STAT 451

Hand-in Homework 4

Problems 1-2 due Wednesday, Oct. 7 by midnight; problem 3 (grad) due Fri, Oct. 9 by midnight.

Problems 1 and 2 are to be done in a single R Markdown document. Problem 3 is for graduate students and is to be done in a separate R Markdown document and submitted separately. Use sentences to report results; don't just give a value. Remember context (your answers should not have "x" or "y" in them). Use inline R code to report values. Remember to include units where appropriate. Submit only the Rmd document.

1. The file Kites.csv contains tail length and wing length measurements (in mm) on 90 hook-billed kites. It also contains the sex of each kite, but please ignore sex for this problem. These data were taken directly from a statistics textbook.

   (a) Plot tail length (y-axis) versus wing length (x-axis). There's an obvious outlier. If you were to do a least squares regression to predict tail length from wing length, would this point be considered high leverage? Would it be influential? Justify your answers based on the plot (not on actually doing any regressions).

   (b) Now reverse the axes and plot wing length versus tail length. If you were to do a least squares regression to predict wing length from tail length, would this point be considered high leverage? Would it be influential? (Note: this data point was obviously a typo.)

2. This problem uses a data set from a package called Stat2Data that you will need to install. It's a package that contains a whole bunch of data sets. Once you install it and load it with the command `library(Stat2Data)` you can see the whole list of data sets by typing `?Stat2Data` in the console and clicking on the Index link at the bottom of the help file.

   The data set you will use is called `ButterfliesBc`. To access it, you run the commands:

   ```
   library(Stat2Data)
   data(ButterfliesBc)
   ```

   `ButterfliesBc` is now a data frame available to you (you'll see it in the list of objects in the upper right window). Read the description of it with the command `?ButterfilesBc`. The goal is to examine the association between average yearly temperature and average butterfly wing length for males and females over the 17-year length of the study. Wing length is the response variable and temperature and sex are the explanatory variables.

   (a) Make a scatterplot of wing length vs. temperature, ignoring sex for now, and briefly describe the relationship. Fit a regression model for predicting wing length from just temperature. Report (using inline code) and interpret the value of the slope coefficient. Report the values of R^2 and s_e (using inline code).

(b) Make a scatterplot like (a) but with sex now included as separate colors. Include the least squares lines. Fit a regression model which has different intercepts, but same slope, for males and females. Interpret the coefficients for temperature and sex in the context of the problem. Report the values of R^2 and s_e. How do they compare to the model in (a)?

(c) Fit a regression model which has different intercepts and different slopes for males and females. What are the resulting slopes for males and females according to this model? Report the values of R^2 and s_e. Does it appear that having different slopes is important?

(d) Make a residual plot and a histogram of residuals (side-by-side)for the model in part (b). Do the regression assumptions appear satisfied?

3. (grad only) The file `Real_Estate.csv` contains the selling price and other information on 894 houses. In this problem, you'll only consider three of the variables: Price (in $), size (in sq. ft.), and location, a categorical variable with 3 levels: `r` for rural, `s` for suburban, and `u` for urban. The goal is to develop an equation for predicting Price from size and location.

(a) Create a scatterplot of Price vs. size with color of the points based on location. Add loess lines. Does the relationship between size and price appear linear for each location?

(b) As a further check, fit a linear regression model with Price as the response variable and size and location as the explanatory variables. Don't include the interaction between size and location. Give the residual plot and a histogram of the residuals. Do the regression assumptions seem satisfied?

(c) Repeat parts (a) and (b), using $\log_{10}$(Price) as the response variable. Do the regression assumptions seem reasonably satisfied?

(d) Interpret each of the coefficients (except the intercept) in the model in (b) in the context of the problem in terms of their effect on predicted Price. Use inline code.

(e) Now fit a model which includes the interaction between size and location. Do the separate slopes appear much different from each other? Are s_e and R^2 much different from the same slopes model?