STAT 463 - Assignment 3

Due Data: Saturday, March 11th, 2023

Data Description and Background

An amateur brewer wishes to better understand how the temperature that the beer ferments at (in degrees Fahrenheit) affects the alcohol content of the beer upon completion of brewing. Fortunately, the brewer has kept copious notes on his many brewing endeavors and has records for each batch on what temperature the beer fermented at and what the final alcohol content was.

You are provided with two datasets, one with data from 200 batches and another with data from 50 batches. Call these datasets the modeling dataset and validation dataset respectively. You are to do the following:

1) Using the modeling dataset, visually display the data in an appropriate graph and comment on anything that may be of note. In particular, are the assumptions needed for fitting the simple linear model met? – 1 point
2) Initially, the brewer would just like to get a rough estimate of what the alcohol content would be if he ferments the batch at a given temperature. Are enough of the regression assumptions satisfied so that simple linear regression can be used towards the prior-mentioned goal? If not, what deviations do you need to address and how do you address them? – 2 points
3) After addressing any necessary issues in part 2, fit the simple linear model to the data. Provide the parameter estimates and the R^2 value. Overlay the estimated regression line on the plot created in part 1. – 1 points
4) Now turn attention to the validation dataset. In order to assess how well the model works, calculate the predicted values using the batch temperatures from the validation dataset and the model from part 3. Plot these predicted values versus the actual values (the actual alcohol content) from the validation dataset in a scatter plot. Additionally, calculate the sample correlation between the predicted values and the actual values. – 2 points
5) As is evident from part 1, for a given fermentation temperature, there is a tremendous amount of variability in the alcohol content of the batch. Consequently, for any given temperature, the brewer would like to get bands that encompass what the final alcohol content of a batch would be with probability 95%. Were the steps taken in part 2 enough to still warrant the use of simple linear regression for this goal, or are there still model deviations that need to be addressed? If so, what are the remaining deviations and how do you address them? – 2 points
6) After addressing any additional issues in part 5, obtain a new model for the data. Describe how you arrived at this model. For a given temperature, x, write out the formula for the predicted alcohol content specified for your model. Overlay the estimated regression curve on the plot created in part 1. – 3 points
7) Repeat part 4, this time using the model you obtained in part 6. – 2 points
8) As mentioned in part 5, for any given fermentation temperature the brewer would like to obtain bands that encompass what the final alcohol content of a batch would be with probability 95%. Write out a formula for the upper and lower endpoints for these bands as a function of the explanatory variable (possibly transformed). Overlay these bands on the plot created in part 1. – 2 points