Homework 7

Due Tuesday, November 12

Show all your work. Data files available from `alr4` package.

1. Consider a model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, with $E(\mathbf{e}) = \mathbf{0}$ and $\text{Var}(\mathbf{e}) = \sigma^2\mathbf{I}$. Let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the hat matrix.
   (a) Show that $\mathbf{I} - \mathbf{H}$ is symmetric.
   (b) Show that $\mathbf{I} - \mathbf{H}$ is idempotent.
   (c) Show that $\mathbf{I} - \mathbf{H}$ and $\mathbf{H}$ are orthogonal.

2. Consider again the same set up as Problem 1. Show that
$$\text{Var}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I} - \mathbf{H}) \quad \text{and} \quad \text{Var}(\hat{\mathbf{Y}}) = \sigma^2\mathbf{H}$$

3. Show that the slopes of the regression of $\hat{\mathbf{e}}$ on $\hat{\mathbf{Y}}$ are 0.

4. Consider the dataset `mantel`. Construct the hat matrix $\mathbf{H}$ and verify the properties of $\mathbf{H}$ (use R; follow Example 1 of Nov 5 Lecture).

5. In the fuel consumption data `fuel2001`, consider the model (see Example 2 of Nov 5 Lecture)
$$Fuel = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 LogMiles + e$$
   (a) Report the regression summary. Produce a residual plot (with fitted versus residual only) and a QQ-plot. Comment.
   (b) Compute h_{ii} (hat diagonals) , D_i (Cook's distance) , and t_i (studentized residual), using matrix algebra. Test for outliers using t_i.
   (c) Verify the numbers above by using `studres()` and `influence.measures()`. Plot the three quantities above as we did in the lecture (also see Chapter 9). Which one would you consider as outlier(s)?

6. The data in the file `florida` has four variables, *County*, the county name, and *Gore*, *Bush*, and *Buchanan*, the number of votes for each of these three candidates. (The data consists of 2000 US presidential data for Florida).
   (a) Draw the scatterplot of Buchanan versus Bush (Y versus X), and test the hypothesis that Palm Beach county is an outlier (with an appropriate regression). Identify another county with an unusual value of the Buchanan vote, given its Bush vote, and test that county to be an outlier. State your conclusions from the test, and its relevance, if any, to the issue of the butterfly ballot.
   (b) Next, repeat the analysis, but first consider transforming the variables in the plot to better satisfy the assumptions of the simple linear regression model. Again test to see if Palm Beach County is an outlier and summarize.