36-617: Applied Linear Models
Fall 2021
HW05 – Due Mon Oct 4, 11:59pm

- Please turn the homework in online as a pdf to Gradescope, using the link provided on the assignment page in canvas.cmu.edu, under Assignments.
- This week we are discussing Ch 6 of Sheather. Next week we will move on to Ch 7.
- Data files (where needed) for these exercises are in the "0 - textbooks" folder in the files area on canvas, unless otherwise noted below.
- There are two major exercises below; each one has "parts".
- Work on Project 01 will begin with hw06.

Exercises

1. Sheather, Ch 6, p 224, #5 (look familiar?).
2. Return to the "beauty" data from Gelman & Hill (2009), p. 51, #5 (data and documentation in the hw05 folder on canvas).

   (a) 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.

   (b) 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 (a). Make a table indicating

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

   in your model.

   (c) 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?

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