Parameters of linear model

Quistian Navarro, Juan Luis
A341807@alumnos.uaslp.mx

January 29, 2024

Ing. Sistemas Inteligentes, Gen. 2021
Machine Learning, Group 281601

Abstract

This document discusses the solution to linear regression, specifically focusing on finding the parameters that describe the linear model. The method of 'Minimum Mean Square Error' is employed, and the mathematical details are explained with a direct link to the implementation.

1 Introduction

A model is a representation of the reality that can describe the world in a simplify form.

In this case, will we analyze the Linear Regression. More in specifically, we will find the parameters of linear model. Linear regression is a statistical modeling technique used to describe a continuous response variable as a function of one or more predictor variables[5].

2 Content

Based on the 'dataset' accidents, which contains the size of the population per state (x-axis) and the number o fatal accidents (y-axis) in United States of America (USA)[2].

Population per state	Number of fatal accidents
493782	164
572059	43
608827	98
...	...
18976457	1493
20851820	3583
33871648	4120

Table 1: dataset: accidents.

Well, the activity that we concern is find a parameters of the linear model. The general equation of a line is:

$$y = mx + b$$

which can be also expressed as :

$$y = a_1 x + a_0$$

The first parameter a_1 describe the height at which it intersects the y-axis and the second parameter a_0 describe graphically what is the slope and conceptually what is the relationship between the input variable x and the output variable y.

The best way to representing the variables of the linear model is the vector form. That is, create a matrix, which we will call 'X', where each column represent a characteristic of the input data. With the output the same way, we can represent with a vector, which we will call 'Y'. And also for the parameters (factors) we can represent them in a vector, which we will call 'W'

If we do it. Now, we can reduce all to a simply vector equation

$$Y = XW$$

Then, the way that i used to find the parameters is with 'minimum root mean square error'. This method consists of obtaining the mean of the errors of the distance between the actual output values and the values predicted by the model, and squaring it. Conceptually, what we do when we square is to penalize more heavily those points farther away from the line, and less heavily those that are closer.

Minimum root mean square error:

$$mean(y_r - y_e)^2$$

Which can be also expressed as vector minimum root mean square error:

$$(Y - XW)^T (Y - XW)$$

If we work on it, we can multiply both parentheses and obtain this expression:

$$Y^T Y - W^T X^T Y - Y^T XW + W^T X^T XW$$

And remember that in mathematics, we can find the minimum of function by calculating its derivative and equaling to zero.

Then, if we derive this function we can obtain this expression:

$$-2X^T Y + 2X^T XW = 0$$

And if we operate and clear, we end up with:

$$W = (X^T X)^{-}1 X^T Y$$

This equation is all we needed to solve our problem. And the only thing we would have to program to find the line we want, and therefore, find the parameters of the line.

In this case, it has been appropriate to use this method to solve the problem, but if we work with other models or cost functions, we may not always be able to find the minimum cost analytically. Additionally, the algorithmic complexity of calculating the inverse of a matrix is cubic $O(n^3)$, so with a larger amount of data, it could become challenging for this method to be useful[1].

Well, so this describe the method used to find the line parameters. I have implemented it in both Python and C.

3 Results

Basically the operation that performs the calculation of the parameters is described in the following[4].

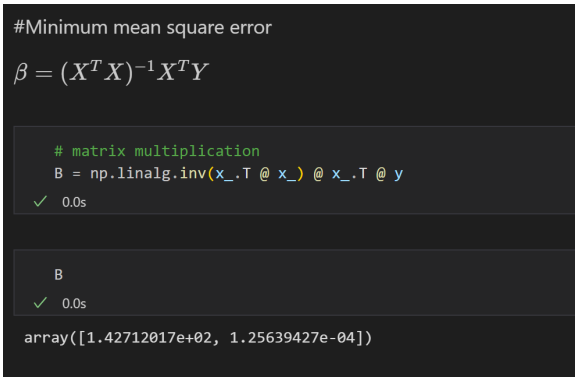

Figure 1: Minimum mean square error

And if we plot the line we get the following.

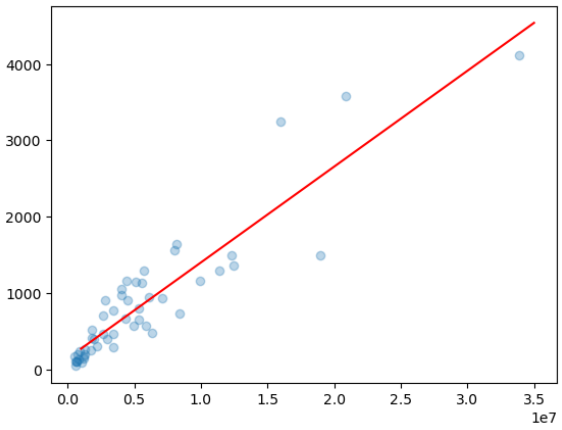

Figure 2: Linear regression

In the case of the code in C, is more difficult show the part most important, because the code is more bigger than the code in Python.

But, your can consult in this direct link [3].
And the results are her.

```
----------------Res Mult (xT * x)^-1 * xT * Y
[142.712017,0.000126]
```

Figure 3: Output of the program in C language

4 Conclusion

In conclusion, understanding various areas of knowledge is crucial for solving problems and finding suitable models. Awareness of algorithmic complexity aids in making informed choices, optimizing resource utilization, and implementing effective solutions with the technology at hand.

References

[1] Dot CSV. Regresión lineal y mínimos cuadrados ordinarios — dotcsv. https://youtu.be/k964_uNn3l0?si=rTet0kCoDIUmcPo2, Dic 2017.

[2] Juan Carlos Cuevas-Tello. Handouts on regression algorithms. https://www.researchgate.net/publication/358279907_Handouts_on_Regression_Algorithms, 09 2020.

[3] Quistian Navarro Juan Luis. Parameters of linear model c. https://github.com/juanQNav/Parameters_of_linear_models_C, 01 2024.

[4] Quistian Navarro Juan Luis. Parameters of linear model python. https://colab.research.google.com/drive/1mnemd8hA8FcGNZK3wFR_Ix40NYZJPJnH?usp=sharing, 01 2024.

[5] MathWorks. What is linear regression. https://www.mathworks.com/discovery/linear-regression.html?s_tid=srchtitle_site_search_1_linear%20regression, s.f.