Linear Regression Tutorial

Sachin Sharma

10/3/2021

In this tutorial, we are going to study about Linear Regression in R Programming with the help of case study using dataframe. . First of all, we will explore the types of linear regression in R and then learn about the least square estimation, working with linear regression and various other essential concepts related to it.

What is Linear Regression?

Regression analysis is a statistical technique for determining the relationship between two or more than two variables. There are two types of variables in regression analysis – dependent variable and independent variable. Independent variables are also known as predictor variables. These are the variables that do not change. On the other side, the variables whose values changes w.r.t. independent variable are known as dependent variables. These variables depend on the independent variables. Dependent variables are also known as response variables.

With the help of linear regression, we carry out statistical procedure to predict the dependent variable or response variable based on the input given to the independent variable or predictor variables.

Linear Regression is of the following two types:

1. Simple Linear Regression – Based on the value of the single explanatory variable, the value of the dependent variable or response variable changes.
2. Multiple Linear Regression – As the name suggest, here value is dependent upon more than one explanatory variables in case of multiple linear regression.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

1. If the goal is prediction, forecasting, or error reduction,[clarification needed] linear regression can be used to fit a predictive model to an observed data set of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
2. If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

1. Simple Linear Regression in R

Simple linear regression is used, when we require to find the relationship between the dependent variable $\boldsymbol{Y}$ *and the independent or predictor variable* $\boldsymbol{X}$*. Both of these variables should be continuous in nature. While performing simple linear regression, we assume that the values of predictor variable X are controlled ie. there are certain inputs which are fixed available to substitute the values of x.*

Furthermore, they are not subject to the measurement error from which the corresponding value of Y is observed.

Given a data set $\{y_i,\, x_{i1},\ldots,x_{ip}\}_{i=1}^{n}\{y_i,\, x_{i1},\ldots,x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors x is linear. This relationship is modeled through a disturbance term or error variable ε an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1,\ldots,n,$

where ^{T} denotes the transpose, so that $x_i^T\beta$ is the inner product between vectors x_i and β .

Often these n equations are stacked together and written in matrix notation as

$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$

where β and ε are coefficients of regression.

Where to use Regression analysis ?

1. When we need to establish a linear relationship between the independent and the dependent variables.
2. When it is required to explain precisely the dependent variable, then we need to identify the independent variables more carefully. This will allow us to establish a more accurate relationship between dependent and independent variable.
3. The input variables $x_1, x_2 2 \ldots . x_n$ is responsible for predicting the value of y.