University of Toronto

Contigency table, Logistic Regression and Poisson Regression

YiFei Gu

1. **Analysis comparing proportions and using contingency tables:**

   a. The 2 by 2 table of `sex` by `like` is shown below, as `t`:

```
        like
sex        0   1
  Female 134 114
  Male    29 122
```

   There are evidence that in fact, sex is not independent of a student's preference for playing video games. The p-value from both `prop.test(t, correct = FALSE)` and `fisher.test(t)` gives significant evidence to reject the hypothesis that sex is independent of a student's preference for playing video games. Equals `6.704e-12` and `2.515e-12` respectively. In practical terms, male tends to like playing video games than female at a statistical significant level, given the odd ratio from `fisher.test(t)`, the odds of preference of playing video games for a male person is `4.9247` times that for a female person.

   b. Construct two tables with different expected grade for male and female respectively,

```
> expect_Aplus

          0  1
  Female 31 26
  Male   11 32
```

   For students expect grade of A+, from `fisher.test(expect_Aplus)` and `chisq.test(expect_Aplus, correct = FALSE)`, both of p-value significant being `0.004462, 0.003861`. Indicating that there are significant evidence that sex is not independent of preference. With an odd ratio of 3.4237, indicating that for students expecting grade of A+, the odds of preference of playing video games for a male person is `3.4237` times that for a female person.

```
> expectOther

           0   1
  Female 103  88
  Male    18  90
```

   For students expect grade others, from `fisher.test(expect_Aplus)` and `chisq.test(expect_Aplus, correct = FALSE)`, both of p-value significant being `1.048e-10, 2.877e-10`. Indicating that there are significant evidence that sex is not independent of preference. With an odd ratio of 5.8175 , indicating that for students expecting other grades, the odds of preference of playing video games for a male person is `5.8175` times that for a female person.

   By calculating the proportion of preference of video games for male and female with expectation of grade A+ is 0.7442 and 0.4561 respectively, and with expectation of other grades, the proportion for male and female likes video games is 0.8333 and 0.4607 respectively. There is an increase of proportion of liking video games for both male and female when do not expect a grade of A+, although not really obvious for female.

2. **Analysis using Logistic Regression:**

a. For model2.1, $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -0.1574 + 1.7668\mathbb{I}_{\texttt{sex\_f}} - 0.0185\mathbb{I}_{\texttt{grade\_f}} - 0.5231\mathbb{I}_{\texttt{sex\_f}} * \mathbb{I}_{\texttt{grade\_f}}$
For model 2.2 without interaction term: $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -0.1189 + 1.6111\mathbb{I}_{\texttt{sex\_f}} - 0.1871\mathbb{I}_{\texttt{grade\_f}}$
where π is is the estimated probability that likes playing video games
$\mathbb{I}_{\texttt{sex\_f}}$:1 if being male and 0 otherwise.
$\mathbb{I}_{\texttt{grade\_f}}$: 1 if the expecting grade of A+ and 0 otherwise.

Conduct LRT and Wald test to see if the additive model is better than interaciton model.
For LRT: H_0: additive model is adequate, and H_1: interaction model is better.
Observe the test statistic $G^2 = 489.37 - 488.41 = 0.96 \sim \chi^2_1$, with $P(\chi^2_1 > 0.96) > 0.10$ **non significant, with p-value in Appendix.**

Wald test: with $H_0 : \gamma^3 = 0$ and $H_1 : \gamma^3 \neq 0$, observe the test statistic $Z = \frac{-0.5231}{0.5297} = -0.9887, Z^2 \approx 0.9775 \sim \chi^2_1$ with $P(Z^2 > 0.9775) > 0.10$, **non significant, with p-value in Appendix**.

Since the p-value is large, we **fail to reject the null hypothesis** and conclude that the data are consistent with the coefficient of the interaction term being 0. Therefore, the interaction does not contributes in a statistically significant way to the explanation of the odds of likes playing video games. **In conclusion, mod2.2 without interaction term should be used.**

b. From estimated coefficient of $\mathbb{I}_{\texttt{sexMale}}, \beta_2 = 1.6111$, the odds ratio of which compares to to female, is $\exp(1.6111) \approx 5.0083$, which is similar to the conclusion drawn from `fisher.test()` in part a.

From estimated coefficient of $\mathbb{I}_{\texttt{gradeA+}}, \beta_3 = -0.1871$, the odds ratio of which compares to whom that do not expect a grade of A+, is $\exp(-0.1871) \approx 0.8294$, which is agrees with the conclusion drawn from part 1 that the lower standards of expectation of grades does increase the proportion of liking video games.

3. **Analysis using Poisson Regression:**

a. First, construct the table, then construct `mod3.1` and `mod3.2` as shown in Appendix. The models are:
mod3.1:
$\log(\mathbb{E}(y_{ijk})) = 4.6347 - 1.2007\mathbb{I}_{\texttt{grade\_p}} - 1.7444\mathbb{I}_{\texttt{sex\_p}} - 0.1574\mathbb{I}_{\texttt{like\_p}} + 0.7083\mathbb{I}_{\texttt{grade\_p}} * \mathbb{I}_{\texttt{sex\_p}} - 0.01850\mathbb{I}_{\texttt{grade\_p}} * \mathbb{I}_{\texttt{like\_p}} + 1.7668\mathbb{I}_{\texttt{sex\_p}} * \mathbb{I}_{\texttt{like\_p}} - 0.5231\mathbb{I}_{\texttt{grade\_p}} * \mathbb{I}_{\texttt{sex\_p}} * \mathbb{I}_{\texttt{like\_p}}$
mod3.2:
$\log(\mathbb{E}(y_{ijk})) = 4.6168 - 1.1256\mathbb{I}_{\texttt{grade\_p}} - 1.6298\mathbb{I}_{\texttt{sex\_p}} - 0.1189\mathbb{I}_{\texttt{like\_p}} + 0.3547\mathbb{I}_{\texttt{grade\_p}} * \mathbb{I}_{\texttt{sex\_p}} - 0.1871\mathbb{I}_{\texttt{grade\_p}} * \mathbb{I}_{\texttt{like\_p}} + 1.6111\mathbb{I}_{\texttt{sex\_p}} * \mathbb{I}_{\texttt{like\_p}}$
where $\mathbb{E}(y_{ijk})$ is the counts classified by i, j, kth status.
$\mathbb{I}_{\texttt{grade\_p}}$: 1 if expecting grade of A+ and 0 otherwise.
$\mathbb{I}_{\texttt{sex\_p}}$: 1 if sex is male and 0 otherwise.
$\mathbb{I}_{\texttt{like\_p}}$: 1 if ith likes video games and 0 otherwise.

b. **Comparison of results by:**
- Deviance: The deviance between model3.1 and model3.2 is the same as the deviance between model2.1 and model2.2, which both result around 0.96
-Wald tests: the Wald tests matches between:

i. interaction of sex and like in model3.2 and sex effect in model 2.1, with same p-value of `2.45e-09`
ii. interaction between grade, sex and like in model 3.1 and grade and sex in model 2.1, with p-value of `0.3230`

iii. interaction between grade and like in model 3.1 and grade effect in model 2.1, with p-value of `0.9510`

- interpretation: Logistic model predicts the log-odd of like with $\log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i})$ while the Poisson model predicts log of mean value of the counts, $\log(\mathbb{E}(y_{ijk}))$. Meanwhile the response variable is clear in Logistic model, a model is built for counts in the Poisson model.

Appendix

Q1:

```
> t = table(sex, like)
> t
        like
sex        0   1
  Female 134 114
  Male    29 122
> fisher.test(t)

        Fisher's Exact Test for Count Data

data:  t
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

> chisq.test(t, correct = FALSE)

        Pearson's Chi-squared test

data:  t
X-squared = 47.112, df = 1, p-value = 6.704e-12

> expect_Aplus <- table(sex[grade==1], like[grade==1])
> expect_Aplus

          0  1
  Female 31 26
  Male   11 32
> expectOther = table(sex[grade==0], like[grade==0])
> expectOther

           0   1
  Female 103  88
  Male    18  90

> fisher.test(expect_Aplus)
```

```
        Fisher's Exact Test for Count Data

data:  expect_Aplus
p-value = 0.004462
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.36147 9.09891
sample estimates:
odds ratio
  3.423749

> fisher.test(expectOther)

        Fisher's Exact Test for Count Data

data:  expectOther
p-value = 1.048e-10
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  3.185573 11.085730
sample estimates:
odds ratio
  5.817501

> chisq.test(expect_Aplus,  correct = FALSE)

        Pearson's Chi-squared test

data:  expect_Aplus
X-squared = 8.3481, df = 1, p-value = 0.003861

> chisq.test(expectOther,  correct = FALSE)

        Pearson's Chi-squared test

data:  expectOther
X-squared = 39.757, df = 1, p-value = 2.877e-10
```

Q2:

```
> grade_f = as.factor(grade)
> sex_f = as.factor(sex)
> mod2.1 = glm(like ~ grade_f * sex_f, family = binomial)
> summary(mod2.1)

Call:
glm(formula = like ~ grade_f * sex_f, 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
grade_f1            -0.0185     0.3030  -0.061    0.951
sex_fMale            1.7668     0.2962   5.965 2.45e-09 ***
grade_f1:sex_fMale  -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

> 1 - pchisq(0.96, 1) # p-value
[1] 0.3271869

> mod2.2 = glm(like ~ grade_f + sex_f, family =  binomial)
> summary(mod2.2)

Call:
glm(formula = like ~ grade_f + sex_f, 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
grade_f1     -0.1871     0.2519  -0.743    0.458
sex_fMale     1.6111     0.2438   6.610 3.85e-11 ***
---
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
> 1-pchisq(0.9775,1)
[1] 0.3228168

> n = 399
> count = rep(1,n)
> new = aggregate(count ~ grade + sex + like,data=a3data, FUN = sum)
> new
```

```
  grade    sex like count
1     0 Female    0   103
2     1 Female    0    31
3     0   Male    0    18
4     1   Male    0    11
5     0 Female    1    88
6     1 Female    1    26
7     0   Male    1    90
8     1   Male    1    32
> grade_p = as.factor(new$grade)
> sex_p = as.factor(new$sex)
> like_p = as.factor(new$like)
> mod3.1 = glm(counts ~ grade_p*sex_p*like_p, family = poisson)
> summary(mod3.1)

Call:
glm(formula = counts ~ grade_p * sex_p * like_p, family = poisson)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.63473    0.09853  47.037  < 2e-16 ***
grade_p1                   -1.20074    0.20486  -5.861 4.59e-09 ***
sex_pMale                  -1.74436    0.25547  -6.828 8.61e-12 ***
like_p1                    -0.15739    0.14516  -1.084    0.278
grade_p1:sex_pMale          0.70827    0.43409   1.632    0.103
grade_p1:like_p1           -0.01850    0.30297  -0.061    0.951
sex_pMale:like_p1           1.76683    0.29621   5.965 2.45e-09 ***
grade_p1:sex_pMale:like_p1 -0.52310    0.52973  -0.987    0.323
---
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: -1.4655e-14  on 0  degrees of freedom
AIC: 59.808

Number of Fisher Scoring iterations: 3

> mod3.2 = glm(counts ~ grade_p + sex_p + like_p + grade_p:sex_p + grade_p:like_p + like_p:sex_p, famil
> summary(mod3.2)

Call:
glm(formula = counts ~ grade_p + sex_p + like_p + grade_p:sex_p +
    grade_p:like_p + like_p:sex_p, family = poisson)

Deviance Residuals:
      1        2        3        4        5        6        7        8
```

```
 0.1812  -0.3220  -0.4170   0.5849  -0.1935   0.3672   0.1940  -0.3171

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.61683    0.09773  47.241  < 2e-16 ***
grade_p1           -1.12555    0.18653  -6.034 1.60e-09 ***
sex_pMale          -1.62975    0.21888  -7.446 9.64e-14 ***
like_p1            -0.11893    0.13968  -0.851    0.395
grade_p1:sex_pMale  0.35467    0.25229   1.406    0.160
grade_p1:like_p1   -0.18713    0.25189  -0.743    0.458
sex_pMale:like_p1   1.61115    0.24375   6.610 3.85e-11 ***
---
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
```