Linear Regression
HW 6
Due 10/6 at 11:59pm

Directions: Submit a .pdf file containing your responses for the homework. The .pdf can be converted from a Latex file, pictures of your handwritten solutions, word files, converted markdown files, jupyter notebooks, etc.

Understanding model selection criteria.

1. Say whether the following statements are true or false and explain why.
   (a) For *any* set of predictor variables, the larger the number of predictor variables in the model, the larger the R^2.
   (b) For model of the same size (fixed p), their C_p, AIC_p, BIC_p values are monotonically increasing in terms of SSE_p.
   (c) For model of the same size, their $PRESS_p$ values are monotonically increasing with SSE_p.
   (d) Compared with AIC, BIC criterion tends to select smaller models because it puts higher penalties on model size.
   (e) The best subsets procedure is guaranteed to find the "best" model under a given criterion.

Coding Questions.

2. Practice model selection on an example data set. Data set "HW6Q2.txt" contains 4 variables with the response variable Y on the first column followed by 3 predictor variables. We consider all first-order models.
   (a) How many first-order models are there?
   (b) Among all the first-order models, report the "best" models according to each of the following criteria: $R^2_{adj,p}$, AIC_p, BIC_p, C_p, $PRESS_p$, as well as their corresponding values according to the criterion.
   (c) Using AIC_p, select the best overall model of any size. Using this model, check for influential points using Cook's Distance. If there are any, print them out.

3. For the data set IceCreamConsumption.csv and consider y=cons with predictors income, price, and temp.
   (a) List all the possible models from this data set (without interactions or higher powers).
   (b) Calculate the adjusted R^2 and C_p for all the models, make a summary table with four columns: Number of predictors, R^2 values, C_p values, Predictors in the model.
   (c) Based on the table above, pick a pool of candidates.
   (d) For the all the candidates in the pool, calculate AIC and BIC values. Based on AIC and BIC, what's your final choice of model?
   (e) Is there a difference in the size of the model selected by AIC and BIC? If yes, state which is more parsimonious and explain why this difference exists.