STAT 426 Assignment 5

Due Tuesday, September 28, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1 (8 pts)

This problem uses the 'bliss' data set from the 'faraway' package in R. It is reproduced here with a different name for convenience. The data are from an early experiment to investigate the effectiveness of different concentrations of a pesticide in killing insects. The response is the number of dead and live insects at each concentration.

```
df = data.frame(
  dead=c(2,8,15,23,27),
  alive=c(28, 22, 15, 7, 3),
  conc=c(0,1,2,3,4)
  )
df
```

```
##   dead alive conc
## 1    2    28    0
## 2    8    22    1
## 3   15    15    2
## 4   23     7    3
## 5   27     3    4
```

(a) Use the `glm` function to fit a binomial logistic regression model to these data, with the numbers dead and alive as the response, and treating 'conc' as a numerical explanatory variable. Display the model summary.

Answer:

```
conc_fit <- glm(cbind(dead, alive) ~ conc,
family=binomial,data=df)
summary(conc_fit)
```

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ conc, family = binomial, data = df)
##
## Deviance Residuals:
##       1        2        3        4        5
## -0.4510   0.3597   0.0000   0.0643  -0.2045
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3238     0.4179  -5.561 2.69e-08 ***
## conc          1.1619     0.1814   6.405 1.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.76327  on 4  degrees of freedom
## Residual deviance:  0.37875  on 3  degrees of freedom
## AIC: 20.854
##
## Number of Fisher Scoring iterations: 4
```

(b) Based on the results you obtained above, find the predicted log-odds, odds, and probability of insect death for `conc=3.5`.

Answer:

```
predic = -2.3238 + 1.1619*(3.5)
predic
```

```
## [1] 1.74285
```

```
odds = exp(predic)
odds
```

```
## [1] 5.713604
```

```
predict.glm(conc_fit, newdata = data.frame(conc = 3.5), type = "response")
```

```
##         1
## 0.8510477
```

(c) The residual deviance reported in the model summary can be used as a goodness-of-fit test of the model when the data are grouped and the group probabilities are not too extreme. Assuming that applies here, compute the p-value for the goodness-of-fit test. Explain whether or not the model adequately fits the data according to the goodness-of-fit test.

Answer:

```
deviance(conc_fit)
```

```
## [1] 0.3787483
```

```
df.residual(conc_fit)
```

```
## [1] 3
```

```
1-pchisq(deviance(conc_fit), df.residual(conc_fit)) #p-value
```

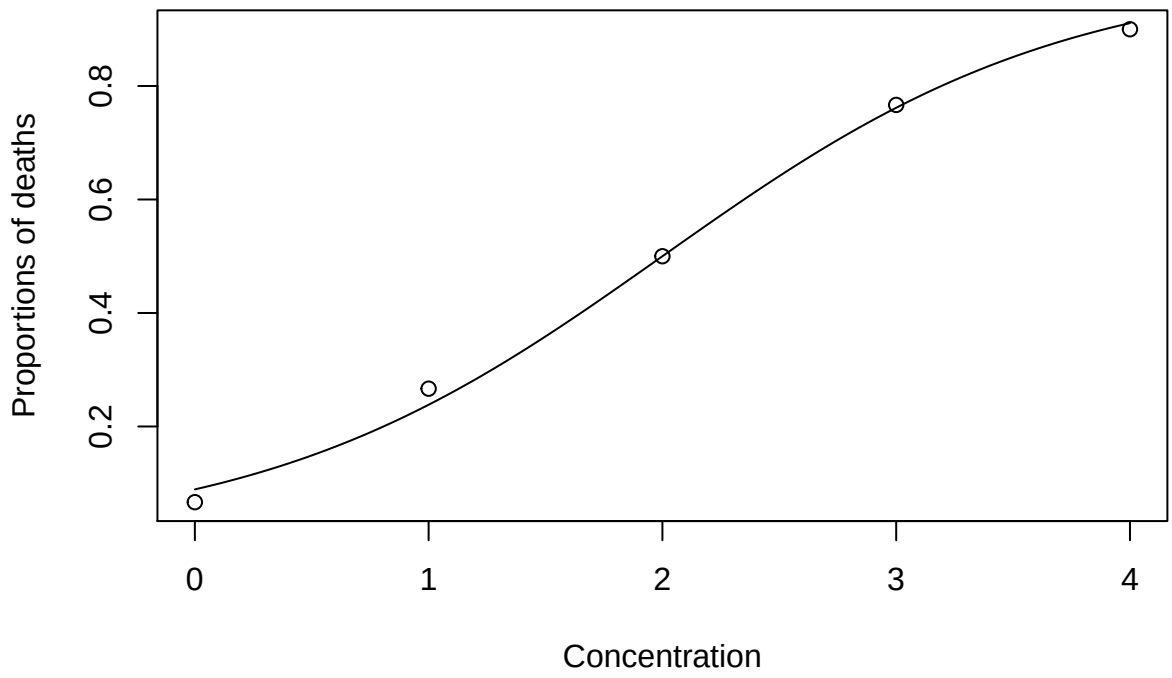
```
## [1] 0.9445968
```

```
#Based off of the test the model adequately fits
```

(d) Make a scatter plot of the proportion of deaths (vertical axis) versus the concentration (horizontal axis). Add the logistic regression curve to the plot. Explain briefly whether the curve appears to fit the data well or not (visual impression). Hint: For a similar plot, see the snoring and heart disease example in "Notes: 4.2 GLMs for Binary and Binomial Response Data".

Answer:

```
with(df, plot(df$conc, df$dead/(df$dead+df$alive), xlab= "Concentration", ylab= "Proport
curve(predict(conc_fit, data.frame(conc=x), type = "response"), add=TRUE)
```



Problem 2 (6 pts)

The following data are from a retrospective study of the relationship between coffee drinking and myocardial infarction (MI) for women aged 30-39. The data show the numbers of nonsmokers with either high coffee consumption (`HiCoffee`) or low coffee consumption (`LoCoffee`) for patients who had suffered MI (`MI=1`) and for control patients (`MI=0`).

```
dfmi = data.frame(
  HiCoffee = c(14, 49),
  LoCoffee = c(75, 381),
  MI = c(1,0)
)
dfmi
```

```
##   HiCoffee LoCoffee MI
## 1       14       75  1
## 2       49      381  0
```

(a) Notice that the first two columns of the data frame form a 2×2 table of counts for Coffee consumption versus MI. Let θ denote the ratio of odds for High versus Low coffee consumption for MI patients versus controls. Using contingency table methods, calculate a

sample estimate of the log-odds, $\ln(\theta)$, along with the lower and upper bounds of its 95% confidence interval for $\ln(\theta)$.

Answer:

```
theta<- (14 * 381) / (49 * 75)
(log_odds = log(theta))
```

```
## [1] 0.3725483
```

```
low = log(theta) - 1.96*(sqrt(1/14+1/49+1/75+1/381))
upper = log(theta) + 1.96*(sqrt(1/14+1/49+1/75+1/381))

(cl = c(low,upper))
```

```
## [1] -0.270961  1.016058
```

(b) Use the `glm` function to fit a logistic regression model with the high and low coffee consumption counts as the response, and MI as a "numerical" explanatory variable (it only has two unique values). Display the model summary. Compare the estimated coefficient for MI with the log-odds estimate you calculated in (a).
Answer:

```
coff_fit <- glm(cbind(HiCoffee, LoCoffee) ~ MI, family=binomial,data=dfmi)
summary(coff_fit)
```

```
##
## Call:
## glm(formula = cbind(HiCoffee, LoCoffee) ~ MI, family = binomial,
##     data = dfmi)
##
## Deviance Residuals:
## [1]  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.0510     0.1518 -13.514   <2e-16 ***
## MI            0.3725     0.3283   1.135    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance:  1.2236e+00  on 1  degrees of freedom
## Residual deviance: -3.2863e-14  on 0  degrees of freedom
## AIC: 13.93
##
## Number of Fisher Scoring iterations: 3
```

```
#The log-odds and estimated coefficient are the same.
```

(c) Calculate the 95% profile likelihood confidence interval for the MI coefficient in the model, and compare with the corresponding 95% Wald type confidence interval.

Answer:

```
confint(coff_fit)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -2.3605080 -1.7642830
## MI          -0.3025497  0.9933076
```

```
#Wald CI
low = 0.3725 - 0.3283*(1.96)
high = 0.3725 + 0.3283*(1.96)
(wald = c(low , high))
```

```
## [1] -0.270968  1.015968
```

```
#They are about the same except for the lower interval.
#It is not within the intervals calculated through Wald.
```

Problem 3 (6 pts)

The following data are from a retrospective study of the relationship between high and low levels of coffee consumption and myocardial infarction (MI) versus control for women aged 30-39 who were nonsmokers. The two variables are coded so that $\{\text{MI} = 1\} = \{\text{myocardial infarction}\}$, $\{\text{MI} = 0\} = \{\text{control}\}$, $\{\text{HiCoffee} = 1\} = \{\text{Coffee consumption} \geq 5 \text{ cups per day}\}$, and $\{\text{HiCoffee} = 0\} = \{\text{Coffee consumption} < 5 \text{ cups per day}\}$.

Answer:

```r
dfmi2 = data.frame(
  MI=c(1,1,0,0),
  HiCoffee=c(1,0,1,0),
  Count=c(14,75,49,381)
)
dfmi2
```

```
##   MI HiCoffee Count
## 1  1        1    14
## 2  1        0    75
## 3  0        1    49
## 4  0        0   381
```

(a) Fit a Poisson log-linear model to the counts, with explanatory variables `MI`, `HiCoffee`, and their product, which is expressed in the model formula as `MI:Coffee`. Display the model summary.

Answer:

```r
coff <- glm(Count ~ HiCoffee * MI, family = poisson, data=dfmi2)
summary(coff)
```

```
##
## Call:
## glm(formula = Count ~ HiCoffee * MI, family = poisson, data = dfmi2)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.94280    0.05123 115.999   <2e-16 ***
## HiCoffee    -2.05098    0.15177 -13.514   <2e-16 ***
## MI          -1.62531    0.12632 -12.866   <2e-16 ***
## HiCoffee:MI  0.37255    0.32832   1.135    0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5.8083e+02  on 3  degrees of freedom
## Residual deviance: 5.9952e-14  on 0  degrees of freedom
## AIC: 32.161
##
## Number of Fisher Scoring iterations: 3
```

(b) For this type of model, the association between the two variables is measured by the coefficient of the interaction term `MI:HiCoffee`. Is there a significant association at level $\alpha = 0.05$?

Answer: **There is not significant association at the leave 0.05 P-value is 0.256**

(c) Compare the estimated coefficient of `MI:HiCoffee` in this problem to the Problem 2a estimate of log-odds ratio and Problem 2b estimated coefficient of MI. According to the mathematical theory, they should all be the same. Does that appear to be correct within numerical rounding error? Confirm, or explain any differences you find.

Answer: **They are all the same 0.3725**