Homework 4

Stat 151A, Fall 2017

Due: November 6

1. (2.5 points) Let $\hat{\beta}$ be the vector of coefficients from the regression of Y on X, with residuals $\hat{e}$. Let $b^{(p)}$ be the slope from the simple linear regression of $Y^{(p)}$ on $X^{(p)}$, with residuals $e^{(p)}$ (See Handout "Regression Diagnostics 4"). Then $e^{(p)} = \hat{e}$.

2. Consider the bodyfat dataset and consider fitting a linear model for the response variable BODYFAT in terms of the explanatory variables AGE, WEIGHT, HEIGHT, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, ANKLE, BICEPS, FOREARM and WRIST.

   (a) Using each of the following methods, perform variable selection to select a subset of the explanatory variables for modeling the response:

      i. (0.5 points) Backward elimination using the individual p-values.
      ii. (0.5 points) Forward Selection using p-values.
      iii. (0.5 points) Adjusted R^2.
      iv. (0.5 points) AIC
      v. (0.5 points) BIC
      vi. (0.5 points) Mallow's C_p.

   (b) (1 points) Let $M_1, \ldots, M_6$ denote the six models selected by each of the six variable selection methods of the previous part. Select one of these models by cross-validation.

   (c) Let M be the model selected in the previous part.

      i. (0.5 points) Fit this model to the data.
      ii. (1.5 points) Perform regression diagnostics.
      iii. (0.5 points) Comment on the validity of the assumptions of the linear model.
      iv. (0.5 points) Identify influential observations and outliers.
      v. (0.5 points) Delete them if necessary and re-fit the model.