STA 214 Homework 2

Due: Friday, January 27, 12:00pm (noon) on Canvas.

Instructions: There are three parts to this assignment. Part I is practice with logistic regression, Part II is practice with maximum likelihood estimation, and Part III is a short (extra credit) problem on debugging in R.

Getting started: Begin by downloading the HW2 template from the course website:

`https://sta214-s23.github.io/homework/hw_02_template.Rmd`

Save this template file to your computer, then open it in RStudio. As you complete the assignment, you will write down your answers to all questions in the R Markdown file, and include all R code in code chunks. *If a question requires code, you will not receive credit if no code is provided.* Refer to the R Markdown instructions on the course website (`https://sta214-s23.github.io/resources/rmarkdown_instructions/`) if you have issues getting started.

Submission: When you have completed the assignment, knit your homework to HTML and submit on Canvas. For Part II, you may submit a separate scan of your written work, if you prefer.

Data

The RMS Titanic was a huge, luxury passenger liner designed and built in the early 20th century. Despite the fact that the ship was believed to be unsinkable, during her maiden voyage on April 15, 1912, the Titanic collided with an iceberg and sank. Of all the passengers and crew, less than half survived. Part of the reason why so few people survived has been attributed to the fact that the Titanic did not carry enough lifeboats for its passengers and crew. This meant that there was competition for space in the boats, and not everyone was able to make it aboard. Communication errors, stress and shock...there were a great many factors that contributed to this tragedy.

The loss of life during the Titanic tragedy was enormous, but there were survivors. Was it random chance that these particular people survived? Or were there some specific characteristics of these people that led to their positions in the life boats? Let's investigate.

We have observations on 12 different variables, some categorical and some numeric:

- `Passenger`: A unique ID number for each passenger.
- `Survived`: An indicator for whether the passenger survived (1) or perished (0) during the disaster.
- `Pclass`: Indicator for the class of the ticket held by this passengers; 1 = 1st class, 2 = 2nd class, 3 = 3rd class.
- `Name`: The name of the passenger.
- `Sex`: Binary indicator for the sex of the passenger.
- `Age`: Age of the passenger in years; Age is fractional if the passenger was less than 1 year old.

- `SibSp`: number of siblings/spouses the passenger had aboard the Titanic. Here, siblings are defined as brother, sister, stepbrother, and stepsister. Spouses are defined as husband and wife.
- `Parch`: number of parents/children the passenger had aboard the Titanic. Here, parent is defined as mother/father and child is defined as daughter,son, stepdaughter or stepson. NOTE: Some children traveled only with a nanny, therefore parch=0 for them. There were no parents aboard for these children.
- `Ticket`: The unique ticket number for each passenger.
- `Fare`: How much the ticket cost in US dollars.
- `Cabin`: The cabin number assigned to each passenger. Some cabins hold more than one passenger.
- `Embarked`: Port where the passenger boarded the ship; C = Cherbourg, Q = Queenstown, S = Southampton

Goal: Our goal is to predict the probability that a passenger survives the Titanic disaster.

Loading the data

The `titanic` data can be loaded into R with the following command:

```
titanic <- read.csv("https://sta214-s23.github.io/homework/Titanic.csv")
```

Here `read.csv` is a function that imports data from a CSV file. We can pass `read.csv` either a local path on our computer, or a URL – in this case, we use the URL where the data is stored online. We have called the data `titanic` in R.

Copy the command to load the data into the setup chunk of your R Markdown file, and run it.

1 Logistic regression

In the first part of the assignment, we will fit a logistic regression model to the Titanic data. If you get stuck, remember that the course codebook may be helpful: `https://sta214-s23.github.io/resources/codebook.html`

1. Before we fit any models, we need to do a bit of data exploration. To keep this assignment manageable, we will only do the first few steps here, but a full analysis would require additional EDA.

   (a) Looking at the available variables, there are some that are not valid choices for explanatory variables, meaning we can not use them as X variables in a parametric model. Which are they, and why can they not be used? Hint: There are three!

   (b) Does the data set have any missing data? If so, remove the missing data and state how many rows you are now left with. (You will need to overwrite the 'titanic' data, so that it does not contain any missing values).

   (c) Create a table showing how many passengers in the data survived and perished.

2. Next, we want to fit a logistic regression model to predict the probability that a patient survives the disaster. For this question, use passenger class, sex, and age as the explanatory variables.

   (a) Write down the population logistic regression model, making sure to include both the random and systematic components. Use proper notation and include all subscripts. (See "writing math in R Markdown" in the course codebook, `https://sta214-s23.github.io/resources/codebook.html`) Note: passenger class is a categorical variable with more than two levels!

   (b) In R, fit your logistic regression model, and write down the equation of the fitted model. (See "writing math in R Markdown" in the course codebook, `https://sta214-s23.github.io/resources/codebook.html`)

   (c) Interpret each estimated coefficient in terms of the log odds.

   (d) Interpret each estimated coefficient in terms of the odds.

3. Finally, let's use the logistic regression model to make some predictions!

   (a) Suppose we have a 20 year old male passenger, but we don't know their passenger class. Which passenger class (first, second, or third) would have the highest chance of survival? You should answer the question without doing any calculations, and explain your reasoning.

   (b) Suppose we have a 30 year old passenger, but we don't know their passenger class or sex. Which combination of class and sex would have the highest chance of survival? You should answer the question without doing any calculations, and explain your reasoning.

   (c) *In the observed data*, what is the highest predicted probability of survival for a female first-class passenger? (Hint: the `range` function in R will give you the highest and lowest values of a variable or vector)

2 Maximum likelihood estimation

This part focuses on practice with maximum likelihood estimation, and is separate from the Titanic data you were working with above in Part I. A brief review of mathematical notation and rules for logs and derivatives can be found in the course codebook.

Suppose we have a random variable Y, which can take on values $y = 1, 2, 3, 4, ...$ (i.e., any positive integer). The probability of each outcome does not depend on any explanatory variables, and we are told that

$$P(Y = y) = (1 - \pi_0)^{y-1}\pi_0$$

where $0 \leq \pi_0 \leq 1$. So, $P(Y = 1) = \pi_0$, $P(Y = 2) = (1 - \pi_0)\pi_0$, etc.

4. We don't know π_0, but we observe 7 observations:

$$Y_1 = 1,\ Y_2 = 1,\ Y_3 = 3,\ Y_4 = 7,\ Y_5 = 2,\ Y_6 = 4,\ Y_7 = 3$$

   We want to use this observed data to estimate π_0.

   (a) Two friends propose different values for π_0: 0.6 and 0.7. Calculate the likelihoods $L(0.6)$ and $L(0.7)$ for each guess. Which friend made a better guess?

   (b) We want to consider other possible values of π_0. Write down the likelihood $L(\pi_0)$ as a function of π_0, using the observed sequence of data.

   (c) Adapting code from class, fill in the `...` in the following code to make a plot of $L(\pi_0)$ as a function of π_0:

```
pi0 <- seq(0, 1, 0.01)
likelihood <- rep(0, length(pi0))
for(i in 1:length(pi0)){
  likelihood[i] <- ...
}
plot(pi0, likelihood, type="l")
```

   Approximately which value of π_0 maximizes the likelihood?

   (d) Use calculus to calculate the maximum likelihood estimate $\widehat{\pi}_0$ from the observed data. You will:

   - differentiate the log likelihood
   - set the derivative equal to 0
   - solve for $\widehat{\pi}_0$

   Make sure to show all your work.

5. What if we observed a *different* sequence of data? Can we calculate the MLE $\widehat{\pi}_0$ as a general function of the observed data, so we don't have to re-derive it every time we get new data?

   In statistics, we represent an arbitrary set of data by

$$Y_1, ..., Y_n$$

   This means we have n observations, but we don't specify what the values are (to make it more general). Then, the likelihood is

$$\begin{aligned} L(\pi_0) &= P(Y = Y_1)P(Y = Y_2) \cdots P(Y = Y_n) \\ &= \prod_{i=1}^{n} P(Y = Y_i) \end{aligned}$$

Here the shorthand $\prod_{i=1}^{n}$ just means "multiply these n things together". The log likelihood is then

$$\begin{aligned}\ell(\pi_0) &= \log L(\pi_0) = \log\left(\prod_{i=1}^{n} P(Y = Y_i)\right) \\ &= \log\left(P(Y = Y_1)P(Y = Y_2)\cdots P(Y = Y_n)\right) \\ &= \log(P(Y = Y_1)) + \log(P(Y = Y_2)) + \cdots + \log(P(Y = Y_n)) \\ &= \sum_{i=1}^{n} \log(P(Y = Y_i))\end{aligned}$$

where the shorthand $\sum_{i=1}^{n}$ just means "add these n things together".

(a) Show that when $P(Y = y) = (1 - \pi_0)^{y-1}\pi_0$, the log likelihood reduces to

$$\ell(\pi_0) = n\log(\pi_0) + \log(1 - \pi_0)\sum_{i=1}^{n}(Y_i - 1)$$

(b) Find the derivative $\dfrac{d\ell}{d\pi_0}$ of the log likelihood.

(c) By setting the derivative equal to 0 and solving for π_0, show that the maximum likelihood estimate is

$$\widehat{\pi}_0 = \frac{n}{\sum_{i=1}^{n} Y_i}$$

Show all work.

(d) Check that your answer to 4(d) agrees with your answer to 5(c): confirm that your answer to 4(d) is equal to

$$\frac{7}{1+1+3+7+2+4+3}$$

3 Extra credit: debugging practice

This part is separate from parts I and II. The purpose of this section is to practice debugging the errors that we sometimes encounter in R and RStudio. This part is *optional*, and a correct submission will earn a small number of extra credit points on the assignment.

6. Your friend wants to analyze the Titanic data from Part I. They have created an R Markdown file in which they import the data and make a bar chart of survival. Here is a link to their file:

   `https://sta214-s23.github.io/homework/hw_02_part3.Rmd`

   However, when they try to knit the file, they get an error!

   (a) Download their R Markdown file to your computer, and try to knit it (don't change anything in the file yet). What error message appears when you try to knit?

   (b) Google this error, and read some of the links that appear. You may need to make the error message a bit more general (not specific to the Titanic data) to get good results. What is causing this error message to appear when you knit?

   (c) Explain to your friend how they can fix the error in their R Markdown document.