Lesson 4.1 Extension & Lesson 4.4 - March 2, 2021

Best Fit Line

- Idea is the same: we want to find the best fit line in a specific scatter plot to examine the relationship between the two variables plotted on the graph
- Questions to keep in mind
  - When is the best fit line useful?
  - How good is our prediction? How can we measure this accuracy?
  - How is the best fit line constructed?
  - *What can go wrong and affect the accuracy of the best fit line?*
- Our goal for this section is to provide a bit more detail
- Recall what r refers to, the correlation coefficient
  - Given any scatter plot, you can calculate r
  - Calculating r by hand is quite complex, so this is a task usually designated to computers
  - Calculated from the z-scores for each element of the sample (for both variables)
- When is the best fit line useful?
  - When the data is linear (best way to check is just by looking at the scatter plot and analyzing it as an object)
  - Look at r^2 as a percentage (after multiplying by 100). basically, you can say $(r^2 \cdot 100)\%$ of the variance in the response variable is explained by the independent variable. Essentially, the larger that r^2 is, the more linear the relationship is
- Equation for best fit line is not $y = mx + b$, but is $y = a + bx$
- $$a = \bar{y} = b\bar{x}$$
- $$b = r \cdot \frac{s_y}{s_x}$$
- $$y = (r \cdot \frac{ss_y}{s_x}) \cdot x + \bar{y} - (r \cdot \frac{s_y}{s_x}) \cdot \bar{x}$$

Regression Line

> ### Rule of Thumb

> Don't use regression lines to extrapolate or predict data from your analysis

- You can only use the regression line for values that are within the set of values that you have

Introduction to Probability Theory

- We normally use data to make a prediction (best fit line), now we work in the opposite direction
- Most common question is about heads/tails and flipping coins
- Each side has an equal chance of coming up (50/50)
- **Sets** contain **elements** which indicate particular outcomes
- **Events** are a subset or subcollection of the elements or outcomes presented in the set
- $P(E)$ represents the probability of the event E
- $P(E) = \frac{\text{Size of E}}{\text{Size of Set of Outcomes}}$
- E^c is the event where E does not occur