Mini-project 2 (due 3/7/2021 by 11:59PM).

This mini-project is entirely based on SAS. Follow the SAS program file submission guidelines to submit a single .sas file that completes the problems below. Answers to queries (such as "comment on the results" etc.) should be included as comments in the .sas file. Any requested output with images should be placed in a single Word or pdf file and uploaded as a separate file.

Problem 1 (1 points)

(2 points) Repeat the SAS steps shown in the notes in order to create the SAS dataset ST514.Emissions_HighSTD as shown in the notes, as it appears on pg. 115 of Section 3 Notes, including all created variables such as ODOMETER_K, CYL2, etc. Run the last chunk of code on pg. 115 and provide the output table, showing that your table matches the table on page 116 of the notes. *Use this dataset for the remaining problems*.

Problem 2 (3 points)

a) Create the new variable LNP_E_HIGH_CO = LOG(E_HIGH_CO+0.005). The idea here is to create a new variable that is similar to LN_E_HIGH_CO that was shown in the notes, but note that instances of E_HIGH_CO=0 now remain in the dataset. (we sometimes call this a "zero-protected log-transformation").

b) Run the code shown on pg. 309, except that the dependent variable should be LNP_E_HIGH_CO for both GLMSELECT and PROC REG. From your output, show the scatterplot of residuals vs. predicted values. Comment on how is this different from the analogous plot for LN_E_HIGH_CO shown in the notes on pg. 320. Do you feel comfortable using this procedure to retain all the values with E_HIGH_CO = 0? Why or why not?

c) Show the Cook's D plot for LNP_E_HIGH_CO. Comment on how it is different from the Cook's D for LN_E_HIGH_CO that had been shown in the notes. The new Cook's D includes new values corresponding to responses that had previously had been missing. However, the Cook's D for the previously most influential value (around observation #135) still shows a peak, but the value of D for that observation is not the same as before. Explain how it is possible that the value changed.

Problem 3 (3 points)

a) Run the stepwise selection as shown on pg. 193, but instead (i) using LNP_E_HIGH_CO as the response, and (ii) using the *additional* predictors E_HIGH_RPM E_HIGH_CO2 E_HIGH_O2 E_HIGH_HC E_HIGH_DCF E_HIGH_HC_LIMIT (total 13 predictors), and (iii) using slstay=0.05 slentry=0.05. Show the final analysis of variance and regression table (just the final table).

b) Run the same analysis as above (same response and predictors), but, instead of stepwise, run all subsets regression, choosing the best predictor model (based on R^2) that has the same number of predictors as were selected in the final model of part a). Are the two models the same? Show any necessary clips of output from part b) to demonstrate whether these final chosen models are the same. If they are different, comment on why they may be different.

c) Following part b), report the model (of any size) with the highest adjusted R^2 among all possible models that use the predictors. For this highest-$R_{\boldsymbol{a}}^{2}$ model, provide the regression coefficient table with t-statistics and coefficient p-values. Are all the predictors significant at alpha=0.05? If not, does this fact contradict the fact that the adjusted R^2 is highest? Why or why not?

Problem 4 (1 point)

From the miniproj2 page, read the file ibm.csv into SAS. This file contains recent stock prices in U.S. dollars for IBM. Using the linear model CLOSE=DAY (i.e., closing price depends on trading day), report the Durbin-Watson statistic and its associated p-value for the residuals.