Assignment 4

1. Materials scientists working for a cement mixing company explored the ingredients in the cement impacted the heat of the cement during hardening. They investigated the impact of 5 composition variables (see below) on the heat during hardening (calories/gram). The data can be found in the file cement.csv on Canvas.

Regressor	Description
x_1	% tricalcium aluminate
x_2	% tricalcium silicate
x_3	% tetracalcium alumino ferrite
x_4	% dicalcium silicate
x_5	% sodium oxide

Use only the lm() command in R for this question. Do not use any variable selection functions or downloaded packages in R (e.g., step()).

a. Use forward selection to identify the "best model". Use a (conditional) t-test with $\alpha_{IN}=0.10$ as the cutoff to decide whether or not to include a variable in the model. At each step in the forward selection, state the variable that was added. At the end of the procedure, state the final model.
b. Use backward elimination to identify the "best model". Use a (conditional) t-test with $\alpha_{out}=0.10$ as the cutoff to decide whether or not to remove a variable from the model. At each step in the procedure, state the variable that was removed from the model. At the end of the procedure, state the final model.
c. In the first step of the forward procedure in part a, both x_4 and x_5 has small p-values. However, in the second step of the forward selection procedure, when both x_4 and x_5 were considered together in the model, neither of the variables had a p-value less than α_{IN}. Why do you suppose this happened?
d. Use stepwise regression to identify the "best model" using t-tests with $\alpha_{IN}=\alpha_{out}=0.10$. At the end of each step, state the variable that was removed or added from the model. At the end of the procedure, state the final model.
e. What is the AIC for the models in a, b and d?

2. Consider the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon.$$

State the appropriate coditional or extra sums of squares (e.g., see section 3.3.2) and F-statistic (with degrees of freedom) for testing the following hypotheses:

a. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
b. $H_0 : \beta_4 = \beta_5 = 0$, given that the first 3 variables are in the model

3. A tube flow reactor is a chemical process in which materials in typically a series of tubes create desired chemical products. A experiment was conducted to study the relationship between the concentration of a desired product, $NbOCl_3$ (y), and the concentration of $COCl_2$ (x_1), reaction time (x_2), molar desnity (x_3), mole fraction (x_4), tube temperture (x_5), ramp-up time (x_6) and product density (x_7). The data can be found in reactor.csv on Canvas.

a. For each $p = 2, \ldots, 5$, find the maximum R^2. Plot the maximum R^2 versus p. Which model would you choose for these data and why?
b. For each $p = 2, \ldots, 5$, find the minimum MS_{Res}. Plot the minimum MS_{Res} versus p. Which model would you choose for these data and why?
c. For each $p = 2, \ldots, 5$, find best model using Mallow's C_p. Plot the best C_p versus p. Which model would you choose for these data and why?