STA 214 Homework 8

Due: Friday, March 31, 12:00pm (noon) on Canvas.

Instructions: In this assignment, you will use count regression to model book purchases from Amazon.

Getting started: Begin by downloading the HW8 template from the course website:

`https://sta214-s23.github.io/homework/hw_08_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.

Data Analysis

In this assignment, you will model the number of purchases of different kinds of books on Amazon. We have a random sample of data from a particular book seller on how many books were purchase from their Amazon store in the last 30 days. Your report will be given to this (imaginary) bookseller who wants you to tell them what kinds of books they might want to stock, or not stock, in their Amazon store for next month. What characteristics might be related to a book that sells a lot of copies? One that sells very few? And so on.

Note: You may assume there are no season buying patterns we need to be aware of (no holiday spending or extraordinary sales). In other words, you can assume there is nothing special about the month of data you have, nor the month you are predicting for, that would impact book buying habits.

Your variables include:

- title: The title of the Book
- author: The author of the book
- rating: An average score the book has received on Amazon.
- purchases: The number of copies of the book purchased in the last 30 days.
- price: The price of the book in US. Dollars.
- publisher: The company that published the book.
- page_count: The number of pages in the book.
- ISBN: a unique numeric identifier for the book.
- published_date: The date the book was published.

- Year: the year in which the book was published
- genre: the book's genre (Fiction, Fantasy, Mystery, Business, General Interest, Comics and Graphic Novels, or Other.

You can load the data into R by

```
books <- read.csv("https://sta712-f22.github.io/homework/books.csv")
```

1. Let's begin with some EDA.
   (a) Create a plot showing the distribution of the number of copies purchased in the last 30 days. Summarize the shape of the distribution, and calculate the mean and variance.
   (b) Create empirical log means plots to explore the relationships between quantitative explanatory variables and the number of purchases. Do you think any transformations are necessary?
   (c) Do you think we need to include an offset in our model? If so, what would the offset be?
   (d) Using your exploratory data analysis, write down an equation for the *systematic* component of a count regression model to predict the number of copies purchased (we will choose a random component in the next question). Include any explanatory variables that you think will be helpful for your client (the bookseller who wants to know which books to stock), and use any transformations you deemed necessary from the empirical log means plots.
2. Next, we need to choose an appropriate model for our response variable (`purchases`). Use a goodness-of-fit test for Poisson data and quantile residual plots to decide between Poisson, quasi-Poisson, and negative binomial models. Use the systematic component from 1(c) for each model.
3. Using your chosen model from question 2, complete model diagnostics for the fitted model:
   - Calculate Cook's distance to check for any influential points (use a threshold of 0.5 or 1 to identify influential points)
   - Calculate variance inflation factors to check for multicollinearity (see the `vif` function in the `car` package, and use a threshold of 5 or 10 to identify high multicollinearity)

   If there are any violations, modify your model to address the violations and report your final fitted model.
4. Using your final fitted model, write a paragraph to your client explaining which factors can help them choose which books to stock. Which characteristics are related to books that sell many copies, and which characteristics are related to books that sell very few?