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?

   $X = [X_1 X_2]$ and the columns are orthogonal, so

   $$H = P = \frac{X_1 X_1^T}{\|X_1\|^2} + \frac{X_2 X_2^T}{\|X_2\|^2}$$

   The first projector has entries all equal to n^{-1}. The second has ij entry equal to $X_2(i)X_2(j)/\sum_{k=-m}^{m} k^2$ Thus the points with highest leverage are $x_2 = \pm m$ and the point with lowest leverage is $x_2 = 0$. Since $\Sigma_{\hat{Y}\hat{Y}} = \sigma^2 H$, the highest and lowest leverage points have highest and lowest variance of fitted values. Since $\Sigma_{\hat{e}\hat{e}} = \sigma^2(I - H)$ the highest and lowest variance residuals are those for which the leverage is lowest and highest. We thus expect the residuals at the extremes to be small relative to those in the middle.

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?

   There is not necessarily a problem. The residuals have to be correlated with the observed values, because the vector of residuals is not orthogonal to the vector of observations.

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

From lecture,

$$\begin{aligned} Y - \hat{Y}_{(i)} &= (Y - \hat{Y}) - (\hat{Y}_{(i)} - \hat{Y}) \\ &= \hat{e} + \hat{e}_i \frac{X(X^T X)^{-1} x_i^T}{1 - p_{ii}} \end{aligned}$$

Let u_i be the ith unit vector. Then

$$\begin{aligned} \hat{e}_{(i)} &= u_i^T (Y - \hat{Y}_{(i)}) \\ &= \hat{e}_i + \hat{e}_i \frac{u_i^T X (X^T X)^{-1} X^T u_i}{1 - p_{ii}} \\ &= \hat{e}_i + \hat{e}_i \frac{p_{ii}}{1 - p_{ii}} \\ &= \frac{\hat{e}_i}{1 - p_{ii}} \end{aligned}$$

4. Let $x = (-20, -19, \ldots, 19, 20)$ 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.

$$R^2 = \frac{\|Y\|^2 - \|Y - \hat{Y}\|^2}{\|Y\|^2}$$

Where Y is the centered data. Divide numerator and denominator by n and observe that in all the scenarios $n^{-1}\|Y - \hat{Y}\|^2$ is approximately equal to the residual variance, $\sigma^2 = 4$. However, the empirical variance of Y is $n^{-1}\|Y\|^2$ and the variances are different in the different scenarios. R^2 is thus increased when the spread of the data is increased, even though the underlying predictive accuracy of the model does not change.