Quiz 3

Prepare your response as a python notebook. The quiz is due Thursday evening before midnight.

1) Print your name, email address, and github id on the first line of your code
2) If you want to submit a handwritten image, include that in the notebook. Do not send separate files or links.
3) Include the code you wrote. No credit will be given to showing only the final results.

1. Load the challenger o-ring data from here https://github.com/bcaffo/ds4bme_intro/blob/master/data/orings.csv
    a. Plot whether there was any o-ring failure (total > 0) versus temperature.
    b. Fit a logistic regression curve of o-ring failure (outcome) versus temperature.
    c. Plot the estimated sigmoid curve from the logistic regression fit on your plot from a.
    d. A new oring has a temperature of 74 degrees. What is the estimated probability of failure from your model?
2. Consider a logistic regression model

$$P(Y_i = 1 \mid X_i) = \frac{exp\{\beta_0(1 - X_i) + \beta_1 X_i\}}{1 + exp\{\beta_0(1 - X_i) + \beta_1 X_i\}}$$

    where X_i is either 0 or 1 depending on subject i's group. We observe data pairs (Y_i, X_i) where each Y_i is either 0 or 1 and each X_i is either 0 or 1.
    a. Write out the log likelihood that we would maximize to obtain estimates of the parameters.
    b. Split the log likelihood into two parts, one where $X_i = 0$ and one where $X_i = 1$.
    c. Argue that the estimates are

$$\hat{\beta_0} = \log\left(\frac{\bar{Y}_0}{1 - \bar{Y}_0}\right) \quad \hat{\beta_1} = \log\left(\frac{\bar{Y}_1}{1 - \bar{Y}_1}\right)$$

    where $\bar{Y}_0$ is the proportion of 1's where $X_i = 0$ and $\bar{Y}_1$ is the proportion of 1's where $X_i = 1$.
3. The project will be to create an colab notebook report of a multivariate regression or logistic regression analysis. Pick out a dataset for analysis and print out the first 10 rows as a dataframe.

Hints:

1 Recall that the likelihood (unlogged) for a logistic regression model is

$$\prod_{i=1}^{n} P(Y_i = 1 \mid X_i)^{y_i} P(Y_i = 0 \mid X_i)^{1-y_i}$$

where y_i are the actual values of the response and $P(Y_i = 1 \mid X_i)$ is the model for the probability of a 1 given that particular value of X.