male-gradenA student to a a female-gradenA student:

$3.286095/3.14307 = 1.045505$ 1

Thus, the odds of a game preference for a male-gradeA student are above 26 times to a female-gradeA student. And, the odds of a game preference for a male-gradenA student is equal likely to a female-gradenA student. This conclution agrees with my conclution in 1(b).

3.Analysis using Poisson Regression:

(a) (10 marks) Model the counts as Poisson variables and fit two models:

- **Model 3.1 with explanatory variables sex, grade and like, the three two-way terms and the three-way interaction, and**

```
## 
## 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)               1.6094     0.4472   3.599  0.00032 ***
## likeyes                  -0.2231     0.6708  -0.333  0.73940    
## sexmale                  -1.6094     1.0954  -1.469  0.14177    
## gradenA                   0.3365     0.5855   0.575  0.56554    
```

```
## likeyes:sexmale           3.2677     1.2238   2.670  0.00758 **
## likeyes:gradenA           1.3683     0.7989   1.713  0.08679 .
## sexmale:gradenA           1.6094     1.2189   1.320  0.18670
## likeyes:sexmale:gradenA  -3.2232     1.3683  -2.356  0.01849 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5.4112e+01  on 7  degrees of freedom
## Residual deviance: 3.1086e-15  on 0  degrees of freedom
## AIC: 47.169
##
## Number of Fisher Scoring iterations: 3
```

Full model: $log(\mu_{ijk}) = 1.6094 - 0.2231I_{likeyes} - 1.6094I_{sexmale} + 0.3365I_{gradenA} + 3.2677I_{likeyes}I_{sexmale} + 1.3683I_{likeyes}I_{gradenA} + 1.6094I_{sexmale}I_{gradenA} - 3.2232I_{likeyes}I_{sexmale}I_{gradenA}$ where μ_{ijk} is expected # of count in each cell, $I_{likeyes}$ is 1 if level of like is yes otherwise 0, $I_{sexmale}$ is 1 if level of sex is male otherwise 0, and $I_{gradenA}$ is 1 if level of grade is nA otherwise 0.

- **Model 3.2 - Model 3.1 with the three-way interaction term removed**

```
##
## Call:
## glm(formula = Count ~ like + sex + grade + like * sex + like *
##     grade + sex * grade, family = poisson)
##
## Deviance Residuals:
##       1        2        3        4        5        6        7        8
##  1.2260  -0.7780  -1.4709   0.9711  -0.9698   0.5005   0.5132  -0.4576
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.0061     0.5122   1.964   0.0495 *
## likeyes           0.8288     0.5586   1.484   0.1379
## sexmale           0.1771     0.5684   0.312   0.7553
## gradenA           1.2201     0.5480   2.227   0.0260 *
## likeyes:sexmale   0.9183     0.5291   1.736   0.0826 .
## likeyes:gradenA  -0.0727     0.5679  -0.128   0.8981
## sexmale:gradenA  -0.8484     0.4819  -1.761   0.0783 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 54.1125  on 7  degrees of freedom
## Residual deviance:  6.8788  on 1  degrees of freedom
## AIC: 52.048
##
## Number of Fisher Scoring iterations: 5
```

Reduced model: $log(\mu_{ijk}) = 1.0061 + 0.8288I_{likeyes} + 0.1771I_{sexmale} + 1.2201I_{gradenA} + 0.9183I_{likeyes}I_{sexmale} - 0.0727I_{likeyes}I_{gradenA} - 0.8484I_{sexmale}I_{gradenA}$ where μ_{ijk} is expected # of count in each cell, $I_{likeyes}$ is 1 if level of like is yes otherwise 0, $I_{sexmale}$ is 1 if level of sex is male otherwise 0, and $I_{gradenA}$ is 1 if level of grade is nA otherwise 0.

(b) (20 marks) Describe how the results from the Poisson regression models compare to the results in part 2 under Logistic regression modelling, in terms of:

i. (5 marks) Deviance

For model 2.1 and 2.2: H_0: Reduced model H_a: Full model

D0,Da:

```
## [1] 92.03091
```

```
## [1] 85.15215
```

$G^2 = D_0 - Da = 92.03091 - 85.15215 = 6.87876 \sim \chi^2_1$

$P(\chi^2_1 > 6.87876) = 0.008722606 < 0.1$

So there is evidence to reject H0. We choose full model.

For model 3.1 and 3.2:

H_0: Reduced model H_a: Full model

D0, Da:

```
## [1] 6.878764
```

```
## [1] 3.108622e-15
```

$G^2 = D_0 - Da = 6.878764 - 0 = 6.878764 \sim \chi^2_1$

$P(\chi^2_1 > 6.87876) = 0.008722586 < 0.1$

So there is evidence to reject H0. We choose full model.

ii. (5 marks) Wald tests

For model 2.1 and 2.2:

H_0:in full model $\beta_3 = 0$

H_a:in full model $\beta_3 \neq 0$

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 7.4, df = 2, P(> X2) = 0.024
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.018
```

Since P-value $= 0.018 < 0.1$, so there is evidence to reject H_0. So we choose full model.

For model 3.1 and 3.2

Wald tests:

H_0:in full model $\beta_7 = 0$

H_a:in full model $\beta_7 \neq 0$

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 33.0, df = 6, P(> X2) = 1.1e-05
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 5.5, df = 1, P(> X2) = 0.018
```

Since P-value = $0.018 < 0.1$, so there is evidence to reject H_0. So we choose full model.

iii. (10 marks) Interpretation

For model 2.1&2.2 and model 3.1&3.2, based on wald test and GOF, we both choose full model.

Poisson regression and logistic regression share a lot of similarity. For example, both can use Wald test and GOF to select model. However, there also exists difference. For example, logistic regresson focuses on the probability for a case to be good or bad (0 or 1) and poisson regression focuses on how much the cases are going to "do" (counts). Besides, in poisson regression, row and coloum variables are treated symmetrically. In contrast, logistic models describe how a categorical response depends on the explanatory variable.

Appendix

1. Analysis comparing proportions and using contingency tables:

(a) (10 marks) Construct a 2 × 2 table of sex by like. Is there evidence that sex is independent of a student's preference for playing video games? Quote 2 different p-values to support your answer. If there is evidence of association between the variables, explain in practical terms, with illustrative numbers, the nature of the association.

2 x 2 table of sex by like:

```r
cvd<-matrix(c(44,8,26,12),nrow=2,byrow=TRUE)
dimnames(cvd)<-list(c("male","female"),c("yes","no"))
names(dimnames(cvd))<-c("Sex","Like           Ruijie Sun 6046")
cvd
```

```r
prop.test(cvd,correct = FALSE)
```

```r
fisher.test(cvd)
```

(b) (15 marks) Examine the sex and like relationship separately for each grade type expected. Is there evidence that the association between sex and student's preference for playing video games changes with grade expected? Quote relevant p-values to support your answers.

For grade A type:

```r
cvd2<-matrix(c(21,1,4,5),nrow=2,byrow=TRUE)
dimnames(cvd2)<-list(c("male","female"),c("yes","no"))
names(dimnames(cvd2))<-c("Sex","Like           Ruijie Sun 6046")
cvd2
```

Fisher's Exact Test:

```
fisher.test(cvd2)
```

For grade nA type:

```
cvd3<-matrix(c(23,7,22,7),nrow=2,byrow=TRUE)
dimnames(cvd3)<-list(c("male","female"),c("yes","no"))
names(dimnames(cvd3))<-c("Sex","Like           Ruijie Sun 6046")
cvd3
```

Fisher's Exact Test:

```
fisher.test(cvd3)
```

2. Analysis using Logistic Regression:

(a) (20 marks) Write the models being fit; clearly define all terms. Which of the two model should you use? Give the results of two tests that support your choice of logistic regression model. Explain clearly what is being tested for each test.

```
data_a3 <- read.csv("video.csv", header=T)
attach(data_a3)
head(data_a3)
```

```
str(data_a3)
```

- **Model 2.1 be the one to include interaction between sex and grade, and**

```
mod1 <-glm(like~sex*grade,family=binomial,data=data_a3)
summary(mod1)
```

- **Model 2.2 be the one without interaction.**

```
mod2 <-glm(like~sex + grade,family=binomial,data=data_a3)
summary(mod2)
```

Wald tests:

```
wald.test(Sigma=vcov(mod1),b=coef(mod1),Term=2:3)
```

```
wald.test(Sigma=vcov(mod1),b=coef(mod1),Term=4)
```

GOF:

```
Da<- deviance(mod1)
D_0<- deviance(mod2)
D0
Da
```

$G^2 = D_0 - Da = 92.03091 - 85.15215 = 6.87876 \sim \chi^2_1$

```
1 - pchisq(6.87876,1)
```

(b) (10 marks) Give practical implications of the model selected in part (a). What do you conclude? Does it agree with your answer to question 1(b)?

```r
exp(-0.2231+3.2677)
exp(-0.2231)
exp(-0.2231+3.2677+1.3683-3.2232)
exp(-0.2231+1.3683)
21.00163/0.8000348
3.286095 / 3.14307
```

3.Analysis using Poisson Regression:

(a) (10 marks) Model the counts as Poisson variables and fit two models:

- **Model 3.1 with explanatory variables sex, grade and like, the three two-way terms and the three-way interaction, and**

```r
Count=c(5,7,1,7,4,22,21,23)
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","nA","A","nA","A","nA","A","nA"))
```

```r
fullmod =glm(Count~like*sex*grade,family=poisson)
summary(fullmod)
```

- **Model 3.2 - Model 3.1 with the three-way interaction term removed**

```r
redmod =glm(Count~like + sex + grade + like*sex + like*grade + sex*grade,family=poisson)
summary(redmod)
```

(b) (20 marks) Describe how the results from the Poisson regression models compare to the results in part 2 under Logistic regression modelling, in terms of:

i. (5 marks) Deviance

For model 2.1 and 2.2:

```r
Da<- deviance(mod1)
D_0<- deviance(mod2)
D0
Da
```

```r
1 - pchisq(6.87876,1)
```

For model 3.1 and 3.2:

```r
Da<- deviance(fullmod)
D_0<- deviance(redmod)
D0
Da
```

```r
1 - pchisq(6.878764,1)
```

ii. (5 marks) Wald tests

For model 2.1 and 2.2:

```
wald.test(Sigma=vcov(mod1),b=coef(mod1),Term=2:3)
```

```
wald.test(Sigma=vcov(mod1),b=coef(mod1),Term=4)
```

For model 3.1 and 3.2

Wald tests:

```
wald.test(Sigma=vcov(fullmod),b=coef(fullmod),Term=2:7)
```

```
wald.test(Sigma=vcov(fullmod),b=coef(fullmod),Term=8)
```

iii. (10 marks) Interpretation