Homework 3

1. Let x be the sequence $-m, -m+1, \ldots, m$ and consider fitting a straight line of the form $y = \beta_0 + \beta_1 x$ to data (y_j, x_j).

   (a) Find the Hat matrix and determine with points have highest and lowest leverage.

   (b) Which fitted value has largest variance? Which has smallest variance?

   (c) Which residual has largest variance? Which has smallest variance?

2. An ecologist friend of yours measures the amount of oxygen, y, emitted from a planted area as a function of temperature, x and fits a straight line to a scatterplot of y versus x. The relationship is positive. He then plots the residuals versus the observed values, y, and finds that they are positively correlated – the residuals for low values of oxygen are negative and those for large values are positive. He finds this puzzling and disturbing and wonders if he is doing something wrong or if this is evidence of model misfit. What would you say to him?

3. Show that the leave-one-out residual, $\hat{e}_{(i)} = \hat{e}_i/(1 - p_{ii})$.

4. Let $x = (-20, -9, \ldots, 9, 20)$ and let $Y_i = 1 + x_i + e_i$, where the e_i are independent, normally distributed random variables with means zero and variance 4. Simulate data of this form and from the simulation,

   (a) Find R^2 and the RMS error. (RMS error is the root-mean-square error, or s. It measures the accuracy of the predictions of the model).

   (b) Find R^2 and the RMS error when the regression is performed on the middle third of the data.

   (c) Leave out the middle third and find R^2 and the RMS error.

   (d) Using only $x = (-20, -19, -18, 18, 19, 20)$, find R^2 and the RMS error.

   Explain the results.