36-617: Applied Linear Models
Fall 2018
HW01 – Due Tue Sep 7, 11:59pm

- Most HW will be due on Mondays at 11:59pm, but since Sep 6 is a holiday, this assignment is due the next day.
- Please turn the homework in online to GradeScope in our course webspace at canvas.cmu.edu. The easiest way to do this is to go to the Assignments area of our Canvas webspace, then click on HW01, then use the link there to submit to Gradescope.
  - Please upload only *one* file (pdf format ONLY) for each homework assignment. If you are an RStudio user, the easiest way create a pdf file to submit is to make an RMarkdown file for your homework solutions, and then "knit" it to pdf. Other approaches such as making an MSWord file or LaTeX file and converting it to pdf, or writing things out by hand and scanning to pdf with your phone, etc., are also acceptable.
  - If you need additional 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:
  - For this week, you should be reading Chapters 1 and 2 in Sheather.
  - For next week, please take a look at the handouts in the week02 folder in the files area on canvas.
  - On Sep 13 we will resume with regression; for this see Chapter 3 of Sheather.
- There are five exercises.

Exercises

1. Sheather, Ch 2, p. 38, #1
2. Sheather, Ch 2, pp. 41–42, #5
3. Sheather, Ch 2, p. 42 #6
4. [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?

5. Sheather, Ch 2, pp 42–43, #7

[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