Int. J. Math. And Appl., 8(3)(2020), 201–206
ISSN: 2347-1557
Available Online: http://ijmaa.in/

International Journal of *Mathematics And its Applications*

Linear Regression With Only One Feature and Applications

Xinran Zheng[1,*]

1 Department of Mathematics, Tianjin University of Technology, Weijin Road Campus, Weijin Road, Nankai, Tianjin, China.

Abstract: We review how to obtain, by elementary means, the equations that determine the slope and y-intercept of the line that best fits a certain given data. We illustrate the use of these equations with an example.

MSC: 47A63.

Keywords: Linear Regression, Elementary Means, Data Science.

1. Introduction

The advancement in technology has lead to the ability of collecting large amounts of data and manipulate that data. This has lead to a new field of science know as Data Science [3], that has as goal the use of data to make predictions and/or increase our understanding of the processes behind the data. This is of interest not only in most fields of sciences and engineering, where the data is being used to increase the understanding of different phenomena, but also in most business, where the data is being used to increase the bottom line. Data Science is a vast field. The subject of this article is linear regression [4], a technique that belongs to the field of Machine learning [2], which in turn is a core subject of Data Science. Broadly speaking, supervised learning [1] refers to techniques that have as goal to predict an output y in terms of an input x. In this article, both the input x and output y will be real numbers, but in others scenarios they could be vectors.

In supervised learning, a set of n training examples are given. This means that we are given a set of n inputs, $x_1, x_2, \ldots, x_n$ and their corresponding outputs $y_1, y_2, \ldots, y_n$. This is our data. In the machine learning language, and as we will explain in more detail in the next section, this data is used to train our model. Once the model is trained, the model is ready to be used to make predictions for new inputs x.

Linear regression is probably the most basic technique in the family of methods that belong to supervised learning. This is the method we will describe in this article. In its simplest form, and as described in this article, linear regression finds the line that best fits the given data. In other words, y is approximated as a linear function of x, i.e. $y = \alpha x + \beta$, and we use the data to find the slope and y-intercept to get the line that best fits the given data. The details of this process is explained in this article.

This article is organized as follows. In Section 2, we give the precise meaning to the phrase *the line that best fits the data*. We obtain the equations that determine the slope and y-intercept of this line in terms of the data of the training examples. In

* *E-mail: 906009880@qq.com*

Section 3, we consider an illustrative example that consists of profit of a food truck over a period of time, and the population of the city where the food truck operates. In Section 4, we finish this article with a small discussion.

2. Linear Regression: Line that Best Fit Data

Linear regression is a sort of statistical analysis technique that is to model the relationship between a scalar response and one or more explanatory variables. In the linear regression, the regression analysis with only one independent variable and one corresponding dependent variable is called the unitary linear recursive analysis. The relationships between them can be approximately expressed by a line. We can model the relationships by using the linear predictor functions to get the parameters of the line, and we can also estimate the unknown model parameters from the data.

Models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and the statistical properties of the resulting estimators are easier to determine. So the linear regression can be applied to many aspects. For example, if we are given a set of data that gives a series of x and the corresponding y, we can use the linear regression to fit a predictive model, and then the model can be used to make a prediction of the value of y even if we are given a new data without an accompanying response value to match with it.

Now we will introduce the linear regression analysis as follows.

Definition 2.1. *Specifically, in the supervised learning, the data set including the above data is called training set. The variate x is called the feature, and the predicted value y is called the target.*

Observation 2.2. *The supervised learning is to output the predicted value of the target through inputting the features.*

Now we are given a set of n vectors, each with two components

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}, \dots, \begin{bmatrix} x_n \\ y_n \end{bmatrix}. \tag{1}$$

From this data we construct the n-vector $\mathbf{x}$ and the n-vector $\mathbf{y}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{2}$$

Definition 2.3. *Given a line $y = \alpha x + \beta$, which is called the hypothesis model. The α in it is called the slope and the β in it is called the y-intercept.*

we can construct the n-vector whose i^{th} component is the y value of this line at x_i, i.e.

$$\begin{bmatrix} \alpha x_1 + \beta \\ \alpha x_2 + \beta \\ \vdots \\ \alpha x_n + \beta \end{bmatrix}. \tag{3}$$

Definition 2.4. *The line that best fits the data is the line $y = \alpha x + \beta$ with the α and β that makes the distance from the vector $\mathbf{y}$ to the vector in Equation (3) as small as possible*

Definition 2.5. *The cost function J is the square of the distance between* $\mathbf{y}$ *to the vector in Equation (3) divided by n to describe how close the predicted value is to the actual value.*

$$J = J(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} (\alpha x_i + \beta - y_i)^2. \tag{4}$$

Observation 2.6. *The α and β of the line $y = \alpha x + \beta$ that best fits the data is the the α and β that minimizes J.*

Observation 2.7. *Let $f(x) = ax^2 + bx + c$ with $a > 0$. Then $f(x)$ attains its minimum at*

$$x = -\frac{b}{2a}. \tag{5}$$

The cost function J can be written in the form

$$J = \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2\right)\alpha^2 + \frac{2}{n}\left(\sum_{i=1}^{n}(\beta - y_i)x_i\right)\alpha + \frac{1}{n}\sum_{i=1}^{n}(\beta - y_i)^2. \tag{6}$$

We think of β as a fixed number. J is a quadratic function of α. The Observation 2.7 implies that if the minimum of J is attained for a pair α, β then, they satisfy the equations

$$\alpha = -\frac{\sum_{i=1}^{n}(\beta - y_i)x_i}{\sum_{i=1}^{n} x_i^2}. \tag{7}$$

Similarly, the cost function J can be written in the form

$$J = \beta^2 + \frac{2}{n}\left(\sum_{i=1}^{n}(\alpha x_i - y_i)\right)\beta + \frac{1}{n}\sum_{i=1}^{n}(\alpha x_i - y_i)^2. \tag{8}$$

We now think of α as a fixed number. J is a quadratic function of β. The Observation 2.7 implies that if the minimum of J is attained for a pair α, β then, they satisfy the equations

$$\beta = -\frac{1}{n}\sum_{i=1}^{n}(\alpha x_i - y_i). \tag{9}$$

After simple algebraic manipulations of Equations (8) and (9), we obtain that the α, β that minimize J are the solution of the system

$$\left(\sum_{i=1}^{n} x_i^2\right)\alpha + \left(\sum_{i=1}^{n} x_i\right)\beta = \sum_{i=1}^{n} x_i y_i \tag{10}$$

$$\left(\sum_{i=1}^{n} x_i\right)\alpha + n\beta = \sum_{i=1}^{n} y_i \tag{11}$$

Observation 2.8. *Assume that $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$ are known numbers.*

$$a_{11} = \left(\sum_{i=1}^{n} x_i^2\right) \tag{12}$$

$$a_{12} = \left(\sum_{i=1}^{n} x_i\right) \tag{13}$$

$$a_{21} = \left(\sum_{i=1}^{n} x_i\right) \tag{14}$$

$$a_{22} = n \tag{15}$$

Then, the system of equations

$$a_{11}z_1 + a_{12}z_2 = b_1 \quad (16)$$

$$a_{21}z_1 + a_{22}z_2 = b_2 \quad (17)$$

has exactly one solution if and only if

$$\Delta = a_{11}a_{22} - a_{21}a_{12} \neq 0. \quad (18)$$

In this case, the solution is

$$z_1 = \frac{a_{22}b_1 - a_{12}b_2}{\Delta} \quad (19)$$

$$z_2 = \frac{-a_{21}b_1 + a_{11}b_2}{\Delta}. \quad (20)$$

3. An illustrative example

Now we illustrate the use of these equations with an example.

The profit of a food truck over a period of time depends on the population of the city where the food truck operates. Assume that we have lots of data of the profit of a food truck and the population of the city, what we want to do is to use the data to fit a predictive model, and then we use the model to make a prediction of the profit of a food truck in terms of the population of the city. Now we will achieve this goal by using the method of the linear regression analysis that belongs to supervised learning. We write a python function that use the equation from Equation (12) to Equation (15) above to get the $a_{11}, a_{12}, a_{21}, a_{22}, b_1$ and b_2. Then we can use them to get z_1 and z_2 in terms of the Equation (19) and Equation (20), which means the slope and y-intercept. So when we input a list, where each elements of the list is a list of two points, x_i and y_i, the output should be α and β, the slope and y-intercept of the line that best fits these data. And we use the correlation function to make a plot in order to show the line that best fits the data, so that we can intuitively see the results.

The code of the python function is as follows:

```
Import matplotlib.pyplot as plt
with open('foodTruck.txt') as file:
    text = file.read()
list = text.split('\n')
list.pop()

x, y = [], []
for word in list:
    a = word.split(',')
    x.append(float(a[0])),y.append(float(a[1]))

a11, a12, b1, b2 = 0, 0, 0, 0
for i in range(len(x)):
    a11 = a11 + x[i] * x[i]
    a12 = a12 + x[i]
```

$b_1 = b_1 + x[i] * y[i]$

$b_2 = b_2 + y[i]$

$a_{21} = a_{12}$

$a_{22} = \text{len}(x)$

$det = a_{11} * a_{22} - a_{21} * a_{12}$

$\alpha = (a_{22} * b_1 - a_{12} * b_2)/det$

$\beta = (-a_{21} * b_1 + a_{11} * b_2)/det$

$x_1 = \min(x)$

$x_2 = \max(x)$

$y_1 = \alpha * x_1 + \beta$

$y_2 = \alpha * x_2 + \beta$

$\text{plt.plot}([x_1, x_2], [y_1, y_2])$

$\text{plt.plot}(x, y, ' m.')$

$\text{plt.show}()$

$\text{print}(\alpha, \beta)$

We import the file foodTruck.txt as the input. The data in this file is: (population of city)/10000, (profit of a food truck)/10000. Then we will get the output as follows, the slope and y-intercept and the plot of the line that best fits the data.

slope = 1.193

y-intercept = -3.896

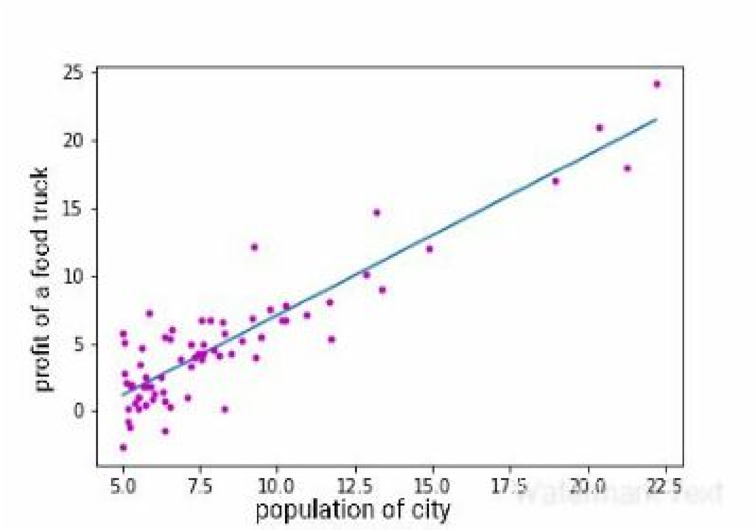

Figure 1. **the line that best fits the data**

The result gives the slope and y-intercept and shows the line that best fits the data. And we can also get the equation of the fitted curve, $y = 1.193x - 3.896$. Therefore, we can easily solve the problem which is to get the values of the profit by inputting the population.

4. Conclusion

The solution of this regression problem in the machine learning, which belongs to the supervised learning, is usually to use a set of known data to predict the continuous values. We will get different target values in terms of different features we input, We can use this method to solve a series of similar problem and apply it to many fields.

For example, we can use the machine learning method to get the approximate price of the house by giving the area of the house and we can get the approximate consumption of a family by giving the income of a family.

The more data we input used to train the model, the more precise predicted values we will get. So what we should do is to train our model by giving the data as many as possible to get the predicted values whose error is so small that we can neglect.

References

[1] Rich Caruana and Alexandru Niculescu-Mizil, *An empirical comparison of supervised learning algorithms*, In Proceedings of the 23rd International Conference on Machine learning, (2006), 161-168.

[2] Donald Michie, David J. Spiegelhalter and C. C. Taylor, *Machine learning*, Neural and Statistical Classiffcation, 13(1994), 1-298.

[3] Wil Van Der Aalst, *Data science in action*, In Process mining, Springer, (2016).

[4] Sanford Weisberg, *Applied linear regression*, Volume 528, John Wiley & Sons, (2005).