**STA674: Regression Analysis and Design of Experiments**
**Midterm Exam**

All work submitted is done with the assumption that it is the work of the person submitting it. Identical submissions or even submissions found to be **"too close to be coincidental" will be flagged and given no credit**. Please keep in mind my tip: whether asked for or not, you must put your answer in the context of the stated problem to get full credit. Computer code, both R and SAS, are included ***at the end of each question***.
The first three problem are worth 30 points each—longer parts are worth more points. The last problem is worth 10 points. Please submit it via electronic submission onCanvas to your instructor **at or before 11:59pm, March 10, 2022, EST, or within 2 hours of starting, whichever is earlier.**

1. When shopping for a car, several factors can affect the price of the car, such as age and milage. A sample of was take of 32 pre-owned Toyota Prius cars and price, age, and milage were recorded. The other attachment contains your dataset, the R and SAS code, and the corresponding R and SAS output that you will need to answer the following questions:

   Let $x_{1i}$ represent age, $x_{2i}$ represent mileage, and $y_{1i}$ represent price for the $i_{th}$ observation.

   a. If we are only considering mileage as a predictor of price and not including age at this point, which is the response variable and which is the explanatory variable?

   b. State the least squares estimates for the coefficients and interpret those estimates (note, we are not using age in the prediction).

   c. Are the assumptions of the least-squares analysis satisfied? State and check all four assumptions, commenting (briefly) on each—whether each is satisfied or not, giving justification for your conclusion.

2. Regardless of your answer to 1.c., we are going to treat the observations circled on the plot with the title including the words "with 'UNUSUAL?' points" as unusual (not a statistical term).

   a. Using statistical terms (and based only on the graph), would we classify these two observations as influential, outliers, both, or neither? Explain your reasoning.

   b. I made a new data set without those two observations and re-fit the model from problem 1; compare/contrast the Analysis of Variance table the F-value, Root MSE, $R^2$, and the t-Value for the explanatory variable. Comment on how their change (or lack of change) agrees with your response to part a.

3. Using the original data set, we now add the remaining variable, age, in the top row of the data print out.

   a. Is the overall linear model useful in predicting the price? If so, how much of the variance in thesludge is explained by the model?

   b. Does introducing the additional variables change whether the assumptions of the least-squares analysis are satisfied? Justify your response with a few brief comments.

   c. Test the overall fit for the model. State the null and alternative hypotheses (you can write out the words for the hypotheses rather than using the symbols), p-value, and conclusion.

   d. For each significant variable, interpret the parameter estimate in the context of the problem.

   e. Which model (the one-variable or the two-variable) model would you choose? Explain your reasoning.

   f. How would you respond to this statement: "It's not necessary or helpful to include both age and miles together in a model for predicting price since they must be related." Explain your reasoning.

4. Suppose that the manager of a local Kroger wishes to collect data on the following variables from a random sample of transactions/receipts at their store:
   - Total amount spent
   - Number of items purchased
   - Day of week
   - Time of day (morning, afternoon, evening)
   - Payment type (credit card, cash, other)

   a. State a research question that could be addressed by applying linear regression analysis to (some of) the data. Be sure to specify the response and explanatory variables. Use only the variables presented above.

   b. State an additional variable for which data could be collected and classify it as categorical or quantitative, being sure not to change the observational units and response variable from part a.

   c. State the linear model for your final decision of predictors from parts a and b.