

Stat 439: Homework 4

Due Thur 3/10/22 by 5pm in D2L

Your name here

Instructions

You are strongly encouraged to use R Markdown to complete your Homework assignments, starting with this file as a template and Knitting to pdf. Submit your homework to [Gradescope](#) as a single pdf file.

Part I: Donner Party

For this problem, we are going to use the built-in **Donner** data set in the **alr4** library. Examine the help file for a description of the data.

Read the data set into your R session:

```
data(Donner)
```

Note that R reads in all variables except **age** as a factor. The **glm** function requires that your binary response be coded as 0's and 1's, so first create an indicator variable for survival:

```
Donner$survive <- as.numeric(Donner$y == "survived")
```

Q1

How many people in the Donner Party survived? What proportion survived?

Q2

Fit a logistic regression model to these data using only **age** as the predictor, and call this model **mod1**. Report a summary of the model. Then answer the following:

- Write out the equation of the fitted model. Define any symbols used.
- Create a scatterplot with **age** on the x-axis and **survive** on the y-axis. Add a nonparametric smoother (e.g., **lowess**), and add the fitted model curve. Ensure your plot is well-labeled.
- Write a sentence interpreting the estimated slope on the odds scale.
- What is the estimated probability of survival for a 20-year-old?
- What is the estimated relative “risk” of survival for a 60-year-old compared to a 20-year-old.
- Calculate and interpret a 95% confidence interval for the odds ratio of survival for a 60-year-old compared to a 20-year-old. Write a sentence interpreting this interval.

Q3

Fit a logistic regression model to these data using only **sex** as the predictor. Call this model **mod2**. What does this model estimate for the odds ratio of survival for males compared to females?

Q4

Create a 2x2 table with **sex** as the rows and **survive** as the columns. Use this table to estimate the odds ratio of survival for males compared to females. Does this value match your answer to part c.?

Q5

Now, fit a logistic regression model using both **age** and **sex** as predictors, with no interaction. Call this model **mod3**. Then answer the following:

- Write out the equation of the fitted model for Males, and the equation of the fitted model for Females.
- Write a sentence interpreting the **sex** coefficient in this model.

Q6

Lastly, fit a logistic regression model using both **age** and **sex** as predictors, and include an interaction term. Call this model **mod4**. Write an interpretation of each coefficient (except the intercept) in this model.

Q7

Again, create a scatterplot with **age** on the x-axis and **survive** on the y-axis. On the same plot, add the following:

- Fitted curve from **mod1**
- Fitted “curve” from **mod2**, one for each sex
- Fitted curve from **mod3**, one for each sex
- Fitted curve from **mod4**, one for each sex

Use different line types for the different models, and different colors for the different sexes. Add a legend. Ensure your plot is well-labeled.

Q8

Choose which model you think fits the data best. Then create an ROC curve for your chosen model. How well does your model predict survival?

Part II: Cite Sources

Write the sources you used to complete this assignment at the end of your homework submission, adhering to the “Guidance on Citing Sources” bullet points in the [collaboration policy section on our course syllabus](#).