STA 214 Homework 4

Due: Friday, February 10, 12:00pm (noon) on Canvas.

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

Getting started: Begin by downloading the HW4 template from the course website:

`https://sta214-s23.github.io/homework/hw_04_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.

1 Maximum likelihood estimation

This part focuses on practice with maximum likelihood estimation. 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 $y = 0, 1, 2, 3, 4, ...$ (i.e., any non-negative integer). The probability of each outcome does not depend on any explanatory variables, and we are told that

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where $\lambda \geq 0$. (Recall that $y!$ is the factorial, with $0! = 1$, and $y! = y(y-1)(y-2)\cdots 1$). So, $P(Y = 0) = e^{-\lambda}$, $P(Y = 1) = \lambda e^{-\lambda}$, etc.

1. We don't know λ, but we get 6 observations:

$$Y_1 = 2,\ Y_2 = 4,\ Y_3 = 6,\ Y_4 = 6,\ Y_5 = 3,\ Y_6 = 1$$

   (a) Write down the likelihood $L(\lambda)$ as a function of λ, using the observed sequence of data.

   (b) Write down the log likelihood $\log L(\lambda)$ as a function of λ, using the observed sequence of data.

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

   - differentiate the log likelihood (hint: the factorials will go away when you differentiate)
   - set the derivative equal to 0
   - solve for $\widehat{\lambda}$

   Make sure to show all your work.

2. Now let's generalize to *any* sample of data, like in HW 2! Let $Y_1, ..., Y_n$ be n observations. We want to calculate the MLE $\widehat{\lambda}$ as a general function of the observed data $Y_1, ..., Y_n$, so we don't have to re-derive it every time. The work for this problem will feel similar to question 5 on HW 2; the same rules for working with sums and products apply, we just have a different distribution.

   (a) Calculate the log likelihood $\ell(\lambda)$ when we observe data $Y_1, ..., Y_n$ and $P(Y = y) = \dfrac{\lambda^y e^{-\lambda}}{y!}$. Your answer should involve the Y_is, and should involve a sum. *Hints*:

   - Start with $\ell(\lambda) = \sum\limits_{i=1}^{n} \log(P(Y = Y_i))$
   - If you get stuck working with sum notation (the $\sum_i$), see the course codebook, and remember that $\sum\limits_{i=1}^{n} a_i$ is just shorthand for $a_1 + a_2 + \cdots + a_n$

   (b) Using the log likelihood from part (a), find the MLE $\widehat{\lambda}$.

2 Data analysis

Here we work with data from a website called ScienceForums.Net (SFN), which has been open since 2002 and hosts conversations on a range of topics from biological and physical science to religion and philosophy. Each row in the data represents one 'thread', which is comprised of a series of posts stemming from an initial post. For each thread, we have some information that SFN collects such as the number of views and the number of authors. The threads present in the data are a random sample of threads from 2002-2014, with the data collected in 2014. SFN moderators are interested in using this data to determine which threads warrant the most attention.

You can load the SFN data into R by

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

The sfn dataset contains the following columns:

- Age: the age of the thread (in days) when the data was collected in 2014, measured from the first post in the thread
- State: sometimes moderators close threads if they are inappropriate. closed indicates the thread has been closed, otherwise State is open
- Posts: the number of posts in the thread
- Views: the total number of views of the thread
- Duration: the number of days between the first and last posts in the thread
- Authors: the number of distinct authors posting in the thread
- AuthorExperience: the number of days the author of the first post in the thread had been registered on SFN when the thread began (0 indicates they registered that day)
- DeletedPosts: the number of posts in the thread that have been deleted by a moderator

- Forum: the forum in which the thread was posted (e.g., Science)
- AuthorBanned: whether the original author of the thread is currently banned from posting on SFN (at the time of data collection, not when the thread was first posted)

Research question: Suppose you have been approached by moderators at SFN. They give you the data, and ask the following question:

- Is there a relationship between the number of Posts in a thread and whether a thread will have *at least one* deleted post, after accounting for the number of Views, the number of Authors, and the Forum?

3. Here you will use logistic regression to answer the moderators' question.

    (a) Which variables should we focus on to answer the moderators' question? Which of these is our response variable, and which will be our explanatory variables, for logistic regression? *Hint: you may need to create a new variable for the response!*

    (b) Before fitting a model, let's see if any of the quantitative variables need transformations.

    - Create empirical logit plots to summarize the relationship between each quantitative predictor and your binary response. See the class activity and the course codebook for details on empirical logit plots.
    - Using the empirical logit plots, discuss whether any transformations are needed on the explanatory variables.

    (c) Based on your empirical logit plots, write down a logistic regression model that will allow you to answer the moderators' question. Describe how you will use the model to answer their question.

    (d) Fit your model from (d), and report the equation of the fitted model. Interpret any estimated coefficients which address the moderators' question.

    (e) Assess the shape assumption for your fitted model: create quantile residual plots to check the shape assumption for quantitative variables (you may use the `qresid` function in the `statmod` package). See the class activity and the course codebook for details on quantile residual plots.

    (f) If there are any violations of the shape assumption in part (e), experiment with further transformations to address these violations. If you made any changes to your model from (e), report and interpret your new fitted model here.

    (g) Carry out a likelihood ratio test to investigate the moderators' question. You should:

    - State the null and alternative hypotheses in terms of one or more βs
    - Calculate a test statistic and p-value
    - Make a conclusion in the context of the original question

    (h) Carry out a Wald test to investigate the moderators' question (in actual practice, we would not run both a Wald test *and* a likelihood ratio test. But the purpose of this question is for you to get experience with both tests). You should:

    - State the null and alternative hypotheses in terms of one or more βs
    - Calculate a test statistic and p-value
    - Make a conclusion in the context of the original question

    (i) Construct and interpret a confidence interval for the change in the odds that a thread will have at least one deleted post, associated with a one-unit increase in the number of posts in the thread, holding Views, Authors, and Forum fixed. You may choose whichever confidence level you like.

3 Thinking about p-values

Very often, when we test hypotheses we calculate a p-value to summarize how "unusual" the observed data is under the null hypothesis. Statistical hypothesis testing and p-values are widely used in a variety of applications. However, there has long been debate in the statistics community about over-use and misinterpretation of p-values, and the many issues surrounding the focus on p-values in the sciences and social sciences.

Several years ago, the American Statistical Association (ASA) released a statement on p-values:

`https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108`

4. Read the ASA statement on statistical significance and p-values, including the six principles for the use of p-values. Then answer the following questions.
   (a) Why are statisticians concerned about how p-values and statistical significance are used in research?
   (b) How are p-values sometimes mis-used and mis-interpreted?
   (c) What information does a p-value *not* provide?
   (d) What can we do to be more careful when using p-values in our research?

4 Extra credit: debugging practice

This part is separate from the previous parts. 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.

5. Your friend is working with the titanic data, and wants to model the relationship between fare and the survival. After doing some exploratory data analysis, they decide that a log transformation on fare is important. They write the following code in R:

   ```
   library(tidyverse)

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

   m1 <- glm(Survived ~ log(Fare), data = titanic, family = binomial)
   ```

   However, when they run their code, they get an error!

   (a) What error appears when you run this code?
   (b) Google this error, and read some of the links that appear. What is causing the error message?
   (c) Explain to your friend how they can fix the error in their code.