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 **Y** and the independent or predictor variable **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.

What is Multiple Linear Regression?

In many cases, there may be possibilities of dealing with more than one independent variable for finding out the value of the response variable. In such cases simple linear models cannot be applied as there is a need for undertaking multiple linear regression for analysing the dependent or predictor variables.

Multiple linear regression is a generalized form of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable. The basic model for multiple linear regression is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

for each observation $i = 1, ..., n$.

In the formula above we consider n observations of one dependent variable and p independent variables. Thus, Yi is the ith observation of the dependent variable, X_{ij} is ith observation of the jth independent variable, $j = 1, 2, ..., p$. The values β_j represent parameters to be estimated, and ε_i is the ith independent identically distributed normal error.

What is Least Square Method ?

The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimizing the sum of the squares of the residuals made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

The errors in the least square estimation are generated due to certain deviations in the observed points which are far from the proposed one. This deviation is residual in nature.

We can calculate the sum of squares of the residuals with the following equation:

$$\textbf{SSR} = \Sigma\, e^2 = \Sigma(y - (a_0 + a_i x))^2$$

Where e is the error, y and x are the variables, and a_0 and a_1 are the unknown parameters or coefficients.

Checking Model Adequacy

While doing predictions, we make use of the Regression Models and to make the correct predictions, the adequacy of these models is first assessed.

We use the R Squared and Adjusted R Squared methods for assessing the model adequacy.

The greater values of R-Square represent a strong correlation between the independent and the dependent variables. On the contrary, a low value represents a weak regression model by which we infer that the model is not useful to get the desired predictions or accuracy of such prediction will be so accurate.

The value of R ranges from 0 to 1. The end-point 0 points out no correlation between sample variables. Whereas, 1 means an exact linear relationship between the two variables.

R Squared can be calculated as follows:

$\boxed{R^2 = 1 - (SSR/SST)}$

Where SSR stands for Sum of Squares of Regression and SST stands for Sum of Squares of Total . These two combine to form the total sums of the squares of errors.

For adding an explanatory variable to the regression model, the adjusted R-Squared is used. **It should be noted that** the number of explanatory variables affects the value of the adjusted R-Squared. However, there is an addition of statistical penalty for each new predictor variable that is present in the regression model.

Just like the R-Squared, the adjusted R Squared can be used for the calculation of the proportion of variation that is caused by the explanatory variables.

Adjusted R Squared can be calculated as:

$R^2 = R^2 - [\frac{k(1-R^2)}{(n-p-1)}]$

Here, n represents the number of observations and p represents the number of parameters.