# Chapter 11 & 2- Simple Linear Regression Model and Correlation

STAT 251

Lecture 33

ANOVA activity - Chapter 10
Scatterplots, Covariance & Correlation - Chapter 11 & 2

Dr. Lasantha Premarathna

# Chapter 11 & 2 - Learning Outcomes

- Scatter plot
- Covariance & Correlation
- Simple linear regression
- Least squares estimates in simple linear regression
- Interpret the parameters in a fitted linear model
- Inference for the slope parameter - Confidence interval & hypothesis testing

# Introduction

- **Explanatory or Predictor Variable**(Independent variable)
  The variable whose value is fixed by the experimenter, denoted by $x$

- **Response Variable**(Dependent variable)
  For fixed $x$, this variable will be random; we denote this random variable and its observered values by $Y$ and $y$ respectively.

  Ex: Blood alcohol level/ number of beers consumed
  Grade on test/ amount of study time

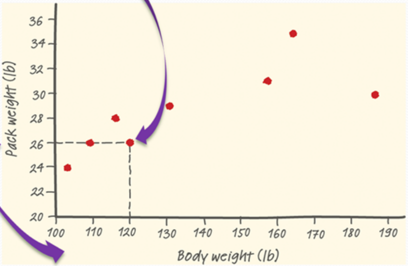
- Let $x_1, x_2, \cdots, x_n$ denote values of the explanatory variable for which observations are made, and $y_i$ denote the observed value associated with $x_i$
- The available bivariate data then consists of the $n$ pairs $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$.

# Scatter Plot

- A scatter plot is a graphical presentation of the relationship between two quantitative variables.
- One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.

**Example:** Make a scatterplot of the relationship between body weight and backpack weight for a group of hikers.

| Body weight (lb) | 120 | 187 | 109 | 103 | 131 | 165 | 158 | 116 |
|---|---|---|---|---|---|---|---|---|
| Backpack weight (lb) | 26 | 30 | 26 | 24 | 29 | 35 | 31 | 28 |



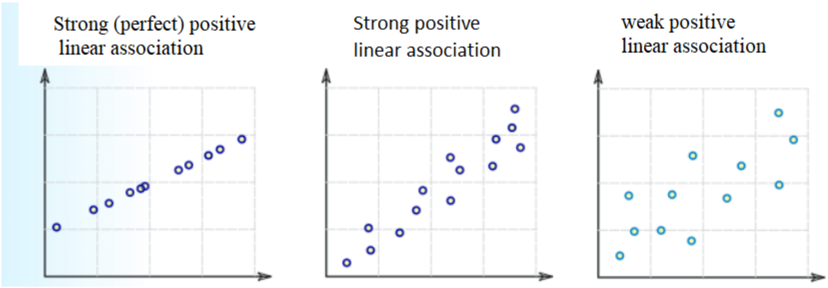
# Interpreting Scatter Plot

Scatter plots help in visualizing statistical relationships between variables

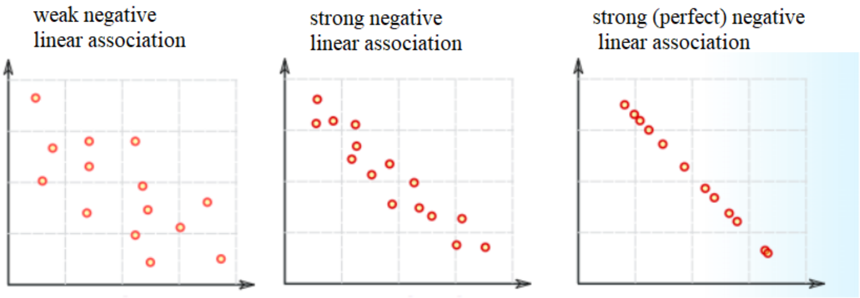
We can describe the overall pattern of a scatter plot by

1. Direction (positive, negative, none)
   - Positively associated when
     high (low) values of $x$ tend to occur with high (low) values of $y$
   - Negatively associated when
     high values of one variable tend to occur with low values of the other variable
2. Form (linear, curved, no clear form)
3. Strength (strong, weak or no relationship)
4. Any outliers?

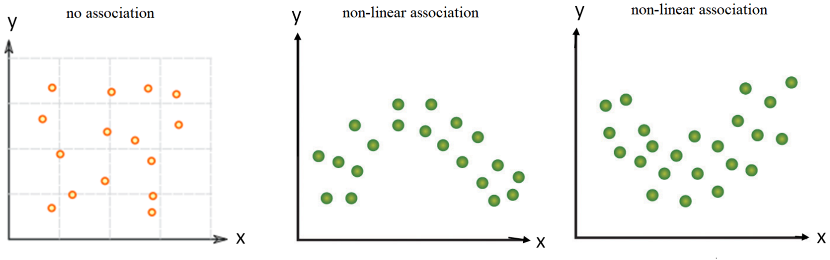
# Interpreting Scatter Plot



# Interpreting Scatter Plot



# Interpreting Scatter Plot



# Examples:

Would you expect a positive or negative association if they are linear , or no association between following variables

- Age of the car and the mileage on the odometer

- Age of the car and the resale value

# Covariance and Correlation Coefficient

- The covariance and the correlation ($r$) quantify the degree of linear association between pairs of variables
- Covariance is a measure of how changes in one variable are associated with changes in a second variable.

**The sample covariance is**

$$\begin{aligned} & \operatorname{cov}(x,y) &=& \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ .\Rightarrow\quad & \operatorname{cov}(x,y) &=& \frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}\right] \\ \Rightarrow\quad & \operatorname{cov}(x,y) &=& \frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}\right] \end{aligned}$$

# Covariance

- If $x$ and $y$ are positively associated, then $\mathrm{Cov}(x, y)$ will be large and positive

- If $x$ and $y$ are negatively associated, then $\mathrm{Cov}(x, y)$ will be large and negative

- If the variables are not positively or negatively associated, then $\mathrm{Cov}(x, y)$ will be small

# Correlation Coefficient ($r$)

- Measures the strength and direction of the linear association between $x$ and $y$
- Sample correlation coefficient is defined by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

$$\text{where} \quad s_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \; \textit{and} \; s_y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

$$\Rightarrow \quad r = \frac{\text{Cov}(x, y)}{s_x s_y}$$

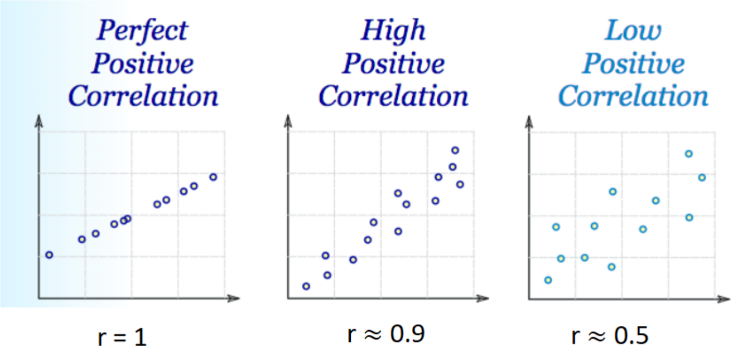
a positive $r$ value indicates a positive association

a negative $r$ value indicates a negative association

$r$ value close to 0 indicates a weak linear association

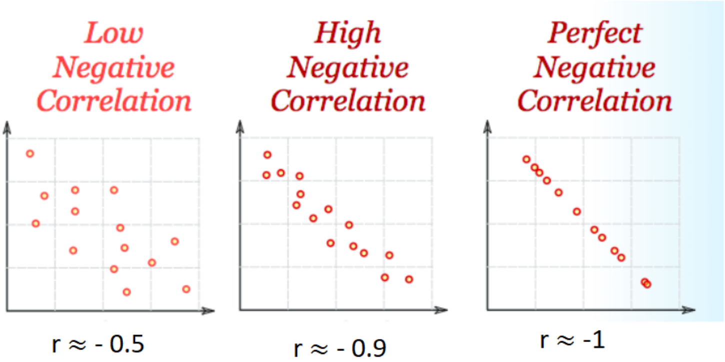
# Correlation

Positive Correlation



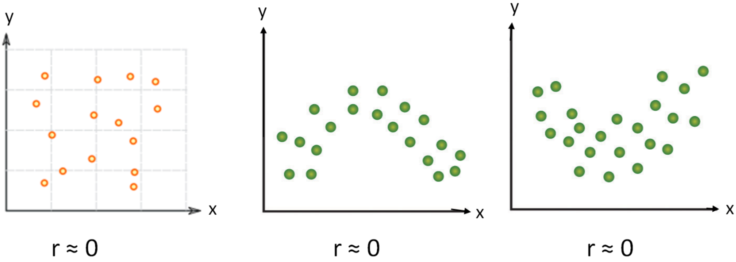
# Correlation

Negative Correlation



# Correlation

No Correlation



# Properties of Correlation

- Always falls between -1 and +1, i.e. $-1 \leq r \leq 1$
- Sign of correlation denotes the direction
  (-) indicates negative linear association
  (+) indicates positive linear association.
- Correlation has no units and does not change when we change the units of measurement of $x, y$ or both
- Two variables have the same correlation no matter which is treated as the response variable.

# Before the next class ...

Visit the course website at canvas.ubc.ca

- Review Lecture 33 and related sections in the text book
- Topic of next class: **Correlation, Simple Linear Regression**