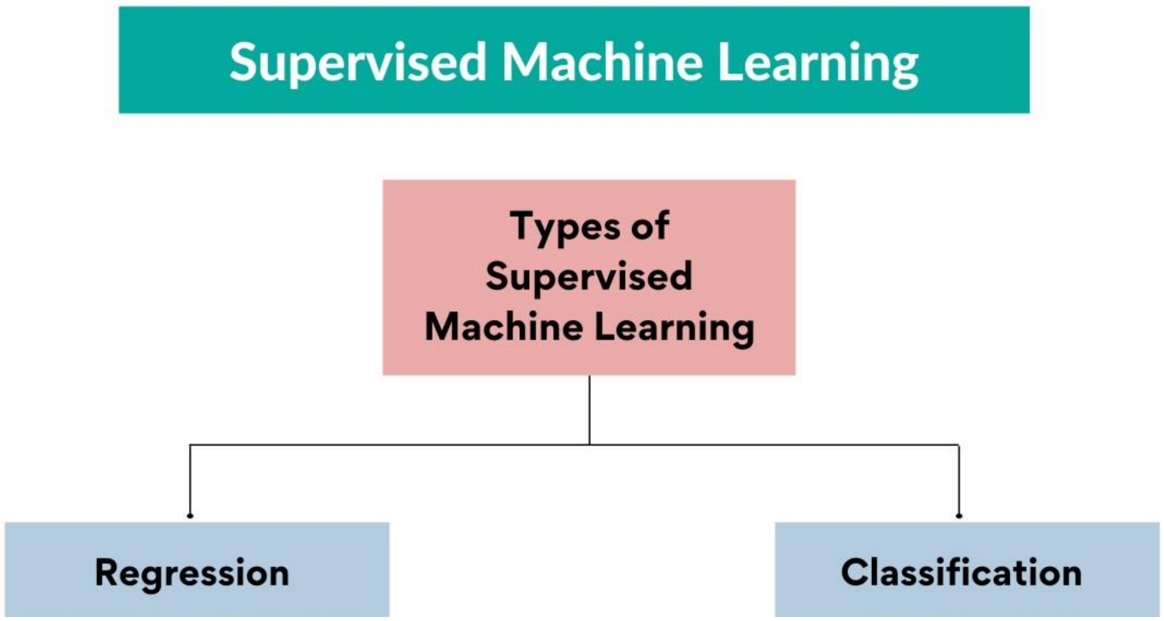
Supervised Machine Learning

Madushanka Wadumulla
PGC in IA
Beng, IIESL, AMEI

Outline

- Types of Supervised Learning
- Supervised Learning
- Classification
- Types of ML Classification Algorithms
- Regression
- Types of ML Regression Algorithms
- Linear Regression
- Simple linear regression
- Random Error (Residuals)
- What is the best fit line?
- Linear Regression - Cost function
- Gradient Descent for Linear Regression
- Simple linear regression (Least Square Method)
- Multiple linear regression

Types of Supervised Learning



Continuous Output

Categorical Output

Examples

Score prediction
Sales prediction
Temperature Prediction

Examples

Fraud detection
Email Spam detection
Diagnostics
Image Classification

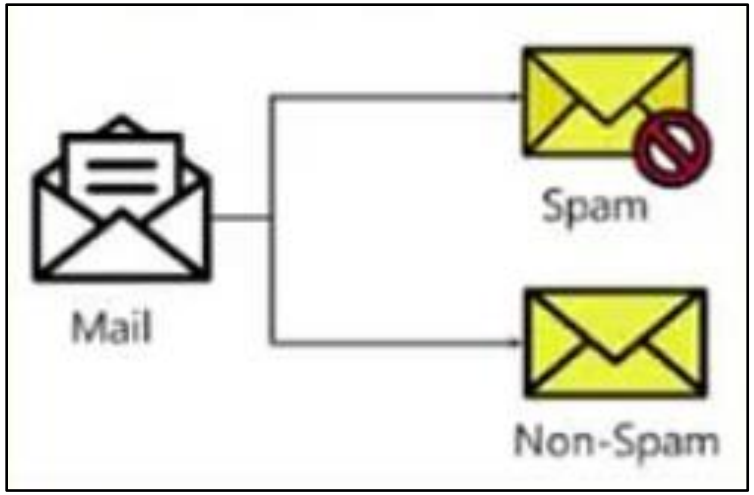
Supervised Learning

- We have training data in the form of d-dimensional points, each with an associated label.
- In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes.
- Unlike regression, the output variable of Classification is a category, not a value, such as “Green or Blue”, “fruit or animal”, etc.
- Simply put, we want to learn a mapping from the data points to labels, using the training data, such that this mapping helps us get reliable estimates of the labels of the test data points.
- If there is a function that captures this relationship perfectly, we would have solved the problem satisfactorily if our learning algorithm can learn that function from the training data.

Classification

Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters.
In Classification, a computer program is trained on the training dataset and based on that training, it categorizes the data into different classes.

Example: Email Spam Detection.
The model is trained on the basis of millions of emails on different parameters, and whenever it receives a new email, it identifies whether the email is spam or not.
If the email is spam, then it is moved to the Spam folder.

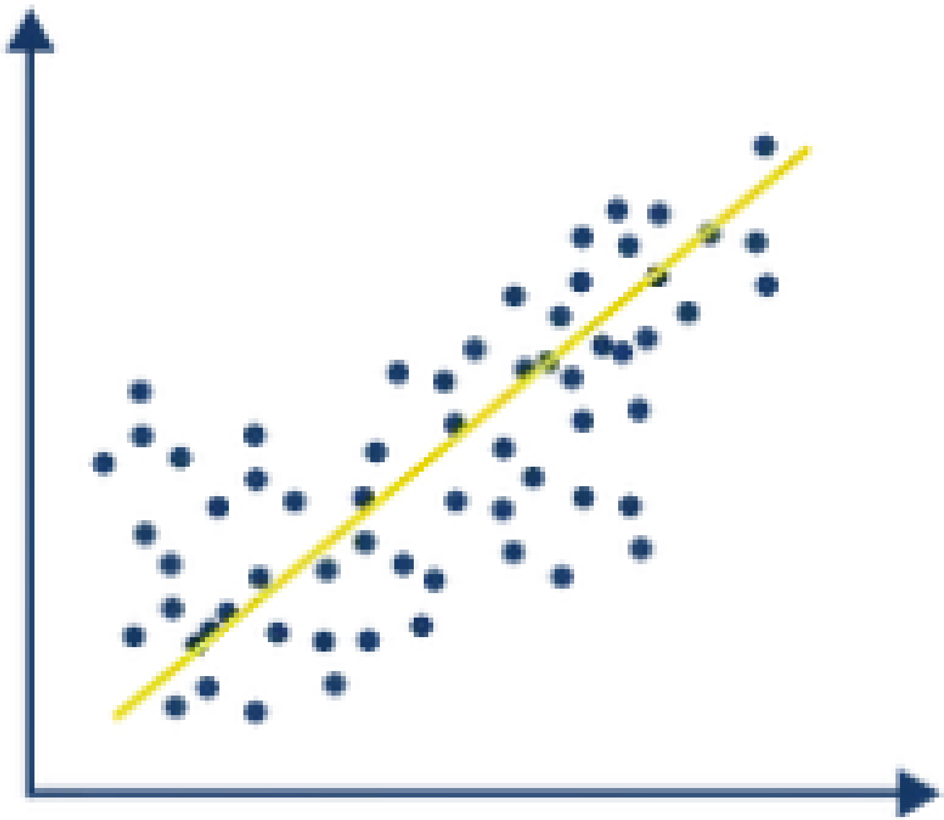


Types of ML Classification Algorithms

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Kernel SVM
- Naïve Bayes
- Decision Tree Classification

Regression

Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices.
Regression algorithms learn a function that maps from the input features to the output value.



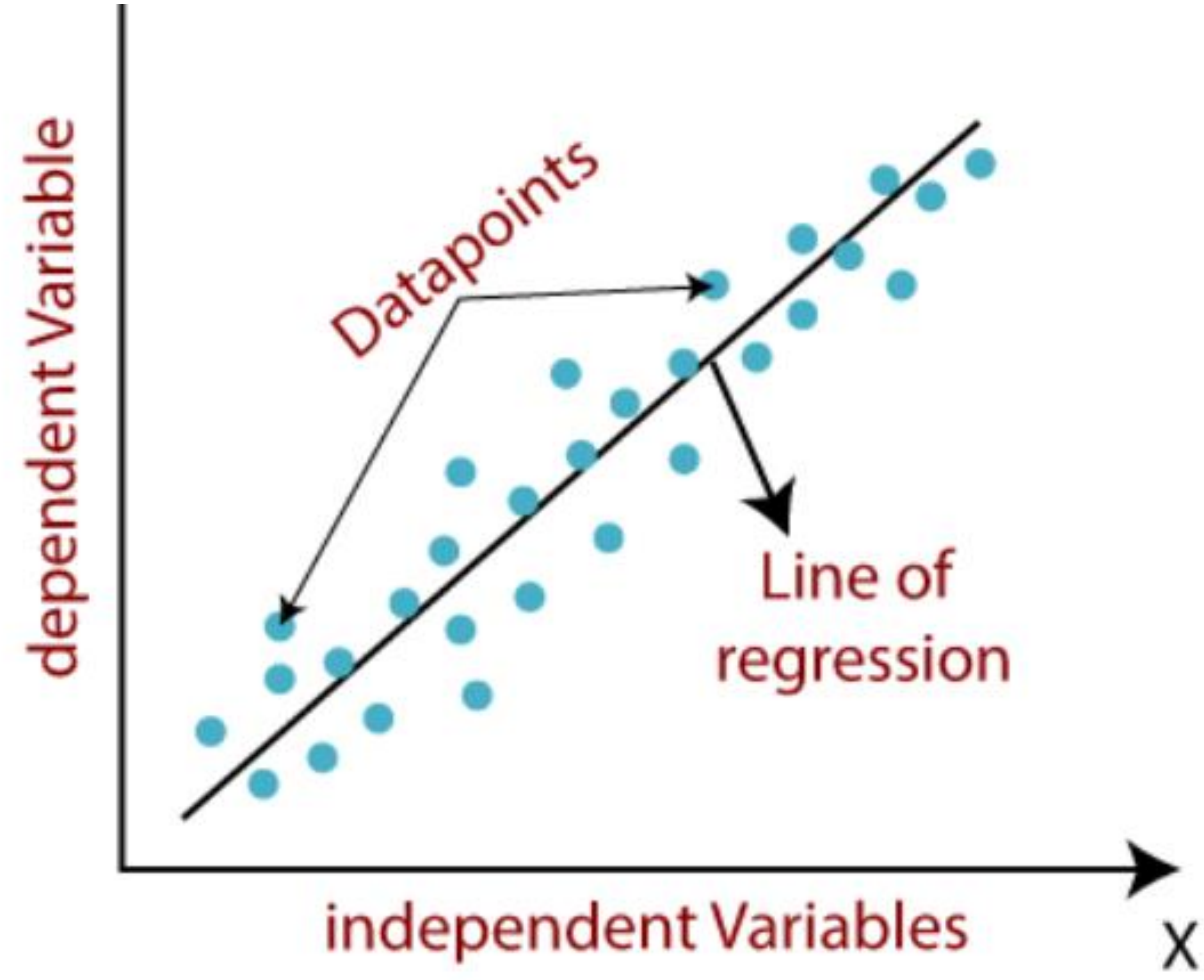
Types of ML Regression Algorithms

- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

Regression

Supervised learning technique
Used when the value that you want to predict is a **continuous** variable
e.g.

- Predicting the **height of a person** based on the **nationality, age, gender**….
- Predicting the **housing price** based on the **floor area, no. of floors, city**……
- Predicting the **value of a share** based on the **company revenue, profit,**……



Regression Line

Linear Regression

The goal of linear regression is to model the **relationship between one or multiple features** and a **continuous target variable**.

The most basic type of linear regression,

Simple Linear Regression:

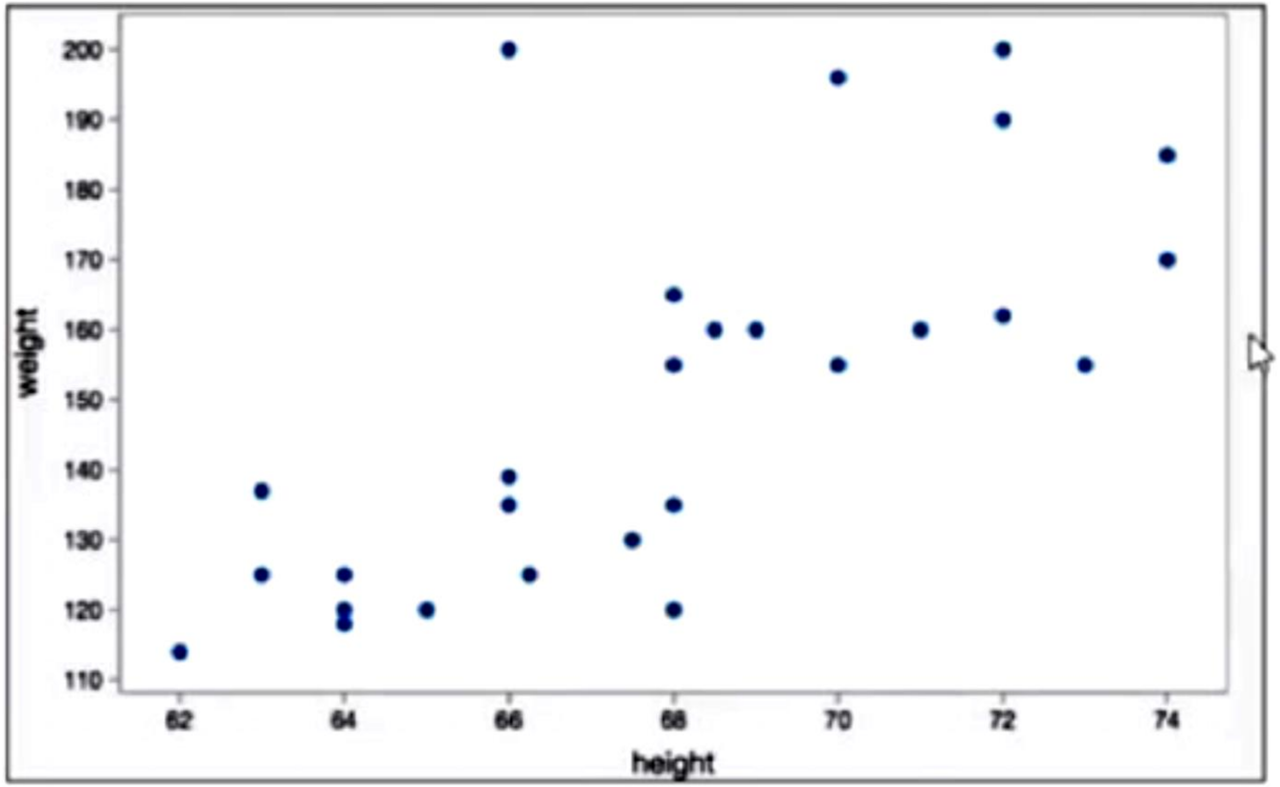
If a single independent variable is used to predict the value of a numerical dependent variable.

Multiple Linear regression:

If more than one independent variable is used to predict the value of a numerical dependent variable.

Simple linear regression

- In a **simple linear regression**, there is **one independent** variable and **one dependent variable**.
- The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables.
- The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.



200
190
180
170
160
150
140
130
120
110
weight
62
64
66
68
70
72
74
height

Simple linear regression

The equation for simple linear regression is:

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

$[h_\theta(x) = \theta_0 + \theta_1 x]$

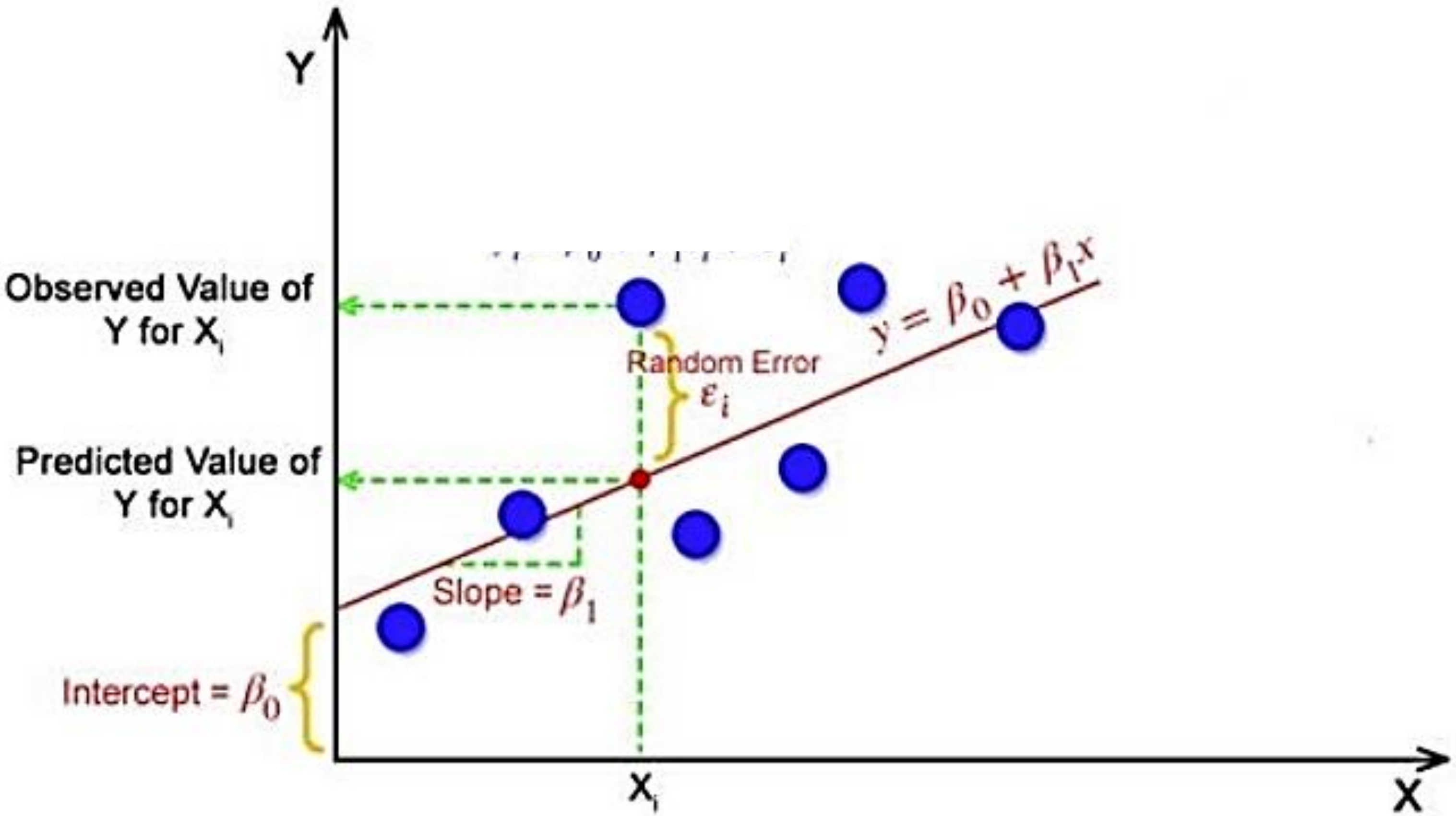
where:

Y - is the dependent variable

X - is the independent variable

β_0 is the intercept

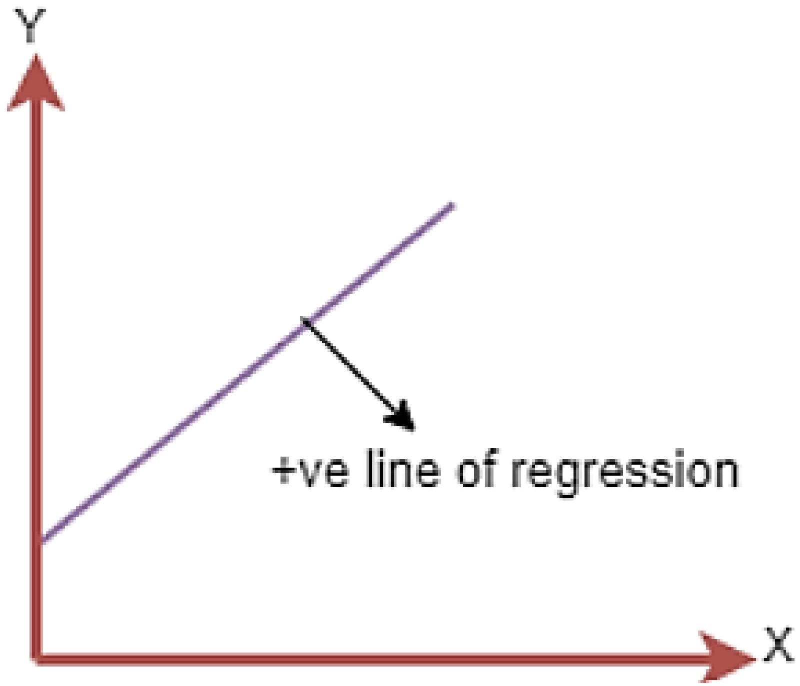
β_1 is the slope



Linear Regression Line

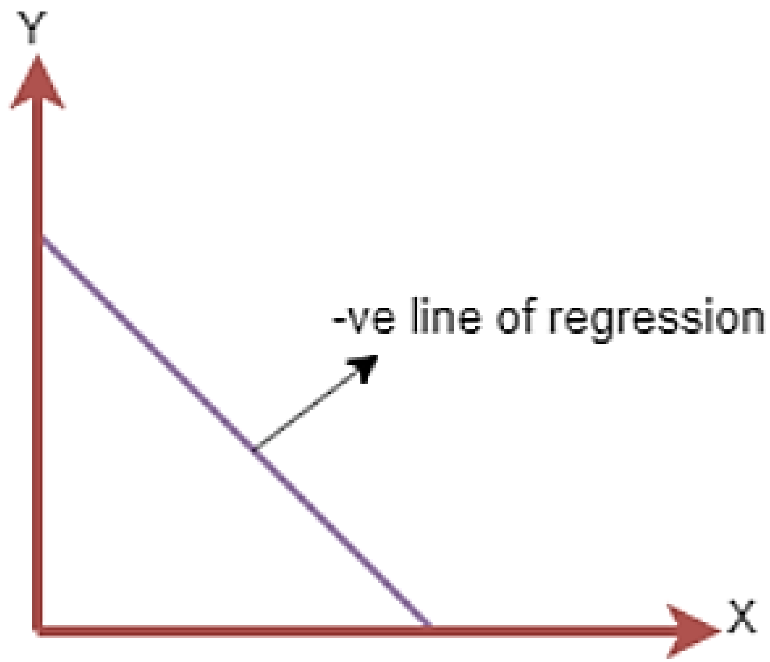
Positive Linear Relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



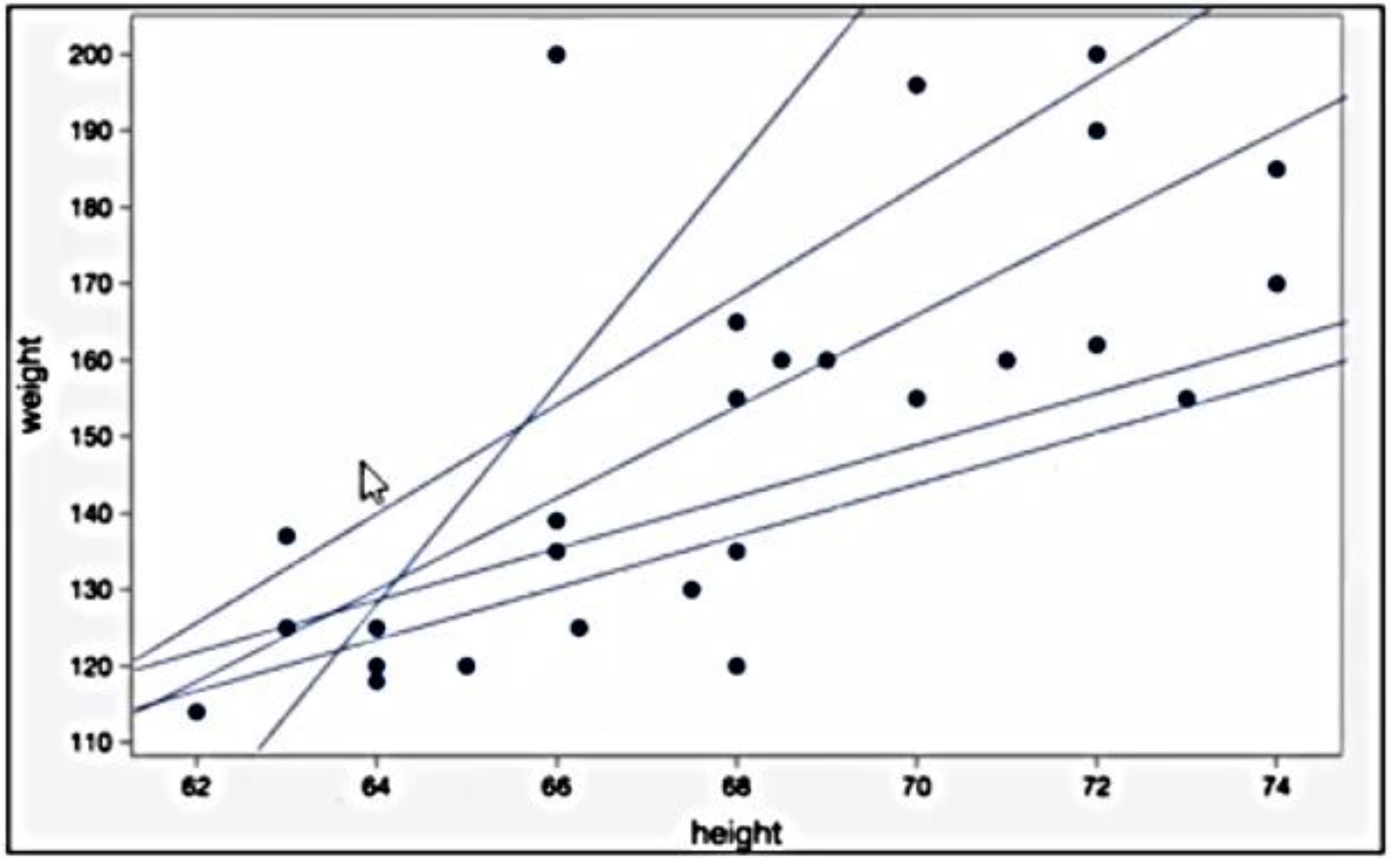
Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



Simple linear regression

The goal of the linear regression algorithm is to get the best values for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ to find the best fit line. The best fit line is a line that has the least error which means the error between actual value and predicted values should be minimum.



Random Error (Residuals)

In regression, the difference between the observed value of the dependent variable $y^{(i)}$ and the predicted value($\hat{Y}^{(i)}$) is called the residuals.

$$\varepsilon_i = \hat{Y}^{(i)} - Y^{(i)}$$

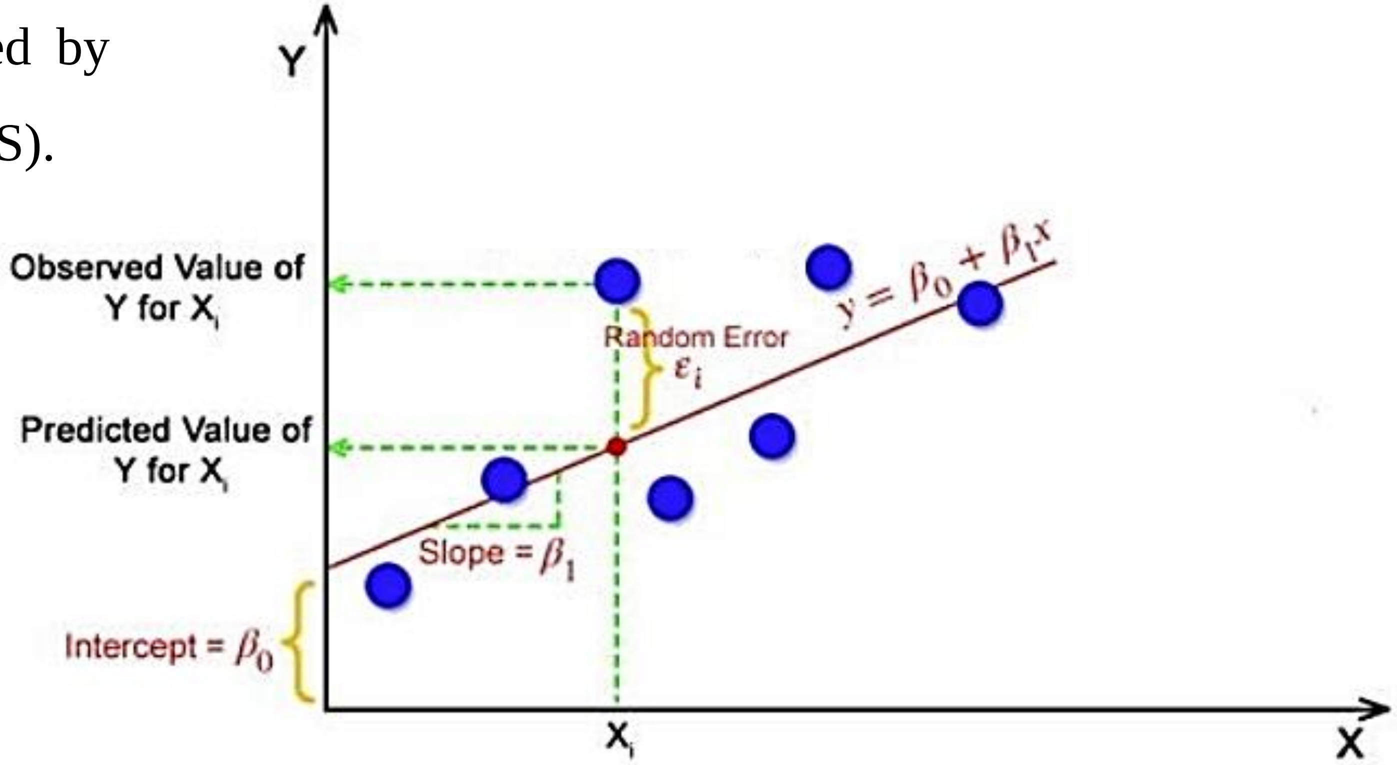
Where;

$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\mathbf{x}$

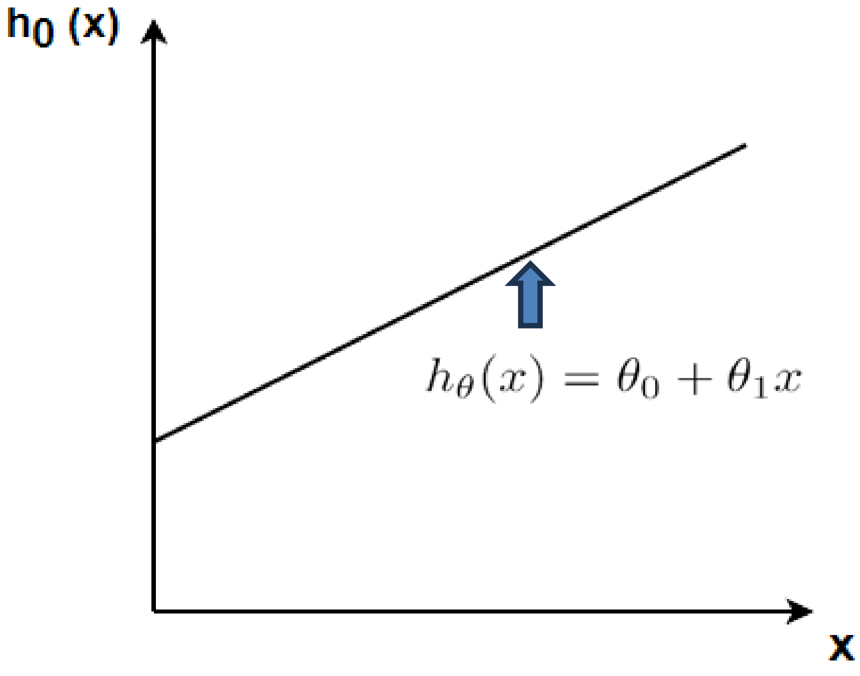
What is the best fit line?

In simple terms, the best fit line is a line that fits the given scatter plot in the best way.

Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares(RSS).



Linear Regression - Cost function



$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

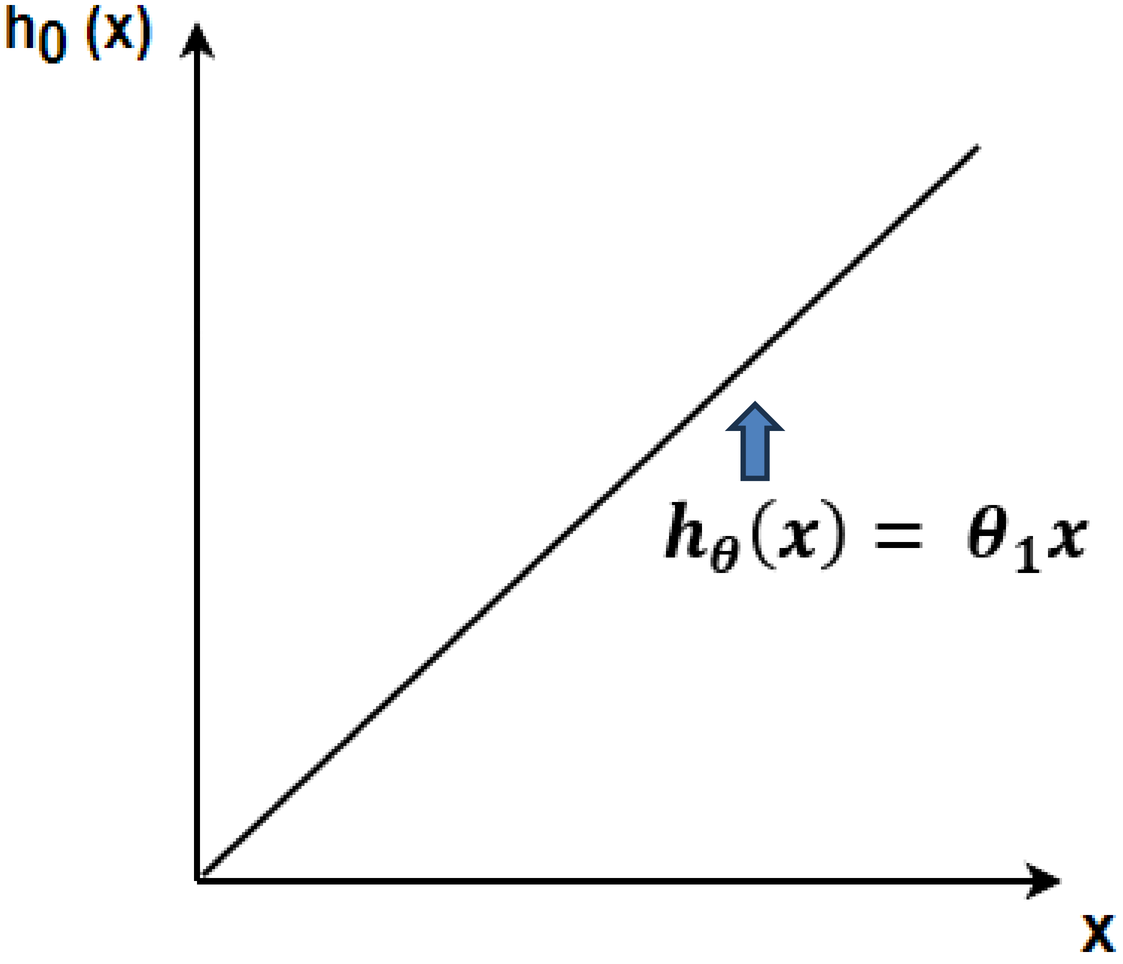
Minimized:

Linear Regression - Cost function

If $\boldsymbol{\beta_0}$=0 ➔ $y = \boldsymbol{\beta_1} x$

Actual Dataset

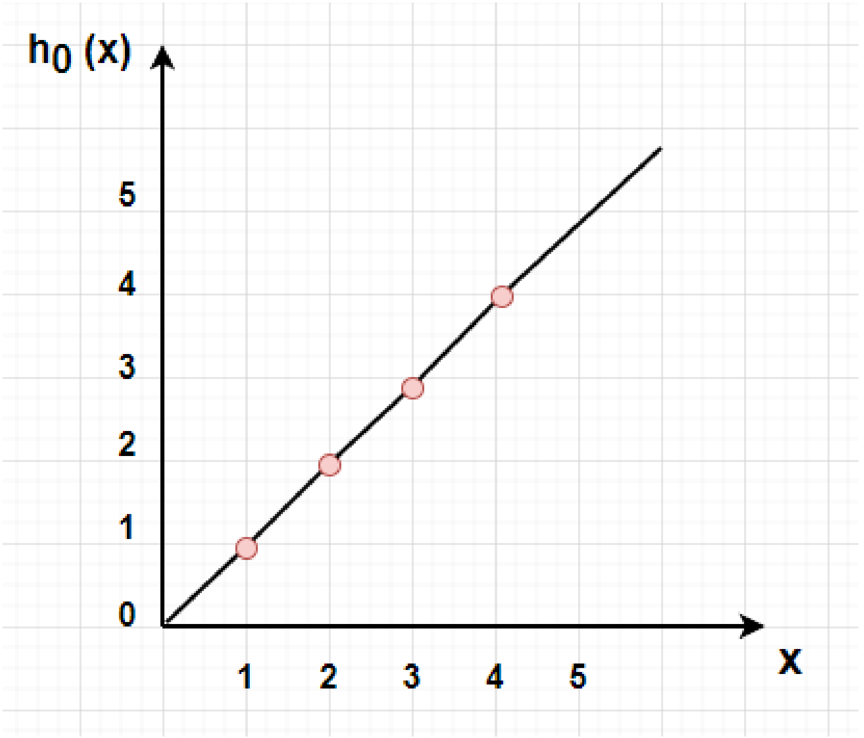
x	***y***
1	1
2	2
3	3
4	4



Linear Regression - Cost function

$h_\theta(x) = \theta_1 x$ and If θ_1 = 1

x	y	$h_\theta(x)$
1	1	1
2	2	2
3	3	3
4	4	4



$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta\,(x)^{(i)} - y^{(i)})^2$; where $[h_\theta(x) = \theta_1 x]$

J

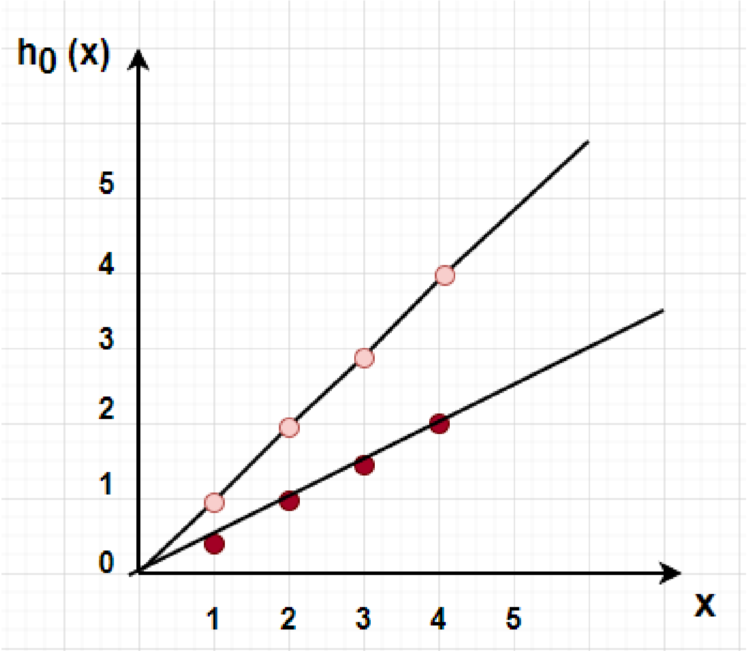
Linear Regression - Cost function

$h_\theta(x) = \theta_1 x$ and If θ_1 = 0.5

x	y	$h_\theta(x)$
1	1	0.5
2	2	1
3	3	1.5
4	4	2

$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta\,(x)^{(i)} - y^{(i)})^2$; where $[h_\theta(x) = \theta_1 x]$

$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\theta_1 (x)^{(i)} - y^{(i)})^2$$



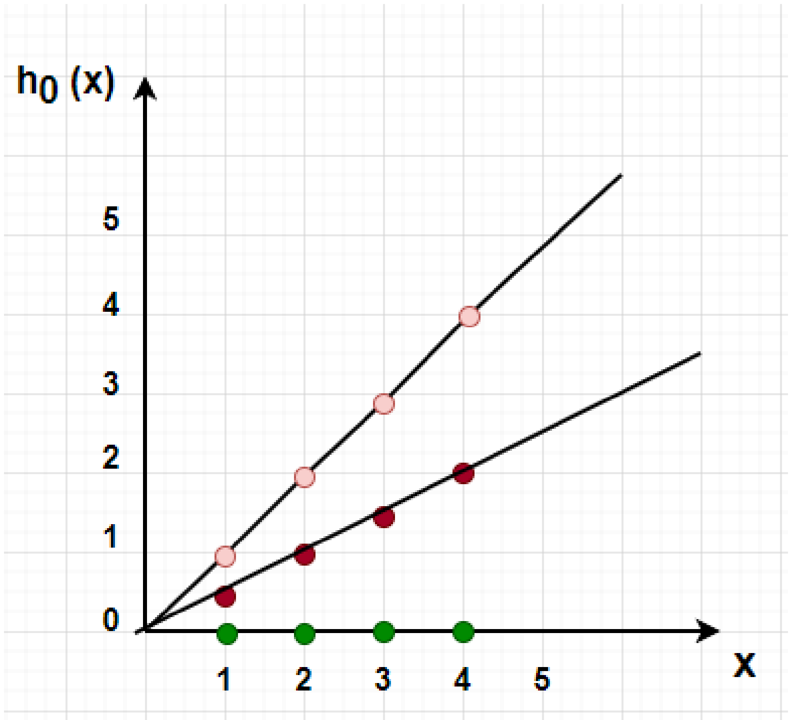
Linear Regression - Cost function

$h_\theta(x) = \theta_1 x$ and If θ_1 = 0

$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta\,(x)^{(i)} - y^{(i)})^2$; where $[h_\theta(x) = \theta_1 x]$

$$J(\theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (\theta_1 (x)^{(i)} - y^{(i)})^2$$

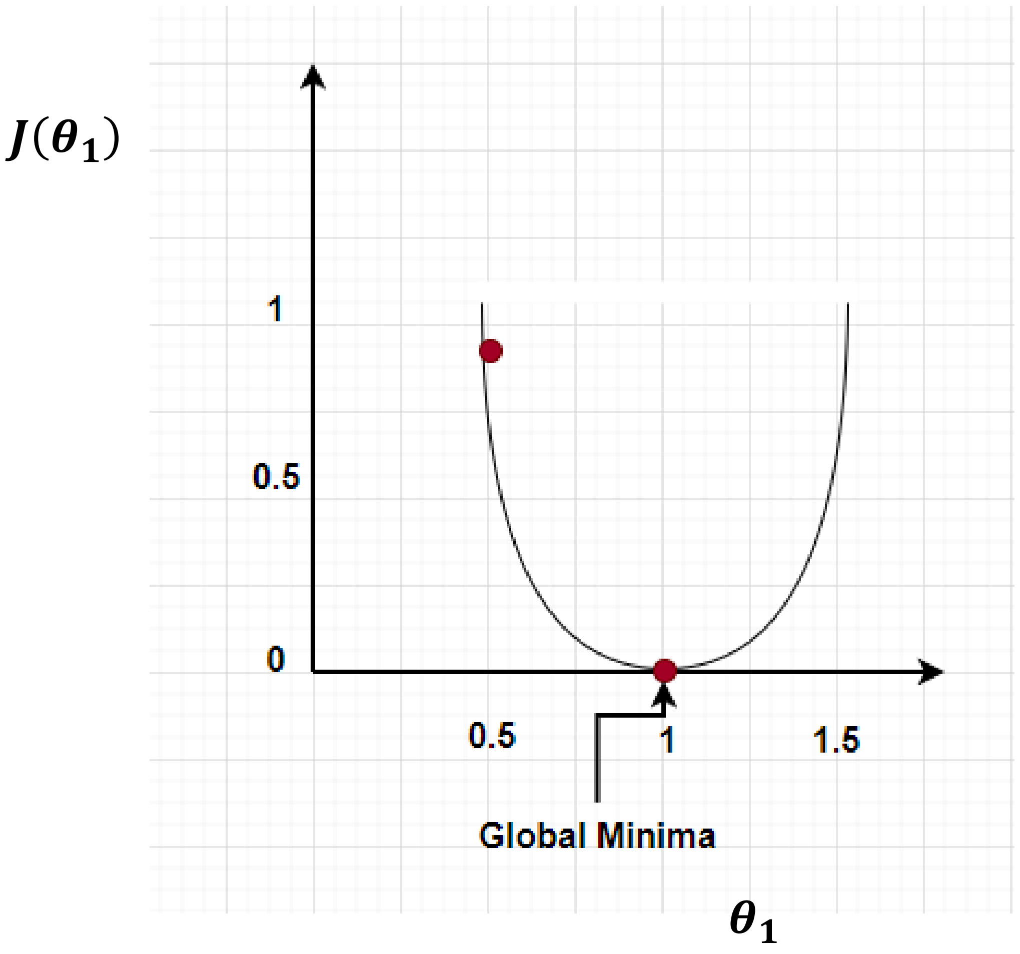
x	y	$h_\theta(x)$
1	1	0
2	2	0
3	3	0
4	4	0



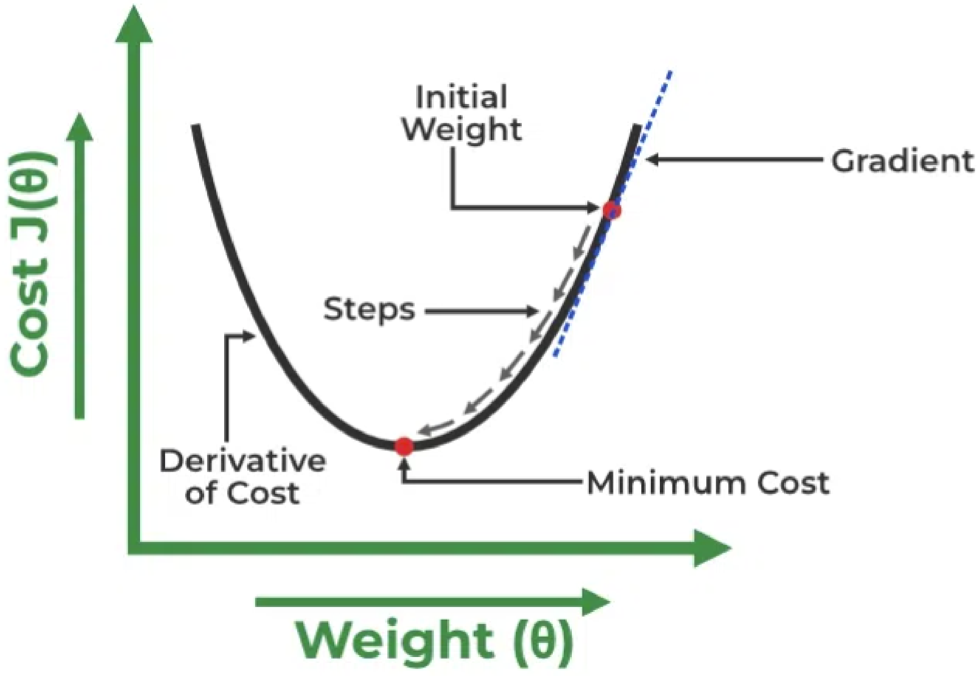
Linear Regression - Cost function

$h_\theta(x) = \theta_1 x$ and If $\theta_1 = 0$

$\boldsymbol{\theta_1}$	$\boldsymbol{J(\theta_1)}$
0	3.75
0.5	0.9375
1	0



Gradient Descent for Linear Regression



Gradient Descent is an iterative optimization algorithm that tries to find the optimum value of an objective function.

It is one of the most used optimization techniques in machine learning projects for updating the parameters of a model in order to minimize a cost function.

This is done by updating the values of θ_0 and θ_1 iteratively until we get an optimal solution.

Simple linear regression (Least Square Method)

As said earlier the equation for this model for the population is as follows,

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

The values 8, and f; 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.

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

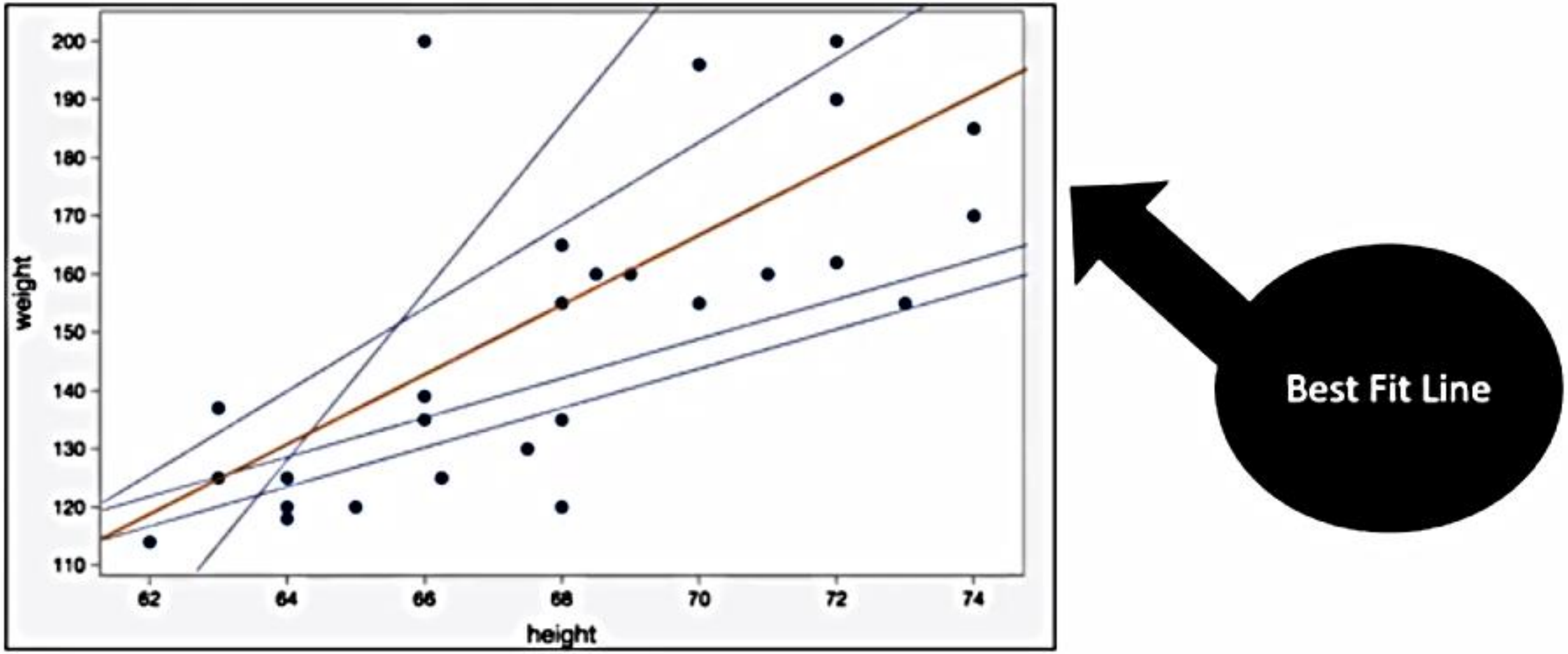
This is called as Residual Sum of Squares (RSS) as well. By minimizing this SSE, we can obtain following parameter estimations.

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

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

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

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})*(y_i - \bar{y})$	$(x_i - \bar{x}^2)$


weight
height
200
190
180
170
160
150
140
130
120
110
62
64
66
68
70
72
74
Best Fit Line

Mean Absolute Error (MAE)

- Mean Absolute Error is an evaluation metric used to calculate the accuracy of a regression model.
- MAE measures the average absolute difference between the predicted values and actual values.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}^{(i)} - y^{(i)}|$$

- i = index of sample
- ŷ = predicted value
- y = expected value
- m = number of samples in the data set

Mean Squared Error (MSE)

MSE represents the average squared difference between the predictions and expected results.

In other words, MSE is an alteration of MAE where, instead of taking the absolute value of differences, we square those differences.

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

- i = index of sample
- ŷ = predicted value
- y = expected value
- m - number of samples in the data set

Root Mean Squared Error (RMSE)

RMSE stands for Root Mean Squared Error.

To obtain the RMSE, simply take the square root of the MSE.

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2}$$

- i = index of sample
- ŷ = predicted value
- y = expected value
- m - number of samples in the data set

R-squared (R^2) Score

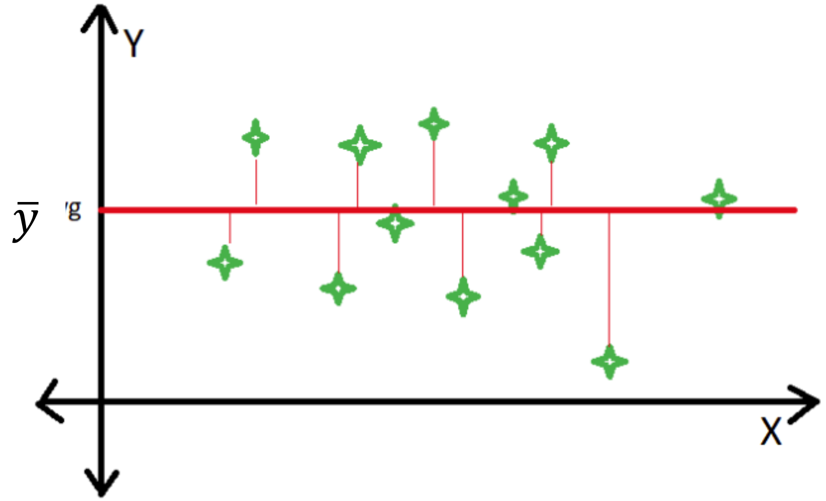
- R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The value of R-square lies between 0 to 1.
- Where we get R-square equals 1 when the model perfectly fits the data and there is no difference between the predicted value and actual value.
- However, we get R-square equals 0 when the model does not predict any variability in the model and it does not learn any relationship between the dependent and independent variables.

How is R-Squared Calculated?

First, calculate the mean of the target/dependent variable y and we denote it by $\bar{y}$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

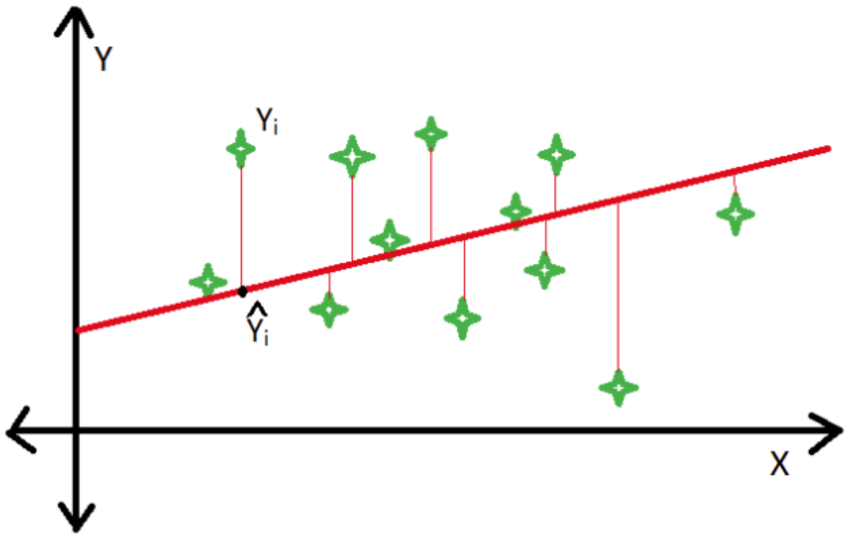
Calculate the total sum of squares [SS_{TOTAL}] by subtracting each observation y_i from $\bar{y}$, then squaring it and summing these square differences across all the values. It is denoted by



How is R-Squared Calculated?

Calculate the total sum of residuals [SS_{RES}] by subtracting each predicted value of y_i from $\hat{y}_i$. It is denoted by

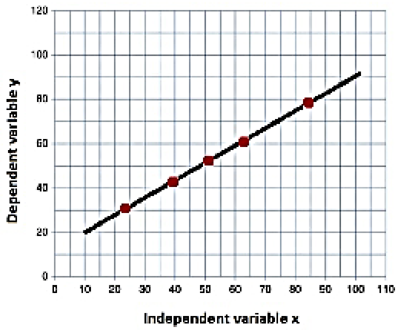
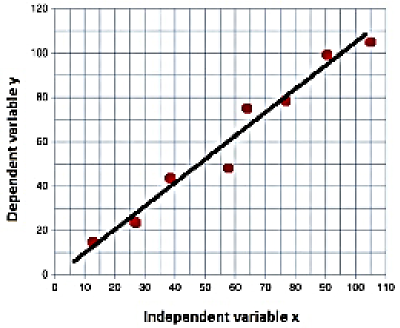
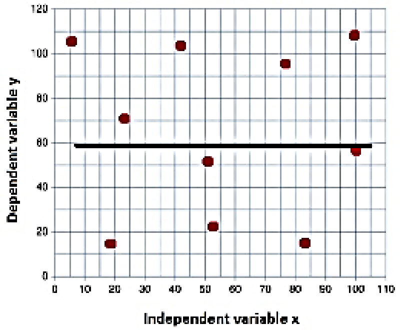
$$SS_{RES} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



R-squared (R^2) can be calculated,

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOTAL}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Interpretation of the R-Squared value

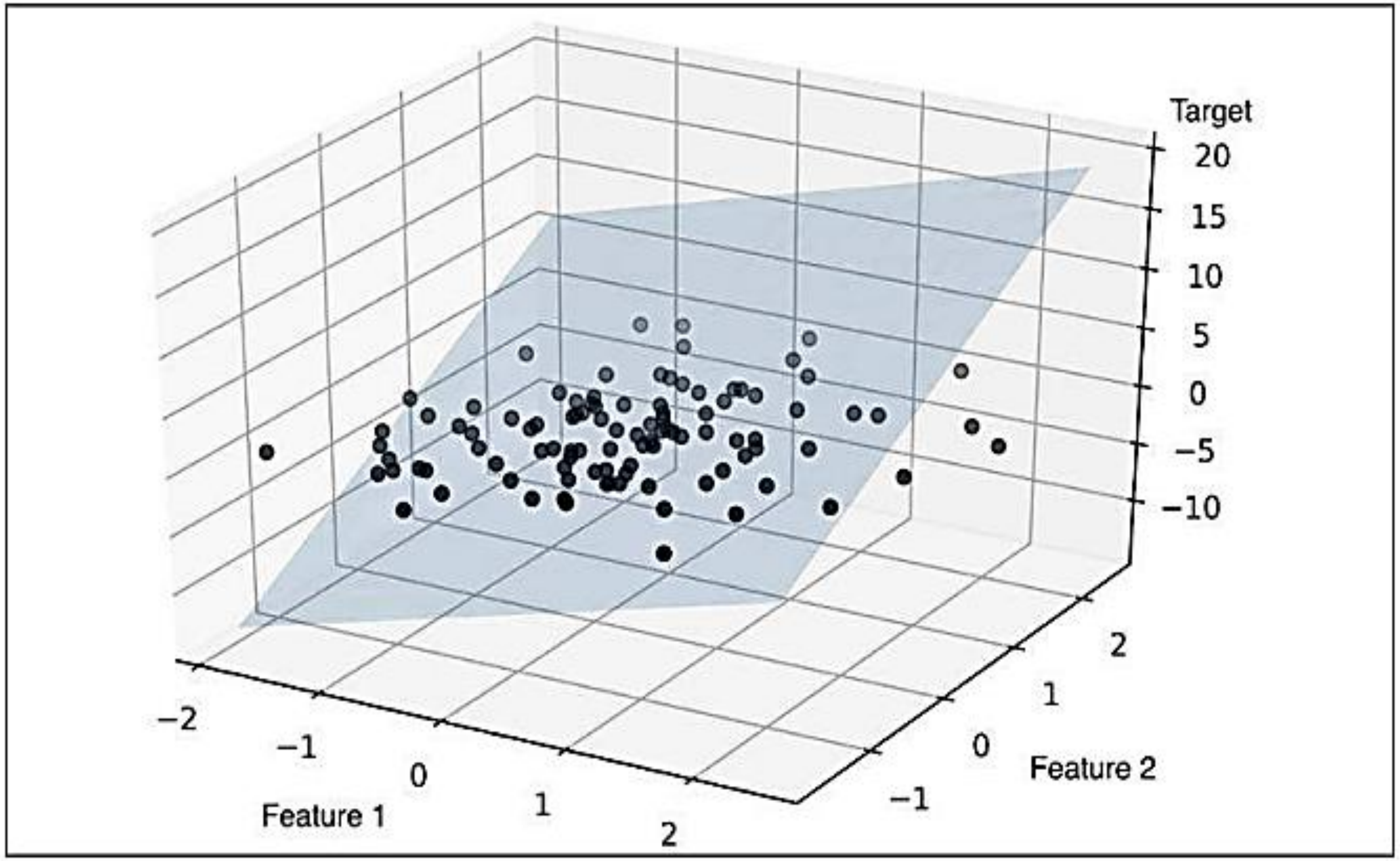
<table>
<tr><th>R^2 Values</th><th>Interpretation</th><th>Graph</th></tr>
<tr><td>$R^2 = 1$</td><td>All the variation in the y values is accounted for by the x values</td><td>
</td></tr>
<tr><td>$R^2 = 0.83$</td><td>83% of the variation in the y values is accounted for by the x values</td><td>
</td></tr>
<tr><td>$R^2 = 0$</td><td>None of the variation in the y values is accounted for by the x values</td><td>
</td></tr>
</table>

Multiple linear regression

Multiple linear regression is a statistical method used to model the relationship between multiple independent variables and a single dependent variable.

$$h_\theta(x) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x_1 + \boldsymbol{\beta}_2 x_2 + \ \ldots\ldots\ldots\ldots\ldots\ldots + \boldsymbol{\beta}_n x_n$$

Multiple linear regression



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 can be obtained through,

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

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

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

Qualitative Predictors (Categorical Independent Variables):

Since one variable can be represented as other variable's O case, one variable can be removed. That is called as the reference level. Consider First as the reference level. This is called as Dummy Trapping.

First	Second	Third	Fourth
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Qualitative Predictors (Categorical Independent Variables):

Since one variable can be represented as other variable's O 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

Qualitative Predictors (Categorical Independent Variables):

Since one variable can be represented as other variable's O 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

Linear Regression Assumptions

There are mainly four assumptions associated with a linear regression model:

« Linearity: The relationship between X and the mean of Y is linear.

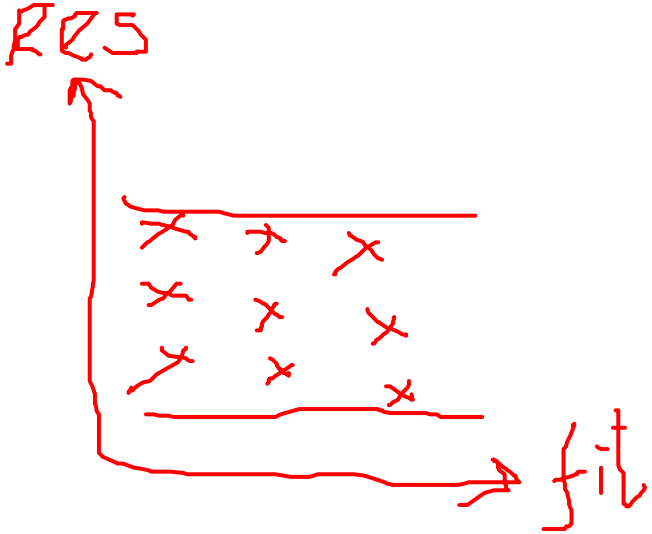
« Scatter plots & partial regression models

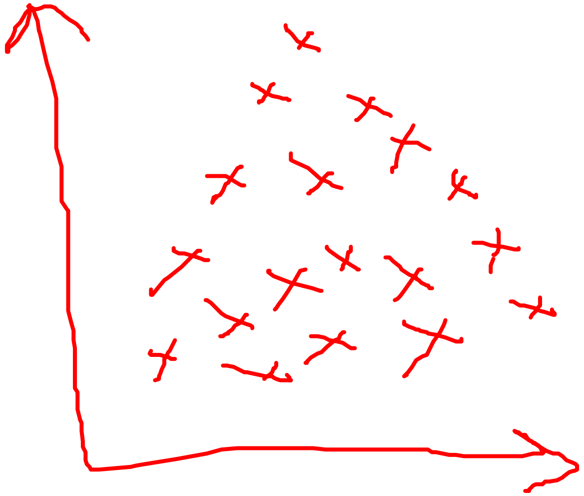
Multicollinearity: Correlation among independent variables.

- VIF - Variance Inflation Factor
- if VIF <10 Variables are independent , If>10 variables are dependent

Homoscedasticity: The variance of residual is the same for any value of X.

* Residual VS Fitted Values plot


Res
fit




Res
fit
Showing pattern

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.

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

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