Linear Regression

Sina Hakimzadeh
Spring 2024
Course: Numerical analysis

Introduction

Nowadays artificial intelligence is a part of our life and lots of tasks can be done quickly and easily with AI powered devices. There are lots of methods to equip machines and softwares to predict an event based on the previous experiences that we provided.

What is Regression?

One of the most useful and powerful tools to train a model is regression. In this method we provide a data set and train a model based on its formulation to equip the program. Its formulation provides a function based on independent variables to predict the dependent variable.

Linear Regression

Linear regression is a subset of Regression methods. It just use the linear relation between dependent and independent variable means if $x_1, x_2, \dots, x_n$ are independent variable and y is dependent there is a relation between them like below:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_n x_n$$

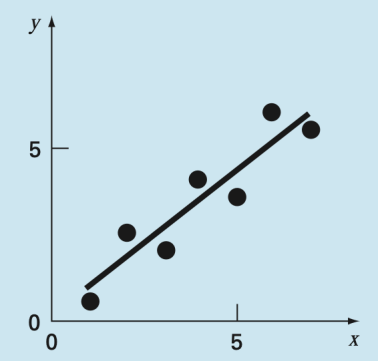
Linear Regression has these types:

1. Simple linear regression
2. Multiple linear regression

1. Single linear regression

Single regression is an analysis for those models that contains only one independent variable and have the form like below:

$$y = a_0 + a_1 x_1$$



This line equation must represent a line with the least amount of $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ that y_i is the exact value and $\hat{y}_i$ is the predicted value of the same input.

Given S as shown below:

$$S = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_1)^2$$

To evaluate the global minimum of S:

$$\frac{\partial S}{\partial a_0} = 0 = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_{1i})$$

$$\frac{\partial S}{\partial a_1} = 0 = -2\sum_{i=1}^{n}[(y_i - a_0 - a_1 x_{1i})x_{1i}]$$

This is a linear system with two variables a_0 and a_1. by solving this system we get this formulation for slope and intercept:

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a_0 = \bar{y} - a_1\overline{x}$$

2. Multiple linear regression

If a data set contains more than one independent variable we must use this method that is an extension of single linear regression. Assume that we have two independent variable, the equation is shown below:

$$y = a_0 + a_1 x_1 + a_2 x_2$$

To find the optimal slopes and intercept use previous procedure as below:

$$\frac{\partial S}{\partial a_0} = 0 = -2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})$$

$$\frac{\partial S}{\partial a_1} = 0 = -2\sum_{i=1}^{n}[(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})x_{1i}]$$

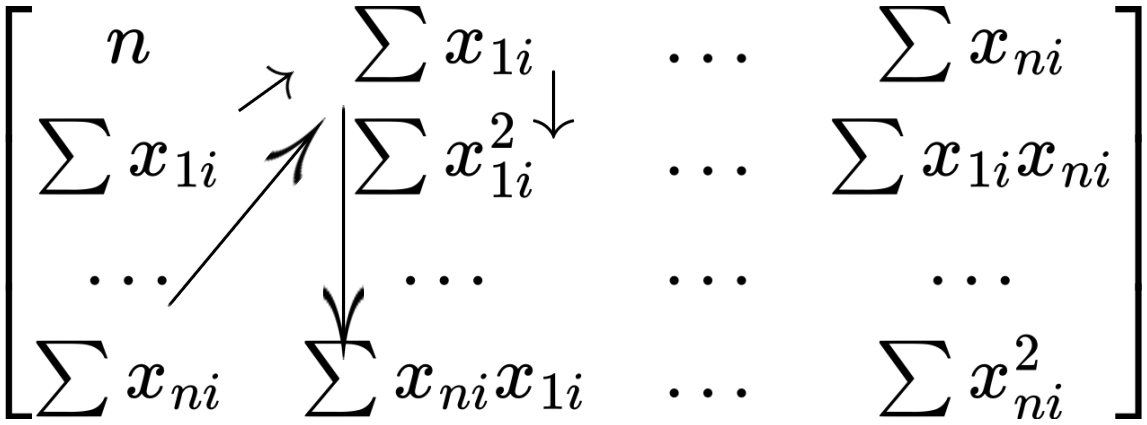
$$\frac{\partial S}{\partial a_2} = 0 = -2\sum_{i=1}^{n}[(y_i - a_0 - a_1 x_{1i} - a_2 x_{2i})x_{2i}]$$

By simplifying this linear system this formulations be yielded:

$$\begin{bmatrix} n & \sum x_{i1} & \sum x_{2i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} \\ \sum x_{2i} & \sum x_{2i}x_{1i} & \sum x_{2i}^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{bmatrix}$$

Now we can use the Gauss-elimination method to find the slopes and intercept.

Coefficient matrix has a sequence that we can use to generate it for any amount of independent variables.



This formulation will work for every possible amount of independent variable and one dependent variable.

Condition for applying linear regression

1. **Linearity**: The relationship between the independent and dependent variable should be linear. It means that if the relation is not linear the model will predict an ineligible data

2. **Independence**: Observations should be independent of each other, meaning that the independent variables must not have any relation with each other.

Conclusion

Linear Regression is one the most powerful methods to predict something depending on different situations, keep in mind that linear regression is a good method when the relation between dependent and independent variables is somehow linear like car price that is dependent on its power.