36-617: Applied Linear Models
Fall 2018
HW02 – Due Mon Sept 14, 11:59pm

- Please turn the homework in online in our course webspace at canvas.cmu.edu.
  - For this assignment you will your work in Gradescope. There is a link to Gradescope in the hw02 description 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.
  - 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.
- Data files (where needed) for these exercises are in the "0 - textbooks" folder in the files area on canvas.
- Reading:
  - You should have read Chapters 1 and 2 in Sheather already. Please read Chapter 3 for next week (week of Sept 14); there will be a reading quiz on Monday as usual.
  - We are skipping Sheather Ch 4 for now, and proceeding to Ch 5, for the following week (the week of Sept 21).
- There are five exercises.

Exercises

1. Sheather, Ch 2, p. 42 #6
2. [Gelman & Hill (2007), Ch 3, #3] In this exercise you will simulate two variables that are statistically independent of each other to see what happens when we run a regression of one on the other.
   (a) First generate 1000 data points from a normal distribution with mean 0 and standard deviation 1 by typing `var1 <- rnorm(1000,0,1)` in R. Generate another variable in the same way (call it `var2`). Run a regression of one variable on the other. Is the slope coefficient statistically significant?
   (b) Now run a simulation repeating this process 100 times. This can be done using a loop. From each simulation, save the z-score (the estimated coefficient of `var1` divided by its standard error). If the absolute value of the z-score exceeds 2, the estimate is statistically significant. Here is code to perform the simulation[1]:

```
z.scores <- rep (NA, 100)
for (k in 1:100) {
  var1 <- rnorm (1000,0,1)
  var2 <- rnorm (1000,0,1)
  fit <- lm (var2 ~ var1)
  z.scores[k] <- coef(fit)[2]/se.coef(fit)[2]
}
```

   How many of these 100 z-scores are statistically significant?
   (c) Is your answer to (b) what you expected? Why or why not?

[1] We have initialized the vector of z-scores with missing values (`NA`s). Another approach is to start with `z.scores <- numeric(length=100)`, which would initialize with a vector of zeroes. In general, however, we prefer to initialize with `NA`s, because then when there is a bug in the code, it sometimes shows up as `NA`s in the final results, alerting us to the problem

3. Sheather, Ch 2, pp 42–43, #7
4. Sheather, Ch 3, pp 112ff, #8
5. In the folder for this hw assignment you will find a pdf called “An IMRAD paper on wine.pdf”, based on Example 1.2.4 in Sheather. This is another data analysis based only on EDA, not on any more sophisticated methods.

   (a) Does the paper appropriately address each of the parts of an IMRAD paper as described on slide 3, lecture 02 from week01 of class? If you need more detail on the sections of an IMRAD paper, see `http://www.jpgmonline.com/documents/author/24/2_Aggarwal_10.pdf`.

   For each section below, either say “yes this section has the right content”, or say “no” and describe what is missing and/or what needs to be moved to another section of the paper or deleted.

   - Abstract
   - Introduction
   - Methods
   - Results
   - Discussion

   (b) As described on slide 5 of the same lecture, an IDMRAD paper is a bit different in that there is a **Data** section between the **Introduction** and the **Methods** sections, devoted to a description of the data used in the paper.

   Please write the appropriate **Data** and **Methods** sections for this paper (just those two sections), as an IDMRAD paper instead of an IMRAD paper (include appropriate text, figures and tables in the two sections). For your convenience the three figures and three tables of this paper are saved as jpg files in the folder for this assignment.