ST512 – Lab 02

These are not exercises to be handed in for a grade. These are just practice problems that should help with completing the homework assignment covering the same material; there may also be relevant conceptual extras not covered in lecture. You're welcome to discuss these problems with your classmates or with the instructor/TA in office hours.

Part 1: Simple Linear Regression.

(Revisit problem from Lab 1.) Software (`lm` in R or `proc reg` in SAS) will give us relevant output from the least-squares fit, as well as some plots to help us assess whether the model assumptions are satisfied. Use the SAS/R code provided on the course website to fit the simple linear regression model to the `grape` data set—with `yield` as the response variable and `nclust` as the predictor variable—and answer the following questions.

1. Find the estimates of the slope and intercept parameters and the corresponding standard errors. Are `yield` and `nclust` related? Formally state your hypotheses, carry out the test, and state your conclusions.
2. Suppose that the number of clusters this year is `nclust` = 102. Estimate the mean grape crop yield for this value of `nclust` and give a 95% confidence interval.
3. Draw a plot of the residuals versus the predictor variable `nclust`. What does this plot say concerning the assumptions of the simple linear regression model?

Part 2: Beyond the Linear Model.

(From Cody & Smith's *Applied Statistics and the SAS Programming Language*, 4th edition, page 116.) Data on the height and weight on $n = 7$ individuals have been measured is provided. The goal is to describe the relationship between the response variable `weight` and the predictor variable `height`.

1. Draw a scatterplot of `weight` versus `height`. Is there a linear relationship?
2. Fit a linear regression model. Draw a plot of the residuals versus `height`. Do you see the subtle quadratic-like pattern?
3. The quadratic pattern observed in the residual plot suggests that perhaps we should use a quadratic instead of linear model. In SAS, create a new data set that contains an additional variable `height2` which is just the square of the original variable `height`. Fit a quadratic model by simply adding a second term to the model statement in `proc reg`, i.e.,

   ```
   model weight = height height2;
   ```

   In R, just define a new variable and call `lm` as follows

   ```
   lm(weight ~ height + height2)
   ```

   Re-draw te scatterplot of `weight` versus `height` and overlay the estimated *quadratic* regression curve. Comment on the fit. Also plot the residuals of this new fit versus `height` and notice that the pattern observed before is gone.

Part 3: Transformations.

(From Cody & Smith's *Applied Statistics and the SAS Programming Language*, 4th edition, page 129.) A particular drug is administered to patients, and it is believed that the dosage of the drug will affect the patient's heart rate. Data on dosage and heart rate on $n = 10$ individuals is available on the course website. The goal is to describe the relationship between the response variable `rate` and the predictor variable `dose`.

1. Draw a scatterplot of `rate` versus `dose`. Is there a linear relationship?
2. A careful inspection of the plot reveals that `rate` increases by about 3 units each time `dose` is doubled. That is,

$$\frac{\text{change in } \texttt{rate}}{\text{change in } \texttt{dose}} \approx \frac{3}{(2 \times \texttt{dose}) - \texttt{dose}} = \frac{3}{\texttt{dose}}.$$

   The right-hand side above is, roughly, the derivative of `rate` with respect to `dose`. This suggests that `rate` is actually a linear function of log(`dose`), the natural log of `dose`, not of `dose` itself. Define a new variable `log_dose` and draw a scatterplot of `rate` versus `log_dose`. Does it look linear?
3. Fit a linear model of `rate` versus `log_dose`. Comment on the quality of the fit.