36-617: Applied Linear Models
Fall 2022
HW04 – Due Mon Sept 26, 11:59pm

- Please turn the homework in online in our course webspace at canvas.cmu.edu.
  - There is a link to Gradescope in the description of this assignment on Canvas.
  - You should submit a single pdf to Gradescope. If you need help with this, please see https://www.cmu.edu/teaching/gradescope/index.html. Also, allow yourself some extra time to create the pdf & upload it in Gradescope.
  - Don't forget to identify which pages each (part of each) problem appears on in your solitions. Gradescope allows the TA to grade all the problem 1's together, then all the problem 2's, and so forth. This leads to more consistent grading and better comments for you.
  - Remember to list who you worked with, on this and every assignment.
- Reading:
  - For this week: Sheather 6.4, 6.5, 6.6, 7.1, 7.2 (Supplemental: ISLR 3.3.3; G&H Ch 4)
  - For next week: Sheather, 7.3, 7.4, 8.1, 8.2 (Supplemental: ISLR 3.3.3, & Ch 6; G&H Ch 4)
- There are 3 exercises below. Except where noted below, data sets will be in the "0-books" folder in the files area of our Canvas site (in the "data-sheather" subfolder).

Exercises

1. Sheather, Ch 6, pp. 216–221, #3. Notes:
   - The data set is available as "cars04.csv".
   - There are some errors in the results for fitting the model (6.37) in the textbook, so I suggest you refit the models for this problem directly from the data in "cars04.csv".
   - Figure 6.57 on p. 221 goes with this problem, not with Sheather's problem #4.
2. Sheather, Ch 7, p. 261, #3. Note:
   - The data is available in "pgatour2006.csv".
3. [Based on Gelman & Hill (2009), p. 51, #5] The subfolder `beauty` in the hw04 folder in the "Files" area for our course on canvas contains data from Hamermesh and Parker (2005) on student evaluations of instructors beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. Various documents in the folder give background and some variable definitions (some variables are defined in the ".log" file there, others' definitions you will have to deduce from pdf's in the subfolder).
   (a) Fit a regression model predicting `courseevaluation` (average student evaluations) from `btystdave` (the average of 6 standardized beauty ratings for each instructor) and `female`. Then fit the same model with the interaction between `btystdave` and `female` added in.
      i. Graph each fitted model on a scatter plot of `courseevaluation` vs `btystdave`. Indicate clearly in the graph what the various parameters in the model represent geometrically.

ii. Display the four standard diagnostic plots in R and comment on their features, for each model. Comment on whether the fit seems adequate from the evidence in these plots, for either model. In case there are problems with the fit, indicate what they are and how you might improve things.

iii. Produce summaries of the two fitted models; comment on the coefficient estimates and their standard errors, and on R^2, for each model Use a partial F test to determine whether the interaction should be kept. *Your comments should include not only technical points ("B" in the "ABA^{-1}" metaphor for applied statistics from the course syllabus), but also what it means for understanding how factors may influence course evaluations ("A^{-1}").*

(b) Now let's look at *all* of the variables in the data set. Should any of the variables in the data set be transformed before being used in a regression model? List each variable that is not a dummy variable, and for each of these,

- Say whether the variable should be transformed (yes or no)
- If yes, indicate what transformation you would make
- Justify these two answers, using both evidence from the data and other considerations

Note: being able to communicate with a client or collaborator matters, so there may be instances where either (a) a transformation might help, but you decide against it since it would be difficult to explain to a client/collaborator, or (b) an automatic method like Box-Cox might suggest one power, but you pick a simpler power "nearby" because it is easier to explain to a collaborator/client.

(c) Fit the model that regresses `courseevaluation` onto all other variables, except for `profnumber`, `multipleclass`, and the 30 `class` variables (`class1` through `class30`). Use the transformations you recommended in part (b). Make a table indicating

- The t-statistics for each variable
- The VIFs for each variable

in your model.

(d) On the basis of this table, and what you know about the definitions of the variables, would you eliminate any variables in your model? Why or why not?

(e) Why might the methods used in parts (c) and (d) not be adequate for deciding which variables to keep, and which ones to eliminate, in a regression model?