Homework 3

Due date: Friday, September 30

Complete the following exercises and submit your assignment via gradescope (linked on the course webpage).

Q1.

One factor affecting water availability in Southern California is stream runoff from snowfall. If runoff could be predicted, engineers, planners, and policy makers could do their jobs more effectively because they would have an estimate as to how much water is entering the area.

The data set compares the stream runoff (in acre-feet) of a river near Bishop, California (due east of San Jose) with snowfall (in inches) at a site in the Sierra Nevada mountains. For each of the following questions, assume that your audience is city water planners with moderate statistical training. You can load the data set using by running the below command:

```r
water <- read.csv("http://aloy.rbind.io/data/water.csv")
```

(a) Create a scatterplot of stream runoff against snowfall. Comment on the apparent relationship (strength, shape, direction).

(b) Find (and write out) the equation of the fitted linear model of the expected runoff given snowfall. (Do not just print the R output; you need to write down the fitted model.) Add the least squares line to the data plot from part (a).

(c) Interpret the slope coefficient in the context of the problem. (Don't forget to specify units.)

(d) Interpret the intercept in the context of the problem. (Don't forget to specify units.)

(e) Interpret the R^2 value in the context of the problem.

(f) Test whether a linear association exists between stream runoff and snowfall. State your hypotheses, p-value, and conclusion in context.

Q2.

For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

(a) What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?

(b) How much do families whose disposable income is \$23,500 spend, on the average, for meals away from home?

(c) How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

Q3.

Consider the data set collected by ETS on 1000 randomly selected students (at an unnamed college). We are interested in predicting first year GPA (`FYGPA`) from the sum of the verbal and math SAT percentiles (`SATSum`). The data can be loaded using the following code:

```r
sat_gpa <- read.csv("http://aloy.rbind.io/data/satGPA.csv")
```

(a) Fit the simple linear regression model predicting first year GPA from the sum of the verbal and math SAT percentiles. Report the fitted regression equation.

(b) Obtain a 95% interval for the average first year GPA for students whose SAT percentiles sum to 120. Interpret your interval

(c) Obtain a 95% interval for the first year GPA of Veronica Davis who who scored an SAT sum of 120. Interpret your interval.

(d) Would you expect the two intervals above to be wider or more narrow for considering an SAT sum of 100? Explain using both the mathematical formula for creating the intervals and using the intuition given by what we've seen in class of the variability of the line.