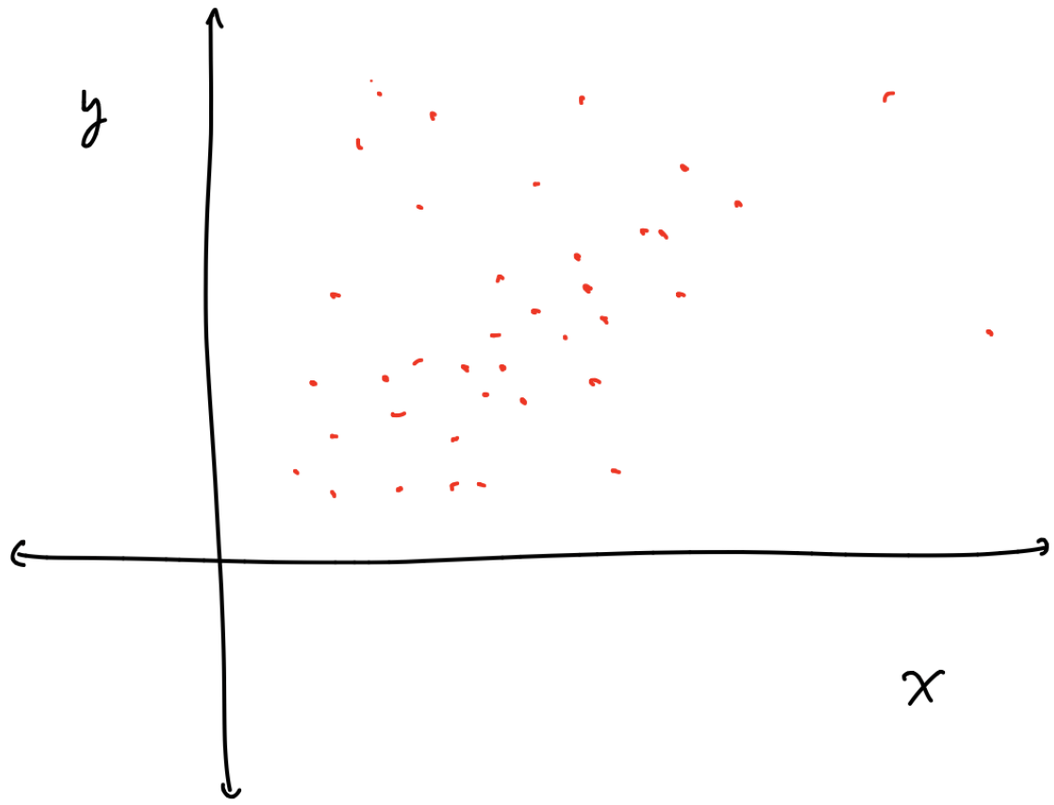
Week 12 – **Linear Regression**

Slides by Suraj Rampure

Fall 2017

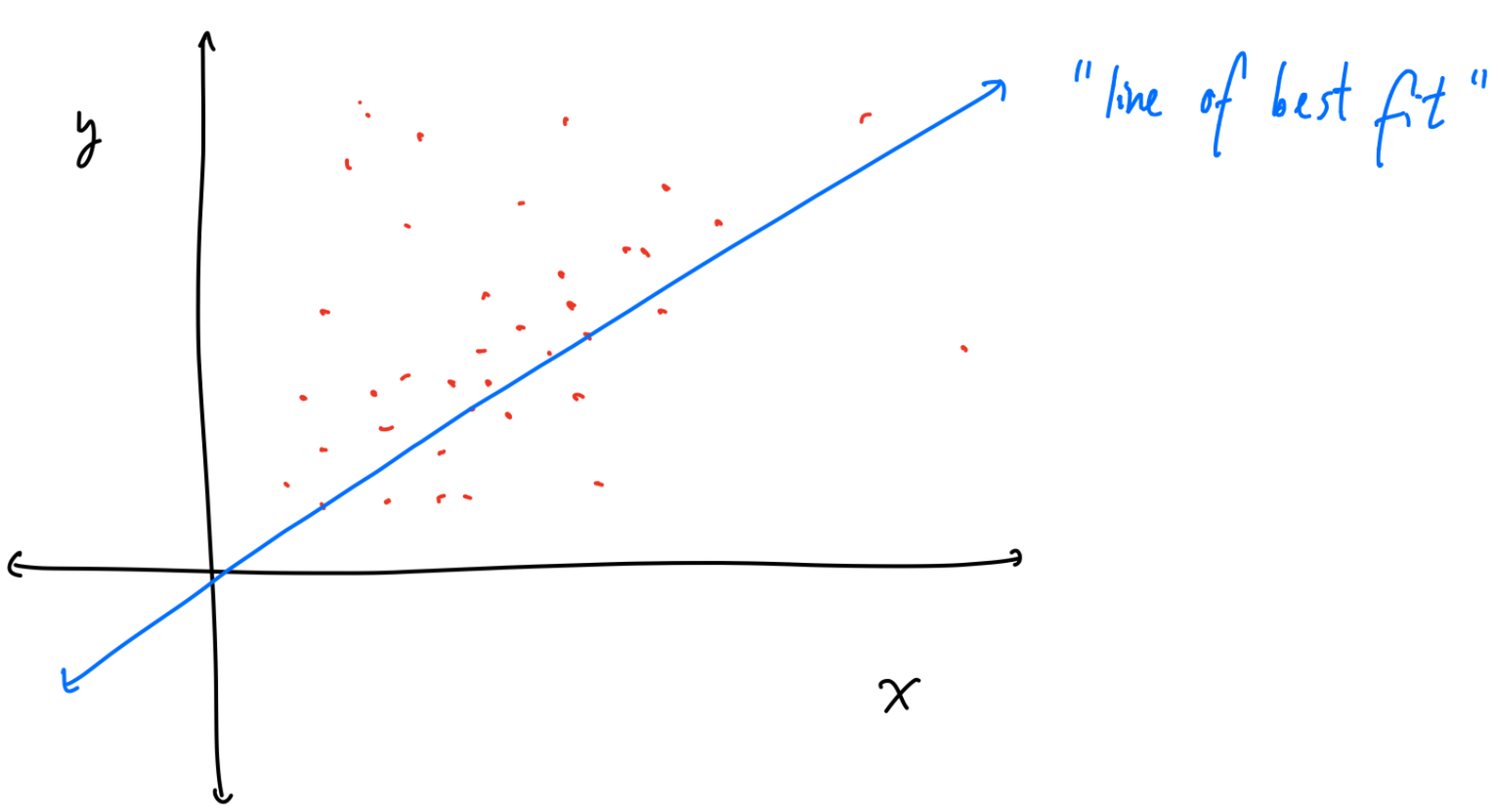
Linear Regression

Given some set of points, we'd like to try and find a **line** that effectively models the relationship between these points. Ideally, we won't find just any line – we want to find the best possible line, or the **line of best fit**.



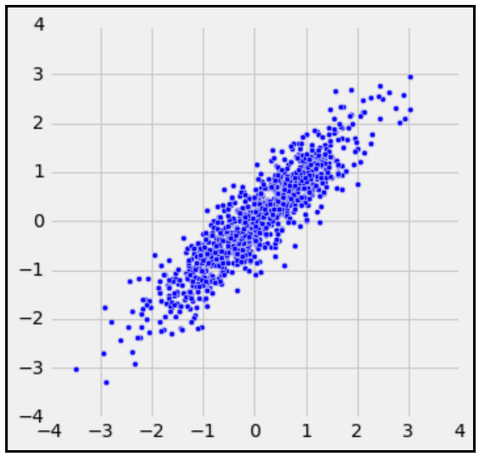
Linear Regression

Given some set of points, we'd like to try and find a **line** that effectively models the relationship between these points. Ideally, we won't find just any line – we want to find the best possible line, or the **line of best fit**.

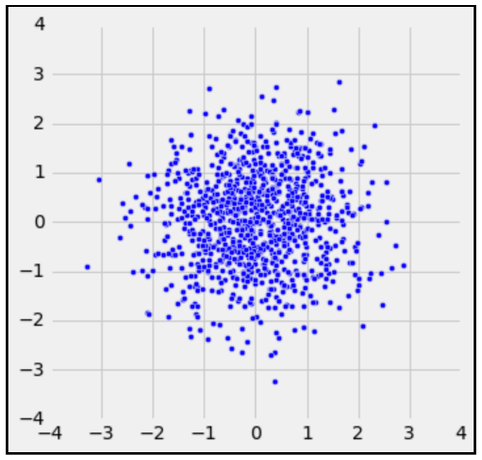


Linear Regression

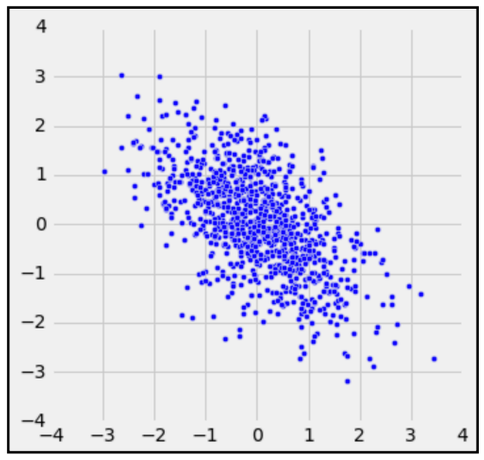
Remember, the correlation coefficient **r** is a number that measures the strength of the linear correlation between two variables.



r = 0.9



r = 0



r = -0.55

It turns out that, given that **x** and **y** are in standard units, the equation of the line of best fit is exactly:

$$y_{standard} = r \cdot x_{standard}$$

Let's derive the **line of best fit** in original units.

$$y_{standard} = r \cdot x_{standard}$$

$$\frac{y_{predicted} - avg(y)}{SD(y)} = r \cdot \frac{x - avg(x)}{SD(x)}$$

$$y_{predicted} - avg(y) = r \cdot \frac{SD(y)}{SD(x)} (x - avg(x))$$

$$y_{predicted} = r \cdot \frac{SD(y)}{SD(x)} (x - avg(x)) + avg(y)$$

$$y_{predicted} = \left(r \cdot \frac{SD(y)}{SD(x)} \right) x + \left(avg(y) - r \cdot \frac{SD(y)}{SD(x)} \cdot avg(x) \right)$$

$$m = \left(r \cdot \frac{SD(y)}{SD(x)} \right)$$

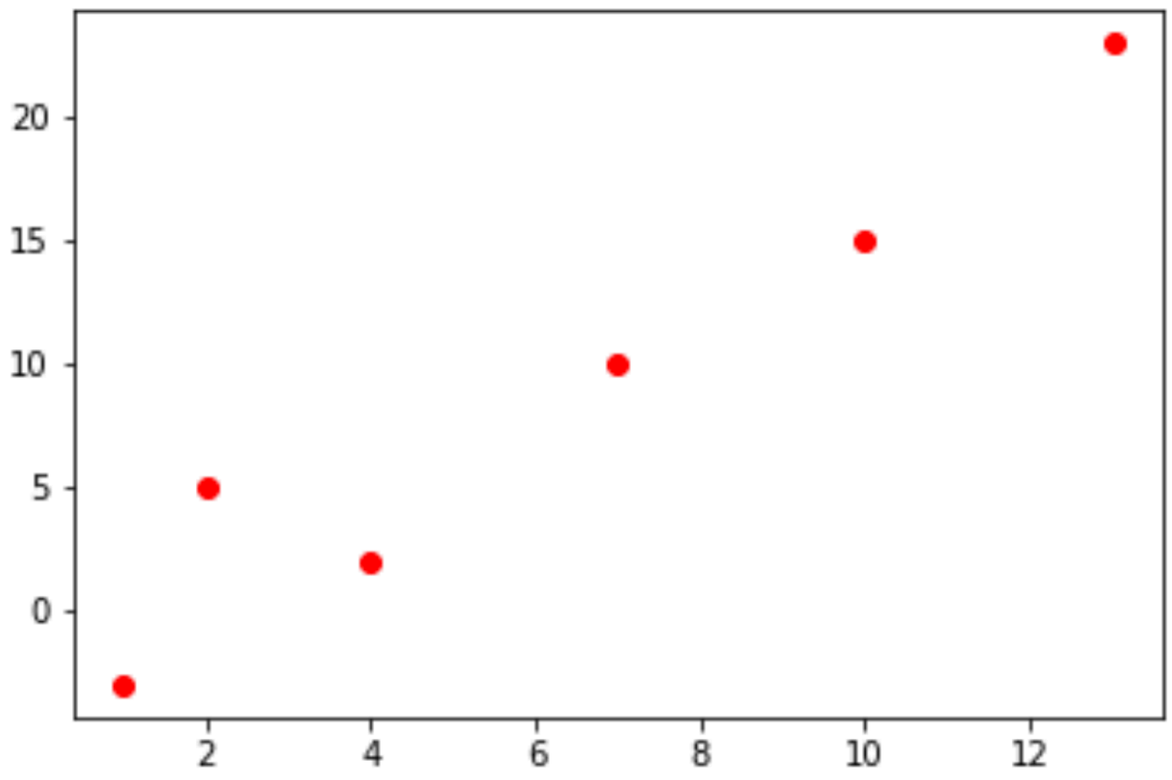
$$b = avg(y) - m \cdot avg(x)$$

$$y = mx + b$$

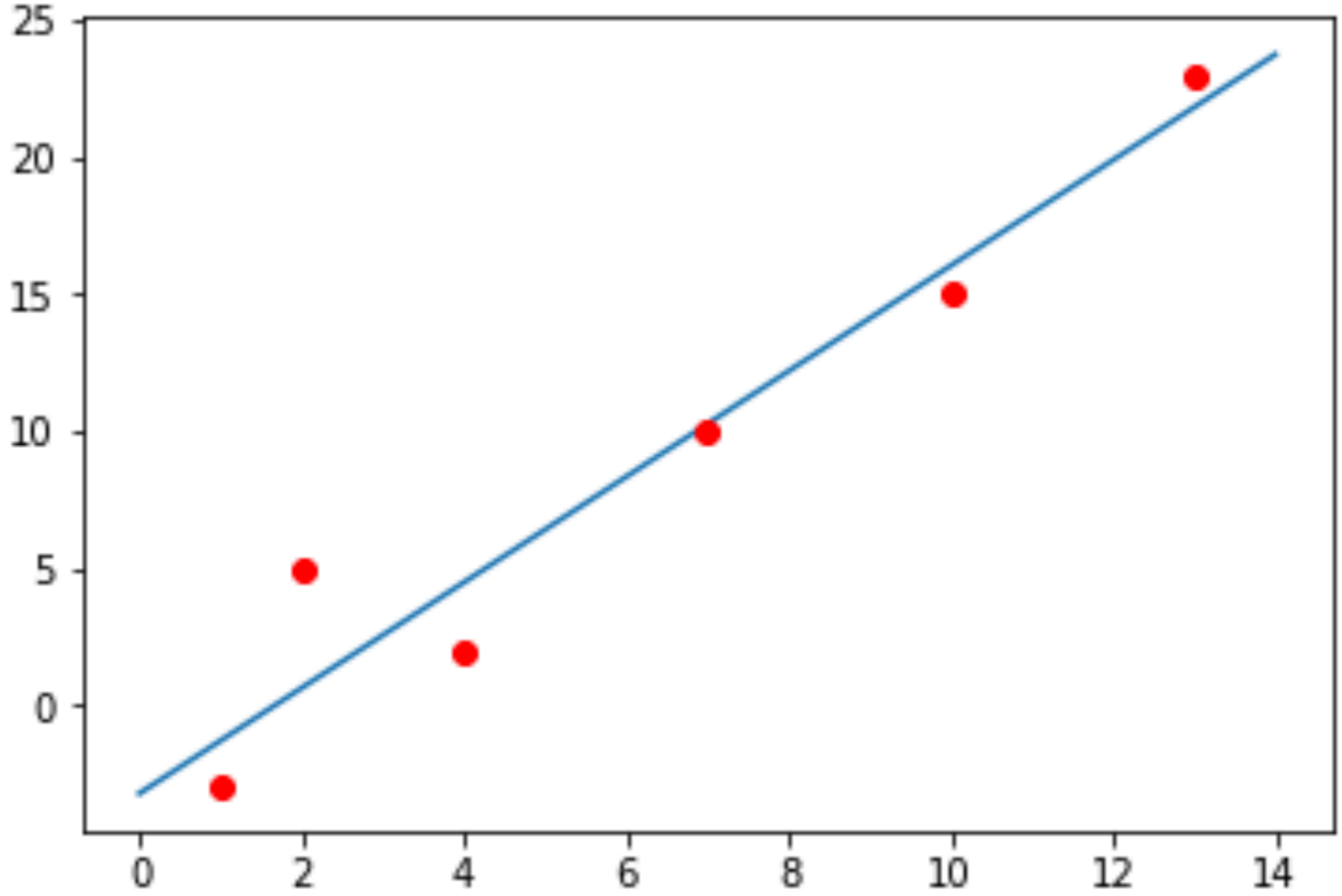
What's the Point?

Consider the following made up set of points:

x	y
1	-3
2	5
4	2
7	10
10	15
13	23



What if I wanted to predict the **y** value for **x = 6**? Or the **x** when **y = 20**?



$$y = 1.9248x - 3.2030$$

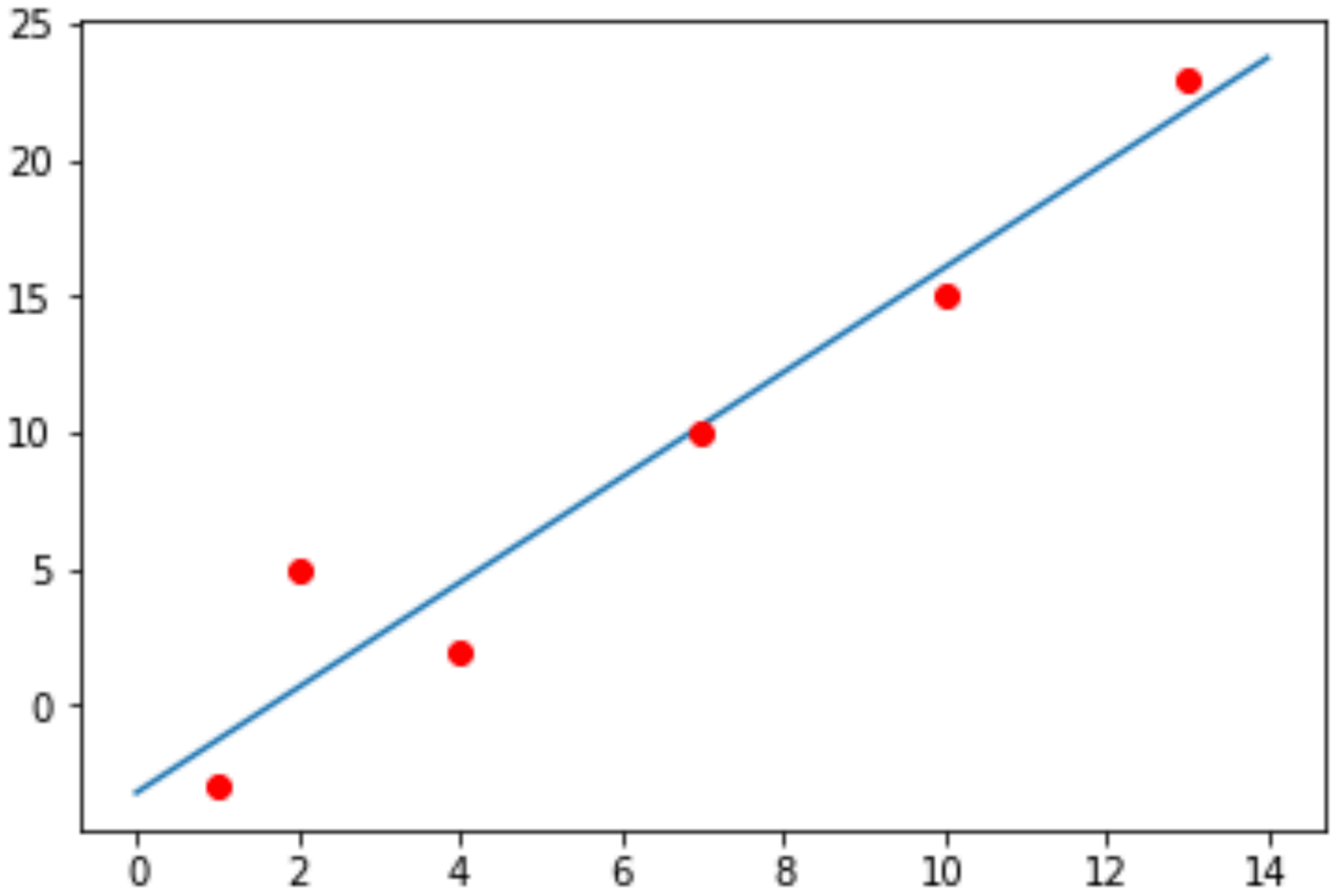
We can use our line of best fit to **predict values we don't have information on**.

For example, to predict **y** when **x = 6**, we can substitute **x = 6** into the equation of the line.

$$y_{predicted} = 1.9248(6) - 3.2030 = 8.35$$

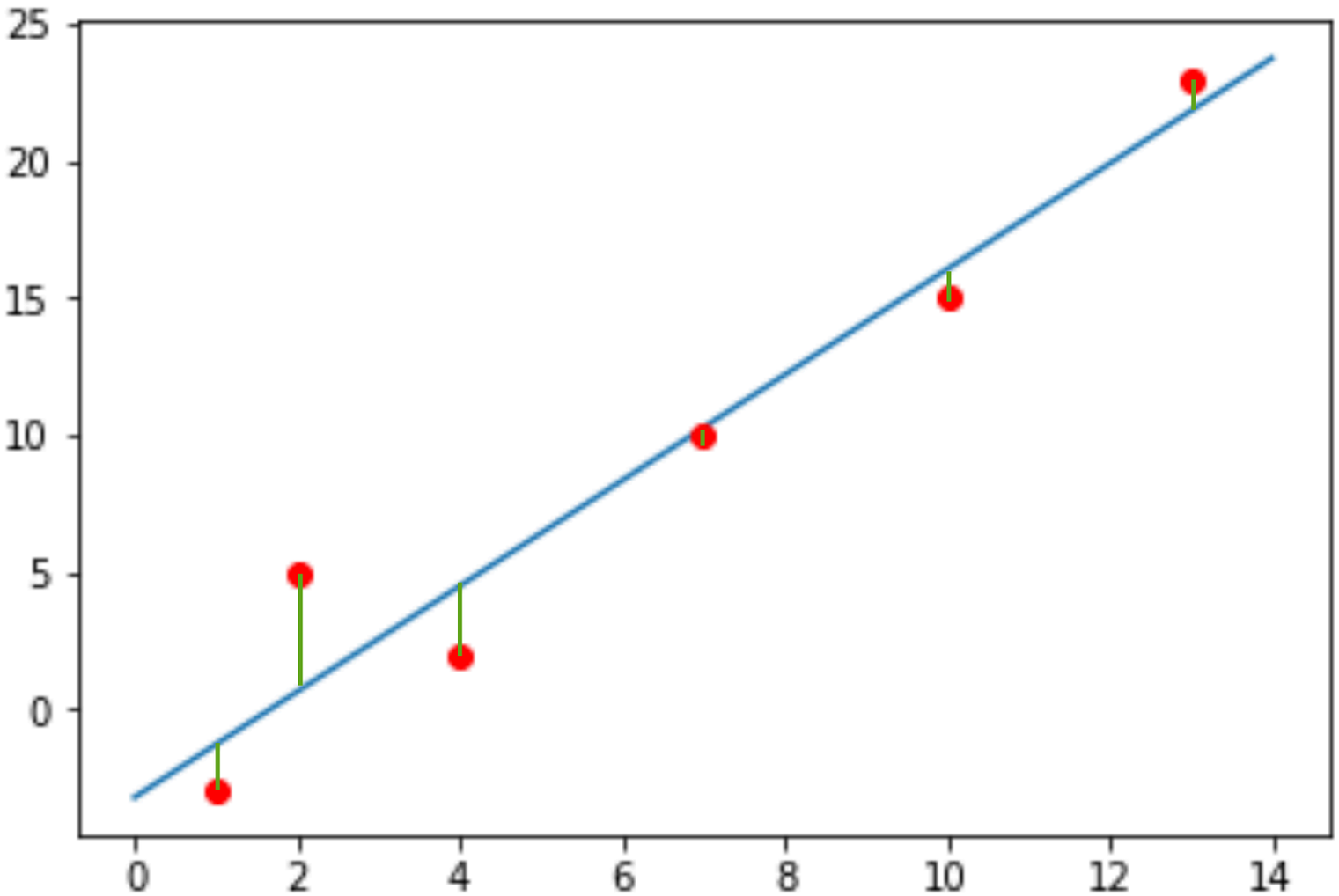
How Good is our Line?

Clearly, this line, **even though it is the best possible line that models these points**, it isn't completely accurate.



How Good is our Line?

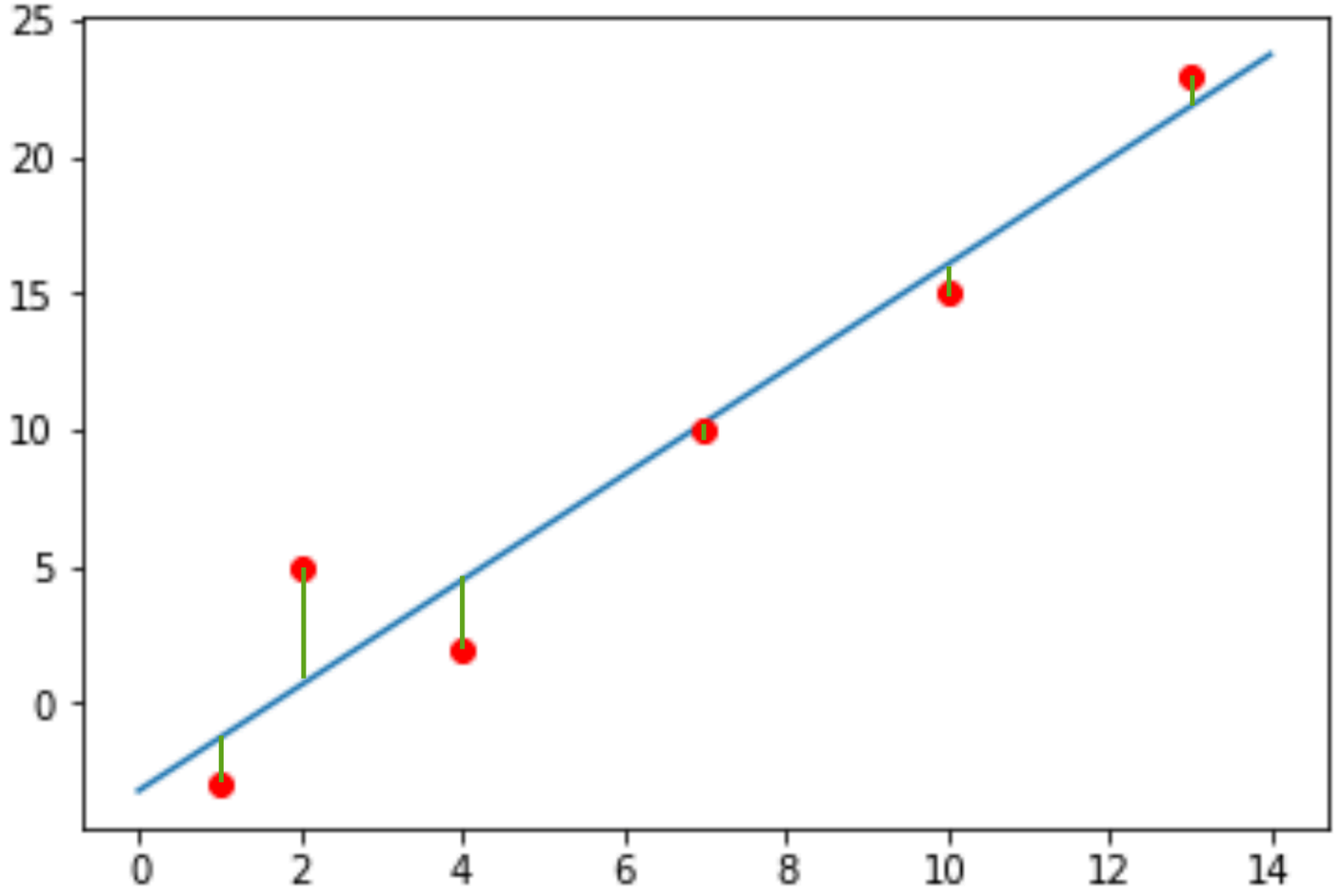
Clearly, this line, **even though it is the best possible line that models these points**, it isn't completely accurate.



We define **residuals** to be the **difference between the actual and predicted values**.

LLSE – Linear Least Squares Estimator

$$\text{residual} = y_{actual} - y_{predicted}$$



The line of best fit is the unique line that **minimizes the sum of the squares of all residuals**. We square each residual so that the negative and positive residuals don't cancel each other out. Notice that the points aren't evenly spread above and below the line, as one may think.

Bootstrapping Returns

Sometimes, we are unsure if there is a linear relationship between two variables, given some large set of data. In these cases, we can **bootstrap** our large random sample, and calculate the **regression slope** for each resample.

We can then calculate a **confidence interval** (surprise) for the slope, and check to see if **0 is in the confidence interval**.

If 0 is in the confidence interval, then there is a chance that there is no linear relationship between the two variables. However, if 0 is **not** the confidence interval, then we are fairly confident that there is a linear relationship between the two variables.