Instructions

- Every group will be assigned a dataset containing at least four predictors.
- All statistical analyses must be performed in SAS.
- Every member is expected to contribute to the analysis of the dataset, and written report.
- The deadline to submit the written report (**one report per group**) is Friday, December 8, 2017 at 3:30 pm. **No late submission will be accepted under any circumstances**.
- **The final report should not exceed 25 pages**.
- Please address the following questions in the final report and presentation.

Part I.

1. Choose a predictor of your choice and perform a piecewise SLR to model the relationship with the response variable. Be smart about choosing the point at which the two pieces change slope, if none of the predictors has curved relationship with the response then the center should be your point where the two pieces meet. Report and plot the model, also determine whether the two pieces are the same.

2. In this question you will illustrate some of the ideas related to the extra sums of squares.

   (a) Create a variable called `SUM`, which equals to the summation of any two predictors and run the following two regression models without the variables you used to create `SUM`:

   i. predict the response using all the explanatory variables;

   ii. predict the response using all the explanatory variables including `SUM`.

   Calculate the extra sum of squares for the comparison of these two analyses. Use it to construct the F-statistic – in other words, the general linear test statistic – for testing the null hypothesis that the coefficient of the `SUM` variable is zero in the model with all predictors. What are the degrees of freedom for this test statistic?

   (b) Use the `test` statement in `proc reg` to obtain the same test statistic. Give the statistic, degrees of freedom, p-value and conclusion.

   (c) Compare the test statistic and p-value from the `test` statement with the individual t-test for the coefficient of the `SUM` variable in the full model. Explain the relationship.

3. Run the regression to predict the response using all predictors except the variable `SUM`. Put the variables in any order you wish in the `model` statement. Use the `SS1` and `SS2` options in the `model` statement. Add the Type I sums of squares for the predictors. Do the same for the Type II sums of squares. Do either of these sum to the model sum of squares? Are there any predictors for which the two sums of squares (Type I and Type II) are the same? Explain why.

4. Run the regression to predict the response using a variety of variables, including `SUM` as an explanatory variable, you should have at least 10 different models. Summarize the results by making a table giving the percentage of variation explained (R^2) by each model.

   (Please do not include the SAS output for all these models. Only R^2 and R^2_{adj} values are needed. Note that you can run `proc reg` with multiple `model` statements to save typing.)

Part II.

1. Report the scatterplot and correlation matrix for this data. Summarize the results.
2. Using techniques you learned in class, determine whether the response variable and any of the predictors need to be transformed. Indicate the reasoning for your decision. If a variable needs to be transformed, transform it and keep it in the full model for the rest of the questions.
3. Use the C_p criterion to select the best subset of variables for your data (i.e. use the options "`/ selection = cp b;`"). Use the original and transformed variables, not `SUM`. Summarize the results and explain your choice of the best model.
4. Use the stepwise option to report the best subset of variables for your data (i.e. use the options "`/ selection = stepwise;`"). Use the original and transformed variables, not `SUM`. Summarize the results and explain your choice of the best model.
5. Check the assumptions of this "best" model using all the usual plots (you know what they are by now). Explain in detail whether or not each assumption appears to be substantially violated.
6. Use the "best" model to predict the response variable. Examine other diagnostics such as (but not necessarily exclusively) studentized and studentized-deleted residuals, Cook's D, hat matrix diagonals, tolerance or vif, and partial residual plots. Explain any problems such as outliers, highly influential observations or multicollinearity that these diagnostics point out. (Do not include in your output any tables of values for all observations. Use plots and verbal summaries instead. You may include values for a few selected individuals if you wish.)
7. For the "best" model report the following:
   (a) Equation of the regression model.
   (b) 90% confidence interval for the mean of the response variable
   (c) 90% prediction interval for individual observations.
   (d) 90% confidence intervals for the regression coefficients.

Grading

Evaluation of the final project is based on the following components:

1. Written Report (15%), I will grade the reasoning and accuracy of addressing the questions.
2. Peer Evaluation (5%); each member of the group will have to evaluate all other members.