Artificial Intelligence Masterclass

# Linear Regression & Supervised Learning Techniques

**H.M. Samadhi Chathuranga Rathnayake**

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

# Linear Regression

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning.

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **Simple Linear Regression**. When there are multiple input variables, literature from statistics often refers to the method as **Multiple Linear Regression**.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

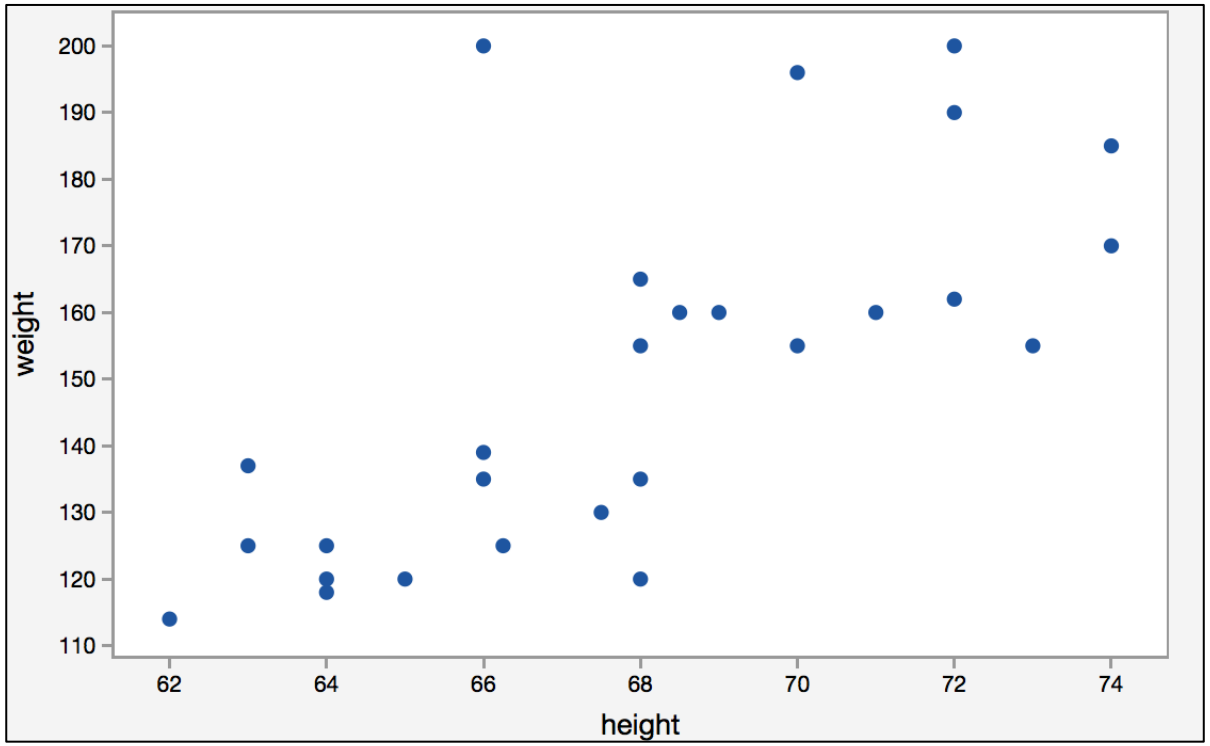
## Simple Linear Regression

Assume that we want to see the relationship between the height and the weight of Sri Lankan university students. Our objective is to fit a model to predict the weight using height of the students.

Since the entire population cannot be accessed, a sample will be taken, and the model will be created.

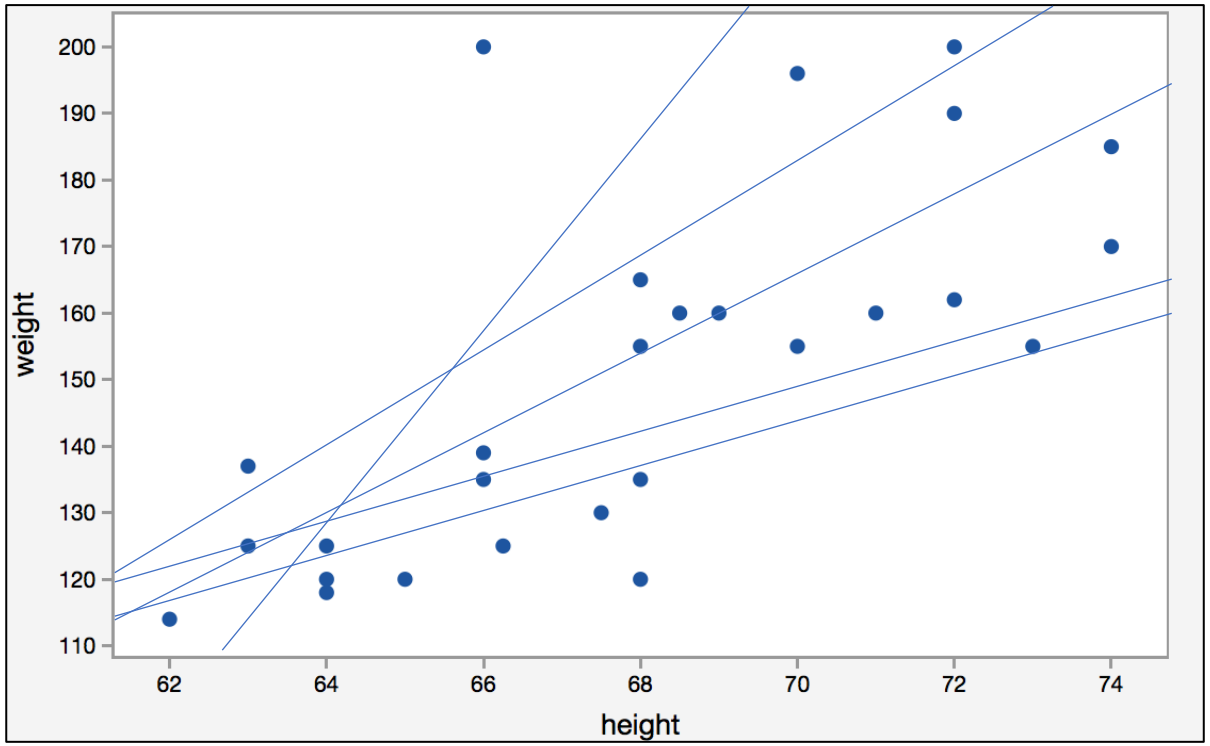
Let’s plot the sample data.

# Simple Linear Regression




Simple linear regression is useful for finding relationship between two variables. One is predictor or independent variable and other is response or dependent variable which is a quantitative variable. For example, relationship between height and weight.

# Simple Linear Regression




Infinitely many number of plots can be created.

## Simple Linear Regression

The equation for this model for the population is as follows,

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The values $\beta_0$ and $\beta_1$ must be chosen so that they minimize the error. If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best estimated model which reduces the error.

$$Sum\ of\ Squares\ of\ Error\ (SSE) = \sum_{i=1}^{n} (Actual\ Output - Predicted\ Output)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

This is called as Residual Sum of Squares (RSS) as well. By minimizing this SSE, we can obtain following parameter estimations.

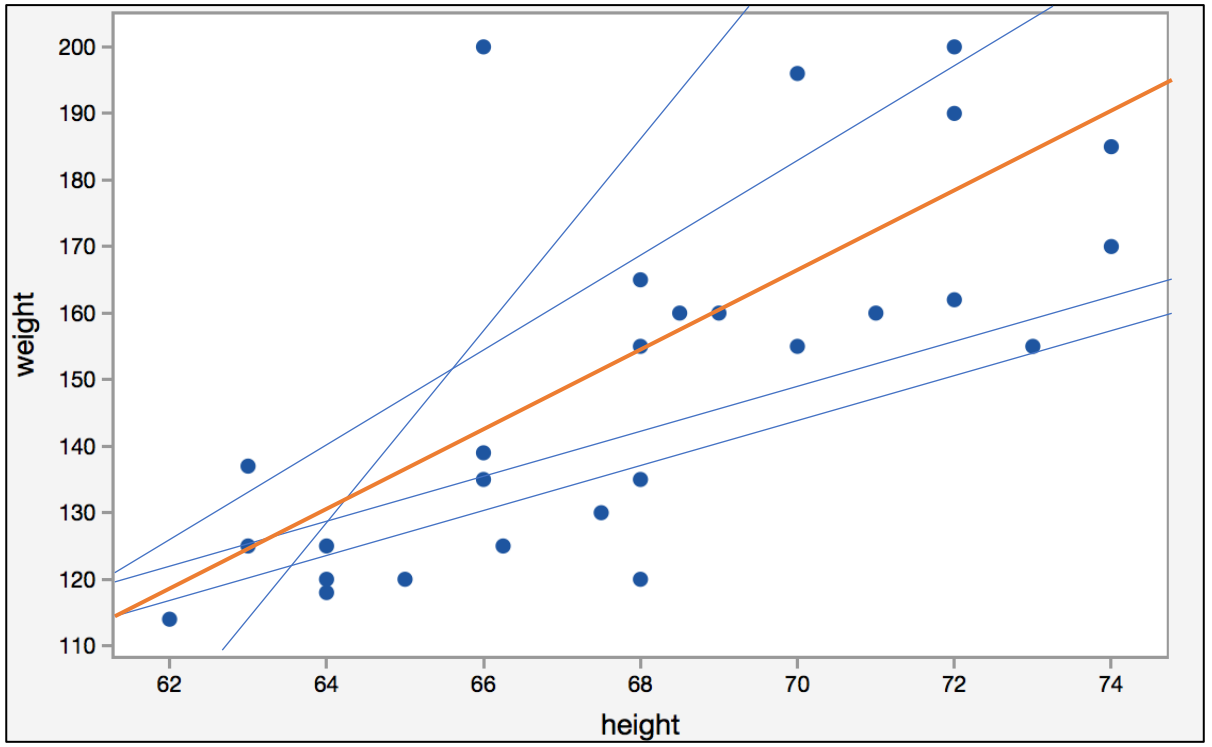
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Then the estimated model is,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

# Simple Linear Regression




Best Fit Line

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible.

## Multiple Linear Regression

The equation for this model is as follows,

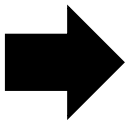
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon$$

Consider here we have k variables. By minimizing this SSE, we can obtain the parameter estimations in here as well. These estimated parameters can be represented as a vector. The parameter vector $\bar{\beta}$ can be obtained through,

$$\underline{\beta} = (X^T X)^{-1} X^T Y$$

Here,

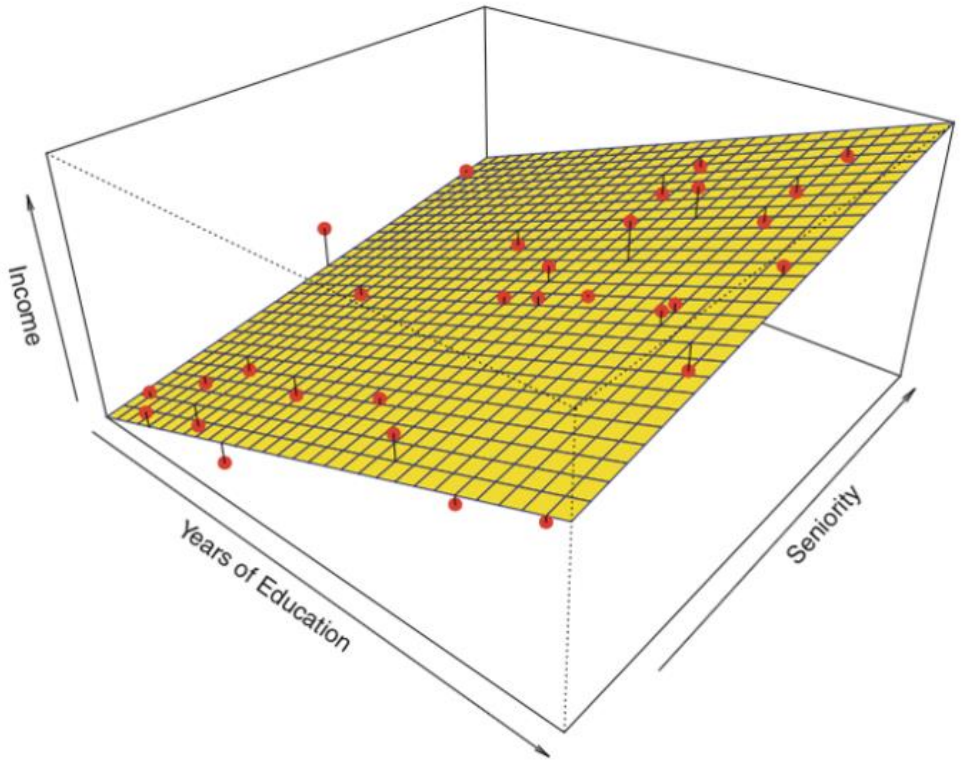
| $X_1$ | $X_2$ | ... | $X_4$ |
|---|---|---|---|
| $X_{11}$ | $X_{12}$ | ... | $X_{1k}$ |
| $X_{21}$ | $X_{22}$ | ... | $X_{2k}$ |
| ... | ... | ... | ... |
| $X_{n1}$ | $X_{n2}$ | ... | $X_{nk}$ |



$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix}$$

## Multiple Linear Regression

How this model is visualized. Consider a 2 variables with one response example. Here the **Income** is the response variable and **Seniority** and the **Years Of Education** are the independent variables.




Higher dimensions cannot be visualized easily. There are several advanced techniques for visualize them.

## Qualitative Predictors (Categorical Independent Variables)

Categorical variables cannot be added to the model as the numerical variables.

Ex- University Year Variable (First, Second, Third, Fourth)

| First | Second | Third | Fourth |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

These are called as **Dummy Variables.**

## Qualitative Predictors (Categorical Independent Variables)

Since one variable can be represented as other variable's 0 case, one variable can be removed. That is called as the **reference level**.

Consider **First** as the reference level.

| First | Second | Third | Fourth |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

This is called as **Dummy Trapping.**

## Qualitative Predictors (Categorical Independent Variables)

Since one variable can be represented as other variable's 0 case, one variable can be removed. That is called as the **reference level**.

| Second | Third | Fourth |
|---|---|---|
| 0 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

This is called as **Dummy Trapping.**

## Dummy Variables

To represent categorical variables in a linear regression model, we use dummy variables. Consider following example.

Ex:- Gender

| Gender | Dummy Variable (G) |
|---|---|
| Male | 1 |
| Female | 0 |

Ex- Temperature (High, Medium, Low)

| Temperature | Dummy Variable (T1) | Dummy Variable (T2) |
|---|---|---|
| High | 1 | 0 |
| Medium | 0 | 1 |
| Low | 0 | 0 |

Ex- Colour (Red, Green, Yellow, Blue)

| Colour | Dummy Variable (C1) | Dummy Variable (C2) | Dummy Variable (C3) |
|---|---|---|---|
| Red | 1 | 0 | 0 |
| Green | 0 | 1 | 0 |
| Yellow | 0 | 0 | 1 |
| Blue | 0 | 0 | 0 |

## Dummy Variables

Now assume that the model to be fitted for a numerical variable (X) and the above discussed categorical variable **Colour**.

| Colour | Dummy Variable (C1) | Dummy Variable (C2) | Dummy Variable (C3) |
|---|---|---|---|
| Red | 1 | 0 | 0 |
| Green | 0 | 1 | 0 |
| Yellow | 0 | 0 | 1 |
| Blue | 0 | 0 | 0 |

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 c_1 + \hat{\beta}_3\ c_2 + \hat{\beta}_4 c_3$$

## R Squared Value (Coefficient of Determination)

$$Total\ Sum\ of\ Squares\ (TSS) = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$Sum\ of\ Squares\ of\ Error\ (SSE) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

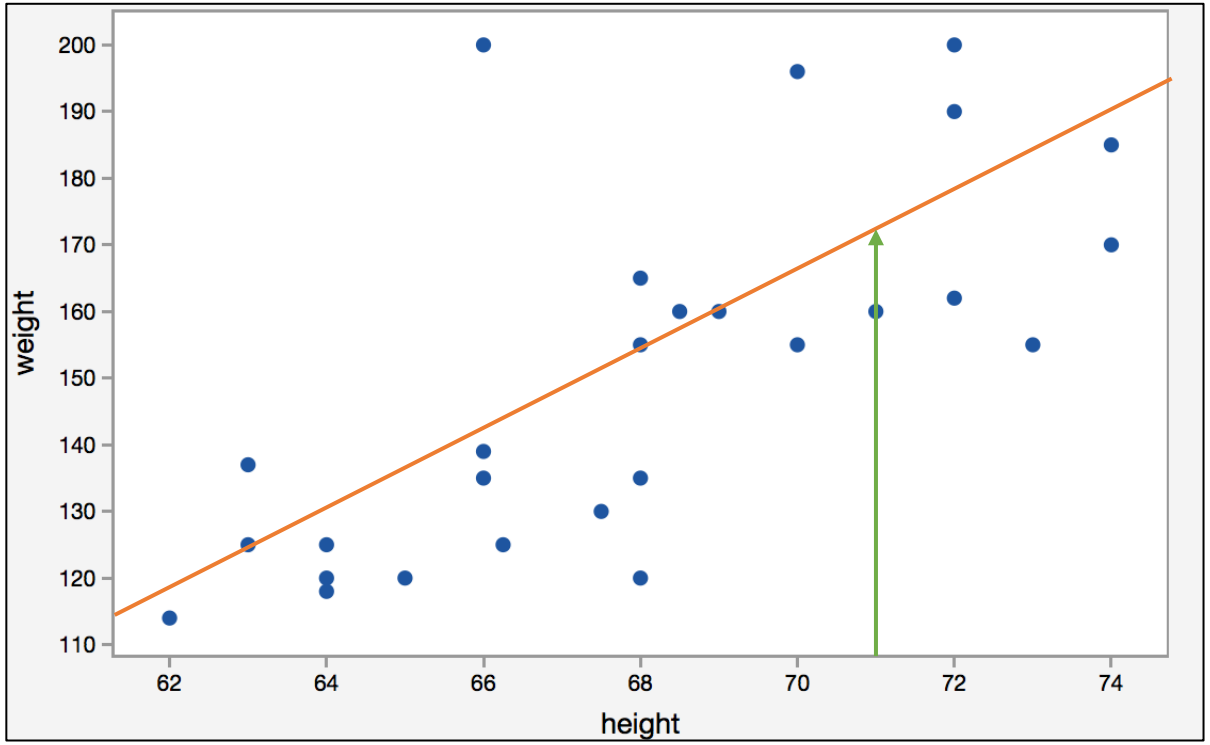
$$Coefficient\ of\ Determination = R^2 = \frac{TSS - SSE}{TSS}$$

This R Squared Value is explaining the fraction of variation explained by the estimated model. In simple words how much of the data captured by this model.

For the Simple Linear Regression case $Coefficient\ of\ Determination = R^2 = (Correlation)^2$

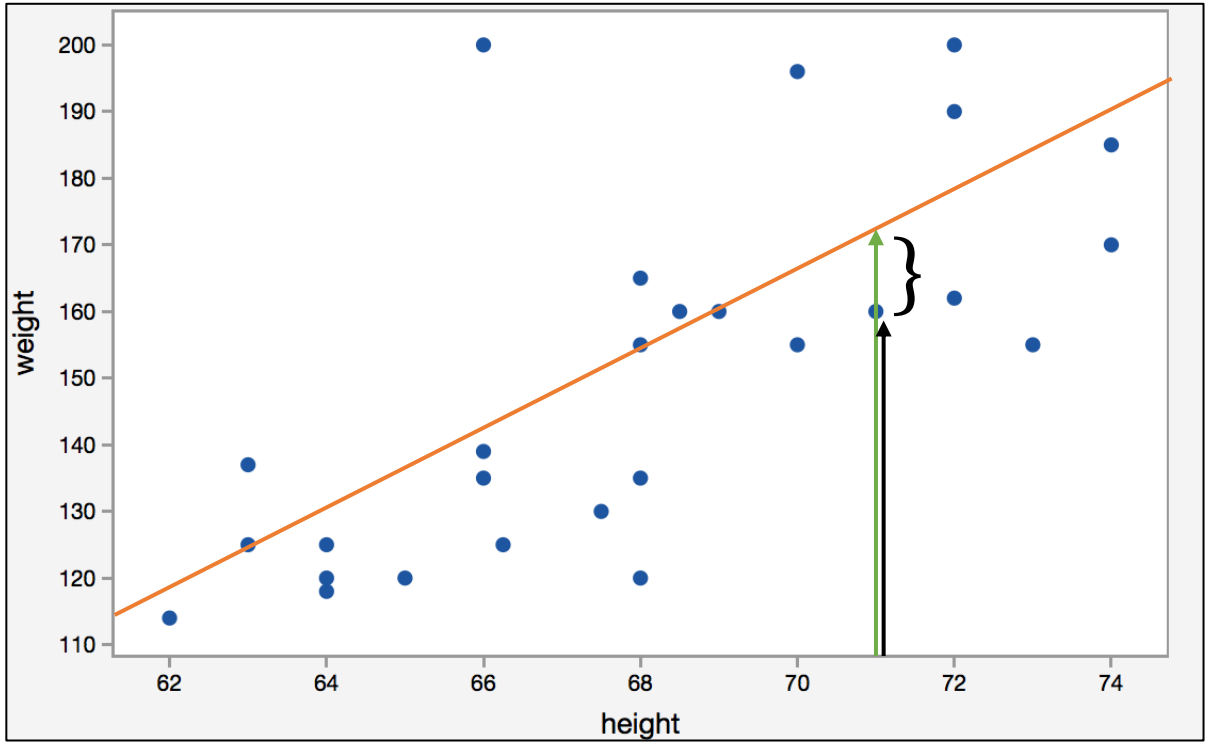
# Predictions

After fitting the model, the goal is to predict the response using independent data.



# Prediction Error

Always there is a prediction error.



## Model Evaluation – Validation Set Approach

Split the dataset into two sets,

- Training dataset (Generally 80% of the data But it can be changed)
- Testing dataset (Rest of the data)

Train the model using the training dataset and then check the accuracy of the model using the testing dataset. MSE of a regression model will be calculated and the model will be evaluated.

$$Mean\ Squared\ Error = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

# Model Evaluation – Cross Validation Approach

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.