Math 140 **Name** _____ **Date:**

R Activity: Regression

Directions: Each student should answer the questions on this worksheet following the **golden rule of write-ups**. You will need to use R code to do so, guided by the complementary Regression Activity page on our course resource site, as we discuss in class. Work in groups! Ask questions!

1. **Exploring the Data**

   (a) Based on the question of interest (see the associated web page), which variable in the data set `df` is our response variable? Which variable is our explanatory variable?

   (b) In R, you made a point plot to visualize association between sugar and calories in Starbucks beverages placing the explanatory variable you identified in Q1a on the x-axis. Write a sentence or two describing the direction, form, and strength of any association you see.

   (c) Based on the plot of Sugars and Calories, does it look like it would be reasonable to consider a linear model for the association between sugar and calories? Explain briefly.

   (d) What is the correlation coefficient between Sugars and Calories? What does this number tell us about the strength and direction of the linear relationship between these two variables?

   (e) Is there also an association between the amount of cholesterol in a Starbucks drink and its calorie content? How does this association compare to the association between sugar and calories? Justify your response with visualizations and a comparison of correlation coefficients.

2. **Build the Model**

(a) Build the model in R using the code provided. No response needed here :)

(b) Find the equation of the least squares line by running the code `fit$coefficients`. Record the equation of the line here.

(c) Use the code provided to fit the least-squares regression line to your scatterplot. Does this line seem like it could be useful for predicting a Starbucks drinks' Calories from its sugar content?

3. **Assess The Model**

(a) What percentage of the variability in Calories is explained by its linear relationship with Sugars? Hint: Think r^2!!

(b) Create the residual plot in R using the provided code. Comment here on whether you think **constant variability** seems to be a reasonable assumption for this model. Explain your answer.

(c) Create a histogram of the residuals by adapting the provided code. Does the histogram of the residuals look like what we would expect for a normal distribution? Explain.

(d) If *constant variability* seems reasonable, the value of r^2 is fairly large (on the scale from 0 to 1), and the histogram of the residuals is reasonably bell-shaped, these are good indications that a linear model is a good model for the relationship between sugar content and calories in Starbucks beverages. Summarize your findings in Q3 - does a linear model seem to be a reasonable model to use here?

4. **Use The Model**

   (a) Interpret the slope of the regression line in the context of the data by completing this sentence:
   "For every 10 grams of additional sugar, the expected increase in calories is ..."

   (b) If a Starbucks drink has 47 grams of sugar, about how many calories would the drink have, as predicted by the linear model? Use the equation of the line you found in Q2 to make this prediction.

   (c) The data set does show a Starbucks drink with exactly 47 grams of sugar (the Frappuccino® Blended Coffee with Soy Milk, of course). Determine the residual value for this drink. Does this drink lie above or below the least squares line?

5. **Grouping by a Category**

   (a) The first variable in the data frame is `Beverage_category`. What are the different beverage categories? List them all here.

   (b) Recreate your plot from Q2 c), but this time color each dot according to the drink's Beverage category. Then explain interesting features this use of color reveals about the association between sugar and calories for Starbucks beverages. For instance, which drink categories does our linear model do the worst job of predicting? Which categories tend to have more calories than what is predicted by the linear fit with sugar? Which categories tend to have fewer calories than what is predicted by the line? Does that seem reasonable?