Homework 7

36-401, Fall 2015, Section B*

Due at the start of class on 5 November 2015

1. Recall the dataset `mobility` from the first DAP. In this problem, we will still predict economic mobility (Y_i) from the proportion of people with short commutes (X_i), considering the following three models:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad (1)$$
$$Y_i = \beta_0 + \beta_1 X_i^2 + \epsilon_i \quad (2)$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i \quad (3)$$

   (a) (15) Fit the three models above, and explore the model diagnostics (no need to include Box-Cox plots or make any transformations). Which would you ultimately choose? Why?

   (b) (10) Find the correlation between X_i and X_i^2. Then center the values of X_i around the mean, i.e. create $Z_i = X_i - \bar{X}$, and recalculate the correlation between the centered Z_i and Z_i^2. What (if anything) changes? Why?

   (c) Refit model (3) using Z_i and Z_i^2. Compare the estimated coefficients and standard errors before and after centering. Does this make sense? Is using centered variables helpful here?

2. The data file `water.txt` contains over four decades of data on precipitation at six locations in the Sierra Nevada mountains of California (APMAM, APSAB, APSLAKE, OPBPC, OPRC, OPSLAKE), and the stream volume at a site near Bishop, California (BSAAM). We want to know how well we can predict the stream volume from the precipitation.

   (a) (10) Create a pairs plot of all seven variables. Run six simple regressions of BSAAM on each of the precipitation variables. Report the regression coefficients and their standard errors, and the root mean square error (= residual standard error) of each regression.

   (b) (5) Regress BSAAM on all six precipitation variables. Report the coefficients, their standard errors, and the root mean squared error.

*There are big differences between the homework in section A and section B this week.

(c) (5) Using the standard errors of the coefficients from problems 2a and 2b, and the root mean squared errors of those regressions, calculate the variance inflation factor for each slope coefficient.

(d) (10) Regress each of the precipitation variables on the other five. (That is, run six regressions.) Report the coefficients, the root mean squared error, and the R^2. Which variable is most predictable from the others?

(e) (5) Create three new variables: AVG = (OPBPC + OPRC + OPSLAKE)/3, DELTA1 = OPBPC - AVG, and DELTA2 = OPRC - AVG. Provide the pairs plot for BSAAM, AVG, DELTA1 and DELTA2.

(f) (5) Regress BSAAM on AVG. Report the coefficients, the standard error, and the root mean squared error. Give an interpretation of the slope coefficient.

(g) (5) Regress BSAAM on AVG, DELTA1 and DELTA2. Report the coefficients, the standard errors, and the root mean squared error. Give an interpretation of the slope coefficient for AVG.

(h) (10) Regress BSAAM on OPBPC, OPRC and OPSLAKE. Plot the fitted values against those of the model from problem 2g. Explain your results.

Rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision. Code is either properly integrated with a tool like R Markdown or knitr, or included as a separate R file. In the former case, please only submit the knitted file. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names. All code is relevant to the text; there are no dangling or useless commands. All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output.