Project 4

We would like to conduct a chemical experiment with twenty predictors, $x_1, ..., x_{20}$, where the first three are 0/1-variables for different conditions and the remaining seventeen are amounts (proportions of some predetermined maximum level) of usage of different substances added to the process. The response y is the chemical yield and higher yield is better. Consider the full linear model with all twenty predictors

$$E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_{20} x_{20}.$$

(a) Explore the data by looking at different plots etc.

(b) Split the data set ($n = 400$) into a training set and a test set.

(c) Fit the full model using least squares on the training set, and report the test error obtained.

(d) Find a suitable linear model by subset / stepwise selection.

(e) Fit a PLS model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

(f) Consider your proposed models. Examine the residuals to look for any systematic effects.

(g) Comment on the results obtained. How accurately can we predict the response variable? Is there much difference among the test errors resulting from the different approaches?

(h) Write a 10 min presentation explaining your model, analysis method (e) and results.