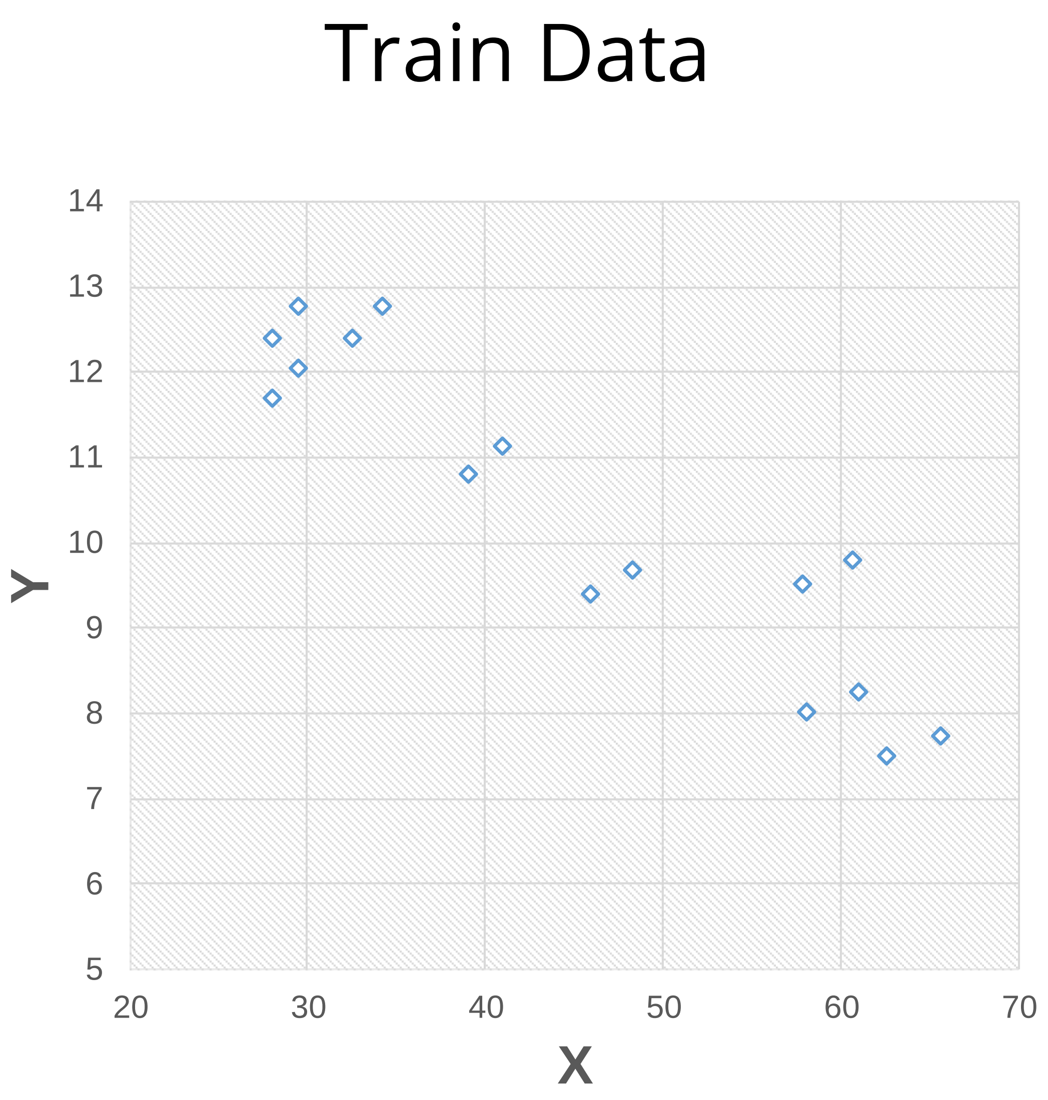
DATA AND
ARTIFICIAL INTELLIGENCE

Regression

©Simplilearn. All rights reserved.

Regression



What is Regression Analysis

It is a statistical technique used to relate a variable of interest(dependent variable) to one or more independent or predictor variables.

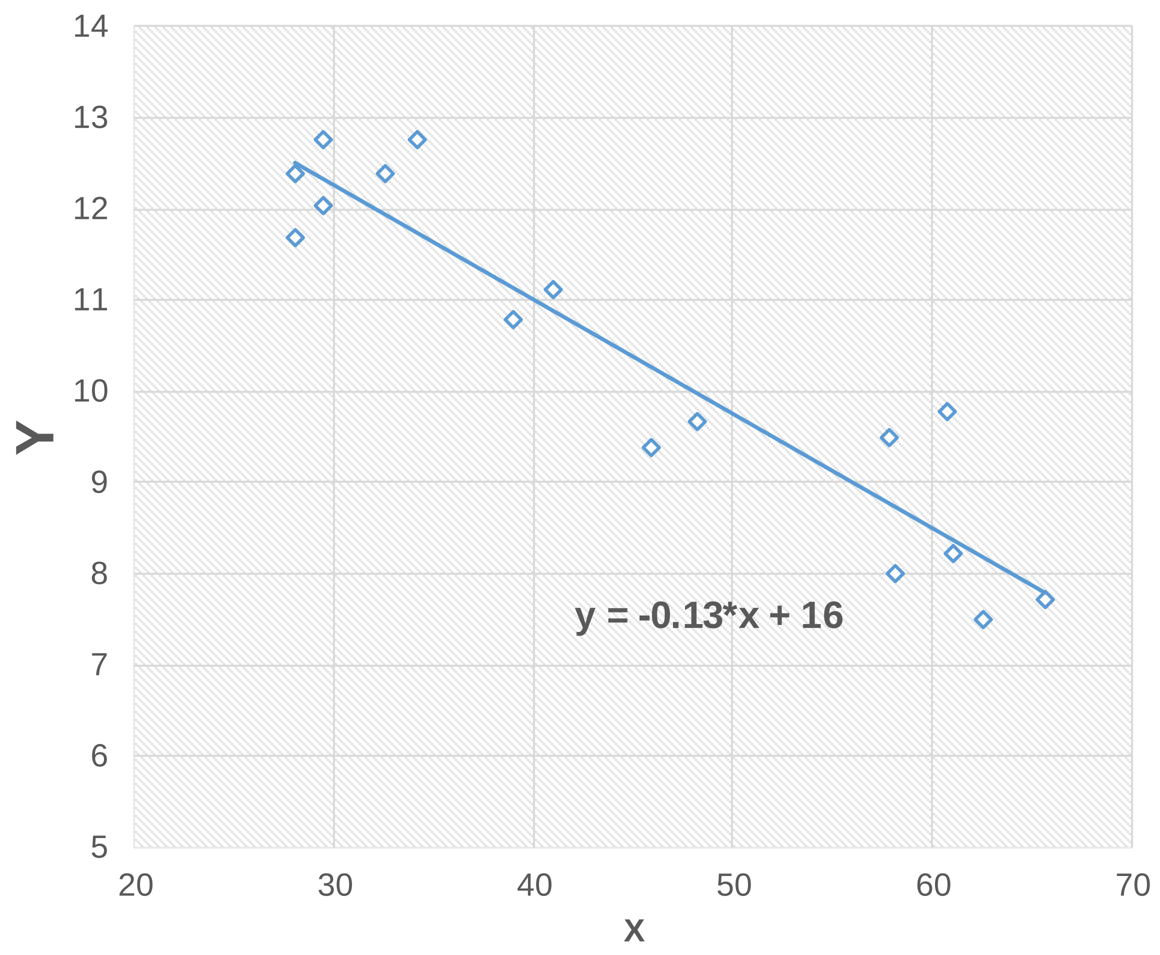
The objective is to build a statistical model to:

- Describe the variable of interest
- Predict the variable of interest
- Control the variable of interest

Simple Linear Regression

Simple linear regression is a linear regression model with a single predictor variable.

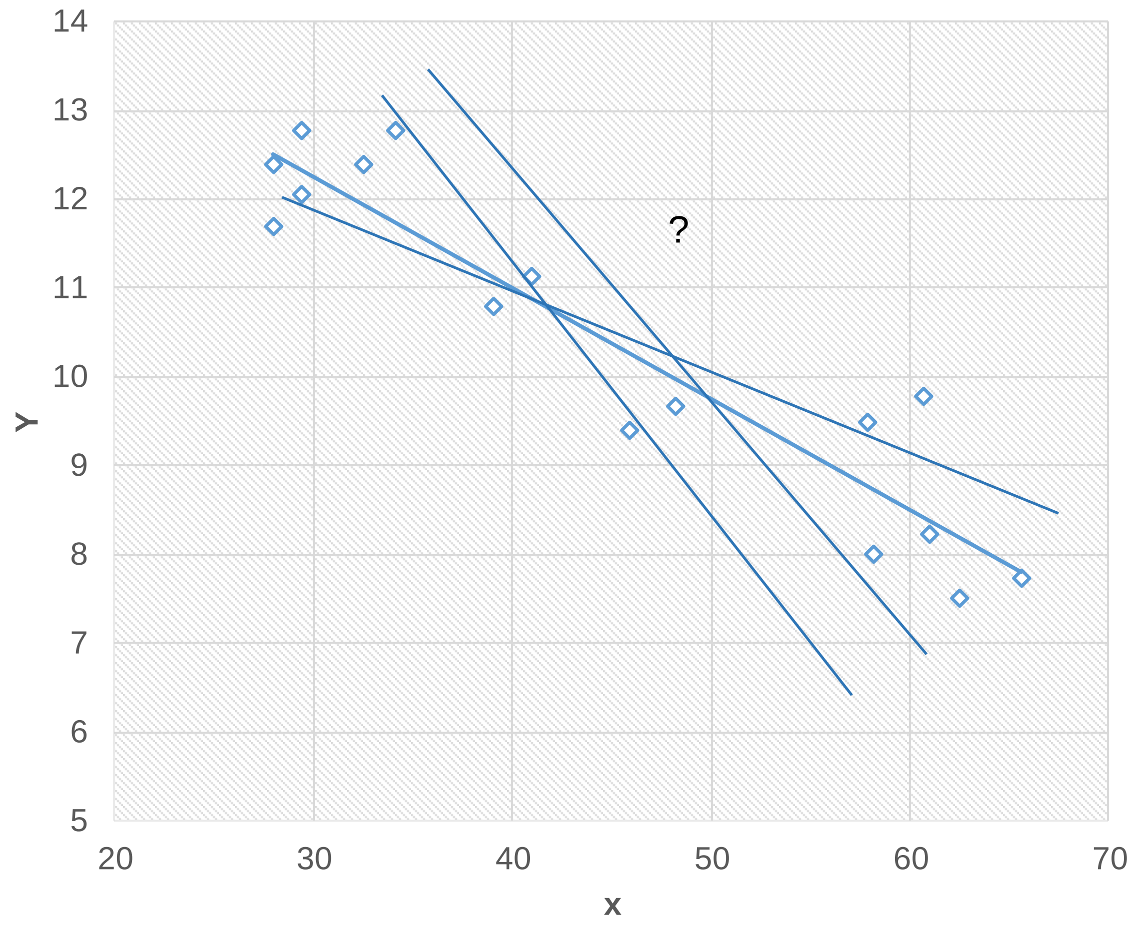
We try to establish a linear relationship between dependent and independent variables.



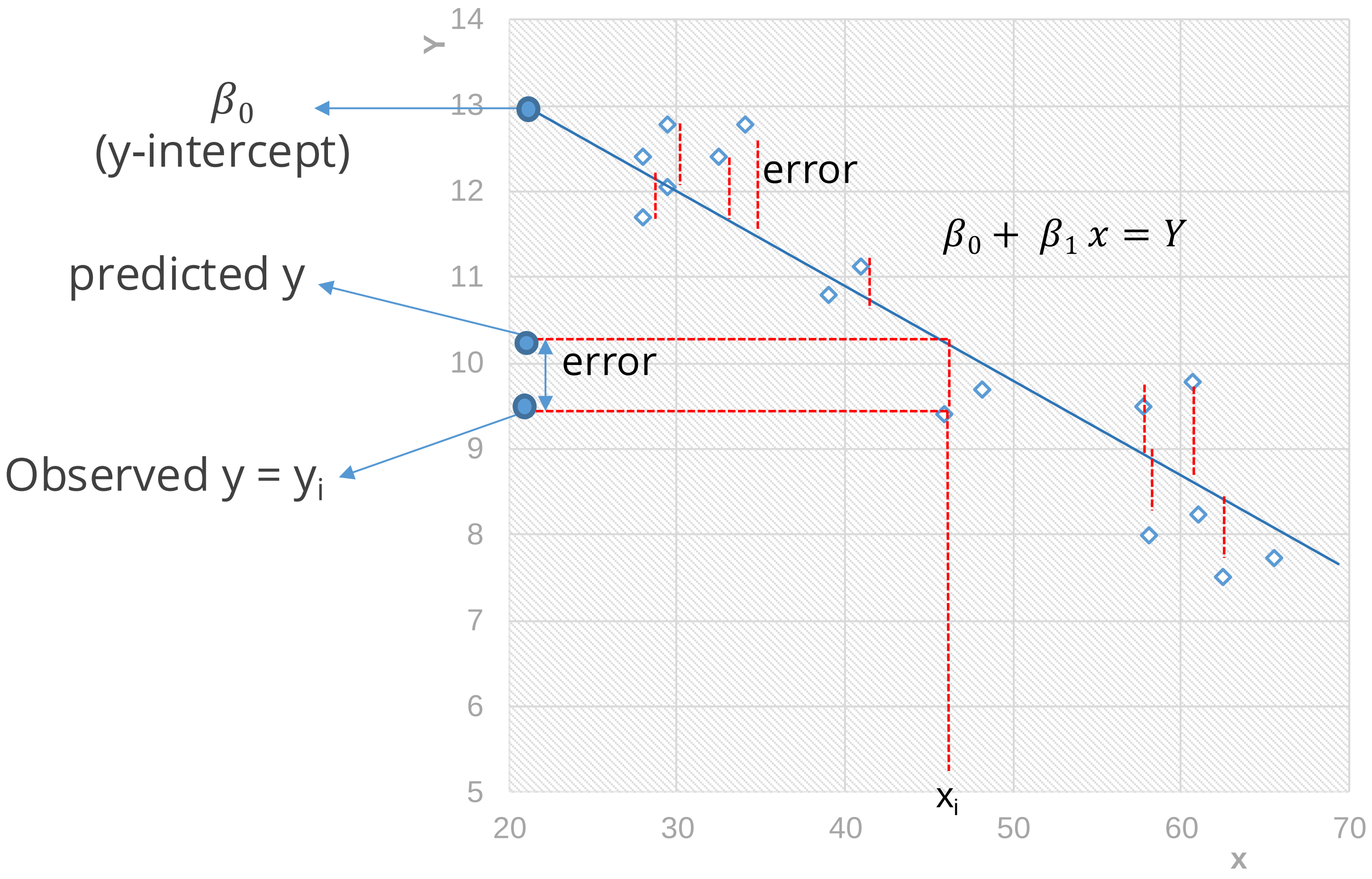
simplilearn

How to Chose the Best Line?

Which line to choose out of so many lines passing through the points?



Ordinary Least Square Regression

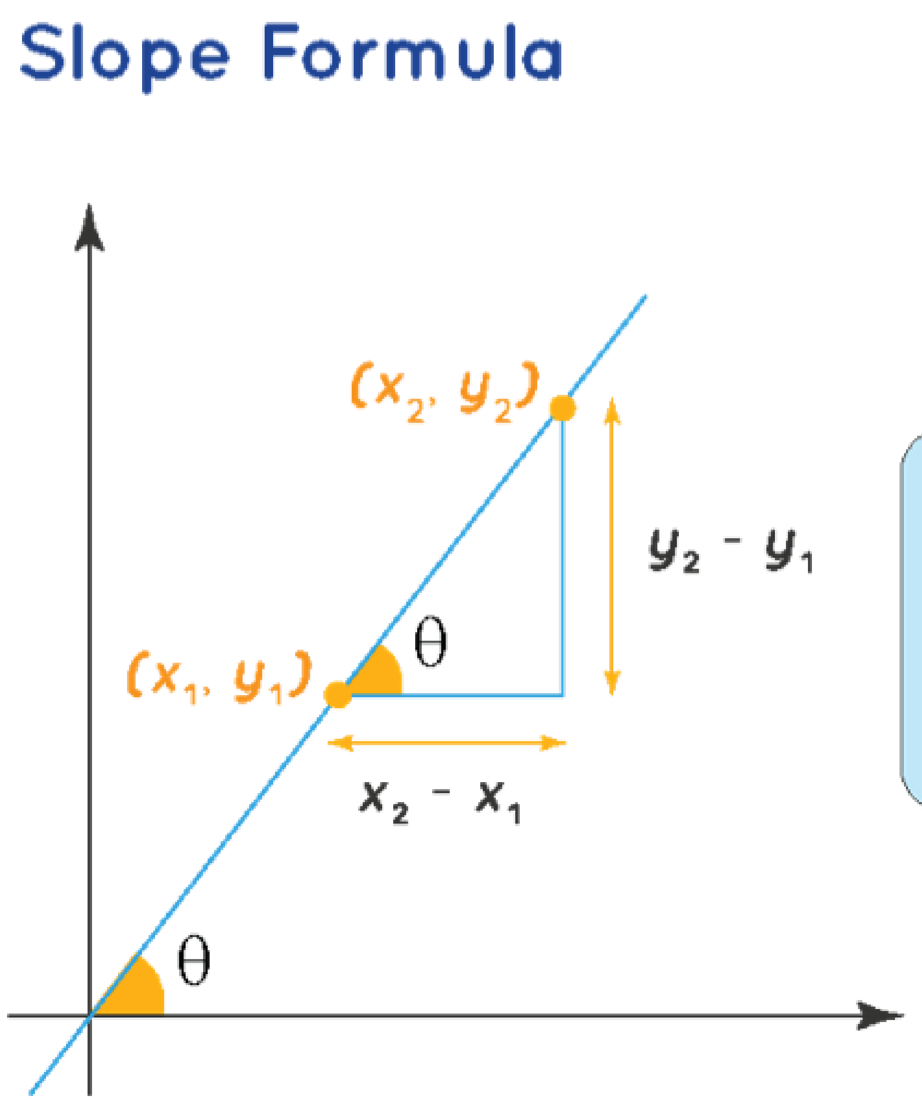


- Assume any line $\beta_0 + \beta_1 x = Y$ passing through the points.
- Here, β_0 is y intercept and β_1 is slope of the line

Find the appropriate values of β_0 and β_1 to get to the best fit line.

simplilearn

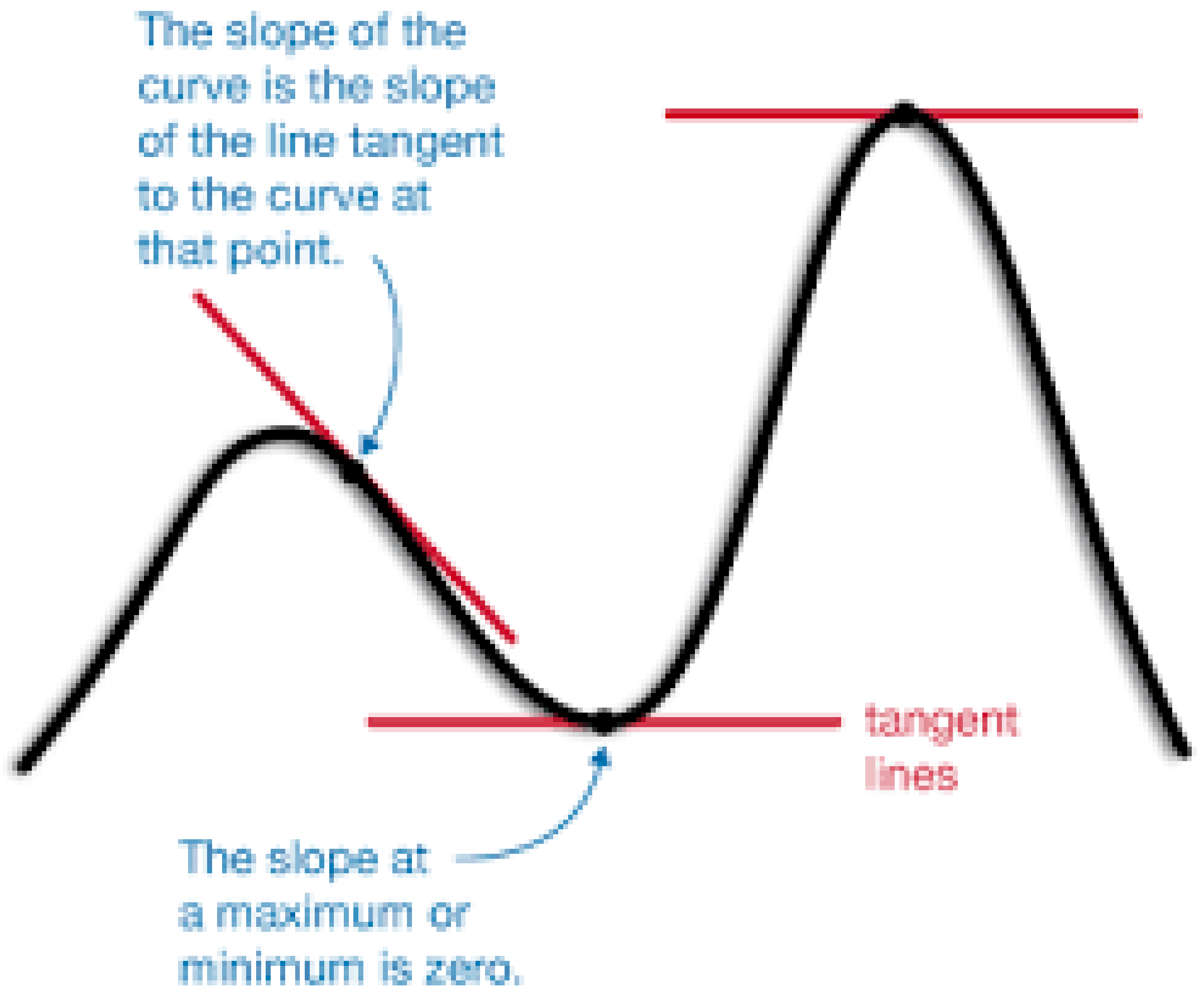
Slope – A little indepth



$$\tan\theta = \frac{y_2 - y_1}{x_2 - x_1}$$

$$m = \tan\theta$$

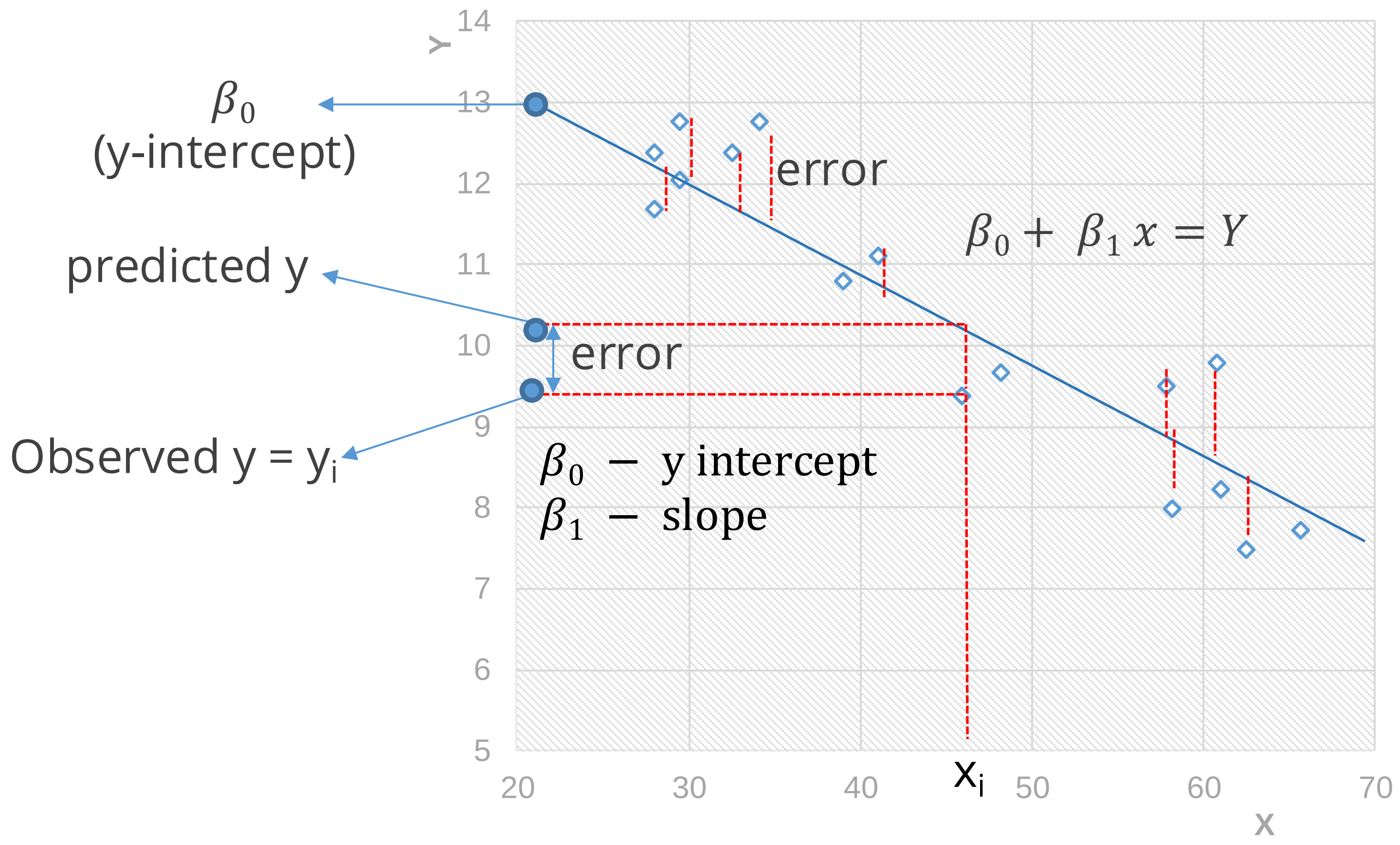
Slope – A little indepth



simplilearn

Ordinary Least Square Regression

The idea is to find a line for which predicted y and observed y are close for all the points.



- Predicted y = $\beta_0 + \beta_1 xi$, find a line and β_0 and β_1 for which $\sum$(predicted y − observed y)^2 is minimum.
- Find β_0 and β_1 *for which* $\sum_{i=1}^{n}((\beta_0 + \beta_1 xi) - y_i)^2$ is minimum.

simplilearn

Loss and Cost of Linear Regression

1. Loss Function (for a single training example):

$$L(y_i, \hat{y}_i) = \frac{1}{2}(y_i - \hat{y}_i)^2$$

Substituting $\hat{y}_i = \beta_0 + \beta_1 x_i$:

$$L(y_i, \hat{y}_i) = \frac{1}{2}(y_i - (\beta_0 + \beta_1 x_i))^2$$

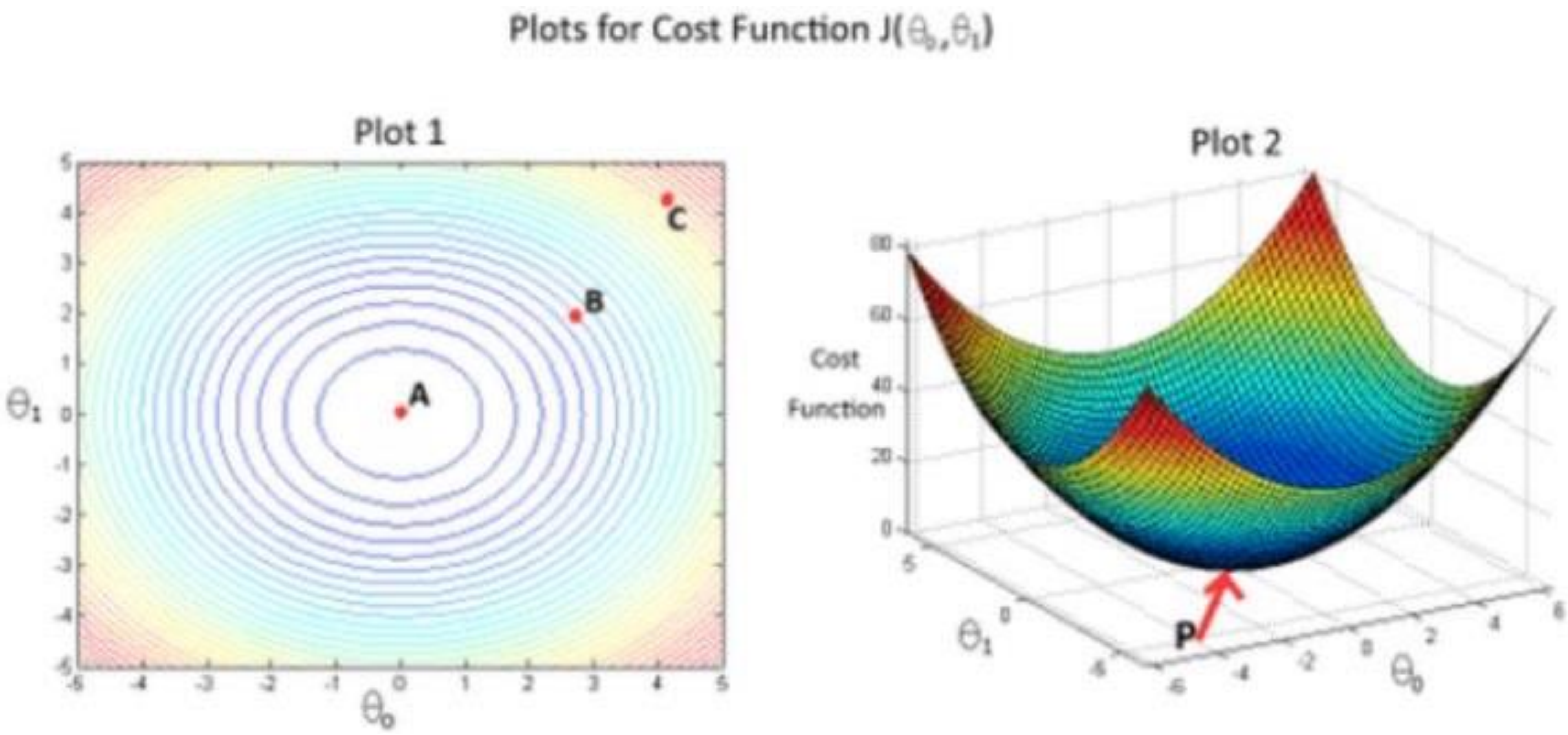
2. Cost Function (for the entire dataset):

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

Substituting $\hat{y}_i = \beta_0 + \beta_1 x_i$:

$$J(\beta_0, \beta_1) = \frac{1}{2m}\sum_{i=1}^{m}(y_i - (\beta_0 + \beta_1 x_i))^2$$

Optimization using Gradient Descent



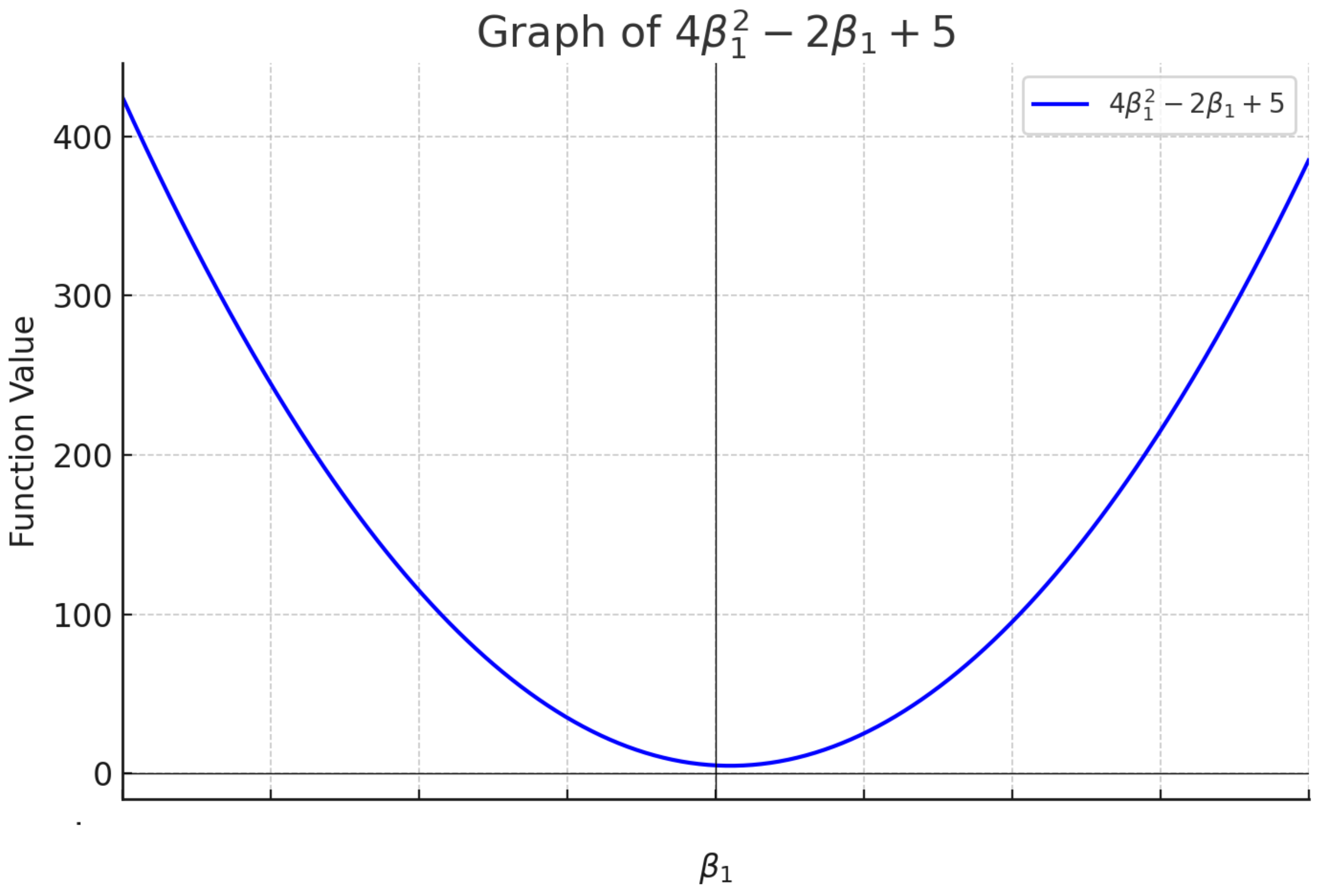
Ultimate Task: Find values of coefficients(parameters or weights & Bias) which makes J minimum = Model Training = Fitting a model

Optimization using Gradient Descent

Let’s do some calculations in excel

simplilearn

Optimization using Gradient Descent



simplilearn

Optimization using Gradient Descent

1. **Initialize** β_0 and β_1 with some values (e.g., 0).
2. **Set learning rate** α.
3. **Repeat until convergence**:
   - Compute predictions:
   $$\hat{y}_i = \beta_0 + \beta_1 x_i$$
   - Compute the gradients (partial derivatives):
   $$\frac{\partial J}{\partial \beta_0} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i)$$
   $$\frac{\partial J}{\partial \beta_1} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) x_i$$
   - Update parameters:
   $$\beta_0 := \beta_0 - \alpha \frac{\partial J}{\partial \beta_0}$$
   $$\beta_1 := \beta_1 - \alpha \frac{\partial J}{\partial \beta_1}$$
4. **Repeat until convergence** (i.e., when the updates become very small).

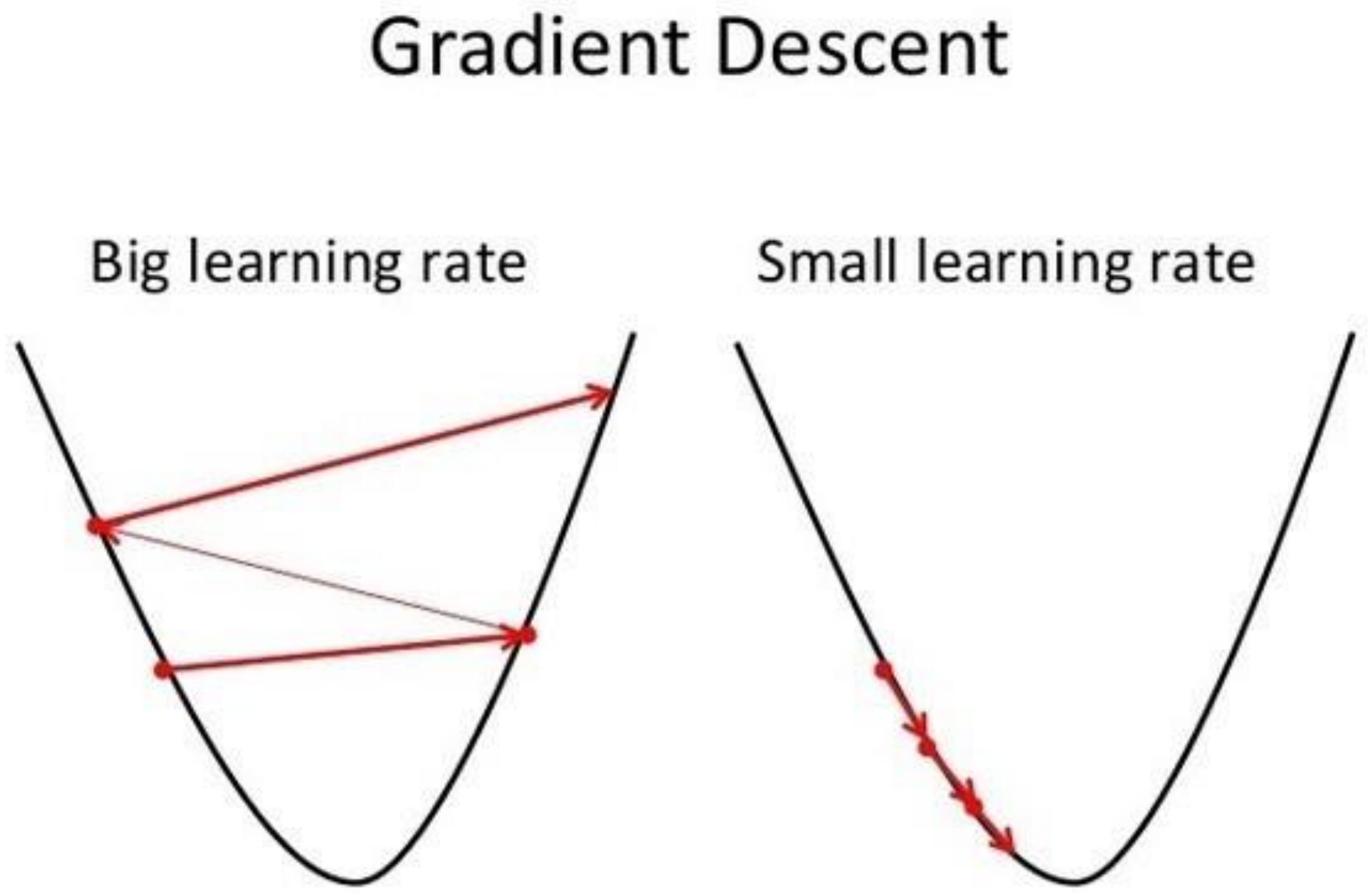
simplilearn

Let’s play

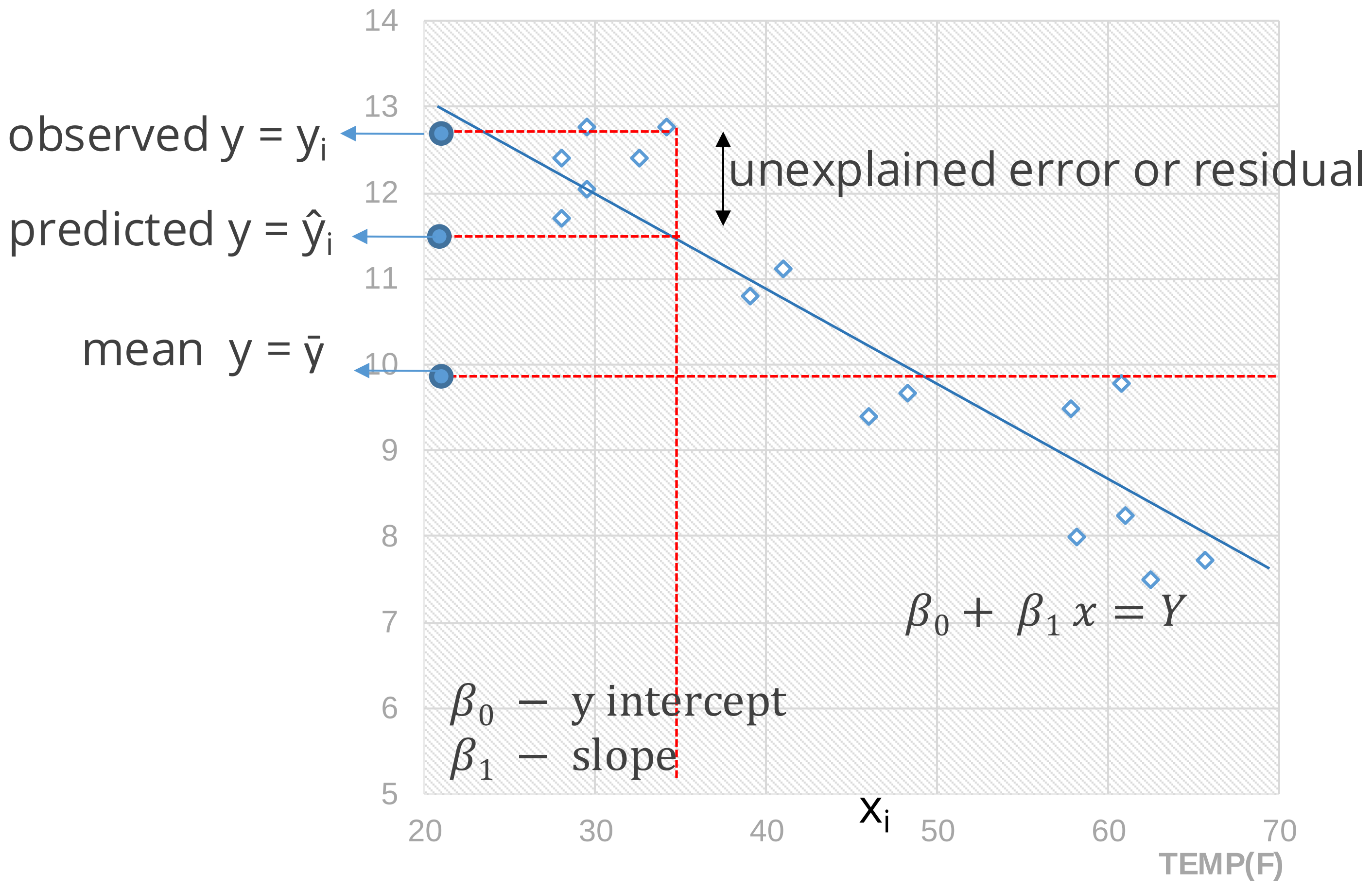
https://uclaacm.github.io/gradient-descent-visualiser/?utm_source=chatgpt.com

simplilearn

Learning rate = Step size



How Good is Regression?



$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Total deviation = Unexplained deviation + Explained deviation

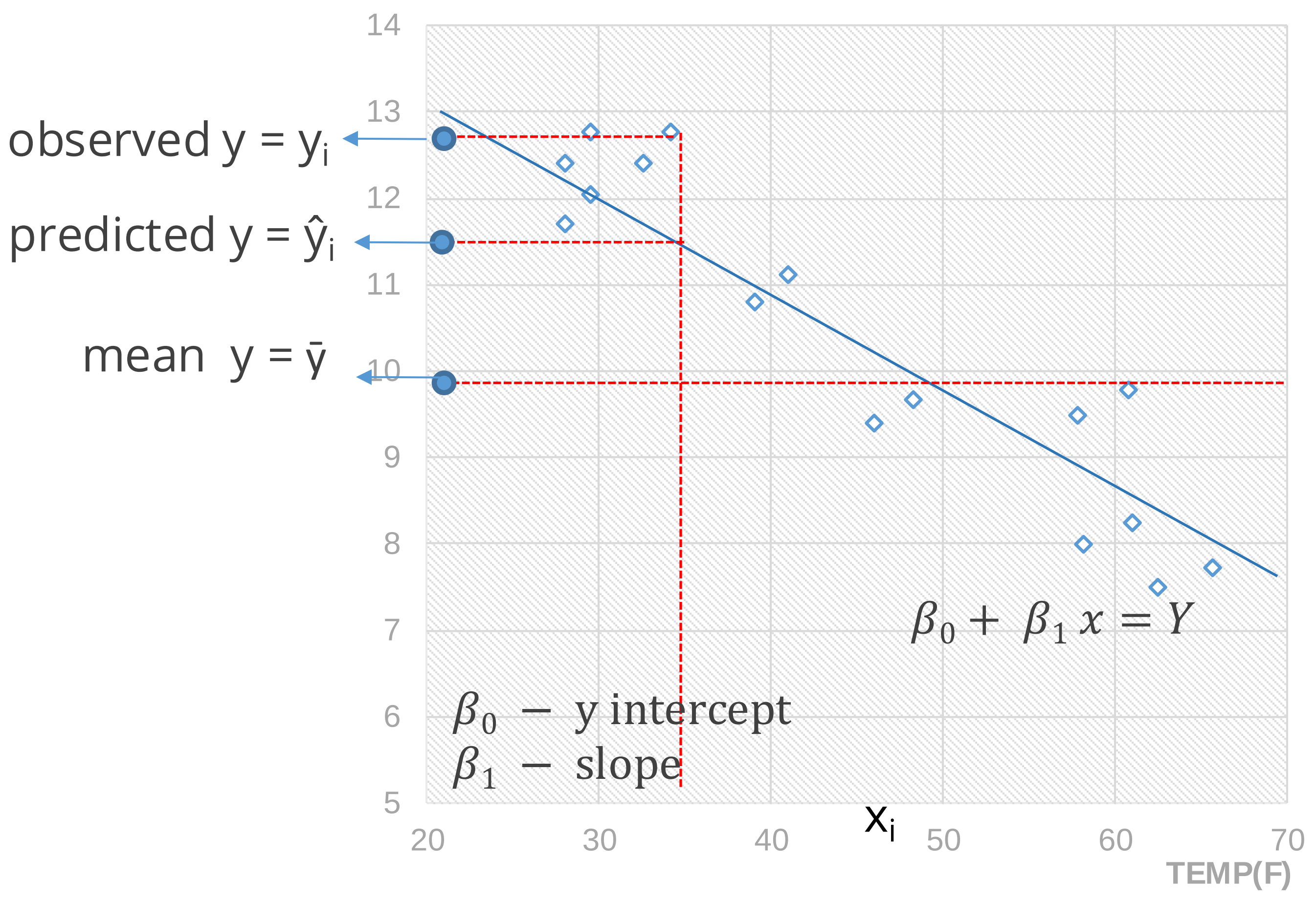
$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$

SST	SSE	SSR
(Total sum of squared)	(sum of Squares of error)	(sum of Squares of regression)

simplilearn

How Good is Regression?

Once the linear relationship is determined, let's analyze how strong is the relationship.



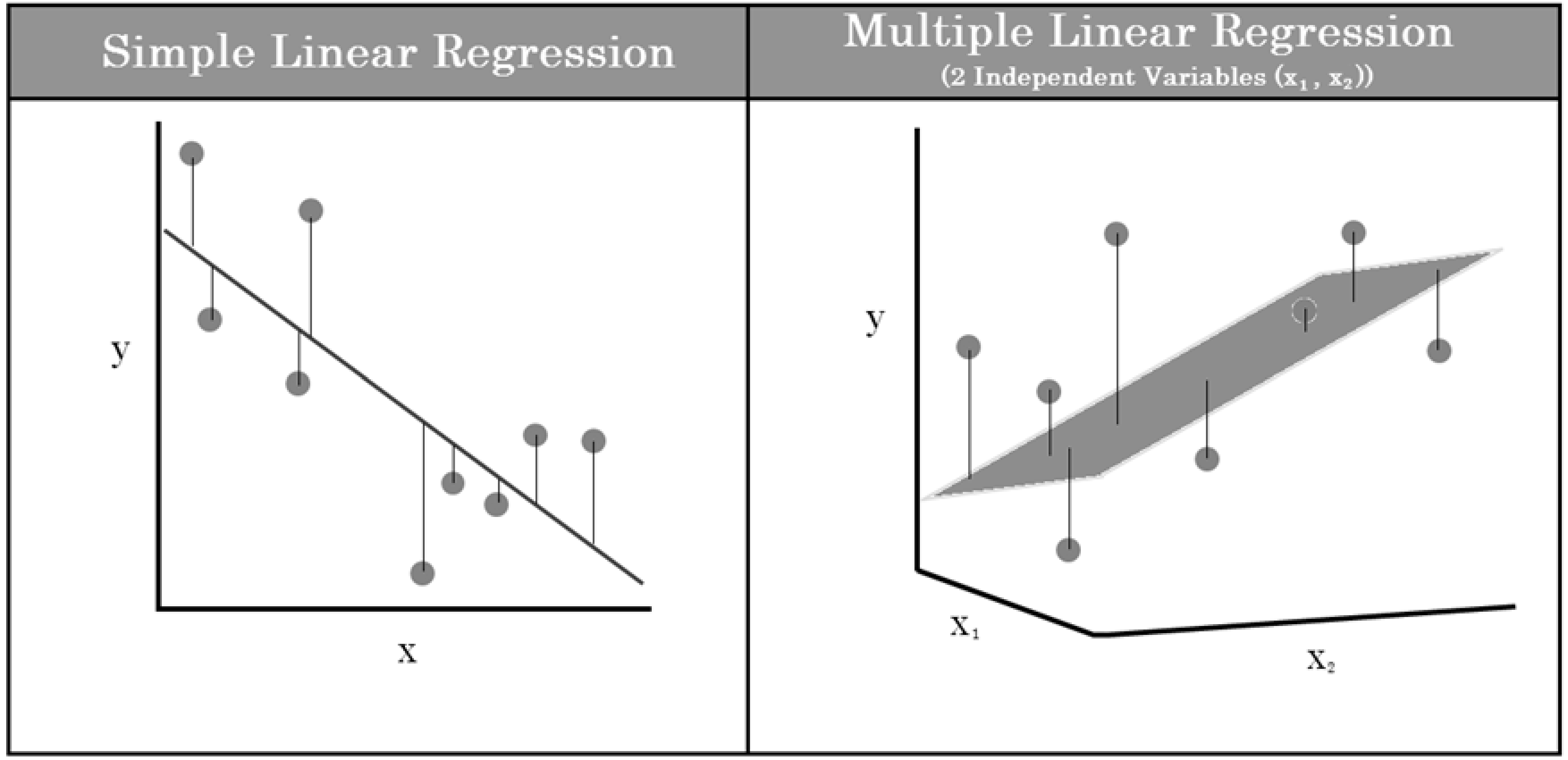
Coefficient of determination = R squared

It is the proportion of the variation in y that is explained by the regression.

It is given by $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

simplilearn

Multiple Linear Regression



$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

simplilearn

Adjusted R^2

Adjusted (or corrected) R^2 is the coefficient of determination corrected for degree of freedom.

- R^2 doesn't always increase as new variables are introduced in the regression model.

- R^2 increases only when a new variable entered the model is adding any additional value.

- It is given by; Adjusted $R^2 = 1 - \frac{SSE/[n-(k+1)}{SST/(n-1)}$

  n = sample size

  k = no of predictors

Evaluation Metrics for Linear Regression

Evaluation metrics are measures of how good a model performs and how well it defines the relationships.

Other than R^2 and Adjusted R^2, other evaluation metrics include:

Metric	Formula
MSE : Mean Squared Error	$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
MAE : Mean Absolute Error	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert\hat{y}_i - y_i\rvert$
RMSE : Root Mean Squared Error	$MSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$

A lower value of these metrics indicates a better model.

simplilearn

Working with Categorical Variables

Some potential predictors are categorical and qualitative.

To accommodate these variables in the Regression model, they should be transformed into Dummy variables.

Original Data

Price	LivingArea	Region
16858	1629	East
26049	1344	West
26130	822	East
31113	1540	East
40932	1320	West
44674	1214	North
44873	882	South
45004	960	North
49564	1363	West

Original Data

Price	LivingArea	East	West	North
16858	1629	1	0	0
26049	1344	0	1	0
26130	822	1	0	0
31113	1540	1	0	0
40932	1320	0	1	0
44674	1214	0	0	1
44873	882	0	0	0
45004	960	0	0	1
49564	1363	0	1	0

simplilearn

Validation

©Simplilearn. All rights reserved.

Creating a Validation Framework

To test the performance of the model on new scenarios, validation frameworks need to be created. Popular validation frameworks include:

Hold-out based Validation

K-fold cross Validation

simplilearn

Creating a Validation Framework

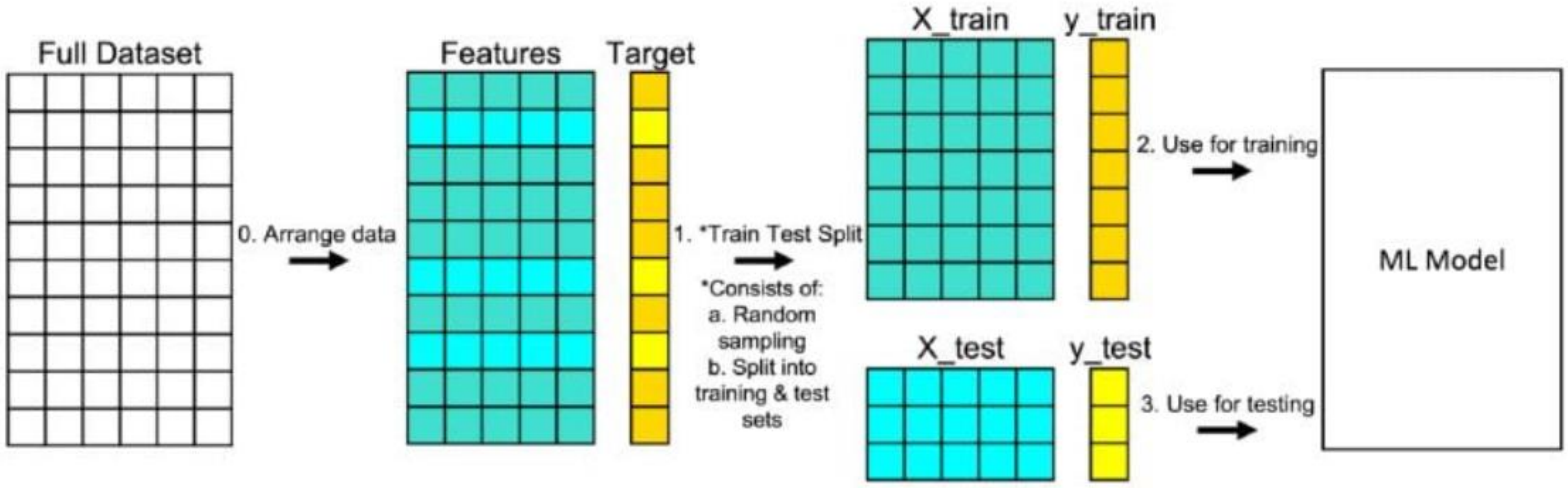
Hold-out based Validation

K-fold cross Validation

Randomly splits the dataset into train and test.

All Data	
Training data	Test data

Hold-out based Validation



simplilearn

Validation Framework: K-fold Cross-Validation

Hold-out based Validation

K-fold cross Validation

The original dataset is equally partitioned into k subparts or folds.

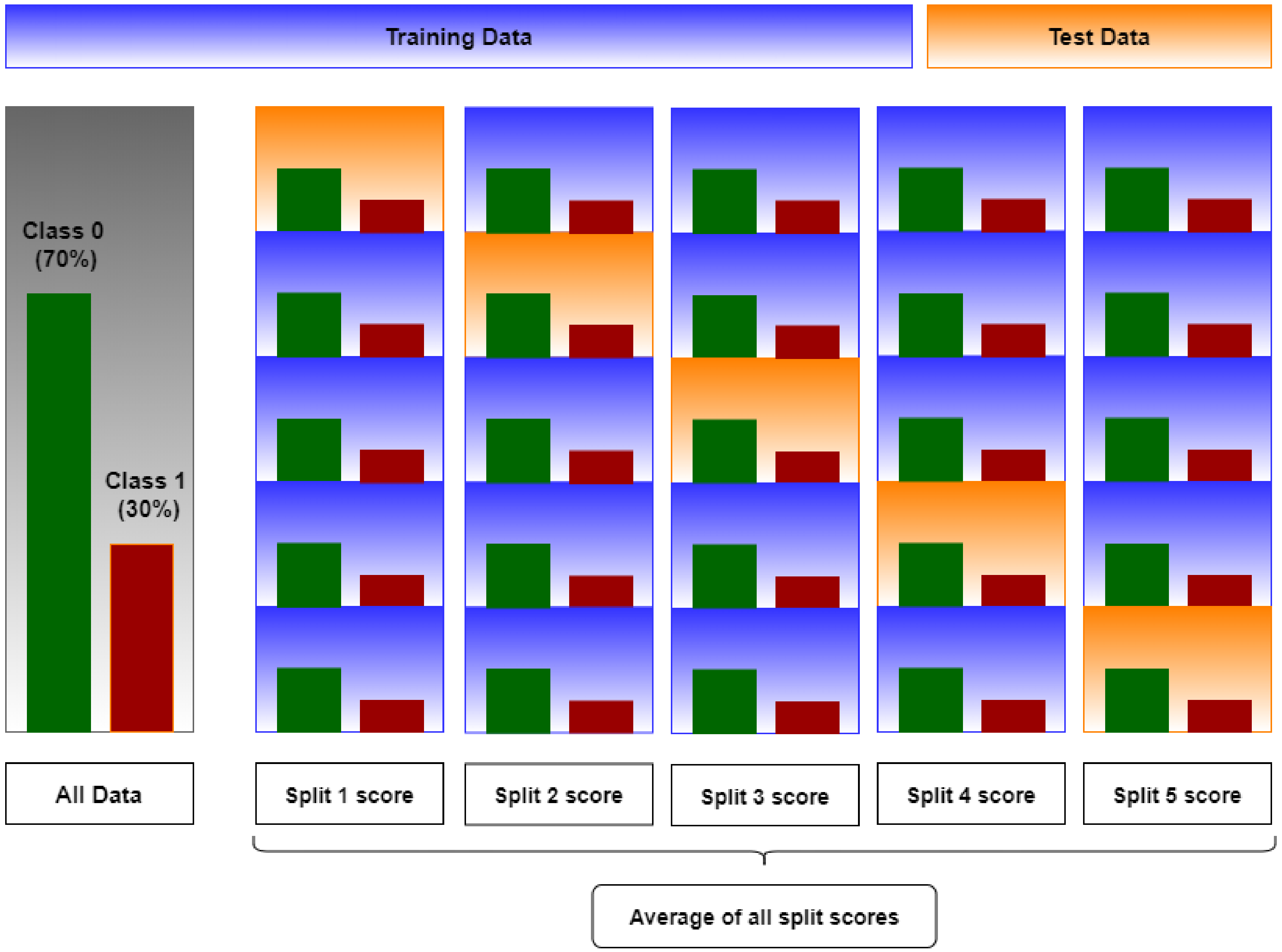
Out of the k-folds, for each iteration, one group is selected as validation data, and the remaining (k-1) groups are selected as training data.

All Data

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

simplilearn

Stratified K-fold cross Validation



Appendix

©Simplilearn. All rights reserved.

DATA AND
ARTIFICIAL INTELLIGENCE

Assumptions of Regression

©Simplilearn. All rights reserved.

simplilearn

Assumption of Linear Regression

Linear Relationship

Independence of Error

Normality of Error Terms

Equality of Variance

simplilearn

Assumption of Linear Regression

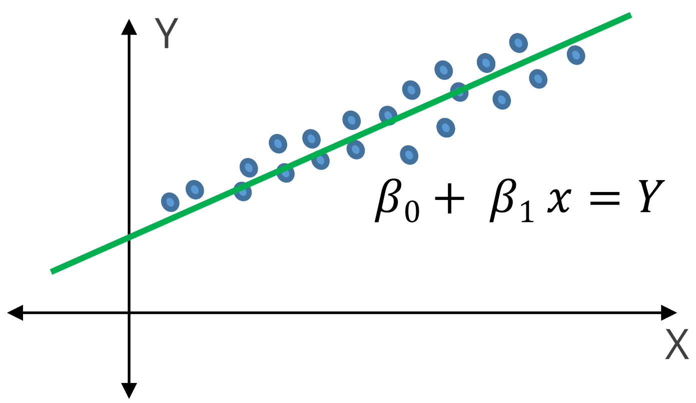
Linear Relationship

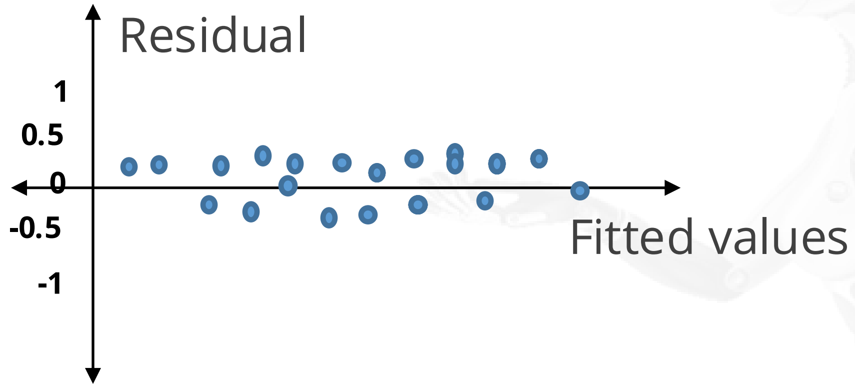
Independence of Error

Normality of Error Terms

Equality of Variance

The relationship between the independent and dependent variables should be linear.





Assumption of Linear Regression

Linear Relationship

Independence of Error

Normality of Error Terms

Equality of Variance

The residuals are independent.

There should be no correlation between consecutive residuals in time series data.

This is assumption is important, when there is longitudinal i.e., time-series dataset, for instance, stock price data.

simplilearn

Assumption of Linear Regression

Linear Relationship

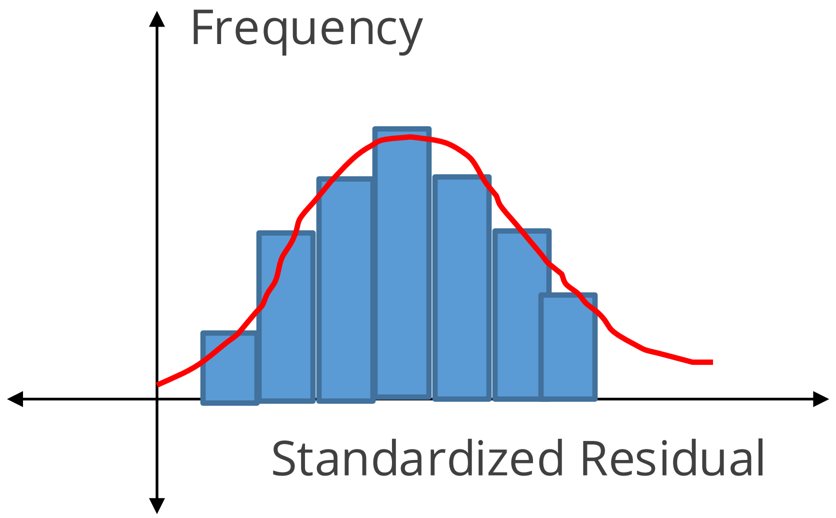
Independence of Error

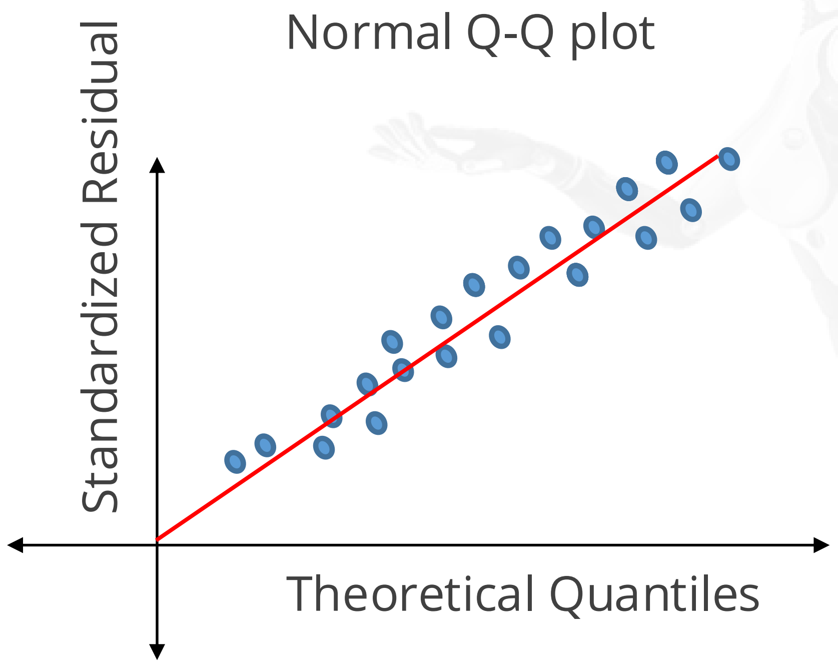
Normality of Error Terms

Equality of Variance

The error terms (residuals) are normally distributed.

Histogram and Quantile-Quantile plots are used to check this.





simplilearn

Assumption of Linear Regression

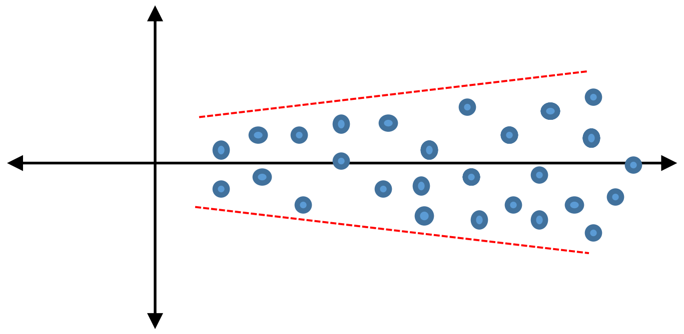
Linear Relationship

Independence of Error

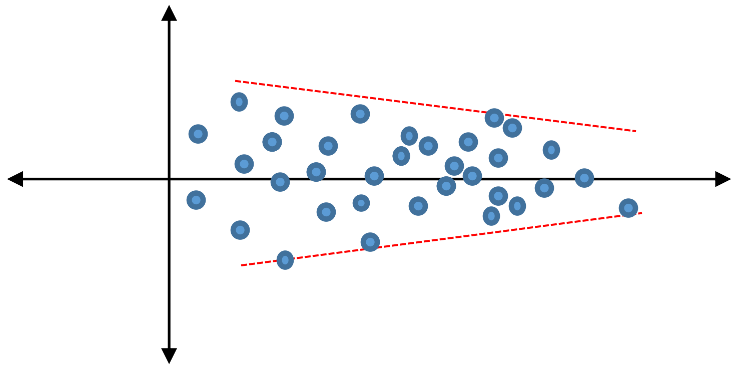
Normality of Error Terms

Equality of Variance

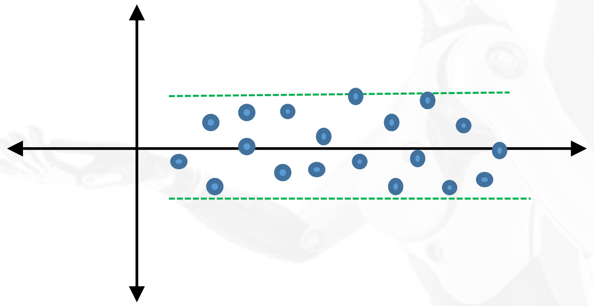
The error terms (residuals) have constant variance at every level of X. It is called Homoscedasticity.



Heteroscedasticity
Increasing error variance



Heteroscedasticity
Decreasing error variance



Homoscedasticity
Constant error variance

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Multicollinearity

©Simplilearn. All rights reserved.

simplilearn

Multicollinearity in Regression

If the independent variables in the Regression model are correlated with one another, it is termed as Multicollinearity.

This problem is detected using:

01 Correlation coefficients through initial inspection.

02 Scatter diagram between independent variables through visual inspection.

03 Variance inflation factor(VIF) is used to diagnose the issue.

Multicollinearity in Regression

Few steps to Remedy:

- Evaluate sample scheme and make changes if required
- Drop colinear variables
- Create new variables using Colinear Variables and form new combination of X variables with are uncorrelated

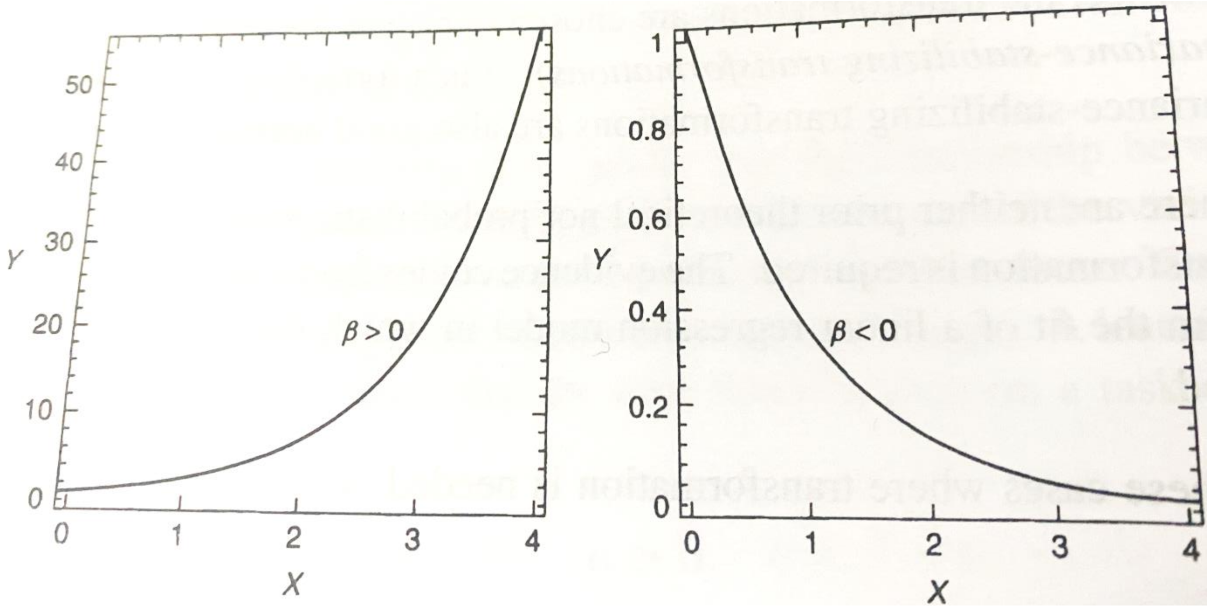
simplilearn

Non-Linear Regression

©Simplilearn. All rights reserved.

Non-linear Relationship and Transformation

Transformation is used to achieve linearity where there is a non-linear relationship between the variables.



Non-linear relationship Y = $\alpha\ e^{\beta X}$
After taking log on both the sides log Y = log α + βX

Transformation steps
Y′ = log Y

Linear form
Y′ = log α + βX