STAT 426 Assignment 4

Due Tuesday, September 21, 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 (6 pts)

The table constructed below shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx.

```
df = data.frame(
    Treatment=c('Surgery', 'Surgery', 'Radiation', 'Radiation'),
    Cancer=c('Controlled', 'Not Controlled',
             'Controlled', 'Not Controlled'),
    Count=c(21, 2, 15, 3))
dftab = xtabs(Count ~ Treatment + Cancer, data=df)
dftab
```

```
##             Cancer
## Treatment    Controlled Not Controlled
##   Radiation          15              3
##   Surgery            21              2
```

(a) Use Fisher's exact test to test the null hypothesis $H_0 : \theta = 1$ versus $H_A : \theta < 1$, where θ is the odds ratio. Report and interpret the p-value, indicating whether or not surgery is significantly better than radiation, at a significance level of 10%. (In the R function `fisher.test` use `alternative='less'`)

Answer:

```r
fisher.test(dftab, alternative = "less", conf.level = .90)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dftab
## p-value = 0.3808
## alternative hypothesis: true odds ratio is less than 1
## 90 percent confidence interval:
##  0.000000 2.457914
## sample estimates:
## odds ratio
##  0.4850293
```

```r
#Based off the p-value I would say the surgery is better than radiation.
```

(b) Use the two-sided alternative in `fisher.test` to obtain an "exact" 95% confidence interval for θ. Interpret what the results tell us about the relative effectiveness of the two treatments in controlling cancer.

Answer:

```r
fisher.test(dftab, alternative = "two.sided")
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dftab
## p-value = 0.6384
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.03629254 4.78671556
## sample estimates:
## odds ratio
##  0.4850293
```

```r
#exact p-value is 0.6384. The confidence interval does not include 1 so it is signific
```

(c) The Fisher exact test treats both the row totals and the column totals as fixed. Under this condition, what is the largest possible value for the count in the upper left cell of the table (Treatment='Radiation', Cancer='Controlled')? Remember that if the count in one cell is increased, then the counts in the other cells have to be adjusted to keep the row and column totals constant.

Answer:

```r
margin.table(dftab, 1)
```

```
## Treatment
## Radiation   Surgery
##        18        23
```

```r
margin.table(dftab, 2)
```

```
## Cancer
##     Controlled Not Controlled
##             36              5
```

```r
#Based off of the the count should stay at 18.
```

Problem 2 (8 pts)

High school seniors were surveyed for their political party identification (columns) and that of their parents (rows). The data are tabulated below

	Democratic	Independent	Republican
Democratic	60	25	7
Independent	13	24	8
Republican	6	18	25

(a) Notice how the cross tabulated table in Problem 1 was constructed in R. Using a similar method or otherwise, enter the party identification data into a 3×3 cross tabulated table in R and display it.

Answer:

```r
q2 = data.frame(
    Parents=c('Democratic', 'Democratic', 'Democratic', 'Independent', 'Independent', 'I
    Student=c('Democratic', 'Independent', 'Republican','Democratic', 'Independent', 'Re
    Count=c(60,25,7,13,24,8,6,18,25))
q2.tab = xtabs(Count ~ Parents + Student, data = q2)
q2.tab
```

```
##              Student
## Parents       Democratic Independent Republican
##   Democratic          60          25          7
##   Independent         13          24          8
##   Republican           6          18         25
```

(b) Calculate the expected counts under the null hypothesis that the party identifications of the students and their parents are statistically independent.

Answer:

```r
q2.X2 <- chisq.test(q2.tab, correct = F)
q2.X2$expected
```

```
##              Student
## Parents        Democratic Independent Republican
##   Democratic     39.07527    33.13978  19.784946
##   Independent    19.11290    16.20968   9.677419
##   Republican     20.81183    17.65054  10.537634
```

(c) Find the p-value of the Pearson Chi-square test of independence.

Answer:

```r
q2.X2
```

```
##
##  Pearson's Chi-squared test
##
## data:  q2.tab
## X-squared = 57.853, df = 4, p-value = 8.192e-12
```

```r
#P-value is 8.192e-12
```

(d) Display the table of standardized Pearson residuals and interpret qualitatively.

Answer:

```r
q2.X2$stdres
```

```
##              Student
## Parents        Democratic Independent Republican
##   Democratic    6.2082047  -2.4866441 -4.5635631
##   Independent  -2.1173648   2.7784217 -0.6990188
##   Republican   -4.9878607   0.1211715  5.8592767
```

Problem 3 (6 pts)

Consider a linear regression model of the form

$$Y_i \sim \text{Normal}(\mu_i, 1) \quad \text{where} \quad \mu_i = E(Y_i) = \eta_i \tag{1}$$

and

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, 2, \ldots, n.$$

(a) What is the link function for the model specified above?

Answer: Answers on speperate sheet of paper. **(b)** Show that the probability density function of Y_i has a natural exponential family form:

$$f(y_i; \mu_i) = a(\mu_i) b(y_i) \exp(y_i Q(\mu_i))$$

and show the forms of the functions $a(\mu_i)$, $b(y_i)$ and $Q(\mu_i)$.

Hint: Recall that, in general, the probability density function for Normal(μ, σ^2) has the form,

$$f(y; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}$$

and expand the square.

Answer:

(c) Show that if $y_1, y_2, \ldots, y_n$ are independent observations from the model specified above in (1), then maximizing the likelihood is the same as minimizing

$$\sum_{i=1}^{n} (y_i - \mu_i)^2$$

Answer: