Unit 5

Statistical analysis

- Statistical analysis in R is performed by using many in-built functions.
- Most of these functions are part of the R base package.
- These functions take R vector as an input along with the arguments and give the result.

Mean

- It is calculated by taking the sum of the values and dividing with the number of values in a data series.
- The function **mean()** is used to calculate this in R.

Syntax

The basic syntax for calculating mean in R is :

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

Following is the description of the parameters used −

- x is the input vector.
- trim is used to drop some observations from both end of the sorted vector.
- na.rm is used to remove the missing values from the input vector.

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find Mean.
result <- mean(x)
print(result)
```

Applying Trim Option

- When trim parameter is supplied, the values in the vector get sorted and then the required numbers of observations are dropped from calculating the mean.
- When trim = 0.3, 3 values from each end will be dropped from the calculations to find mean.
- In this case the sorted vector is (−21, −5, 2, 3, 4.2, 7, 8, 12, 18, 54) and the values removed from the vector for calculating mean are (−21,−5,2) from left and (12,18,54) from right.

```
# Create a vector.

x <- c(12,7,3,4.2,18,2,54,-21,8,-5)

# Find Mean.

result.mean <-  mean(x,trim = 0.3)

print(result.mean)
```

Applying NA Option

- If there are missing values, then the mean function returns NA.
- To drop the missing values from the calculation use na.rm = TRUE. which means remove the NA values.

```
# Create a vector.
x <- c(12,7,3,4.2,18,2,54,-21,8,-5,NA)

# Find mean.
result.mean <-  mean(x)
print(result.mean)
```

```
# Find mean dropping NA values.
result.mean <-  mean(x,na.rm = TRUE)
print(result.mean)
```

Median

- The middle most value in a data series is called the median.
- The **median()** function is used in R to calculate this value.

Syntax

The basic syntax for calculating median in R is −

median(x, na.rm = FALSE)

Following is the description of the parameters used −

- x is the input vector.
- na.rm is used to remove the missing values from the input vector.

Mode

- The mode is the value that has highest number of occurrences in a set of data.

Analyzing the CSV File

```
data <- read.csv(“data.csv")

print(is.data.frame(data))

print(ncol(data))

print(nrow(data))
```

Get the maximum salary

Get the details of the person with max salary

Get all the people working in IT department

Get the persons in IT department whose salary is greater than 600

Get the people who joined on or after 2014

Linear Regression

Regression

Regression is a well-known statistical technique to model the predictive relationship between several independent variables (DVs) and one dependent variable.

Regression

- Linear Regression.
- Multiple regression.

Introduction to Linear Regression

Linear regression may be defined as the statistical model that analyzes the linear relationship between a dependent variable with given set of independent variables.

Mathematically the relationship can be represented with the help of following equation :-

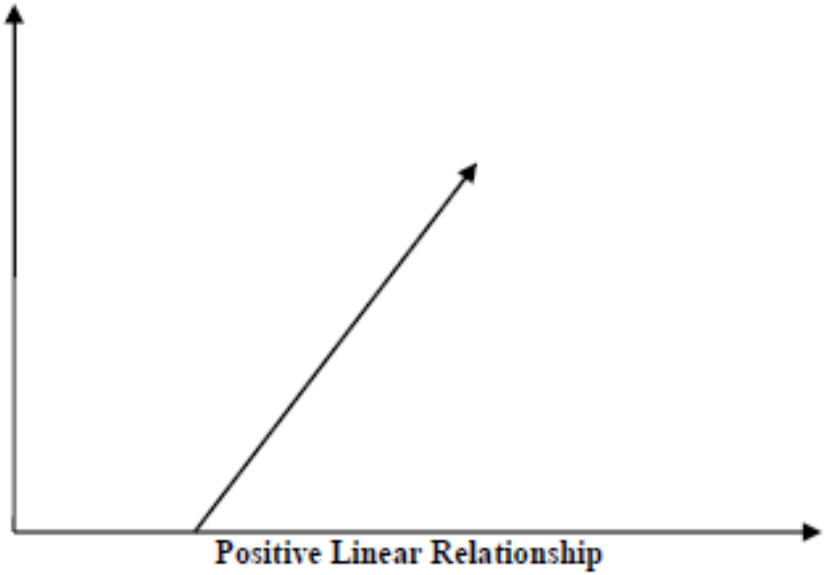
$$Y=mX+b$$

Here,

- Y is the dependent variable we are trying to predict.
- X is the independent variable we are using to make predictions.
- m is coefficient and b is intercept

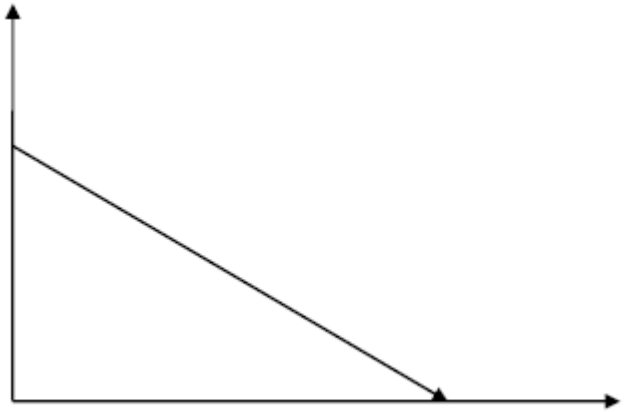
Positive Linear Relationship

- A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



Negative Linear relationship

A linear relationship will be called Negative if independent increases and dependent variable decreases. It can be understood with the help of following graph −



Negative Linear Relationship

Linear Regression

The general mathematical equation for a linear regression is:

$$y = mx+b$$

- Y is dependent variable
- X is Independent variable
- **m is coefficient**
- **b is intercept**

Input Data

Below is the sample data representing the observations −

Values of height

151, 174, 138, 186, 128, 136, 179, 163, 152, 131

Values of weight.

63, 81, 56, 91, 47, 57, 76, 72, 62, 48

lm() Function

This function creates the relationship model between the predictor and the response variable.

Syntax

The basic syntax for lm() function in linear regression is −

lm(formula,data)

Following is the description of the parameters used −

- ***formula*** *is a symbol presenting the relation between x and y.*
- ***data*** *is the vector on which the formula will be applied.*

```
x <- c(151, 174, 138, 186, 128, 136, 179, 163, 152, 131)
y <- c(63, 81, 56, 91, 47, 57, 76, 72, 62, 48)

# Apply the lm() function.
relation <- lm(y~x)

print(relation)
```

predict() Function

Syntax

The basic syntax for predict() in linear regression is −

predict(object, newdata)

Following is the description of the parameters used −

- *object is the formula which is already created using the lm() function.*
- *newdata is the vector containing the new value for predictor variable.*

```
# Find weight of a person with height 170.

a <- data.frame(x = 170)

result <-  predict(relation,a)

print(result)
```