STA 4364 Midterm 1

due Monday October 24 by 11:59PM on Webcourses

Submission Format: Please submit your midterm as both 1) a HTML or pdf document, and 2) also submit the source file in either R Markdown or Jupyter notebook format (at most one of each type of file).
Note: All data in this exam is synthetic data.

Problem 1: (25 points) This problem will involve linear regression on the dataset `midterm_data_1.csv`. The response column is `response` and all other columns are features.

(a) (5 points) Load the dataset. Remove any unnecessary columns. Remove any rows that have an `NA` value. Format the columns for `feat.c` and `feat.g` as categorical variables. Make pairwise plots showing the relations between all columns. Compute the pairwise correlations between all numerical columns. Split the dataset into a training set (75% of observations) and validation set (25% of observations).

(b) (5 points) Make a linear model using all features. How can you interpret the coefficients of `feat.c`? What does R^2 signify? How can you interpret the value of the residual standard error? What does the F-statistic say about your model? Make a residual plot of the residuals vs. fitted values and comment on what this says about validity of the linearity and constant-variance error assumptions of the model.

(c) (5 points) Make a linear model that includes all interactions between features and all quadratic terms for numerical features. From this model, identify a reduced set of coefficients that are the most relevant predictors. Look at the residual plot of your reduced model and comment on any observed differences between this plot and the residual plot from part b).

(d) (5 points) Calculate the MSE value on the validation set using your full quadratic model and your reduced model. Comment on the degree of overfitting compared to the model performance on training data and the adequacy of your reduced model compared to your full model.

(e) (5 points) Using your reduced model, calculate a 95% confidence interval for each validation set prediction (you can do this using the `predict` function in R). Calculate the percentage of true observations from your validation set that fall within your prediction interval. (For this problem, you don't need to print all of the confidence intervals. Please only print the final value of the number of true observations that fall within your confidence interval).