THE METHOD OF LEAST SQUARES

VERONIKA TITARCHUK AND YOUNHA LEE

CONTENTS

1. INTRODUCTION

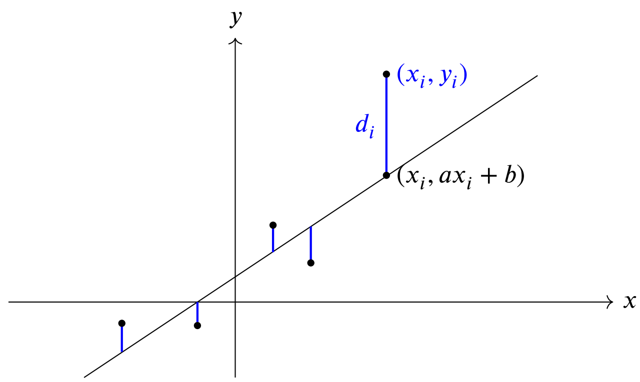
Imagine you are a TV advertiser. You have some data that the more you advertise TV, the more sales increase. How did you get the data? As an advertiser, you have to investigate the sales volume about the advertising time. Then, you can predict whether the sales increase or not as you want to advertise TV. In this case, using Linear regression is a method to predict it.

Now, we introduce the Linear Regression. First of all, Linear Regression is a regression analysis that models the linear correlation between the dependent variable y and one or more independent variables x. The dependent variable is the variable we want to predict and independent variables is used for us to predict the dependent variable. In order to find the line $y = ax + b$ which represents a lot of data, the error between the predicted value fitting to the line and the actual value must be minimized.

2. THE MAIN RESULT

In our article we are going to present the method of least squares, which helps us to find the best fitting line $y = ax + b$ to the given data $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ in the sense that

$$\sum_{i=1}^{n} [(y_i - (ax_i + b)]^2 \text{ is minimal.}$$



This means that we are trying to find the line which has the least sum of errors (sum of all blue vertical lines in the picture). That is, we need to find a and b that minimizes the function below:

$$f(a, b) = \sum_{i=1}^{n}[(y_i - (ax_i + b)]^2.$$

Firstly, we need to find a and b in $y = ax + b$. To do this, we need to find critical points (particularly minimum) which we can find using derivative of formula 1. As it is a 2 variables function, we need to find partial derivatives of a and b. We will start with derivative of a. To find critical points of $f(a, b)$, we compute the partial derivatives of $f(a, b)$. The partial derivative of f with respect to a is

$$\begin{aligned}\frac{\partial f}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}[(y_i - (ax_i + b)]^2 &= \sum_{i=1}^{n} 2[(y_i - (ax_i + b)] \cdot (-x_i) \\ &= \sum_{i=1}^{n}(-2x_i y_i + 2x_i^2 a + 2x_i b) \\ &= -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i.\end{aligned}$$

and the partial derivative of f with respect to b is

$$\begin{aligned}\frac{\partial f}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n}[(y_i - (ax_i + b)]^2 &= \sum_{i=1}^{n} 2[(y_i - (ax_i + b)] \cdot (-1) \\ &= -2\sum_{i=1}^{n} y_i + 2a\sum_{i=1}^{n} x_i + 2bn.\end{aligned}$$

The result is:

$$\frac{\partial f}{\partial a} = -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i \text{ and } \frac{\partial f}{\partial b} = -2\sum_{i=1}^{n} y_i + 2a\sum_{i=1}^{n} x_i + 2bn.$$

Secondly, we need to equate our results of partial derivatives of a and b to 0 in order to find critical points. We shall solve the following system of linear equations with unkowns a and b

$$\begin{cases} \left(\sum_{i=1}^n x_i^2 \right) a + \left(\sum_{i=1}^n x_i \right) b & = \sum_{i=1}^n x_i y_i \\ \left(\sum_{i=1}^n x_i \right) a + \qquad nb & = \sum_{i=1}^n y_i \end{cases}$$

Solving the second equation for b, we find that

$$b = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n}$$

Substituting b into the first equation, we have

$$a \sum_{i=1}^n x_i^2 + \sum_{i=1}^n x_i \left(\frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n} \right) = \sum_{i=1}^n x_i y_i$$

which becomes

$$a \sum_{i=1}^n x_i^2 + \frac{1}{n} \sum_{i=1}^n x_i \cdot \sum_{i=1}^n y_i - \frac{a}{n} \left(\sum_{i=1}^n x_i \right)^2 = \sum_{i=1}^n x_i y_i \implies a \left(\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$

We find that

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

and

$$b = \frac{1}{n} \left[\sum_{i=1}^n y_i - \sum_{i=1}^n x_i \left(\frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \right) \right].$$

Example 1. We are given a set of data points $(x_1, y_1) = (1, 2)$ and $(x_2, y_2) = (2, 4)$. We will find a and b in $y = ax + b$ that best fits the data.

(Solution) Let's insert data set into formulas a and b.

$$a = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{2(1 \cdot 2 + 2 \cdot 4) - (1+2)(2+4)}{2(1^2 + 2^2) - (1+2)^2} =$$

$$= \frac{20 - 18}{10 - 9} = 2 \implies a = 2$$

and

$$b = \frac{1}{n} \left[\sum_{i=1}^n y_i - \sum_{i=1}^n x_i \left(\frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \right) \right] = \frac{\sum_{i=1}^n y_i - a \sum_{i=1}^n x_i}{n} =$$

$$= \frac{(2+4) - 2(1+2)}{2} = 0 \implies b = 0$$

We find that $y = 2x$ best fits to the data set.

3. Local minimum and absolute minimum

The first partial derivatives of f are

$$\frac{\partial f}{\partial a} = -2\sum_{i=1}^{n} x_i y_i + 2a\sum_{i=1}^{n} x_i^2 + 2b\sum_{i=1}^{n} x_i \quad \text{and} \quad \frac{\partial f}{\partial b} = -2\sum_{i=1}^{n} y_i + 2a\sum_{i=1}^{n} x_i + 2bn.$$

and the second partial derivatives of f are

$$\frac{\partial^2 f}{\partial a^2} = 2\sum_{i=1}^{n} x_i^2, \qquad \frac{\partial^2 f}{\partial b^2} = 2n \qquad \text{and} \qquad \frac{\partial^2 f}{\partial b \partial a} = 2\sum_{i=1}^{n} x_i$$

We got $\frac{\partial^2 f}{\partial a^2} = 2\sum_{i=1}^{n} x_i^2 > 0$ and

$$\begin{aligned} D &= \frac{\partial^2 f}{\partial a^2} \cdot \frac{\partial^2 f}{\partial b^2} - \Big(\frac{\partial^2 f}{\partial b \partial a}\Big)^2 \\ &= 4\Big[n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big] \end{aligned}$$

The Second Derivative Test says that

- if $D > 0$ and $f_{aa} > 0$, then $f(x, y)$ is a local minimum
- if $D > 0$ and $f_{aa} < 0$, then $f(x, y)$ is a local maximum
- if $D < 0$ and (x, y) is a saddle point
- if $D = 0$, no conclusions may be drawn.

Now we have to analyze whether D is greater, less or equal to zero. We can do this using the Cauchy-Schwarz inequality. The general form of it is

$$(x_1 y_1 + x_2 y_2 + ... + x_n y_n)^2 \leq (x_1^2 + x_2^2 + ... + x_n^2)(y_1^2 + y_2^2 + ... + y_n^2)$$

Let's assume that

$$\begin{cases} x_i = x_i \\ y_i = 1 \end{cases}$$

then $(x_1 + x_2 + ... + x_n)^2 \leq (x_1^2 + x_2^2 + ... + x_n^2) \cdot n$.

We can rewrite this as $\left(\sum_{i=1}^{n} x_i\right)^2 \leq n\sum_{i=1}^{n} x_i^2$. After transfering everything to the right side and multiplying both the sides by 4, we get

$$D = 4\Big[n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big] \geq 0.$$

This result is not enough for us, because we have only 3 options greater, less or equal to 0. Let's assume that $D = 0$. Then all x_i must be constant and our data plot is a straight line which can not have any critical points. So, we assume that x_i are not constant. Furthemore, we have at least 2 points in our data set. These conditions make the case when $D = 0$ impossible. That's why

$$D = 4\Big[n\sum_{i=1}^{n} x_i^2 - \Big(\sum_{i=1}^{n} x_i\Big)^2\Big] > 0.$$

We got the results $f_{aa} > 0$ and $D > 0$, which makes this point (a, b) a local mininum by The Second Derivative Test.

Proposition 2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a twice differentiable function. Assume that

i) there exists exactly one critical point (x_0, y_0),
ii) $f(x_0, y_0)$ is a local minimum, and
iii) $\lim_{x^2+y^2\to\infty} f(x, y) = \infty$.

Then $f(x, y)$ has the absolute minimum at (x_0, y_0).

Proof. Choose a point $(x_1, y_1) \in \mathbb{R}^2$ such that $f(x_1, y_1) > f(x_0, y_0)$ and let $M = f(x_1, y_1)$. There exists a real number $R > 0$ such that

$$x^2 + y^2 \geq R \implies f(x, y) \geq M$$

because $\lim_{x^2+y^2\to\infty} f(x, y) = \infty$. Let

$$D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq R\}.$$

The function $f : D \to \mathbb{R}$ has the minimum value by the extreme value theorem. The minimum value of f on D occurs either at critical points inside D or on the boundary $\partial D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = R\}$ of D. However, the minimum value of f on D cannot be obtained on ∂D because $f(x, y) \geq M > f(x_0, y_0)$ for all $(x, y) \in \partial D$ and (x_0, y_0) is the only critical point inside D. Therefore $f(x_0, y_0)$ is the minimum value of f on D. Since $f(x, y) \geq M > f(x_0, y_0)$ for all $(x, y) \notin \partial D$, we conclude that $f(x_0, y_0)$ is the minimum value on $\mathbb{R}^2$. □

This propositions helps us define that our local minimum point (a, b) is not only a local minimum, but also an absolute minimum on $\mathbb{R}^2$. This is because our data has only one critical point and it is a local minimum. Furthermore, $\lim_{\substack{a\to\infty \\ b\to\infty}} f(a, b) = \infty$ which means that $\lim_{a^2+b^2\to\infty} f(a, b) = \infty$.

Example 3. Let's find a fitting line for the data with 10 different points $(x_i, y_i) = (3, 4)(4, 6)(6, 5)(7, 7)$ $(7.5, 8)(8, 6)(9, 7)(9, 8)(9.5, 9)(10, 8)$

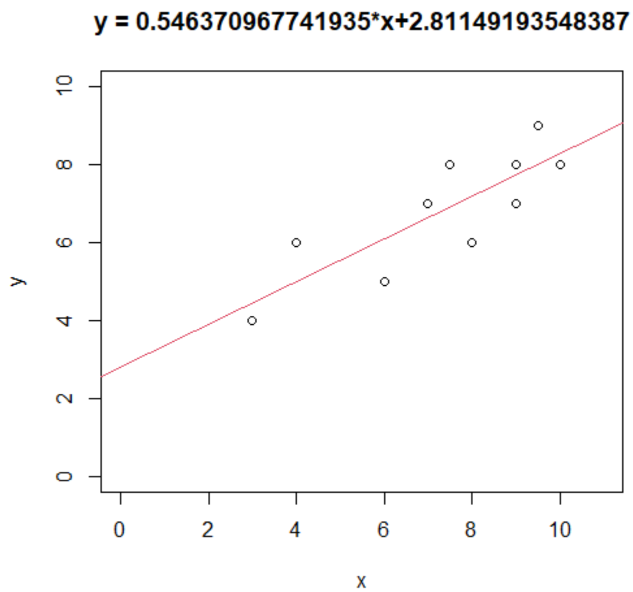
(Solution)

Let's build our model using R language. The benefit of it is that we can not only find the slope, but also check whether our Linear regression model is correct by using stantard R functions.

We found a and b using our formulas and got the result:

Values	
a	0.546370967741935
b	2.81149193548387

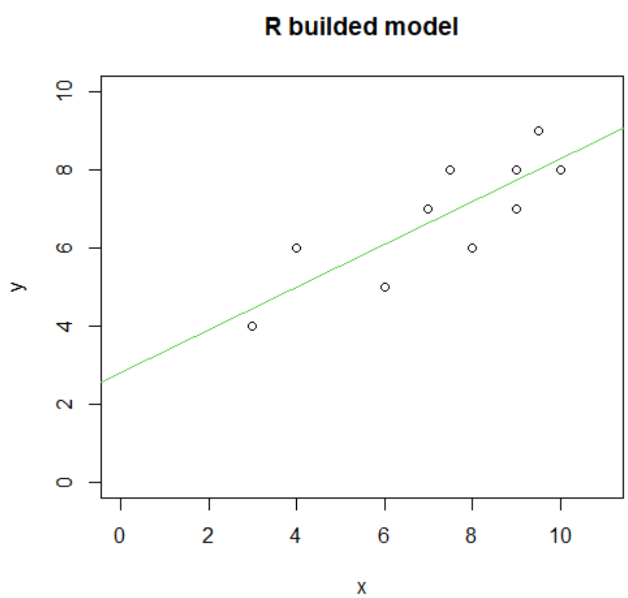
Our graphic with all the data points and a fitting line is shown below.



After this, let's use a standart R function, which do the same calculations and we will get the *a* and *b*

```
> model$coefficients
(Intercept)           x
   2.811492    0.546371
```

And we get the graphic



4. Difference between "vertical distance", "horizontal distance" and "distance between a point and a line" for finding sum of errors

Usually, working with linear regressions we use "vertical distance". But why not 'horizontal distance" or "distance between a point and a line"? Let's talk about the first one. Horizontal and vertical distances are actually symmetric to each other. The difference is that in vertical distance, we assume that x data is correct and y data can have some errors, and in horizontal distance that the y data is correct and x data can have some errors. The basic linear equation is $y = ax + b$, so it is easier to assume that x data is stable so that we can calculate y using x.

Let's assume that we can use "distance between a point and a line" instead of "vertical distance". The basic equation for the distance between the point and the line is

$$d = \frac{|ax + by + c|}{\sqrt{a^2 + b^2}}$$

for the line $ax + by + c = 0$. Our line is $ax_i - y_i + b = 0$, so we get

$$d_i = \frac{|ax_i - y_i + b|}{\sqrt{a^2 + 1}}.$$

In order to find sum of errors and the local minimum of it, we have to sum all d_i, and find the $f(a, b)$ where

$$f(a, b) = \sum_{i=1}^{n} \frac{|ax_i - y_i + b|}{\sqrt{a^2 + 1}} \text{ is minimal.}$$

Let's find it using partial derivatives. We will start with the partial derivative of a.

$$\frac{\partial f}{\partial a} = \frac{\partial}{\partial a} \sum_{i=1}^{n} \frac{|ax_i - y_i + b|}{\sqrt{a^2 + 1}} = \sum_{i=1}^{n} \frac{(ax_i - y_i + b) \cdot x_i \cdot \sqrt{a^2 + 1}}{|ax_i - y_i + b| \cdot (a^2 + 1)} - \sum_{i=1}^{n} \frac{|ax_i - y_i + b| \cdot a}{\sqrt{a^2 + 1} \cdot (a^2 + 1)}.$$