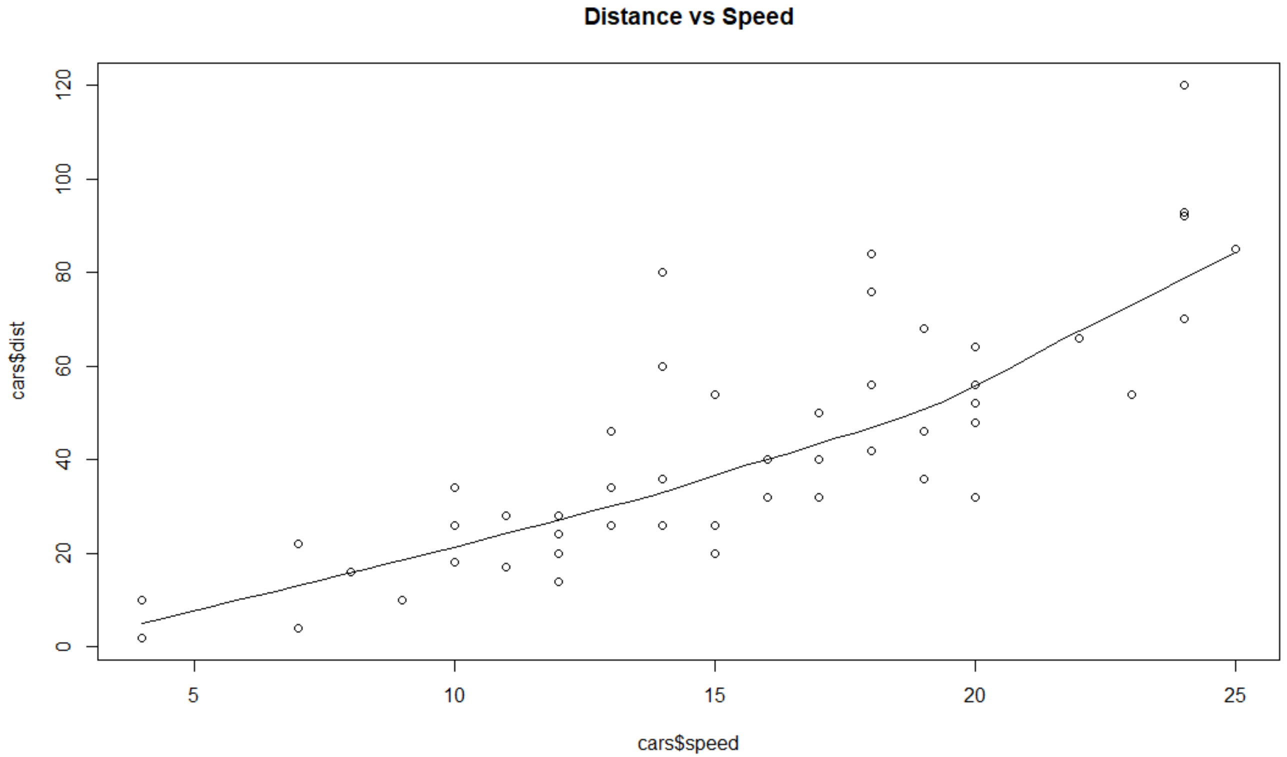
THE CONCEPT OF LINEAR REGRESSION

LET’S LEARN MACHINE LEARNING WITH PYTHON

What is Regression



© Amit Mishra

Course Outline

- Introduction to Linear Regression
- Example Problem
- Graphical Analysis
- Correlation Analysis
- Applying the Model
- Linear Regression Diagnostics
- How to know which regression model is best fit for the data?
- Predicting Linear Models
- K - Fold Cross Validation

Definition

Linear regression is statistical tool for modeling the relationship between a scalar **dependent variable** y and one or more **explanatory variables** (or independent variables) denoted by X.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

Regression Analysis

Regression analysis is used to:

- Predict the value of a dependent variable based on the value of at least one independent variable.
- Explain the impact of changes in an independent variable on the dependent variable.

Dependent variable: the variable we wish to explain
Independent variable: The variable used to explain the dependent variable

Let us take an example:

The goal here is to establish a mathematical equation for distance as a function of speed, so you can use it to predict distance when only the speed of the car is known.

So it is desirable to build a linear regression model with the response variable as distance and the predictor as speed.

Speed(X)	Distance(y)
4	2
4	10
7	4
7	22
8	16
9	10

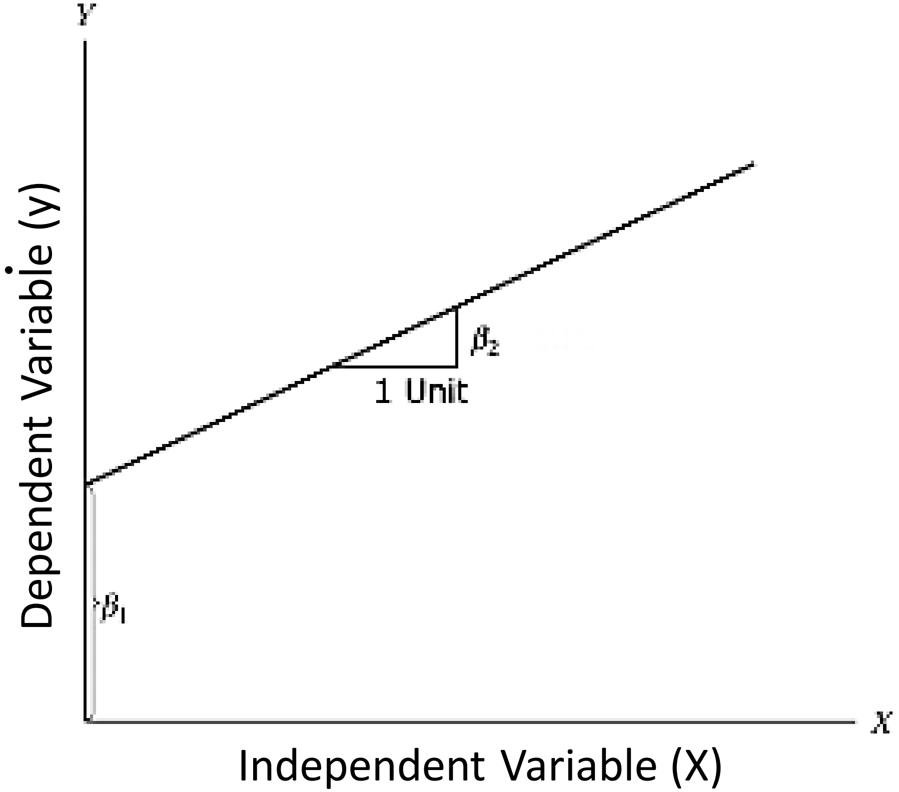
© Amit Mishra

Mathematical Equation

The linear regression model fits a linear function to a set of data points. Mathematical equation can be generalized as:

$$\mathbf{Y=\beta 1+\beta 2X}$$

where, β1 is the intercept and β2 is the slope.



Linear Regression Line

© Amit Mishra

Population Linear Regression

The population regression model representation

Dependent Variable

Population y intercept

Population Slope Coefficient

Independent Variable

Random Error term, or residual

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}}$$

© Amit Mishra

Linear Regression

Mathematical Equation

The linear regression model fits a linear function to a set of data points. The form of the function for multiple linear regression is:

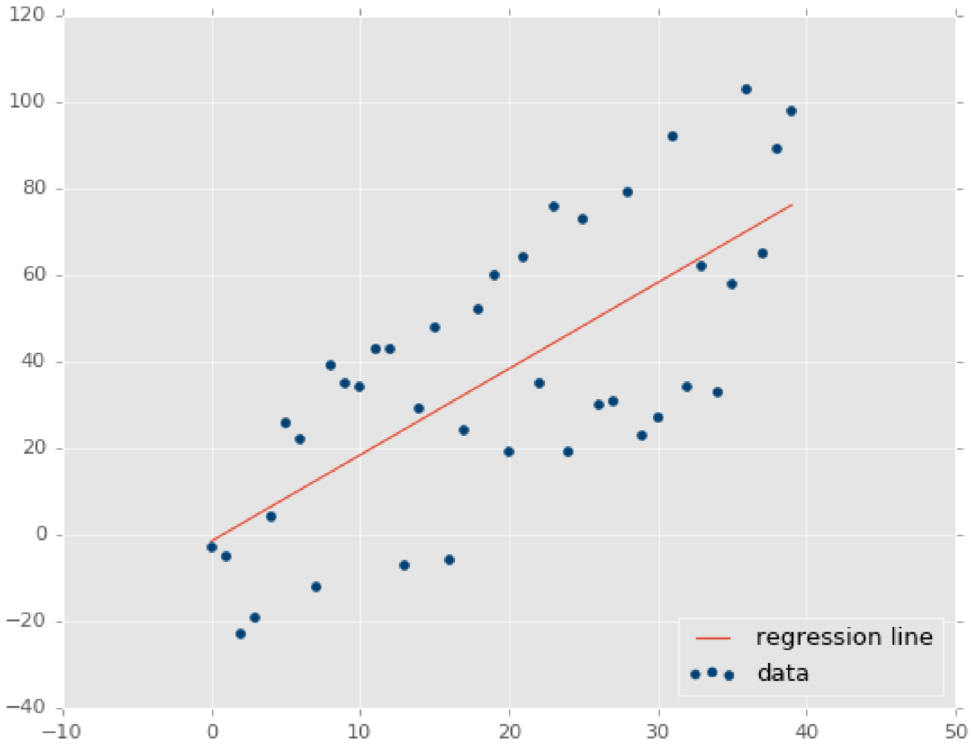
$$\mathbf{Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \ldots + \beta n * Xn}$$

Where Y is the target variable, and X1, X2, ... Xn are the predictor variables and β1, β2, … βn are the coefficients that multiply the predictor variables. β0 is constant.

© Amit Mishra

Linear Regression

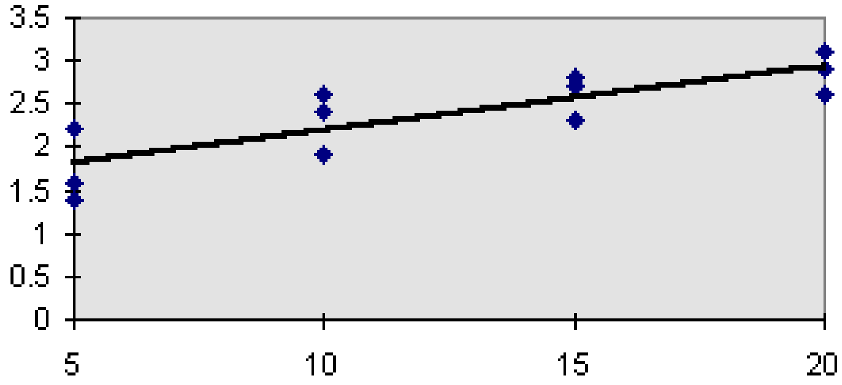
The goal is to take continuous data, find the equation that best fits the data, and be able forecast out a specific value. With simple linear regression, you are just simply doing this by creating a best fit line:



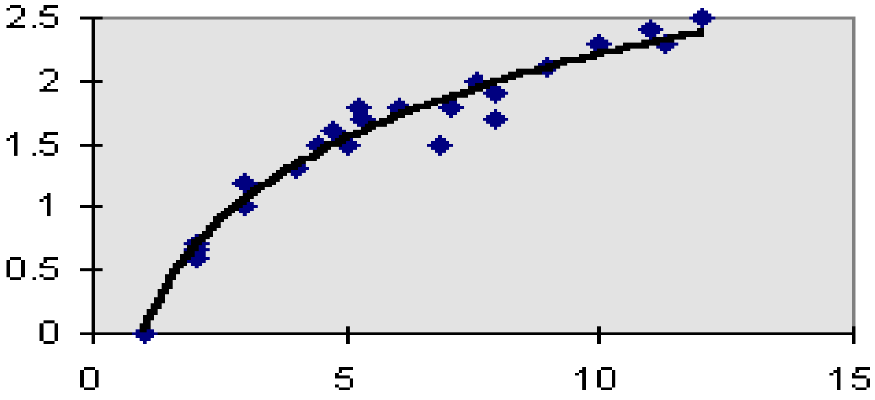
© Amit Mishra

Types of Regression

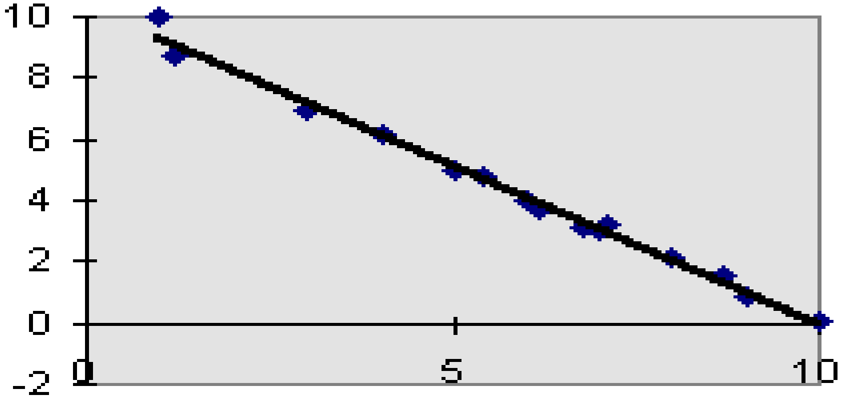
Positive Linear Relationship



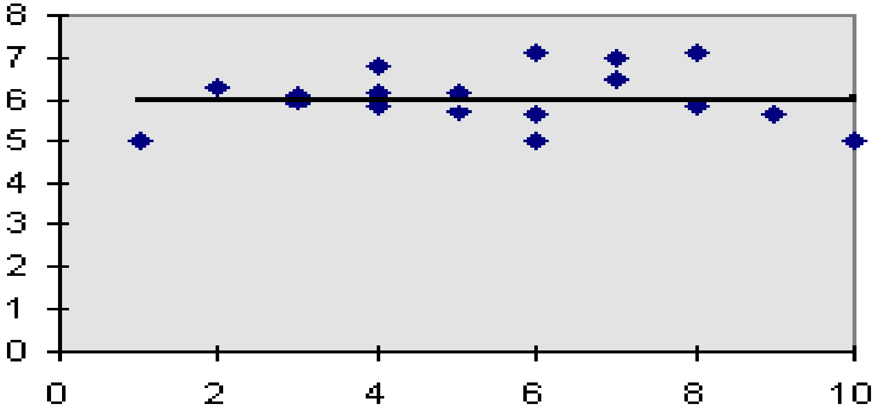
Relationship NOT Linear



Negative Linear Relationship



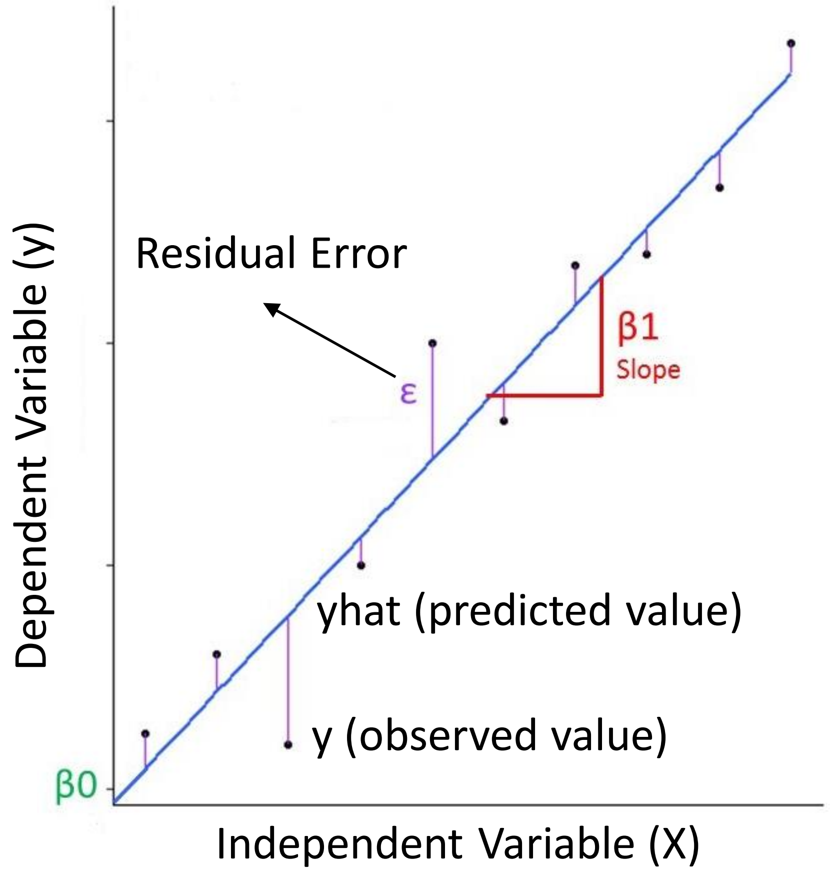
No Relationship



© Amit Mishra

The Linear Regression Line

The Least square regression line is the unique line such that the sum of squared vertical error (y) distances between the data points and the line is the smallest possible.



Linear Regression Line

© Amit Mishra

Linear Regression Equations

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

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

© Amit Mishra

Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

© Amit Mishra

Correlation

Correlation analysis studies the strength of relationship between two continuous variables. It involves computing the correlation coefficient between the two variables.

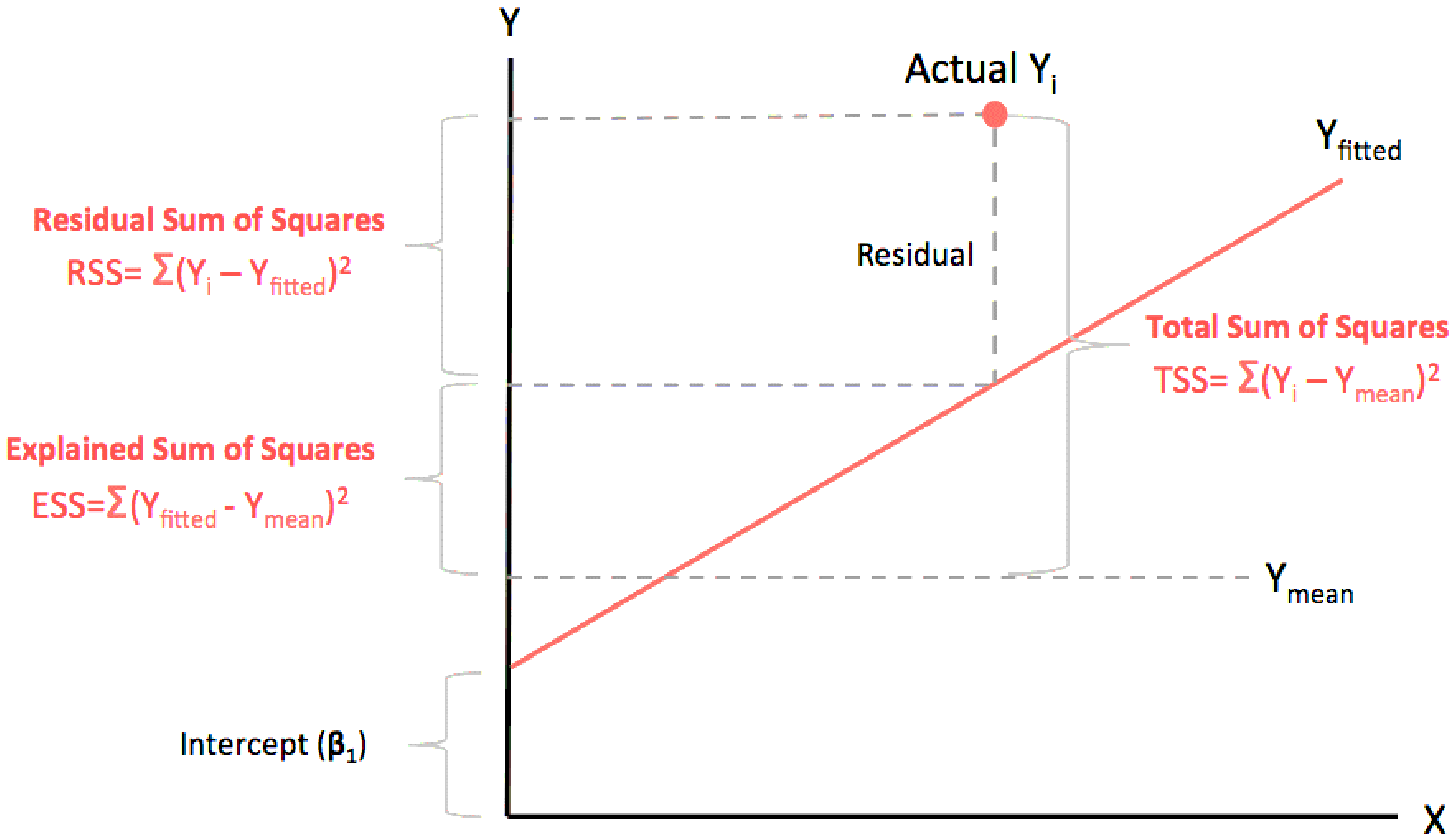
Correlation is a statistical measure that shows the degree of linear dependence between two variables.

Correlation can take values between -1 to +1. If one variables consistently increases with increasing value of the other, then they have a strong positive correlation (value close to +1).
If one consistently decreases when the other increase, they have a strong negative correlation (value close to -1).

Correlation

Correlation is only an aid to understand the relationship.
Simply use the cor() function with the two numeric variables as arguments.

R-Squared and Adj R-Squared



© Amit Mishra

R Squared

Total information in a variable is the amount of variation it contains.

R Sq.=1 – RSS/TSS

Where, RSS is the Residual Sum of Squares given by

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

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

lm ( ) Linear Model

data

an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.

Regression Model Summary

Statistic	CRITERION
R-Squared	Higher the better
Adj. R-Squared	Higher the better
Std. Error	Closer to zero the better
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better

© Amit Mishra

Least Squares Regression

The principle of least squares is one of the methods for finding a curve fitting a given data. Say (x1, y1), (x2, y2), …. (xn, yn) be n observations from an experiment.

We are interested in finding a curve.

$$y = f(x)$$

Closely fitting the given data of size 'n'. Now at x=x1 while the observed value of y is y1, the expected value of y from the curve (above) is f(x1). Let us define the residual by ..

$$e1 = y1 - f(x1)$$

Likewise, the residuals at all other points are given by ..

$$e2 = y2 - f(x2)$$

$$\ldots\ldots..$$

$$en = yn - f(xn)$$

© Amit Mishra

lm ( ) Linear Model

Some of the residuals may be positive and some may be negative. We would like to find the curve fitting the given data such that the residual at any xi is as small as possible. Now since some of the residuals are positive and others are negative and as we would like to give equal importance to all the residuals it is desirable to consider sum of the squares of these residuals, say and thereby find the curve that minimizes E.

Thus, we consider …

$$E = \sum_{i=1}^{n} \left(e_i^2\right)$$

HAPPY LEARNING