Lesson 4.2 - February 25, 2021

Review

- The closer that a set of points are to being a straight line, the closer the correlation coefficient is to 1 and -1
- If the data points are in an exactly straight line, the correlation coefficient will be either 1 or -1
- Certain plots will have no pattern and the correlation coefficient will be exactly 0
- Describing relationships and trends using the trend, shape, and strength

---

Correlation Coefficient Calculation

- $$r = \sum \frac{Z_x \cdot Z_y}{n-1}$$

- There are two variables: (x, y)
- For instance, x could be height while y could be weight
- $Z_x = \frac{x - \bar{x}}{s_x}$, where $\bar{x}$ refers to the mean of the x values
- Ultimately, the formula for the Z-score is generalized as $Z_n = \frac{n - \bar{n}}{s_n}$ where n is some arbitrary set of numbers
- You have to calculate the individual Z-score for each one of the values in the data set
- Unfortunately, this calculation is very tedious, thus this is not usually done by hand unless the dataset happens to be within reasonable limits

---

Correlations

- > The number that you get from your correlation coefficient calculation determines how "strong" the correlation is between the two variables that you are measuring

- When is the product of the two Z-scores positive?
  - When both values are positive

    - When both values are negative
- Positive correlation means that as one variable increases, the other variable seems to increase as well ($y = x$)
- The Z-score is negative when the two values have opposite signs
- Negative correlation means that as one variable increases, the other variable decreases ($y = -x$)
- No correlation means that as one variable increases, it has no effect on the value of the other variable ($y = 0$ or $y = n$ where n is any constant)

Modelling Linear Trends

- We can use the best fit line to make predictions from the data that we've been given
- Given a value for one variable, you can predict the value of the other variable using the best fit line
- It is often easier to use x and y as these are included naturally in the cartesian plane, so it can be easier to follow trends and understand how this relationship looks
- Given a value for y, you can solve for x
- Unless the correlation coefficient is 1 or -1 (indicating a perfect correlation), your prediction is going to be slightly off from the real answer
- A sample table which could be represented perfectly by a best fit line is shown below:
Value	Age
30,000	0
30,000 - 4,000	1
30,000 - 8,000	2

Algebra versus Statistics

Algebra	Statistics
$y = m \cdot x + b$	$y = a + b \cdot x$
$x \rightarrow \text{Independent Variable}$	$x \rightarrow \text{Predictor/Explanatory}$
$y \rightarrow \text{Dependent Variable}$	$y \rightarrow \text{Predicted/Response}$