Linear Regression

First of all, my purpose in writing this article is to explain my journey to learning data science, and I will write another article soon, so let's get started. Hello, everyone. My name is Fernanda Subekti, and I want to share what I learned about linear regression and non-linear least squares some time ago. Firstly, linear regression is a statistical method to predict some data points (not only about maths) from data points that we have in a linear function. In maths, the general function for simple linear regression is $y = mx + c$, with m as slope and c as intercept. Look at this example.

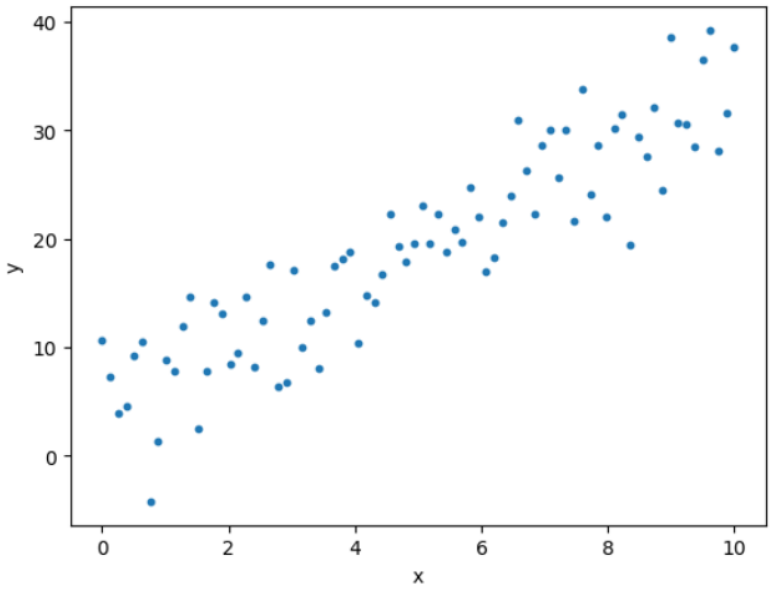


Figure 1. Data points

We have some data points, like in the figure. So, our purpose is to determine the best value for m and c so that our prediction has the best value. Let's look at this plot again, but this time, we use $y = 3x + 1$ as an example.

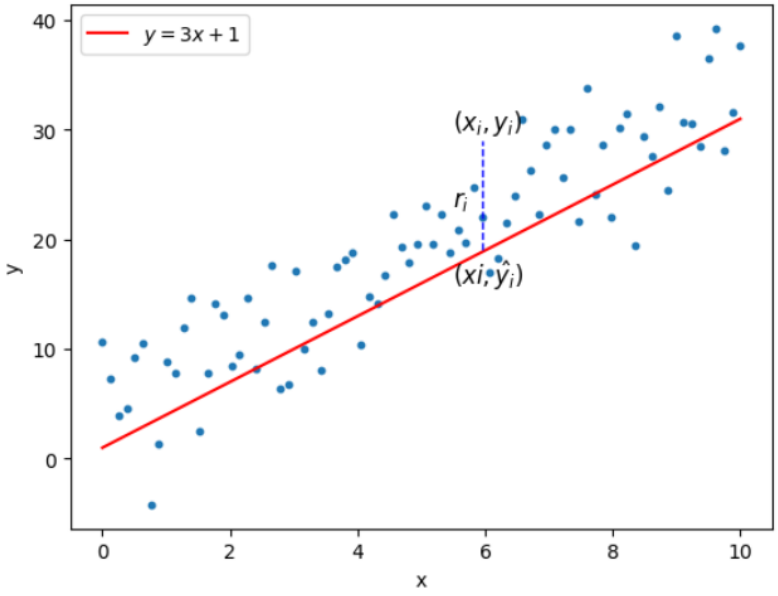


Figure 2. Data points with linear regression

The data points have an attribute (x_i, y_i), we have the labels that are y_i. Meanwhile, our prediction is $\hat{y_i} = mx_i + c$. So, we have the residual that is $r_i = y_i - \hat{y_i}$ or $r_i = y_i - mx_i - c$. After that, the sum of ${r_i}^2$ we call chi-squared, which

$$\chi^2 = \sum_{i=1}^{n} {r_i}^2$$

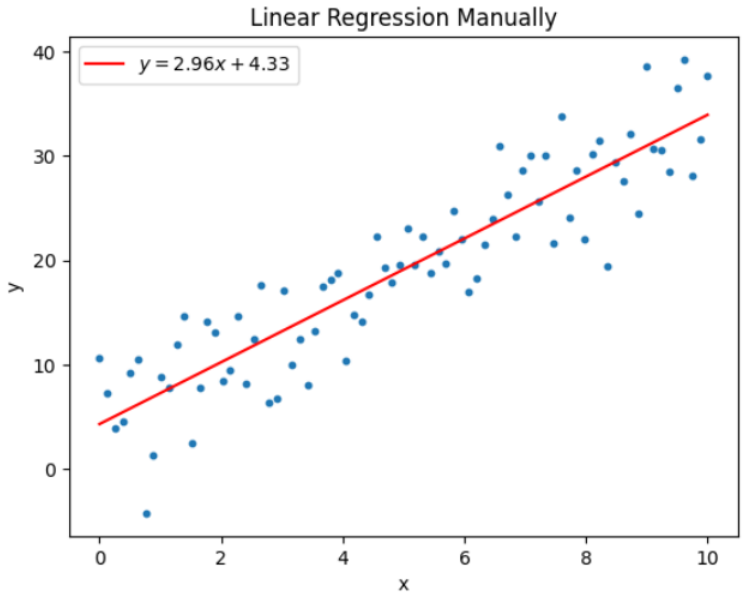
Our purpose is to make the distance between y and $\hat{y}$ minimum. In other words, we want to chi-squared minimum, and we can do that by partially derivating the chi-squared with respect to parameters m and c and equaling zero. From that, we got,

$$c = \bar{y} - m\bar{x}$$

and

$$m = \frac{\sum(x - \bar{x})y}{\sum(x - \bar{x})^2}.$$

We will get the m and c values by inputting the x_i value to the $mx_i + c$. Also, we can predict that each value will form a linear function. Let's do that computationally. We will use Jupyter Notebook as a tool to write the Python code. Check my Jupyter Notebook for the computational solution, which we will do manually and use the library Scipy separately. From the data points in Figure 1, we will have the m and x best fit to that figure, which you can see in Figure 3.



From Figure 3, we got an m value of 2.96 and a c value of 4.33, from those values, we can create the plot corresponding to each data point, and the output is $\hat{y_i}$ and form linear regression as Figure 3. From this intuition, we can also build another case with linear regression, such as the linearity of GDP per capita against life expectancy or other cases. We also can do machine learning modeling from training labels and testing labels by this intuition. The final word. Thank you. You can repost this article if you want it or correct the article if there are wrong explanations.