Math 739/839
Homework Assignment #6
Total Points: 33 points

1. The data set BelleAyr_MPV.csv contains results from a study of coal liquefaction using a hydrogenation process. There are seven regressors and the response is CO2 emission from the conversion process. Lower levels of CO2 are preferred.
    a. (4 pts.) Fit a full seven regressor MLR model (main effects only). State the model and discuss the Effect Leverage Plots and Residual Plots. In your discussion, include any issues that you find with either the model adequacy or possible data issues such as outliers and influential points.
    b. (3pts.) Examine the scatterplot matrix for the seven regressors and comment on the possibility of multicollinearity between some of the regressors. Which regressors are involved with the most severe multicollinearity. Is there a possible simple way to mitigate that multicollinearity in this case? Note, scatterplots can also be useful in spotting outliers and other data anomalies.
    c. (1 pt.) Generate the VIF values in the Fit Model Parameter Estimates table for the full model. Is there evidence of significant multicollinearity? Explain.
    d. (1 pt.) Obtain the RStudent, Hats, Cook's DF, DFFITS, and COVRATIO. Display the information in a data set or table.
    e. (5 pts.) Using the information, you have collected for parts a through d, identify and discuss any points that are potentially highly influential. For this observation or these observations, give a physical interpretation of what the RStudent, Hat, Cook's D, DFFITS, and COVRATIO mean i.e., what do these statistics measure for each observation?
    f. (2 pts.) Obtain the DFBETAS and identify which coefficients are most influenced by that observation or observations identified in part e.
    g. (1 pt.) From the analyses you have performed so far, state the observation that has been identified as most influential and has outlier potential. For sake of time, if you have found multiple influential observations and potential outliers, then only exclude the most influential one.
    h. (1 pt.) With the observation excluded, refit and state the model.
    i. (5 pts.) Generate the statistics called for in part d. Discuss any new highly influential observation that you have identified and comment on the possibility of being an outlier(s) – do not assume that new highly influential observations necessarily exist. Also, discuss changes in the parameter estimates, standard errors and predicted values. Are these changes surprising given your influence diagnostic values prior to excluding the observation in part h?
    j. (3 pts.) Generate the VIF values in the Parameter Estimates table with the observation excluded. Have the VIF values changed significantly after excluding the observation? Based upon your experience with this data set, comment on the effects that an influential point can have on the ability to detect multicollinearity.

Math 739/839
Homework Assignment #6
Total Points: 33 points

2. Hmwk6.2 presents data on the quality of Pinot Noir wine.

    a. (2 pts.) Build a regression model relating quality, y, to flavor x4 that incorporates the region information given in the last column. Does the region have an impact on wine quality?

    b. (2 pts.) Perform a residual analysis for this model and comment on model adequacy.

    c. (2 pts.)Are there any outliers or influential observations in this data set?

    d. (3 pts.) Modify the model in part a to include interaction terms between flavor and the region variables. Is this model superior to the one you found in part a?