STAA 552: HW 3

YOUR NAME HERE

See Canvas Calendar for due date.
40 points total, 4 points per problem unless otherwise noted.
Content for Q1-Q6 is from section 04 or earlier.
Content for Q7-Q10 is from section 06 (slide 11) or earlier.
Add or delete code chunks as needed.
For full credit, your numeric answers should be clearly labeled, outside of the R output.

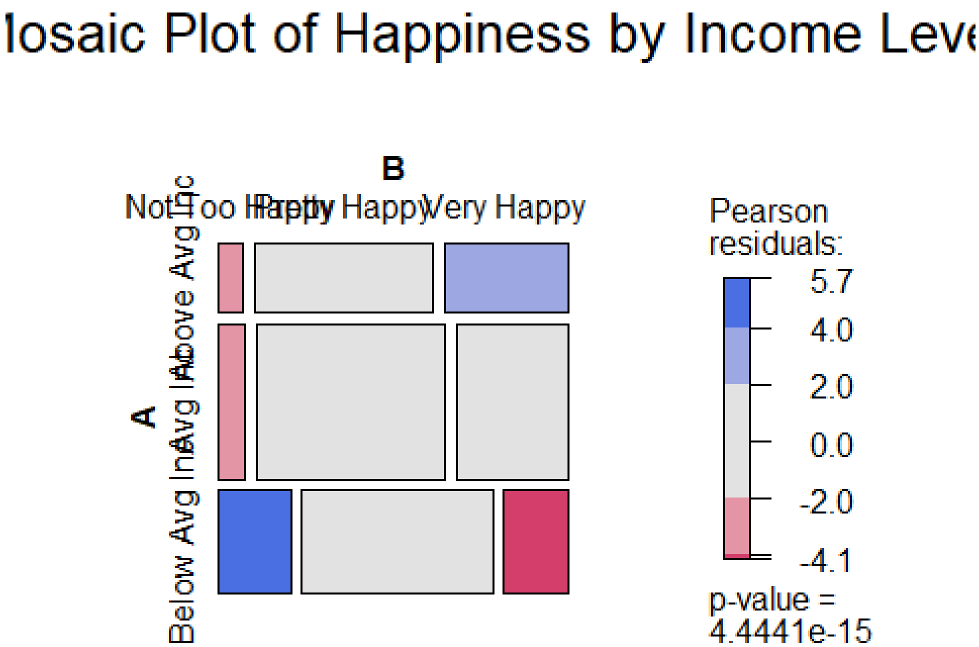
Happiness and Income (Q1 - Q6)

We consider data from a General Social Survey (GSS) cross-classifying a persons perceived happiness (not too happy, pretty happy, very happ) with their family income (above average, average, below average). A total of n = 1362 subjects participated in the survey. This data is taken from Introduction to Categorical Data Analysis, 3rd Edition.

```
##               Not Too Happy Pretty Happy Very Happy
## Above Avg Inc            21          159        110
## Avg Inc                  53          372        221
## Below Avg Inc            94          249         83
```

Q1 (2 pts)

Create a visual summary of the data using a mosaic plot.



Q2

Create a numeric summary of the data by calculating estimated conditional probabilities **by income level**. Which income category has the **highest** estimated probability of being “very happy”? Which income category has the **lowest** estimated probability of being “very happy”?

```
##                Not Too Happy Pretty Happy Very Happy
## Above Avg Inc     0.07241379    0.5482759  0.3793103
## Avg Inc           0.08204334    0.5758514  0.3421053
## Below Avg Inc     0.22065728    0.5845070  0.1948357
```

Response

Q3

State the hypotheses corresponding to the chi-square test for this research scenario. In this case, it is easiest to state the hypotheses in words.

Response

Hypotheses for the Chi-Square Test:

- Null Hypothesis: There is no association between income level and perceived happiness. That is, income level and happiness are independent.
- Alternative Hypothesis: There is an association between income level and perceived happiness. That is, income level and happiness are not independent.

Q4

Use `chisq.test()` to run the chi-square test. Use the results to make a conclusion in context.

```
##
##  Pearson's Chi-squared test
##
## data:  HappyData
## X-squared = 73.352, df = 4, p-value = 4.444e-15
```

Response

Q5 (6 pts)

Show the expected cell counts under the null hypothesis. Use these to discuss:
(a) whether the chi-square test is appropriate here based on the "rule of thumb" from the notes.
(b) one income-happiness category that has a large difference between observed and expected cell counts. For the cell you chose, mention whether the observed count is higher or lower than expected.

Note: For (b) there are several possible cells that can be discussed.

```
##               Not Too Happy Pretty Happy Very Happy
## Above Avg Inc      35.77093     166.0793   88.14978
## Avg Inc            79.68282     369.9559  196.36123
## Below Avg Inc      52.54626     243.9648  129.48899

##               Not Too Happy Pretty Happy Very Happy
## Above Avg Inc     -14.77093    -7.079295   21.85022
## Avg Inc           -26.68282     2.044053   24.63877
## Below Avg Inc      41.45374     5.035242  -46.48899
```

Response

All Expected cell counts are well above 5 so the chi squared is a appropriate test.

The “Below Average Income” group shows the most signifficant deviation from expected behavior with 41 people more than expected being not to happy, and 46 less than expected being very happy.

Q6

Use `GTest()` from the `DescTools` package to run a likelihood ratio test. Use the results to make a conclusion in context.

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  HappyData
## G = 71.305, X-squared df = 4, p-value = 1.199e-14
```

Response

Snoring (Q7 - Q10)

An epidemiological survey was done to investigate snoring as a risk factor for heart disease. n = 2484 subjects were classified based on how much they snored (based on spouse report) and whether they reported having heart disease. Snore score is given as 0 = “never”, 2 = “occasionally”, 4 = “nearly every night” and 5 = “every night”. This data appears in CDA Table 4.2. We will fit, visualize and interpret a logistic regression model.

```
##   Snore NoHD YesHD
## 1     0 1355    24
## 2     2  603    35
## 3     4  192    21
## 4     5  224    30
```

Notes:

- For these questions, it will be helpful to look at the Beetles example (Sec06_Examples sections 1.1 – 1.4).
- We will treat snore score as continuous, but it feels strange! We will do several more examples of logistic regression analysis.
- I specifically chose this data because you can check your own work with information from CDA (p119-120): $logit[\hat{p}(x)] = -3.87 + 0.40x$

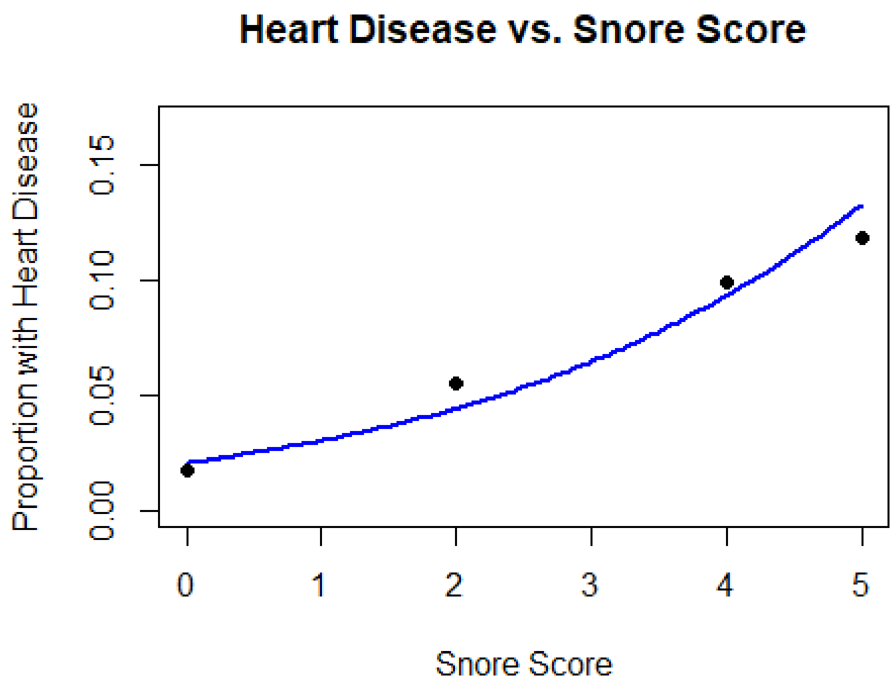
Q7

Fit an appropriate logistic regression model. Use YesHD as the event of interest. You can just show the R output for full credit.

```
##
## Call:
## glm(formula = HD ~ Snore, family = binomial, data = SnoreData_long,
##     weights = Count)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.86625    0.16620 -23.263  < 2e-16 ***
## Snore        0.39734    0.05001   7.945 1.93e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 900.83  on 7  degrees of freedom
## Residual deviance: 837.73  on 6  degrees of freedom
## AIC: 841.73
##
## Number of Fisher Scoring iterations: 6
```

Q8

Create a plot of the observed proportion of subjects with heart disease versus snore score. Overlay a smooth curve corresponding to the model based estimated probabilities.



Q9

Calculate and interpret the estimated odds ratio corresponding to score snore in context.

Response

```
##    Snore
## 1.487857
```

Q10

We will "verify" the estimated odds ratio from above empirically. Specifically, we will calculate the model based estimated probabilities of heart disease at snore values in 1 unit increments. Then we will use those values to calculate odds ratios corresponding to a 1 unit increase. While we are at it, we will calculate risk ratio values.

Notes:

- Code using tidyverse is provided. Just show the resulting output for full credit.
- Remove the comments (#) and change "SnoreModel" to whatever you called the logistic regression model in Q7.
- The `lag()` function extracts the previous value in the vector.

```
##   Snore       Prob       Odds OddsRatio RiskRatio
## 1     0 0.02050742 0.02093678        NA        NA
## 2     1 0.03020986 0.03115092  1.487857  1.473119
## 3     2 0.04429511 0.04634811  1.487857  1.466247
## 4     3 0.06451072 0.06895934  1.487857  1.456385
## 5     4 0.09305411 0.10260161  1.487857  1.442460
## 6     5 0.13243885 0.15265650  1.487857  1.423246
```

Appendix

```
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
HappyData <- matrix(c(21, 159, 110,
                      53, 372, 221,
                      94, 249, 83), byrow = TRUE, nrow = 3)
rownames(HappyData) <- c("Above Avg Inc", "Avg Inc", "Below Avg Inc")
colnames(HappyData) <- c("Not Too Happy", "Pretty Happy", "Very Happy")
HappyData
#Q1

library(vcd)

mosaic(HappyData, shade = TRUE, legend = TRUE, main = "Mosaic Plot of
Happiness by Income Level")

#Q2
row_totals <- rowSums(HappyData)

cond_probs <- sweep(HappyData, 1, row_totals, FUN = "/")

cond_probs

#Q4
chi_test <- chisq.test(HappyData)

chi_test
#Q5
expected_counts <- chi_test$expected
expected_counts

differences <- HappyData - expected_counts
differences

#Q6
library(DescTools)
g_test <- GTest(HappyData)
g_test
```

```r
SnoreData <- data.frame(Snore = c(0, 2, 4, 5),
                        NoHD = c(1355, 603, 192, 224),
                        YesHD = c(24, 35, 21, 30))
SnoreData
#Q7
library(tidyr)
library(dplyr)

SnoreData_long <- SnoreData %>%
  gather(key = "HeartDisease", value = "Count", NoHD, YesHD) %>%
  mutate(HD = ifelse(HeartDisease == "YesHD", 1, 0))

SnoreModel <- glm(HD ~ Snore, weights = Count, data = SnoreData_long, family
= binomial)

summary(SnoreModel)
#Q8

SnoreData$Total <- SnoreData$NoHD + SnoreData$YesHD
SnoreData$Prop_HD <- SnoreData$YesHD / SnoreData$Total

Snore_seq <- seq(min(SnoreData$Snore), max(SnoreData$Snore), length.out =
100)

Predicted_Probs <- predict(SnoreModel, newdata = data.frame(Snore =
Snore_seq), type = "response")

plot(SnoreData$Snore, SnoreData$Prop_HD, xlab = "Snore Score", ylab =
"Proportion with Heart Disease",
     main = "Heart Disease vs. Snore Score", pch = 16, ylim = c(0,
max(SnoreData$Prop_HD) + 0.05))

lines(Snore_seq, Predicted_Probs, col = "blue", lwd = 2)

#Q9

# Extract the coefficient for Snore from the model
beta <- coef(SnoreModel)["Snore"]

# Calculate the odds ratio
odds_ratio <- exp(beta)
odds_ratio

library(tidyverse)

PredValues <- data.frame(Snore = seq(from = 0, to = 5))
```

```r
PredValues <- PredValues %>%
              mutate(
                Prob = predict(SnoreModel, newdata = PredValues, type =
"response"),
                Odds = Prob / (1 - Prob),
                OddsRatio = Odds / lag(Odds),
                RiskRatio = Prob / lag(Prob)
              )

PredValues
```