Fitting a Curve to Points via Ordinary Least Squares

by: Jabari Dash

Contents

1 What Is Curve Fitting?

1.1 What is it?

Suppose we are given a set of data points that take the form (x, y). The points appear to have a *trend* to them - such as a linear or parabolic relationship. But, at the moment, the only thing that we know about them is their coordinates. **Curve fitting** is the process of analyzing these data points, performing some operations on them, and extracting an *equation* or *curve* that explains the relationship of the points in our data set.

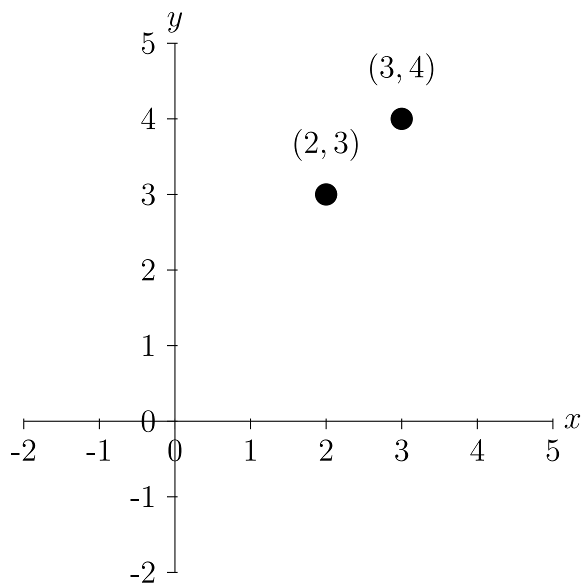

Figure 1: Two arbitrary points

There are several ways to fit a curve to points. For example, in a high school algebra class we are given questions such as "Given points $(2, 3)$ and $(3, 4)$, find a line that passes through each point." Naturally, we do the following:

$$y = mx + b \tag{1}$$

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} \tag{2}$$

$$m = \frac{3-2}{4-3} = \frac{1}{1} = 1 \tag{3}$$

$$b = y - mx \tag{4}$$

$$b = 3 - (1 \cdot 2) \tag{5}$$

$$b = 3 - 2 = 1 \tag{6}$$

$$\boxed{y = x + 1} \tag{7}$$

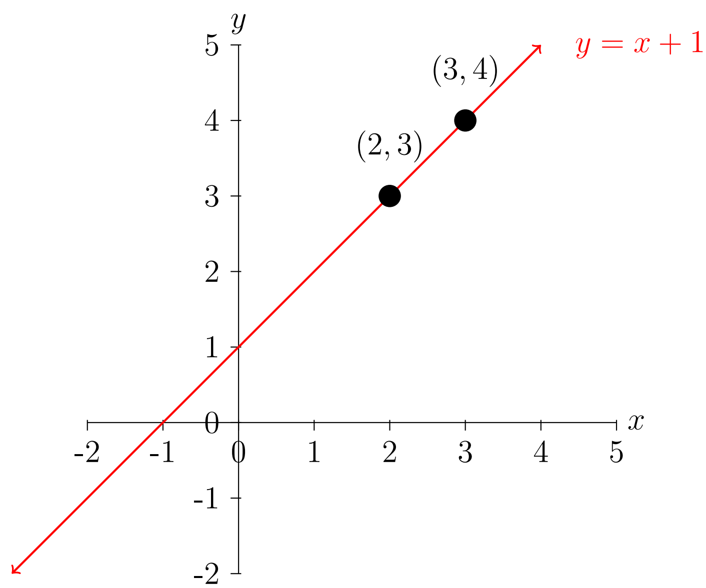

Figure 2: Linear fit to the 2 points

1.2 Why do it?

The high school approach is a perfectly valid approach. However, it has some limitations and makes some assumptions. For example, if we were asked to find a a line that intersects the points $(2, 3)$, $(3, 4)$, and $(5, 3)$, we would not be able to solve this problem. That is because the line does not exist. We would be search forever via the high school method because it *assumes* that the line exists. If the line does exist, there is a perfect match for the points, and the equation that we get back.

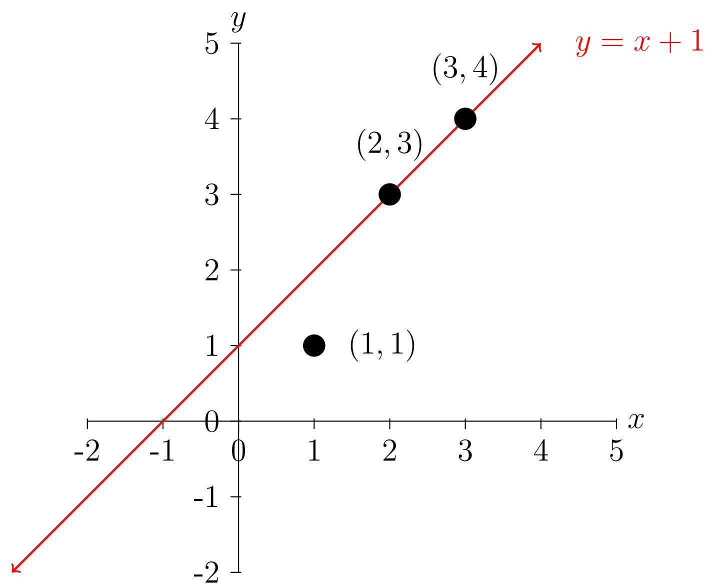

Figure 3: The 3rd point is not intersected by our line

But what do we do when we do not have a *perfect* match? We can try connecting the different pairs, and *see* which line *looks* like that best fit?

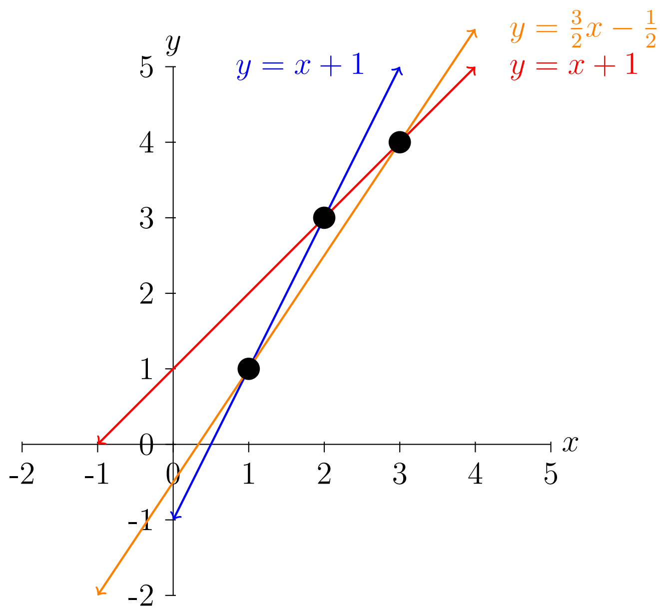

Figure 4: We find lines via the high school method

We see here that the high school method is already getting out of hand because we have 3 potential best fit lines, and we only have 3 points. If we were given a data set of ten, or one-hundred data points, it would be impossible (or would take a really long time) to find all of the lines, and choose the best one. Not to mention, that we also have not designed a method of measuring which one is the *best fit*. This is the reason why mathematicians have designed more *general* curve fitting solutions. It turns out the the *best fit* equation is $y = \frac{3}{2}x - \frac{1}{3}$ and was computed via the **Least Squares** method.

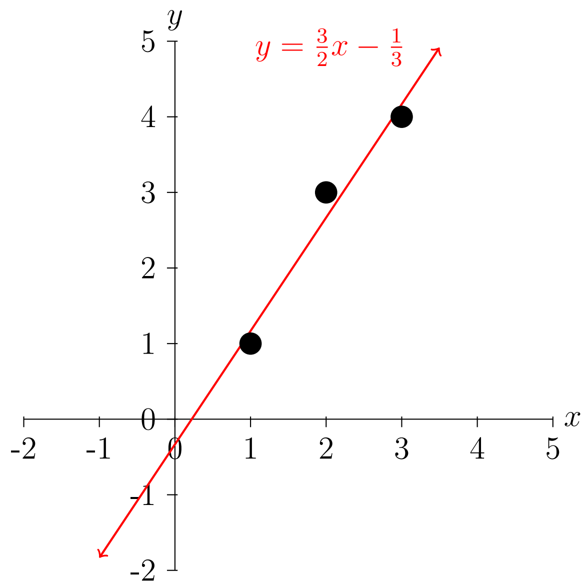

Figure 5: We use Least Squares to find a best fit line

2 What is Least Squares?

2.1 Linear Regression

2.2 Non-linear Regression

3 Linear Fit

4 Polynomial Fit

5 Exponential Fit

6 Logarithmic Fit

7 Sinusoidal Fit

8 References