SDGB 7844 HW 1: Chocolate & Nobel Prizes

Instructor: Prof. Nagaraja

Due: 2/7

Submit two files through Blackboard: (a) .Rmd R Markdown file with answers and code and (b) Word document of knitted R Markdown file. Your file should be named as follows: "HW[X]-[Full Name]-[Class Time]" and include those details in the body of your file.

Complete your work individually and comment your code for full credit. For an example of how to format your homework see the files posted with Lecture 1 on Blackboard. **Show all of your code in the knitted Word document.**

Read the *New England Journal of Medicine* article, "Chocolate Consumption, Cognitive Function, and Nobel Laureates" (Messerli, F.H., Vol. 367(16), 1562-1564; 2012) which is posted with this assignment. We will be using a reconstruction of Messerli's data. The variables in the data set you will use are (file: "nobel_chocolate.txt" on Blackboard) are "country", "nobel_rate", and "chocolate".

The information gathered in the data set you will be using is from several different sources. The number of Nobel prize winners is from Wikipedia and includes winners through November 2012, population information (used to compute the "nobel_rate" variable) is from the World Bank, and chocolate market size is from the Euromonitor International's Passport Database.

Goal: In this assignment, you will be replicating Messerli's analysis.

1. According to Messerli, what is the variable "number of Nobel laureates per capita" supposed to measure? Do you think it is a reasonable measure? Justify your answer.
2. Are countries without Nobel prize recipients included in Messerli's study? If not, what types of bias(es) would that introduce?
3. Are the number of Nobel laureates per capita and chocolate consumption per capita measured on the same temporal scale? If not, how could this affect the analysis?

4. Create a table of summary statistics for the following variables: Nobel laureates per capita, GDP per capita, and chocolate consumption. Include the statistics: minimum, maximum, median, mean, and standard deviation. Remember to include the units of measurement in your table.

5. Create histograms for the following variables: Nobel laureates per capita, GDP per capita, and chocolate consumption. Describe the shape of the distributions.

6. Construct a scatterplot of Nobel laureates per capita vs. chocolate consumption. Label Sweden on your plot (on the computer, not by hand). Compute the correlation between these two variables and add it to the scatterplot. How would you describe this relationship? Is correlation an appropriate measure? Why or why not?

7. What is Messerli's correlation value? (Use the correlation value that includes Sweden.) Why is your correlation different?

8. Why does Messerli consider Sweden an outlier? How does he explain it?

9. Regress Nobel laureates per capita against chocolate consumption (include Sweden):
   (a) What is the regression equation? (Include units of measurement.)
   (b) Interpret the slope.
   (c) Conduct a residual analysis to check the regression assumptions. Make all plots within one figure. Can we conduct hypothesis tests for this regression model? Justify your answer.
   (d) Is the slope significant (conduct a hypothesis test and include your regression output in your answer)? Test at the $\alpha = 0.05$ level and remember to specify the hypotheses you are testing.
   (e) Add the regression line to your scatterplot.

10. Using your model, what is the number of Nobel laureates expected to be for Sweden? What is the residual? (Remember to include units of measurement.)

11. Now we will see if the variable GDP per capita (i.e., "GDP_cap") is a better way to predict Nobel laureates.
    (a) In one figure construct a scatter plot of (i) Nobel laureates vs. GDP per capita and (ii) Nobel laureates vs. log(GDP per capita). Which plot is more linear? Label Sweden on both plots. On the second plot, label the two countries which appear on the bottom left corner.
    (b) Is Sweden still an outlier? Justify your answer.
    (c) Regress Nobel laureates against log(GDP per capita). Provide the output and add the regression line to your scatterplot. (In practice, we would do a residual analysis here, but we will skip it to reduce the length of this assignment.)

(d) The log-y model is a multiplicative model: $log(y) = \beta_0 + \beta_1$ is $y = e^{\beta_0+\beta_1 x}$. For such a model, the slope is interpreted as follows: a unit increase in x changes y by approximately $(e^{\beta_1} - 1) \times 100\%$. For your regression, model interpret the slope (remember to include units of measurement).

12. Does increasing chocolate consumption cause an increase in the number of Nobel Laureates? Justify your answer.