# Sri Lanka Institute of Information Technology

## Faculty of Computing

IT2120 - Probability and Statistics

Ms. K. G. M. Lakmali

Year 02 and Semester 01

Lecture 11

# REGRESSION ANALYSIS

Numerical Variables??

- Weight
- Height
- Temperature etc.

Paired Numerical Variables??

**Paired Variables**

| ID_No (Females) | Age | Systolic BP |
|---|---|---|
| 001 | 45 | 151 |
| 002 | 25 | 138 |
| 003 | 48 | 143 |
| 004 | 37 | 140 |
| 005 | 24 | 136 |

**Unpaired Variables**

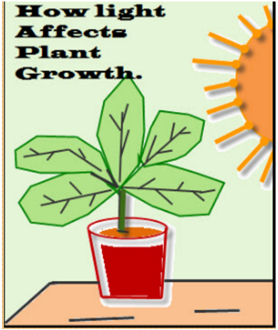
| Age of Females | Systolic BP of Males |
|---|---|
| 45 | 149 |
| 25 | 150 |
| 48 | 138 |
| 37 | 142 |
| 24 | 139 |

**Dependent Variable??**

The variable we wish to explain

**Independent Variable??**

The variable we use to explain the dependent variable



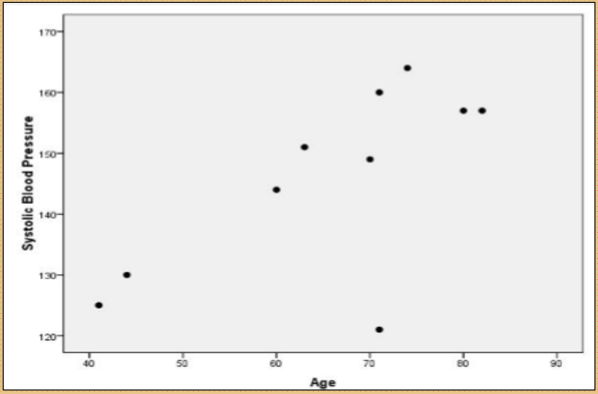
# How to identify Relationships??

Basically, we will learn three main methods. They are,

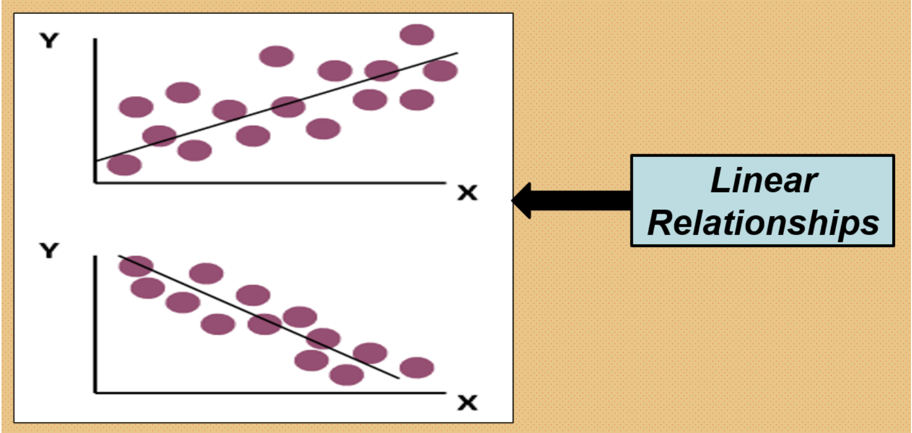
- Scatter plot **(Graphical Method)**
- Correlation
- Regression Analysis

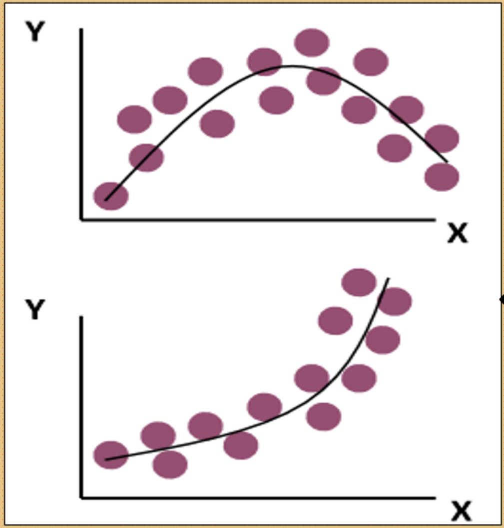
# Scatter Plot

| Age | Systolic BP |
|---|---|
| 63 | 151 |
| 70 | 149 |
| 74 | 164 |
| 82 | 157 |
| 60 | 144 |
| 44 | 130 |
| 80 | 157 |
| 71 | 160 |
| 71 | 121 |
| 41 | 125 |

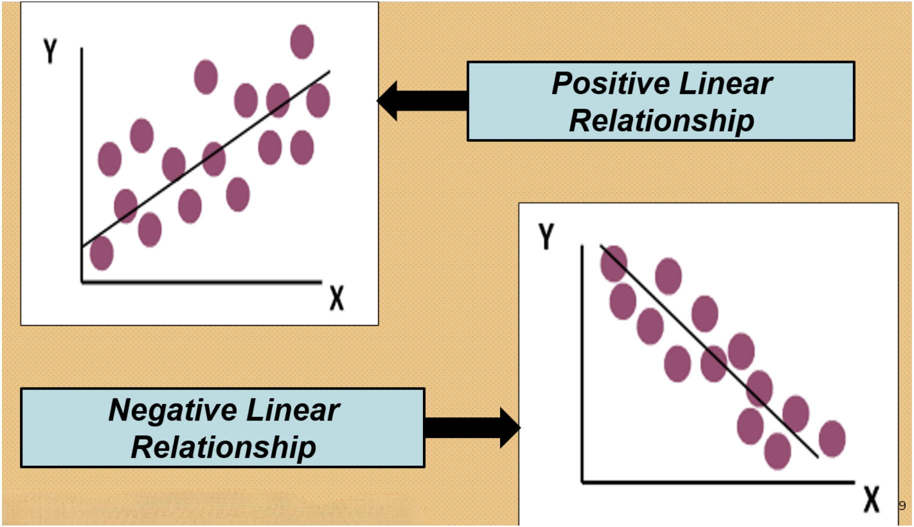


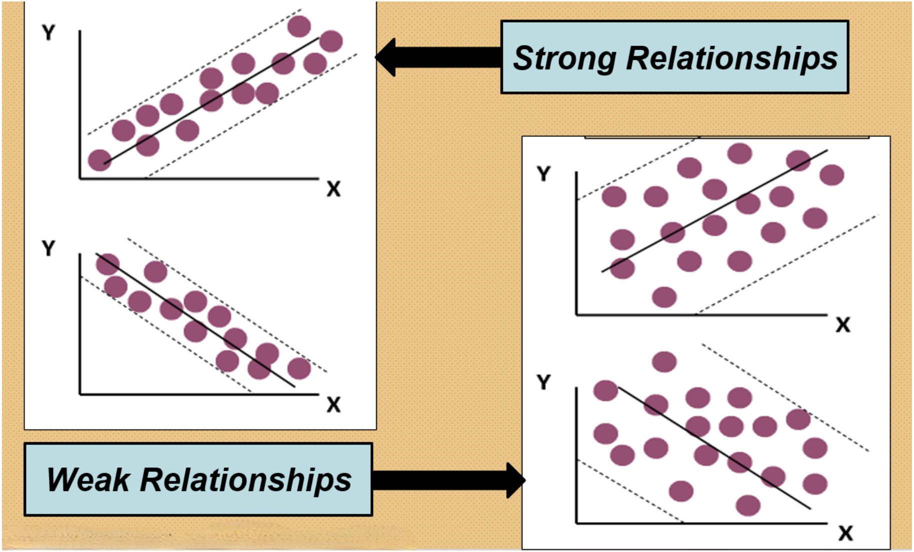
# Types of Relationships

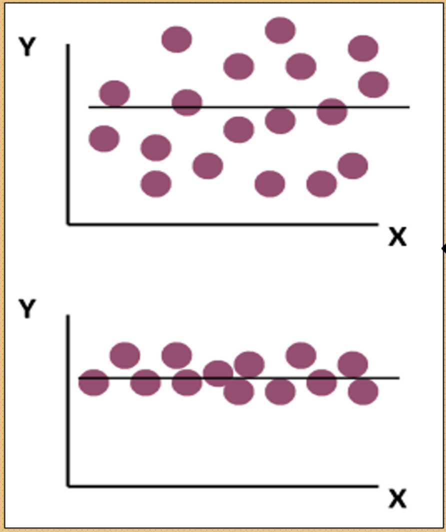





**Non-Linear Relationships**


Y
X
Positive Linear Relationship
Negative Linear Relationship
Y
X
9


Strong Relationships
Weak Relationships
Y
X
Y
X
Y
X
Y
X




**_No Relationships_**

# Correlation??

# Correlation

- This measures **strength** and the **direction** of the **linear relationship** between two numerical variables.
- Correlation is a **value** in between **-1 & +1.**

***Population*** $\longrightarrow$ $\rho$

***Sample*** $\longrightarrow$ $r$

This is also known as **Pearson product-moment correlation coefficient.**

# Sample correlation coefficient (r)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2]\,[n(\sum y^2) - (\sum y)^2]}}$$

$r = -1$ → Perfect Negative Linear Relationship

$r = 0$ → No Linear Relationship

$r = +1$ → Perfect Positive Linear Relationship

# Exercise:

In the pursuit of finding whether the age is related with systolic blood pressure of females, the following data were observed from 10 randomly selected females between ages 40 and 82.


| Age | Systolic BP |
|---|---|
| 63 | 151 |
| 70 | 149 |
| 74 | 164 |
| 82 | 157 |
| 60 | 144 |
| 44 | 130 |
| 80 | 157 |
| 71 | 160 |
| 71 | 121 |
| 41 | 125 |

# Correlation – Hypothesis Testing

- A hypothesis test can be carried out to find whether the population correlation is zero.
- $H_0 : \rho = 0$ Vs. $H_1 : \rho \neq 0$
- Under $H_0$,

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

Regression??

# Regression

- The process of **finding a mathematical equation** that **best fits the noisy data** is known as **regression analysis**.
- In this chapter, only **Simple Linear Regression model and Multiple Linear Regression model** will be discussed.
- The **primary usage** of a regression model is **prediction**.

# Simple Linear Regression Model

$$\boxed{Y = \alpha + \beta X + \varepsilon}$$

- $\alpha$ - y Intercept
- $\beta$ - Regression Coefficient (Slope)
- $\varepsilon$ - Random Error

- This model is defined for population data.
- Should be careful when making predictions outside the observed range.

- $\alpha$ and $\beta$ in the regression model are population characteristics which cannot be measured straightaway.
- Therefore, they should be estimated by using sample data.
- Estimated regression model would be as follows.

$$\hat{y} = \hat{\alpha} + \hat{\beta} X$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\alpha} = a = \bar{y} - b\bar{x}$$

# Significance of Regression Coefficient

- A hypothesis test can be carried out to find whether the true slope ($\beta$) is actually zero (This is same as testing whether the regression model is significant).
- An **ANOVA** table is used to evaluate the **test statistic** for this test.

# ANOVA Table

| Model | Df (Degrees of Freedom) | Sum of Squares (SS) | Mean Sum of Square (MSS) | F Statistic | P Value |
|---|---|---|---|---|---|
| Regression | 1 | SSR | MSSR | F Statistic | |
| Error / Residual | n-2 | SSE | MSSE | | |
| Total | n-1 | SST | | | |

- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $SST = SSR + SSE$
- $MSSR = SSR/1$
- $MSSE = SSE/(n-2)$
- $F\ Statistic = MSSR/MSSE$
- $P\ value = Pr(F > F_{Cal})$

# Coefficient of Determination ($R^2$)

One way to measure the strength of the relationship between the response variable (y) and the predictor variable (x) is to calculate coefficient of determination.

This refers to the proportion of the total variation that is explained by the linear regression of y on x. In other words, $R^2$ is percentage of variation of Y explained by the X variable in the fitted model.

$$R^2 = \frac{SSR \; * 100}{SST}$$

FACULTY OF COMPUTING

# Regression Assumptions

- The model is linear in parameters
- $E(\varepsilon_i) = 0$ (Mean of residuals is zero)
- $V(\varepsilon_i) = \sigma^2$ (Variance of residuals is constant)
- The residuals $(\varepsilon_i)$ are normally distributed.
- The residuals $(\varepsilon_i)$ are independent.

# Important


Remember that, neither correlation nor regression imply any causation between variables.

# Thanks!

Any questions?