DSC 423: Data Analysis and Regression
Assignment 04: Model Building

Your submission must include your name and student ID. Your submission must include the honor statement: "I have completed this work independently. The solutions given are entirely my own work." Your submission must be submitted as a PDF.

1. WATEROIL. In the oil industry, water that mixes with crude oil during production and transportation must be removed. Chemists have found that the oil can be extracted from the water/oil mix electrically. Researchers at the University of Bergen (Norway) conducted a series of experiments to study the factors that influence the voltage (y) required to separate the water from the oil (Journal of colloid and interface science, Aug. 1995). The seven independent variables investigated in the study are listed in the table. (Each variable was measured at two levels - a "low" level and a "high" level.) Sixteen water/oil mixtures were prepared using different combinations of independent variables; then each emulsion was exposed to a high electric field. In addition, three mixtures were tested when all independent variables were set to 0. The variables are given in the table below.

   Experiment number
   y: voltage (kw/cm)
   x1: disperse phase volume (%)
   x2: salinity (%)
   x3: temperature (^{0}C)
   x4: time delay (hours)
   x5: surfactant concentration (%)
   x6: span:triton
   x7: solid particles (%)

Use R to perform a regression analysis on the WATEROIL dataset (found on the D2L). Consider interaction terms and second-order terms.

a. (10 pts.) Paste your final model into your submission (just the R output).
b. (10 pts.) Describe the model building process through which you generated this model.
c. (10 pts.) What significant second-order terms did you find, if any? Did you try all second-order terms? Did you look at scatter plots to determine which second-order terms to evaluate? Discuss the benefits and drawbacks of these two strategies.
d. (10 pts.) What significant interaction terms did you find, if any? Did you try all combinations of interaction terms? Do you think that is an appropriate strategy? What happens to the number of interaction terms as the number of independent terms increases?
e. (10 pts.) Discuss your final model. Evaluate the t-tests, F-Test and adj-R^2 accordingly. Do you think this is a “good” model? Explain.