11 Lecture 12: Correlation and Regression

Failure is not an option. Success is long process - Olubowale Victor Akintimehin

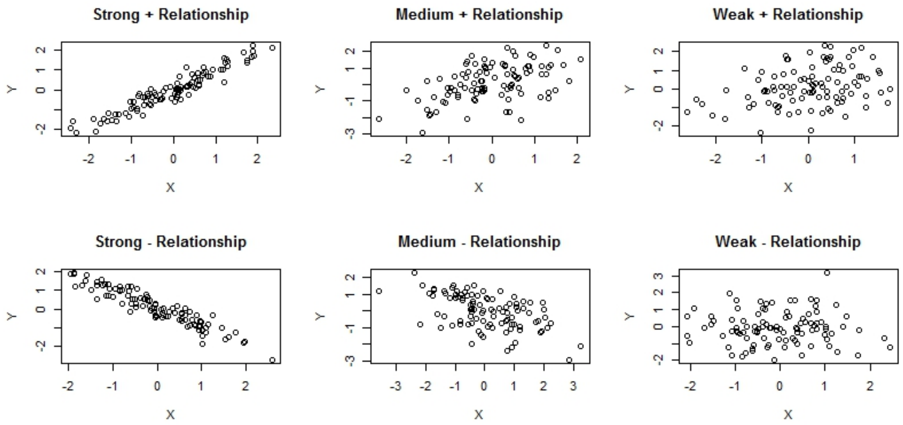
Lets do a recap

11.1 Theory Behind Correlation

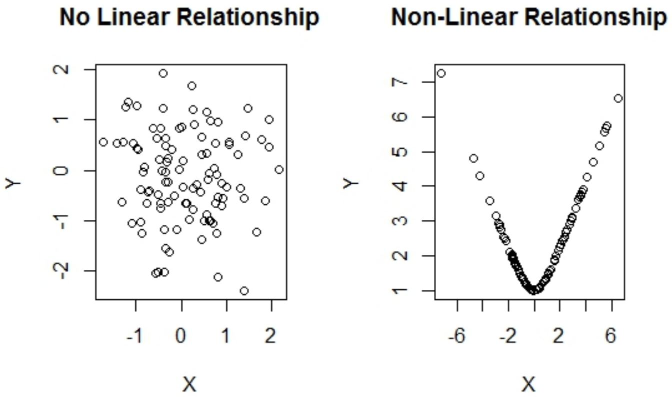
Definition of correlation - two variable are associated with each other

We can visualize this relationship using a scatter-plot

Lets look at a few:



Some irregulars:



We can quantify this relationship by calculating the linear correlation coefficient r, this is also called **Pearson product moment correlation coefficient**.

- $r=$ linear correlation coefficient for the sample
- ρ linear correlation coefficient for the population

A few requirements:

1. Both X and Y are random
2. Visualize data to ensure linear relationship
3. Outliers that are based on errors should be removed

Always remember many ways to the top of mountain $\cdots$

$$\begin{aligned} r &= \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \\ &= \frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y} \end{aligned}$$

Properties of Correlation Coefficient

1. $-1 \leq r \leq 1$
2. r does not change, if the scale changes
3. Switching x and y does not effect r
4. r is measures the strength of a linear association

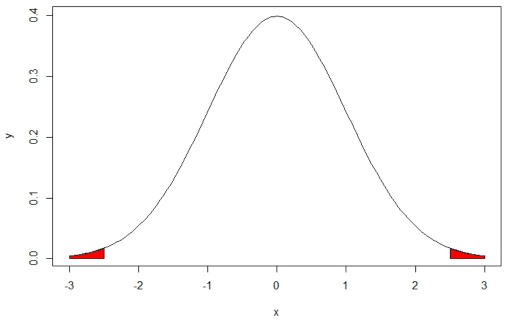
r^2 denotes the proportion explained by the linear association (model)

Common errors involving correlation:

- Correlation does not imply causality (lurking variables)
- When data averages are used correlation can be inflated
- A low correlation values does not imply a lack of relationship between two variables. The relationship can be nonlinear as in the graph shown before.

Remember the Statistical Process **Process for Hypothesis Testing for this class:**

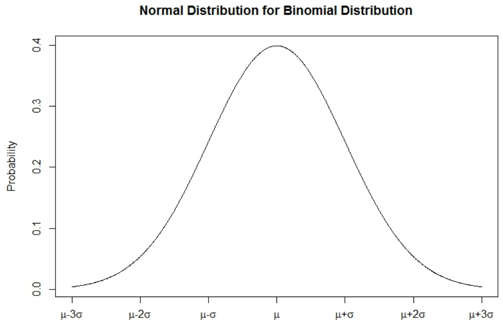
1. Identify and State the Statistical Question
   - Determine the variable(s) of interest
   - Determine the type variable(s) (i.e., quantitative or qualitative)
   - Identify and state the hypotheses (Null and Alternative Hypotheses) based on the question at hand
2. Identify and state level of significance α (the probability of rejecting the H_0 when H_0 is true)



Really IMPORTANT:
   - α:
   - $df = n - 2$
   - Critical Value:

3. Perform Statistical Test and Interpret Results

$$\begin{aligned} TS = t &= \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \\ &= \frac{r\sqrt{n-2}}{1-r^2} \end{aligned}$$



- Test Statistic:
- p-value:

4. State the sample, null hypothesis, test that was used, and conclusion with non-statistical terms

Example 1: Listed below are systolic blood pressure measurements (in mm Hg) obtained from the same woman (based on data from "Consistency of Blood Pressure Differences Between the Left and Right Arms," by Eguchi, et al., Archives of Internal Medicine, Vol. 167). Is there sufficient evidence to conclude that there is a linear correlation between right and left arm systolic blood pressure measurements?

						Mean	SD
Left Arm	102	101	94	79	79	91	11.3798
Right Arm	175	169	182	146	144	163.2	17.254

Example 2: Height and Pulse Rate The heights (in inches) and pulse rates (in beats per minute) for a sample of 40 women were measured. Using STATDISK with the paired height/pulse data, the linear correlation coefficient is found to be 0.202 (based on data from the National Health Examination Survey). Is there sufficient evidence to support the claim that there is a linear correlation between the heights and pulse rates of women? Explain.