# Data 8 Spring 2022

Lab 09: Regression

---

**Correlation**

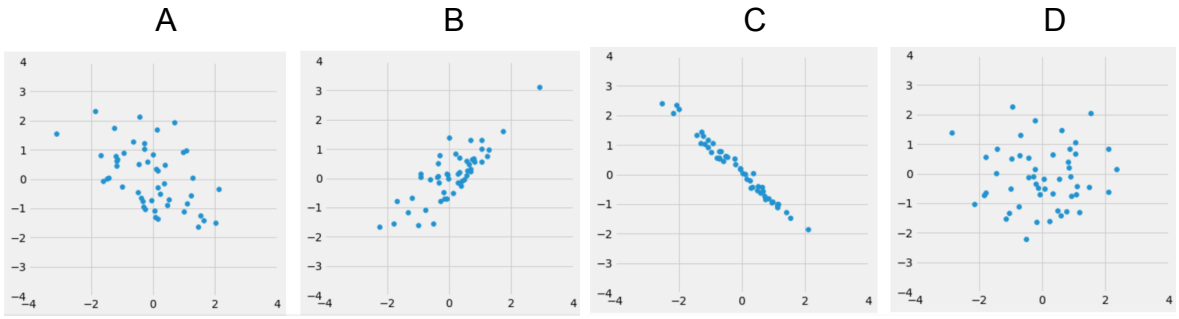
An important aspect of data science is using data to make *predictions* about the future based on the information that we currently have. A question one might ask would be "Given the US GDP of every year of the previous decade, how can we predict the US GDP for next year?" In order to answer this question, we will investigate a method of using one variable to predict another by looking at the *correlation* between two variables.

**Question 1. Function:** Let's start by writing a function called `correlation_coefficient` that takes in two arrays `x` and `y` of the same length and returns the correlation coefficient between the two.

*Hint: Assume you have a function called `convert_su` defined, that converts an array to standard units (we did this last week).*

```
def correlation_coefficient(x, y):
    x_su = convert_su(x)
    y_su = convert_su(y)
    return np.mean(x_su * y_su)
```

**Question 2. Comparing Correlation:** Look at the following four datasets. Rank them from weakest correlation to strongest correlation. Remember that a *strong* correlation has |r| close to 1.




Ranking: D, A, B, C

- D has almost no visible negative or positive trend as it is basically a blob, so its correlation is near 0

- A has a negative correlation, but the points are not very tightly clustered around a straight line, so the strength of its correlation is smaller
- B has a positive correlation, and the points are more tightly clustered around a positive sloping line, so its correlation strength is stronger than A
- C has a negative correlation, and the points are almost all along a straight line, so it has a very strong correlation in magnitude

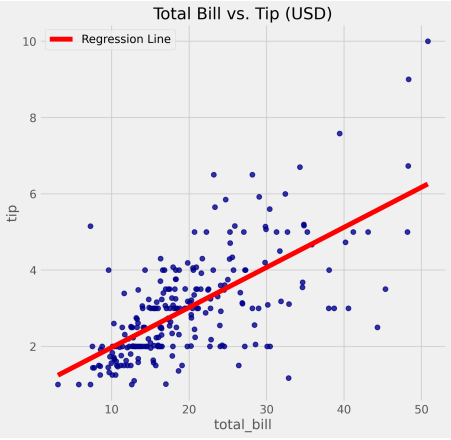
We have introduced correlation as a way of quantifying the *strength* and *direction* of a linear relationship between two variables. However, the correlation coefficient can do more than just tell us about how clustered the points in a scatter plot are about a straight line. It can also help us define the straight line about which the points (in original units) are clustered, also known as the *regression line*.

The formulae for the *slope* and *intercept* for the regression line are shown below. In fact, by a remarkable fact in mathematics, the line uniquely defined by the slope and intercept below is *always* the best possible straight line we could use.

$$\textbf{slope of the regression line} = r \cdot \frac{\text{SD of } y}{\text{SD of } x}$$

$$\textbf{intercept of the regression line} = \text{average of } y - \text{slope} \cdot \text{average of } x$$

**Question 3. Restaurants:** Suppose you are given the scatter diagram shown below that shows the relationship between the total bill versus tip at American restaurants. You have calculated the line of best fit (shown in red). Suppose your friend Alice goes out to dinner and tells you her total bill was $35. Based on the regression line, what would we predict Alice's tip to be?




**Answer:** Simply find the value of the line at x=35. In this case, it is roughly 4.75. Thus, we predict her to tip $4.75

**Question 4. Meggy's Coffee:** We want to investigate the correlation between the daily ounces of coffee consumed by an individual and the number of hours the individual stayed awake. It is our intention to use the ounces of coffee consumed to predict the number of hours the individual stayed awake. The data from our sample of 500 people has the following characteristics:

- The number of ounces of coffee consumed has a mean of 12 ounces and SD of 4
- The number of hours stayed awake has a mean of 16 and an SD of 2
- The correlation between the number of ounces of coffee consumed and number of hours spent awake is 0.5.
- Suppose the scatter plot is roughly linear.

a) What is the slope of the regression line?

Slope = r*(2/4) = 0.5*0.5 = 0.25

b) What is the intercept of the regression line?

Intercept = 16 - 0.25*(12) = 13

c) Suppose your friend Matthew is in this population (and not in the sample). He told you that he consumed 16 ounces of coffee that morning. Use your line of best fit to predict how many hours Matthew will stay awake today.

Using the slope of 0.25 and the intercept of 13 that we calculated earlier, we can substitute these values in to the general form of the line of best-fit and compute: 0.25(16) + 13 = 17

d) Your other friend, Meghan, is also in the population and not in the sample. She confesses that she drank 80 ounces of coffee that day (wow!). Based on the information above, would the regression line we computed in parts (a) and (b) be appropriate to predict the number of hours Meghan stayed awake? Explain.

If we were to use the equation y = 0.25x + 13, we would predict that Meghan stayed up 0.25(80) + 13 = 33 hours! While we are technically able to make an estimate on the number of hours she'd stay away using our regression line, since 80 oz is so far away from the mean in our sample, this method is invalid.

Meghan's datapoint is 17 SDs above the mean, making it an outlier.