STA 108 Exam I R Problem
Due Monday, October 17^{th} on Gradescope by 8am.

Directions

- **You are not allowed to discuss the questions with anyone other than the instructor or TA.**
- **Any outside help beyond that from the instructor or TA is considered plagiarism. This including asking a tutor, your classmates (for example, comparing answers), posting the questions to homework help sites, etc. Should we believe you have sought outside help, you will be reported to the Student Judicial Affairs office.**
- **You are allowed to use or modify your previous R code, or the instructors' R code that are posted on Canvas.**
- **Do not share answers, or specific values for calculations, particularly on Piazza**.
- **You may ask clarifying questions about code and general approach through a private message on Piazza.**

Problem: Explore the dataset with Simple Linear Regression.

The data is found in the file `hospital.csv`, with the following columns:

Column 1: `Days`: The number of days a patient has been in the hospital.

Column 2: `Infect`: The estimated probability of an infection for the patient.

This data was collected as part of the SENIC project, and the overall goal is to build a model to predict a patient's risk of infection based on how many days they are in the hospital.

It is requested that you attempt the following predictions:

- Predict the infection risk for a patient who stayed 8 days.
- Predict the average infection risk for a patients who stay 15 days.
- Predict the infection risk for a patient who stayed 40 days.

Format

Submit a short report. This means you should write in full sentences, and have the following sections for each question, while being **as specific as you can** about your results. **There should not be any "copy and pasted" R code in this report. Create an R Markdown document and submit the html (or pdf) file on Gradescope.**

I. Introduction. State the question you are trying to answer, why it is a question of interest (why might we be interested in the answer), and what approach you are going to take (just the name of the approach).

II. Summary of your data. This should include things like plots (histograms, boxplots) including the interpretation of the plots, and summary values such as sample means and standard deviations. You should have an idea about the trend of the data from this section. Display your plots, summary statistics, and scatterplot in this section.

III. Diagnostics. You should discuss the assumptions here. Remove outliers if necessary.

IV. Analysis. Report back the model fit, confidence intervals, test-statistic/s, and p-value/s, nulls and alternatives. Include your predictions here. You may use tables here, but be sure that you organize your work. Remember to write your results in full sentences where possible.

V. Interpretation. State your conclusion, and what inference you may draw from your corresponding tests or confidence intervals. These should all be in terms of your problem.

VI. Conclusion. Summarize briefly your findings. Here you do not have to re-iterate your numeric values, but summarize all relevant conclusions.