Project 3 Guidelines

This project is focused on linear regression. You will either choose one of the datasets included in the folder or can find/collect data yourself.

The available datasets include census data about Massachusetts or the US broadly, baseball team performance data, data about life expectancy around the world, and a crime rates dataset. If you use one of these datasets, **look at the variable descriptions linked below before beginning to do the questions**.

Some examples of data you could collect yourself: how grams of sugar or fat in snack foods are related to calories, how athletes' pay is related to their performance, critic vs audience ratings for a sample of movies, etc.

This project is due on **Wednesday, May 4th.**

To do: Write a short report answering the following questions. Include any relevant plots/graphs.

1. Start by choosing a variable that you want to understand and make predictions about. This will be your response variable.

2. What are some variables (whether you have them in your dataset or not) that could be predictors of your chosen response variable?

3. Choose one of those variables that you do have access to as an explanatory variable. Make sure it is numerical. (It is possible to use categorical variables for prediction, but we haven't talked about methods for it.) Why do you think that this quantity might predict your response variable?

4. Make a scatterplot showing the relationship between these two variables. Describe the relationship: strong, moderate, or weak? Linear or nonlinear? Positive, negative, or neither?

5. Find the correlation coefficient *R*. Does its value agree with your assessment of the relationship in Question 3?

6. Find the coefficient of determination R^2 and complete this sentence to interpret it: The variation in ___________ explains _____ percent of the variation in ________________.

7. Do you think your explanatory variable is a useful predictor of your response variable using a linear model? Why or why not?

8. If it **is** a useful predictor, do you think your explanatory variable causes your predictor variable, the other way around, or do you think something else results in the correlation?

9. If it is **not** a useful predictor, do you think there is a relationship that isn't linear, or do you think there isn't a strong relationship between the two variables at all? Explain your answer.

Included Datasets (in the folder)

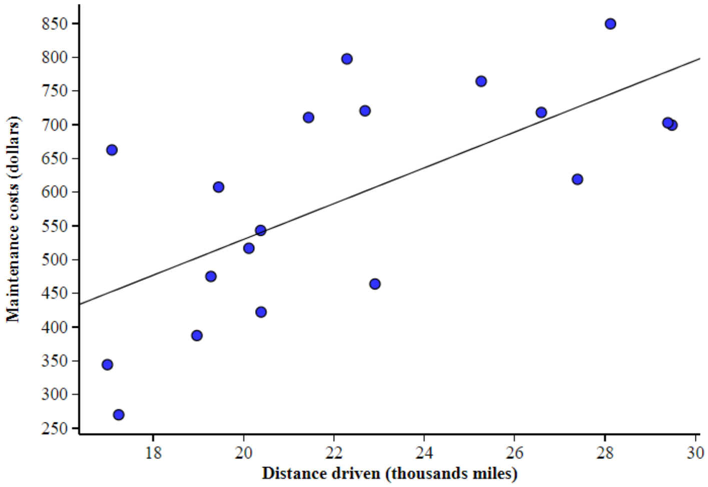
Note: Not all variables are included in the data in the folder. Please check the data first!

- Massachusetts County Data – Focused on language, age, income, race/ethnicity, housing type, etc. in counties in MA. More information about the variables here.
- US County Data – Focused on population, housing, and income in counties across the US. More information about the variables here.
- MLB Team Data – Statistics of MLB teams from 1876 to 2019. More information about the variables here.
- US Crime Rates 1960-2019 – National data on various crimes for each year from 1960 to 2019. more information about the variables here
- Life Expectancy (2015) Data – Life expectancy and factors around health and quality of life in countries around the world. More information about the variables in the 2nd tab of the data spreadsheet.

Learning Target 7: Linear Regression

Question 1

A study is conducted examining how far a car has driven (in thousands of miles) compared to how expensive the annual maintenance costs are. The researcher makes the following scatterplot with a best fit line:



(a) What is the explanatory variable, and what is the response variable?
(b) Describe the relationship between the two variables (strong/weak, linear/nonlinear, positive/negative).
(c) The equation of the regression line is ŷ = -0.55 + 26.5x. Identify the slope, including its units, and explain what it means in the context of the study.
(d) Predict the maintenance costs of a car that has been driven for 25,000 miles. (Be careful with units here!)

Question 1

A nutrition researcher collects data on a variety of restaurant burgers, and based on that data, they find the following least squares line: ŷ = 7 + 0.9x, where ŷ is the predicted amount of fat in the burger, measured in grams, and x is the amount of protein, also measured in grams.

(a) Is this relationship positive or negative? What does that mean in this context?
(b) What is the y-intercept in this regression line equation? What does it mean in this context?
(c) How much fat would you predict to be in a burger with 14g of protein?
(d) A restaurant introduces a meat-free burger with 14g of protein and 10g of fat. What is the residual of that data point? How would you explain why the prediction was not accurate?

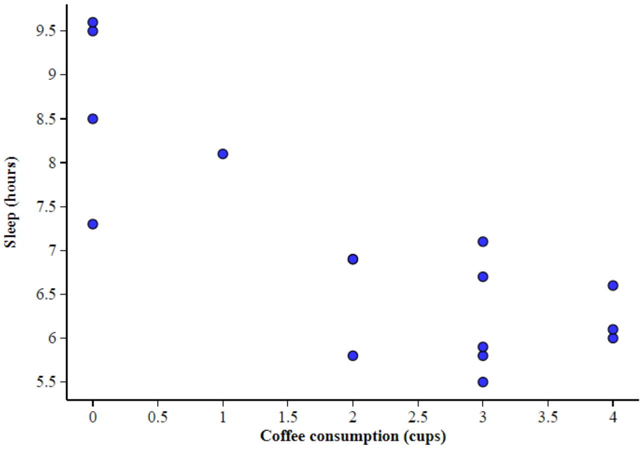
Question 1

A student at Elmhurst University takes data from a random sample of accepted students and correlates student family income with non-loan financial aid offered.

(a) Which variable should the student use as the explanatory variable? Which should be the response variable?
(b) The student calculates the least squares regression line and finds a slope of -0.043 and a y-intercept of 24327. What are the units of each of these? Interpret what they mean in context.
(c) Use the model to predict the amount of aid that a student with a family income of $50,000 would receive.

Question 2

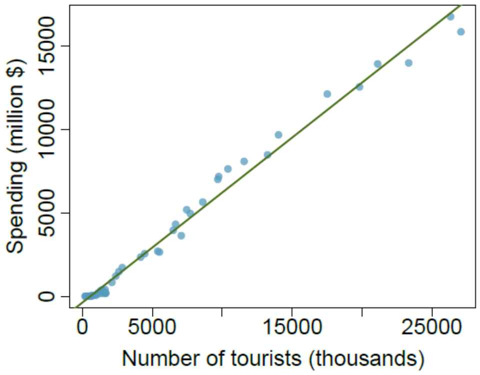
A professor observes how many cups of coffee they drink per day and how many hours they sleep per night, and they make the following scatterplot.



(a) Describe the relationship between the two variables (weak/strong, linear/nonlinear, positive/negative).
(b) The R^2 value for this relationship is 0.657. Calculate the correlation coefficient *R*. (Make sure it has the correct sign based on your answer to (a)!)
(c) What percent of the variability in the professor's hours of sleep is predicted by their coffee consumption?

Question 2

The Association of Turkish Travel Agencies reports the number of foreign tourists visiting Turkey and tourist spending by year. Below is a scatterplot produced from this data.



(a) What is the explanatory variable here?

(b) Describe the relationship between the two variables (weak/strong, linear/nonlinear, positive/negative).

(c) The R^2 value for this relationship is about 0.94. Calculate the correlation coefficient R. (Make sure it has the correct sign based on your answer to (a)!)