STAA 552: HW 4

YOUR NAME HERE

See Canvas Calendar for due date.
64 points total, 4 points per problem unless otherwise noted. **The first 4 items will not be graded. They are included for extra practice.** Content for all questions is from section 06 or earlier.
Add or delete code chunks as needed.
For full credit, your numeric answers should be clearly labeled, outside of the R output.

For this assignment we will use data from the Statistical Sleuth (3rd Edition) about the Donner party.

From Wikipedia: “The Donner Party, sometimes called the Donner–Reed Party, was a group of American pioneers who migrated to California in a wagon train from the Midwest. Delayed by a multitude of mishaps, they spent the winter of 1846–1847 snowbound in the Sierra Nevada mountain range.”

The DonnerData.csv file includes the Age (years) and Sex (Female or Male) of n = 45 adults (over 15 years). **Survived (0,1) should be used as the response for logistic regression analyses.** Status (Died or Survived) may be used elsewhere or ignored.

These data were previously used by an anthropologist (Grayson, 1990) to study the “theory that females are better able to withstand harsh conditions than are males”.

Note: We essentially have a census, but we will perform statistical analysis anyway.

Q1 (not graded)

Create a summary table including the following information **by Sex**: n (sample size), proportion that survived, minimum Age, mean Age, max Age. Hint: `group_by()` and `summarise()` from tidyverse/dplyr are helpful here. Or, use `table1()` form the `table1` package.

Q2 (not graded)

Calculate the odds ratio of survival for Males vs Females and corresponding 95% (Wald) confidence interval. Some example code has been provided, but other approaches are possible.

OR =
95% CI:

Q3 (not graded)

Run a chi-squared test using Status (or Survived) and Sex. Report the p-value and use it to make a conclusion in context. Be sure to mention the direction of the association. Hint: This result appears in the `oddsratio()` output.

p-value =
Conclusion:

Q4 (not graded)

The chi-squared test is based on a large sample approximation. Is the sample size large enough for the chi-squared test to be appropriate? Use the rule of thumb from the notes to justify your response.

Response

Model 1 (Q5 - Q8)

For this group of questions, fit an appropriate logistic regression model using Survived (0,1) as the response and **Sex** as the predictor.

Q5 (2 pts)

Fit the model and show the coefficients table.

```
##
## Call:
## glm(formula = Survived ~ Sex, family = "binomial", data = donner_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6931     0.5477   1.266   0.2057
## SexMale      -1.3863     0.6708  -2.067   0.0388 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 57.286  on 43  degrees of freedom
## AIC: 61.286
##
## Number of Fisher Scoring iterations: 4
```

Q6

Calculate the (model based) estimated odds ratio of survival for Males vs Females and corresponding 95% (profile) confidence interval. Use this information to make a conclusion in context. Be sure to mention the direction of the association.

```
## OR = 0.25

## 95% CI: 0.06 - 0.9

## [1] "The Odds Ratio of .25 tesls us that the odds of survival for males is
25% of the odds of survival for females. We also see that the 95% confidence
interval does not include 1 which indicates that atleast 95% of the time our
directional conclusion is correct."
```

Q7

Using this model, we will test for a Sex effect.

Q7a (2 pts)

Do this using a Wald test. Report the Z test statistic and p-value.

---

log(odds_ratio) + 1.96 * SE = log(.9) SE = (log(.9) - log(.25))/ 1.96 = .653 Z = Coef/SE = -1.386/.653 = -2.12

p = .017

Q7b (2 pts)

Do this using a likelihood ratio test “by hand” (using summary information from the model). Report the chi-square test statistic and p-value.

```r
null_model <- glm(Survived ~ 1, data = donner_data, family = binomial)
logLik_full <- logLik(log_model1)
logLik_null <- logLik(null_model)
```

```
lr_stat <- 2 * (as.numeric(logLik_full) - as.numeric(logLik_null))
p_value <- pchisq(lr_stat, df = 1, lower.tail = FALSE)

cat("X2 =", lr_stat, "\n")

## X2 = 4.540267

cat("p = ", p_value)

## p =  0.03310643
```

Q7c (2 pts)

Confirm your likelihood test results using Anova() from the car package. Just show the output for full credit.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Survived
##     LR Chisq Df Pr(>Chisq)
## Sex   4.5403  1    0.03311 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q8

Use emmeans() from the emmeans package to (a) calculate (model based) estimated probabilities of survival by Sex and (b) corresponding odds ratio. Just show the output for full credit. Hint: Remember to use type = “response”. Note: The estimated probabilities should exactly match the simple proportions from Q1.

```
##  Sex     prob     SE  df asymp.LCL asymp.UCL
##  Female 0.667 0.1220 Inf     0.406     0.854
##  Male   0.333 0.0861 Inf     0.190     0.516
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale

##  contrast      estimate    SE  df asymp.LCL asymp.UCL z.ratio p.value
##  Female - Male     1.39 0.671 Inf    0.0715       2.7   2.067  0.0388
##
## Results are given on the log odds ratio (not the response) scale.
## Confidence level used: 0.95
```

Model 2 (Q9 - Q15)

For this group of questions, fit an appropriate logistic regression model using Survived (0,1) as the response and **Sex and Age** as predictors. This should be an additive model (no interaction).

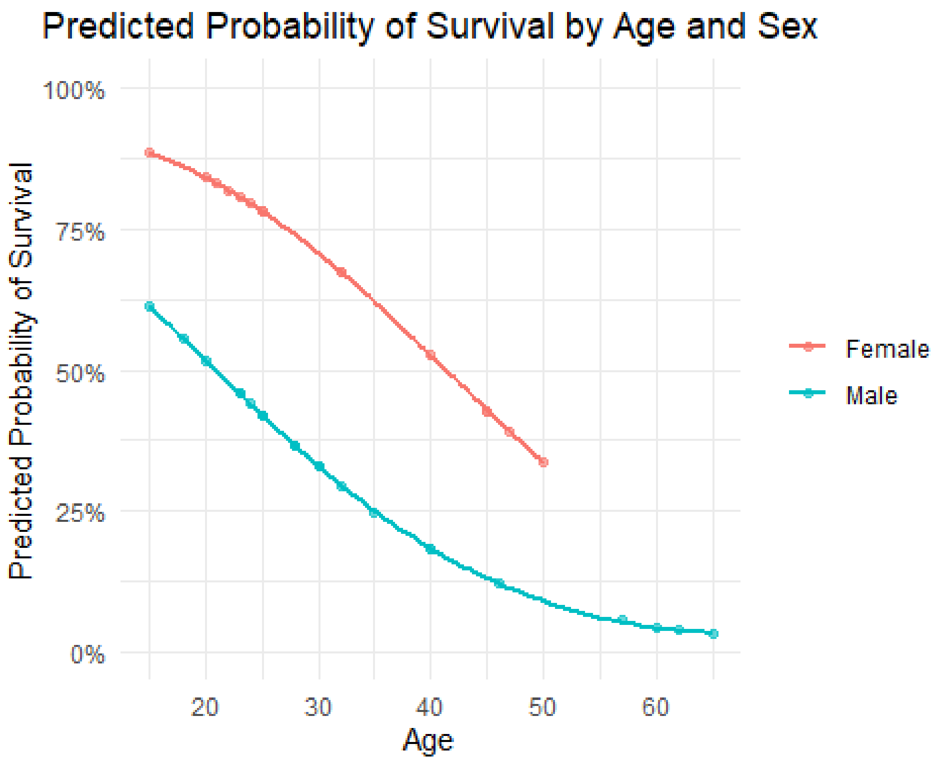
Q9 (2 pts)

Fit the model and show the coefficients table.

```
##
## Call:
## glm(formula = Survived ~ Sex + Age, family = "binomial", data =
donner_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.23041    1.38686   2.329   0.0198 *
## SexMale     -1.59729    0.75547  -2.114   0.0345 *
## Age         -0.07820    0.03728  -2.097   0.0359 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 51.256  on 42  degrees of freedom
## AIC: 57.256
##
## Number of Fisher Scoring iterations: 4
```

Q10

Create a visual summary of this model. Specifically, graph fitted curves showing the probability of survival for females and males as a function of age. Hint: See Crabs Example section 3.3.



Q11

Calculate **and interpret** the (model based) estimated odds ratio corresponding to **Sex.**

OR:
Interpretation:

```
## Estimated Odds Ratio for Sex (Males vs Females): 0.2
```

The Estimated Odds Ratio of .2 implies that the odds of survival for men are 80% lower than the odds of survival for women when we control for age. *****

Q12

```
## Estimated Odds Ratio for Age: 0.92
```

Calculate **and interpret** the (model based) estimated odds ratio corresponding to **Age.**

OR: .92 Interpretation: for each year of age, the odds of survival decrearse by 8%

Q13

Using this model, test for the effect of **Sex** (controlling for Age).

Q13a (2 pts)

Do this using a Wald test. Report the Z test statistic and p-value.

Z = Coef/SE = -1.59729/0.75547 = -2.1143 p = .017258.

Q13b

Do this using a likelihood ratio test "by hand" (using summary information from the model). Report the chi-square test statistic and p-value.

```
#Q13b
reduced_model <- glm(Survived ~ Age, data = donner_data, family = "binomial")

logLik_full <- logLik(log_model2)
logLik_reduced <- logLik(reduced_model)

lr_stat <- 2 * (as.numeric(logLik_full) - as.numeric(logLik_reduced))

# Calculate the p-value
p_value_lr <- pchisq(lr_stat, df = 1, lower.tail = FALSE)

cat("X2 =", lr_stat, "\n")

## X2 = 5.034437

cat("p = ", p_value_lr)

## p =  0.02484816
```

Q13c (2 pts)

Confirm your likelihood test results using `Anova()` from the car package. Just show the output for full credit.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Survived
##     LR Chisq Df Pr(>Chisq)
## Sex   5.0344  1    0.02485 *
## Age   6.0300  1    0.01406 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q14

Use emmeans() from the emmeans package to (a) calculate (model based) estimated probabilities of survival by Sex and (b) corresponding odds ratio. Just show the resulting output for full credit. Hint: Remember to use type = "response".

```
##  Sex     prob     SE  df asymp.LCL asymp.UCL
##  Female 0.678 0.1310 Inf     0.394     0.872
##  Male   0.299 0.0916 Inf     0.153     0.501
##
## Confidence level used: 0.95
## Intervals are back-transformed from the logit scale

##  contrast      estimate    SE  df asymp.LCL asymp.UCL z.ratio p.value
##  Female - Male      1.6 0.755 Inf     0.117      3.08   2.114  0.0345
##
## Results are given on the log odds ratio (not the response) scale.
## Confidence level used: 0.95
```

Q15

Calculate the (model based) estimated probabilities of survival for Males and Females with (mean) age = 31.8 years. Note: These results should match the previous question.

Q15a

Do this "by hand" (using summary information from the model).

Females:
Males:

```r
#Q15a
beta_0 <- coef(log_model2)["(Intercept)"]
beta_SexMale <- coef(log_model2)["SexMale"]
beta_Age <- coef(log_model2)["Age"]

# Mean age
mean_age <- 31.8
```

```
eta_female <- beta_0 + beta_SexMale * 0 + beta_Age * mean_age
prob_female <- exp(eta_female) / (1 + exp(eta_female))

# For Males (SexMale = 1)
eta_male <- beta_0 + beta_SexMale * 1 + beta_Age * mean_age
prob_male <- exp(eta_male) / (1 + exp(eta_male))

# Print results
cat("Females:", round(prob_female, 2), "\n")

## Females: 0.68

cat("Males:", round(prob_male, 2), "\n")

## Males: 0.3
```

Q15b

Confirm your results using the `predict()` function.

```
#Q15b
new_data <- data.frame(
  Sex = c("Female", "Male"),
  Age = c(31.8, 31.8)
)

predicted_probs <- predict(log_model2, newdata = new_data, type = "response")

print(predicted_probs,2)

##    1    2
## 0.68 0.30
```

Model 3 (Q16 - Q18)

For this group of questions, fit an appropriate logistic regression model using Survived (0,1) as the response and **Sex, Age and Sex:Age interaction** as predictors.

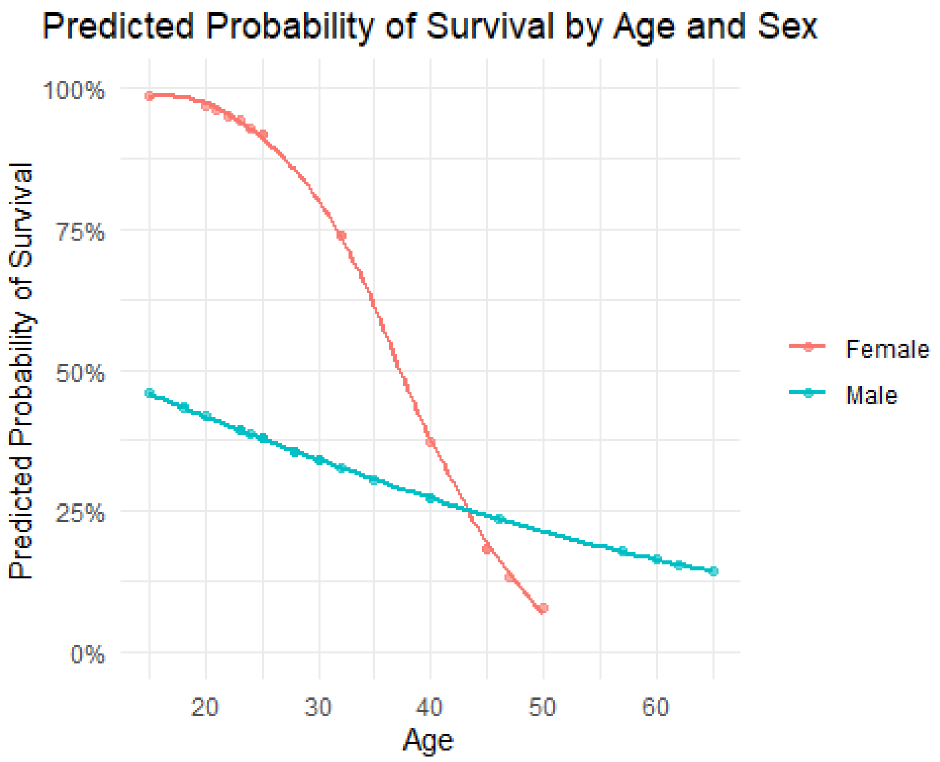
Q16 (2 pts)

Fit the model and show the coefficients table.

```
##
## Call:
## glm(formula = Survived ~ Sex + Age + Sex:Age, family = "binomial",
##     data = donner_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.24638    3.20517   2.261   0.0238 *
## SexMale     -6.92805    3.39887  -2.038   0.0415 *
## Age         -0.19407    0.08742  -2.220   0.0264 *
## SexMale:Age  0.16160    0.09426   1.714   0.0865 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 61.827  on 44  degrees of freedom
## Residual deviance: 47.346  on 41  degrees of freedom
## AIC: 55.346
##
## Number of Fisher Scoring iterations: 5
```

Q17

Create a visual summary of this model. Specifically, graph fitted curves showing the probability of survival for females and males as a function of age. Hint: See Crabs Example section 4.2.



Q18

The interaction model allows the effect of Age to depend on Sex (and vice versa).

Q18a

Calculate **and interpret** the (model based) estimated odds ratio corresponding corresponding to Age for **Females**.

OR:
Interpretation:

```
## Estimated Odds Ratio for Age (Females): 0.8236
```

For Females, each one year increase in age is associated with a 16% decrease in the odds of survival *****

Q18b

Calculate **and interpret** the (model based) estimated odds ratio corresponding corresponding to Age for **Males**.

OR:
Interpretation:

```
## Estimated Odds Ratio for Age (Males): 0.968
```

For males, each one year increase in age is associated with a 3.2% decrease in the odds of survival. *****

Appendix

```r
#Retain this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
#install.packages("emmeans")
library(car)
library(emmeans)
library(ggplot2)
#Q1

#Q2
library(epitools)
#DonnerTable <- table(DonnerData$Sex,DonnerData$Status)
#oddsratio(DonnerTable, method = "wald")
#Q4

#Q5
donner_data <- read.csv("Data/DonnerData.csv")
donner_data$Survived <- as.factor(donner_data$Survived)

log_model1 <- glm(Survived ~ Sex, data = donner_data, family = "binomial")

summary(log_model1)
#Q6
odds_ratio <- exp(coef(log_model1)["SexMale"])

# Calculate the 95% confidence interval for the odds ratio
confint_model <- confint(log_model1, "SexMale")
confint_odds_ratio <- exp(confint_model)

# Print results
cat("OR =", round(odds_ratio, 2), "\n")
cat("95% CI:", round(confint_odds_ratio[1], 2), "-",
round(confint_odds_ratio[2], 2), "\n")

print("The Odds Ratio of .25 tesls us that the odds of survival for males is
25% of the odds of survival for females. We also see that the 95% confidence
interval does not include 1 which indicates that atleast 95% of the time our
directional conclusion is correct.")
null_model <- glm(Survived ~ 1, data = donner_data, family = binomial)
logLik_full <- logLik(log_model1)
logLik_null <- logLik(null_model)
```

```r
lr_stat <- 2 * (as.numeric(logLik_full) - as.numeric(logLik_null))
p_value <- pchisq(lr_stat, df = 1, lower.tail = FALSE)

cat("X2 =", lr_stat, "\n")
cat("p = ", p_value)
#Q7c
anova_results <- Anova(log_model1, test="LR")
print(anova_results)
#Q8

#estimated probabilities of survival by Sex
emmeans_prob <- emmeans(log_model1, ~ Sex, type = "response")
print(emmeans_prob)

#odds ratio
emmeans_odds <- contrast(emmeans(log_model1, ~ Sex), method = "pairwise",
infer = TRUE)
print(emmeans_odds)
#Q9
donner_data <- read.csv("Data/DonnerData.csv")
donner_data$Survived <- as.factor(donner_data$Survived)

log_model2 <- glm(Survived ~ Sex + Age, data = donner_data, family =
"binomial")

summary(log_model2)
#Q10

donner_data$predicted_prob <- predict(log_model2, newdata = donner_data, type
= "response")

ggplot(donner_data, aes(x = Age, y = predicted_prob, color = Sex)) +
  geom_point(alpha = 0.6) +  # Points for each individual's predicted
probability
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +  # Smooth
curve for each sex
  labs(
    title = "Predicted Probability of Survival by Age and Sex",
    x = "Age",
    y = "Predicted Probability of Survival"
  ) +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 1), labels = scales::percent) +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.title = element_blank()
  )
```

```r
#Q11
# Extract and exponentiate the coefficient for Sex to get the odds ratio
odds_ratio_sex <- exp(coef(log_model2)["SexMale"])

# Print the odds ratio
cat("Estimated Odds Ratio for Sex (Males vs Females):", round(odds_ratio_sex,
2), "\n")


#Q12

odds_ratio_age <- exp(coef(log_model2)["Age"])
cat("Estimated Odds Ratio for Age:", round(odds_ratio_age, 2), "\n")

#Q13b
reduced_model <- glm(Survived ~ Age, data = donner_data, family = "binomial")

logLik_full <- logLik(log_model2)
logLik_reduced <- logLik(reduced_model)

lr_stat <- 2 * (as.numeric(logLik_full) - as.numeric(logLik_reduced))

# Calculate the p-value
p_value_lr <- pchisq(lr_stat, df = 1, lower.tail = FALSE)

cat("X2 =", lr_stat, "\n")
cat("p = ", p_value_lr)

#Q13c
anova_results <- Anova(log_model2, test="LR")
print(anova_results)
#Q14
emmeans_prob <- emmeans(log_model2, ~ Sex, type = "response")
print(emmeans_prob)

emmeans_odds <- contrast(emmeans(log_model2, ~ Sex), method = "pairwise",
infer = TRUE)
print(emmeans_odds)
#Q15a
beta_0 <- coef(log_model2)["(Intercept)"]
beta_SexMale <- coef(log_model2)["SexMale"]
beta_Age <- coef(log_model2)["Age"]

# Mean age
mean_age <- 31.8


eta_female <- beta_0 + beta_SexMale * 0 + beta_Age * mean_age
prob_female <- exp(eta_female) / (1 + exp(eta_female))
```

```
# For Males (SexMale = 1)
eta_male <- beta_0 + beta_SexMale * 1 + beta_Age * mean_age
prob_male <- exp(eta_male) / (1 + exp(eta_male))

# Print results
cat("Females:", round(prob_female, 2), "\n")
cat("Males:", round(prob_male, 2), "\n")

#Q15b
new_data <- data.frame(
  Sex = c("Female", "Male"),
  Age = c(31.8, 31.8)
)

predicted_probs <- predict(log_model2, newdata = new_data, type = "response")

print(predicted_probs,2)
#Q16
donner_data <- read.csv("Data/DonnerData.csv")
donner_data$Survived <- as.factor(donner_data$Survived)

log_model3 <- glm(Survived ~ Sex + Age + Sex:Age, data = donner_data, family
= "binomial")

summary(log_model3)
#Q17
donner_data$predicted_prob <- predict(log_model3, newdata = donner_data, type
= "response")

ggplot(donner_data, aes(x = Age, y = predicted_prob, color = Sex)) +
  geom_point(alpha = 0.6) +  # Points for each individual's predicted
probability
  geom_smooth(method = "loess", formula = y ~ x, se = FALSE) +  # Smooth
curve for each sex
  labs(
    title = "Predicted Probability of Survival by Age and Sex",
    x = "Age",
    y = "Predicted Probability of Survival"
  ) +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 1), labels = scales::percent) +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.title = element_blank()
  )
#Q18a
beta_Age <- coef(log_model3)["Age"]
```

```r
odds_ratio_age_female <- exp(beta_Age)

cat("Estimated Odds Ratio for Age (Females):", round(odds_ratio_age_female,
4), "\n")
beta_Age <- coef(log_model3)["Age"]
beta_SexMale_Age <- coef(log_model3)["SexMale:Age"]

beta_Age_male <- beta_Age + beta_SexMale_Age
odds_ratio_age_male <- exp(beta_Age_male)
cat("Estimated Odds Ratio for Age (Males):", round(odds_ratio_age_male, 4),
"\n")
```