Recitation #1
Simple Linear Regression
MSAN 601

Install the `car` and the `alr4` packages in `R`.

Question 1

The data frame `Heights` has two variables, `mheight` and `dheight`, which represent the a mother's height and a daughter's height in inches. This data was collected by Karl Pearson from 1893-1898 in an attempt to determine the relationship between mother (under the age of 65) and daughter's (over the age of 18) heights.

1. How many observations are there in the data set?
2. What are the maximal, minimal, mean, median, standard deviation and variance of each of the variables in the data set? Generate a histogram for each variable.
3. What is the correlation between `mheight` and `dheight`? Does it matter the ordering of the variables in the `corr` function when it is called? why or why not?
4. Which of the two variables in the `Heights` data frame would you consider to be the explanatory variable? Why? Generate a scatter plot, plotting the response (y) on the explanatory (x). Don't forget to label both axes and include a title for the graph. What do you observe?
5. Create the simple linear regression model, regressing the response variable on the explanatory variable. Once again, create the scatter plot, but this time, overlay the fitted regression line. Don't forget to label both axes and include a title for the graph. What do you observe?
6. What is the equation of the fitted regression line?
7. Interpret the estimated coefficient b_1 in a well-articulated, complete English sentence.
8. Is the coefficient of the predictor variable, β_1, statistically significantly different from 0? Write out the null and alternative hypotheses, provide the critical t value (indicate degrees of freedom, assume $\alpha = 0.05$), compute the test statistic and articulate your conclusion.
9. Is the y-intercept, β_0, statistically significantly different from 0? Instead of writing out the null and alternative hypotheses, determining the critical t-value and test statistic, what shortcut can you use in `R` to come to the correct conclusion?
10. If the y-intercept β_0 is determined **not** to be statistically significantly different from 0, then what?
11. How would you verify that the assumptions underpinning the theoretical Simple Linear Regression model we developed in class are valid in this instance, i.e., for the linear model generated from the data set `Heights`? What are those assumptions?
12. **Manually** compute `SSR`, `SSE` and `SSTO`. Confirm that $SSTO = SSR + SSE$.
13. Compute $\sum e_i$. Is it equal to what you expected? If not, what did you expect and why do you believe it differs?
14. **Manually** compute R^2. How do you feel about the magnitude of R^2? Interpret the value of R^2 in a well-articulated, complete English sentence.

15. Test whether the model has any parameters that are statistically significantly different from 0. Write out the null and alternative hypotheses, manually compute the test statistic F^*, provide the critical F value (indicate all degrees of freedom, assume $\alpha = 0.05$). Is this a one- or two-tailed test? Articulate your conclusion.
16. **Manually** construct a confidence interval for $E\{Y_h\}$ at $X_h = 69.432$. Interpret your result in a well-articulated, complete English sentence.
17. What value of X_h will generate the smallest (shortest, tightest) confidence interval for $E\{Y_h\}$? Explain mathematically.
18. **Manually** construct a prediction interval for $\hat{Y}_{h(new)}$ at $X_h = 69.432$. Interpret your result in a well-articulated, complete English sentence. Compare and discuss its width to the width of the confidence interval above. Explain mathematically why the relationship you have just described always holds.
19. Are you, on the whole, convinced that the simple linear model we have created approximates the data well? Justify your conclusion with a well-reasoned explanation.

Question 2

In 1857, James Forbes discussed a series of experiments where he collected paired observations of atmospheric pressure (`pres`) and boiling point (`bp`) in degrees Fahrenheit. Both variables are located in the data frame `Forbes`.

Using the two aforementioned variables, create a linear regression model. Be sure to evaluate the model and the assumptions. Is the model valid? If not, why? What assumption(s) is/are violated? If the model is in violation of some assumptions, can you correct for those violations? How? Is the new model appropriate? Is the new model linear? Is the new model first-order?