How Do We Define a *Best Fit Line*?

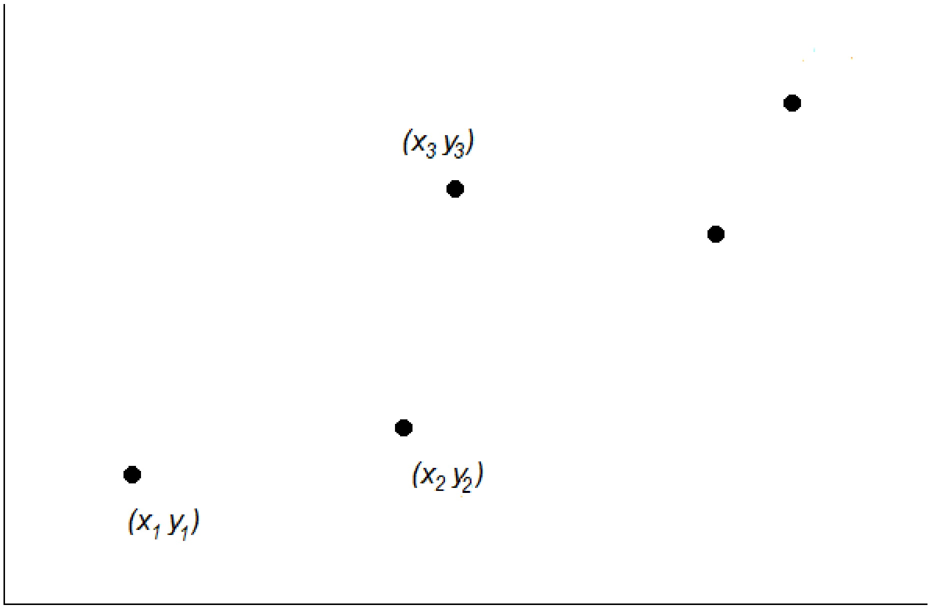
Quite often in science we design an experiment so that we measure two types of data for which we would like to establish some correlation (in the hope of either supporting or refuting our hypothesis.) A graph of the data can be useful as visual representation of the correlation. The simple *xy-scatter* plot is a popular option for graphing two sets of data, and a linear correlation is a popular expectation (i.e. we can design our experiment and the data we collect so that we expect the data to form a straight line.)

The parameters of the line formed by our data, i.e. the *slope* and *y-intercept* that define the line, are typically relevant to the analysis of our data. So we are interested in the straight line that is the best representation of the data.

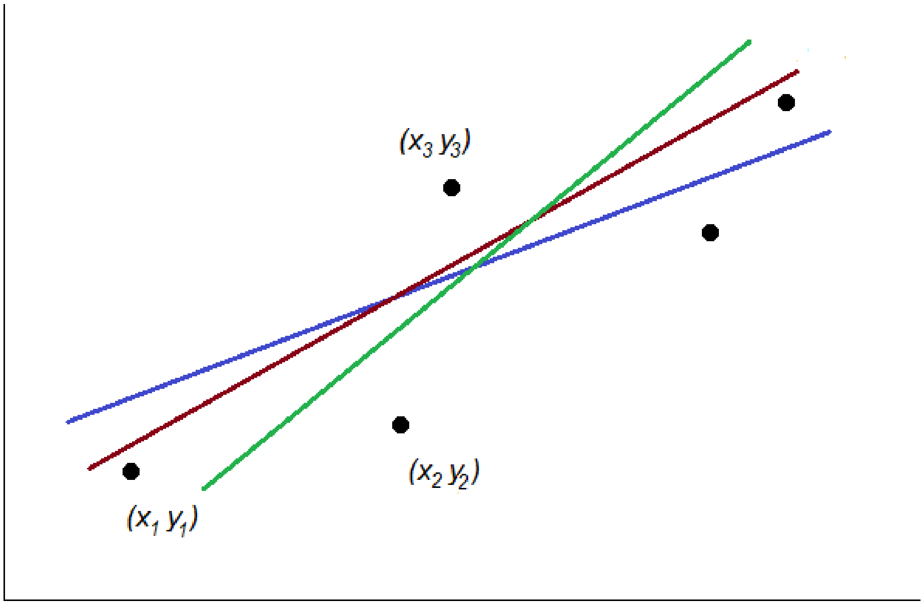
Unfortunately, real data rarely forms a perfectly straight line, so the straight line that best "fits" our data can seem a bit ambiguous. Science relies on definitions to dispel such ambiguity, and so…

we will define the *best-fit line* for the data of an *xy-scatter* plot.

A typical *xy-scatter* plot might look something like this:



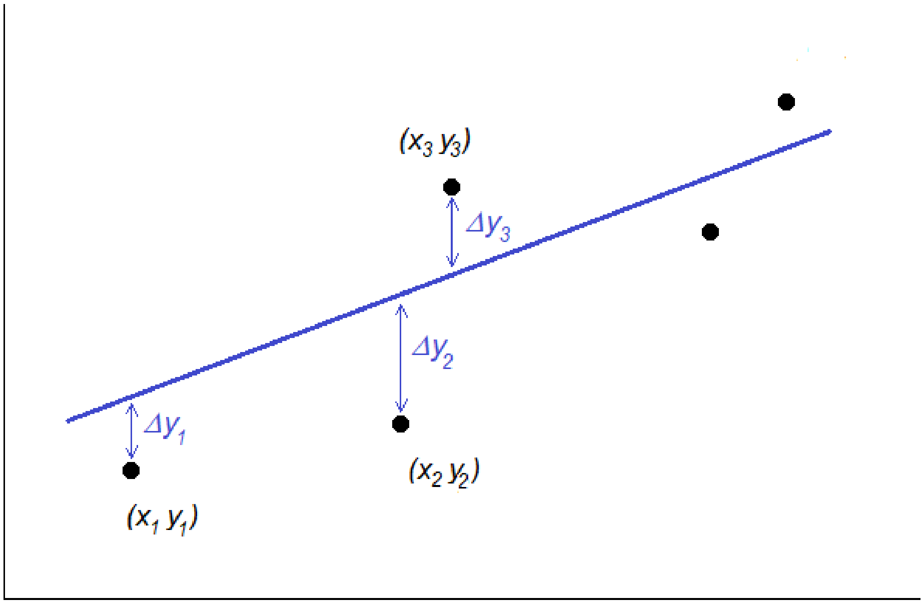
A quick and simple way to determine a *best-fit line* might be to place a straightedge on the data and use our best judgment to draw a line that seems to fit the data well. Unfortunately, this method is arbitrary… three different people might have three different opinions as to which line is "best".



This method is not *scientific*… because science requires that our methods produce results that can be reliably repeated. For scientific purposes, we need to create a definition of a *best-fit line* that we can all agree on, and then that definition can be used by anyone to reliably determine the parameters (i.e. the slope and y-intercept) of the best fit line. Using this definition, different people will always get the same result for a given set of data.

Where do we start? We have to decide what "best-fit line" means. To do this, we consider that if our data were "perfect", they would form a straight line. But we know our data cannot be perfect… so we could ask "how far are our data from a perfectly straight line?"

Consider the line drawn through the set of data here:



How far are our data from the line? To answer this, we define the vertical “gap” between each data point and the line, and we label this “gap” as Δy (which just means “difference in y”.)

Ideally, we would like our Δy gaps to each be zero, but they will not be zero because our data are not perfect. What is the next best thing?

We would like our Δy “gaps” to be *as small as possible*.

Or more specifically… we would like to find the line that creates the smallest possible Δy “gaps” with our data points. This implies that we want to calculate the gap for each data point and add the results to get the total.

But there is a technical problem with this idea: some gaps will be positive (point above the line?) and some gaps with be negative (point below the line?) If we simply add the gaps, the total could be deceptive, as positives and negatives could cancel and the result would not accurately represent how well the line “fits” the data.

There is another very subtle consideration: we have to treat our data points equally, i.e. we are not allowed to show favoritism to one point over another. In science, all data in a set are assumed to be equally valid.

For simplicity, consider just three data points. One straight line, a candidate for our “best-fit line”, might result in gaps of 1, 3, and 8 for a total of 12. Another candidate straight line might result in gaps of 4, 4, and 4… also for a total of 12. Which line is a “better fit”?

> **We decide that the line that is the best fit is not only the line that minimizes the gaps, but also the line that treats the data most equally… i.e. creates gaps that are as equal as possible.**

By this definition, our line with gaps of 4, 4, and 4, is a better “fit” than the line with gaps 1, 3, and 8. We now need a way to turn this concept into a calculation… how can we reliably calculate the gaps in a way that uses this definition?

A very clever solution: *we will* ***square*** *the measurement of each gap, and then add the squares of the gaps.* How does this express our definition?

Consider the gaps created by the two lines described above. We want to show that 4, 4, 4 is a better “fit” (i.e. lower total) than 1, 3, 8. They both sum to 12, but consider what happens when we instead calculate the sum of the *squares* of the gaps:

$$1^2 + 3^2 + 8^2 = 1 + 9 + 64 = \boldsymbol{74}$$

$$4^2 + 4^2 + 4^2 = 16 + 16 + 16 = \boldsymbol{48} \quad \textbf{\textit{a lower total!}}$$

It can be proven that the smallest sum of the squares will occur when the numbers are equal… and the further from equal the numbers get, the greater the sum of the squares will be.

We now have our definition of a *best-fit line*: **the line that best minimizes and equalizes the ∆y "gaps" for our data**. We have a way to calculate how well a line "fits" our data: by taking the sum of the squares of the ∆y "gaps".

We can now move forward with an equation to express this idea, do a bit of algebra, and derive an equation that allows us to calculate the slope and y-intercept of the *best-fit line*.

To start, we define the sum of the squares of the ∆y "gaps" for any candidate line (i.e. pick a line, any line, and we can calculate the gaps). We can abbreviate this sum as "s":

$$s = (\Delta y_1)^2 + \ (\Delta y_2)^2 + \ (\Delta y_3)^2 + \cdots$$

Where the 1, 2, 3, etc subscripts are for our first, second, third, etc data points.

We can now consider how we write ∆y for any one of our data points. It is…

the y-coordinate of the line above the data point minus the y-coordinate of the data point

Note that the x-coordinate of the point on the line is the same as the x-coordinate of our data point… because the point on the line is directly above the data point. If we write the equation of the line as:

$y = mx + b$ (*equation of our straight line… "m" is slope, "b" is y-intercept*)

then the y-coordinate of the line directly above Data Point #1 can be written as:

$$y = mx_1 + b$$

We can then write ∆y for our first data point as:

$$\Delta y_1 = mx_1 + b \ - \ y_1$$

We can write ∆y for our remaining data points the same way… the only difference is the numerical subscripts:

$$\Delta y_2 = mx_2 + b \ - \ y_2$$

$\Delta y_3 = mx_3 + b \ - \ y_3$ etc, etc…

We can now rewrite our equation for "*s*", substituting these expressions:

$$s = (mx_1 + b \ - \ y_1)^2 + \ (mx_2 + b \ - \ y_2)^2 + \ (mx_3 + b \ - \ y_3)^2 + \cdots$$

Obviously this expression is going to get very large, but fortunately we will be able to take advantage of patterns to simplify it very quickly. It is worth noting that the right side of the expression contains only our data, i.e. x_1, x_2, x_3, … and y_1, y_2, y_3…, along with *m* and *b*, the slope and intercept of the line. This provides us a clue that this equation will allow us to calculate *m* and *b* using our measured data.

But first we need to simplify the right side. If we expand the first term, i.e. square the expression in parentheses, we get:

$$(mx_1 + b - y_1)^2 = m^2{x_1}^2 + b^2 + {y_1}^2 + 2mbx_1 - 2mx_1y_1 - 2by_1$$

We can do the same for the expression for the second data point, the third data point, etc (and of course we get exactly the same expression, but with different subscripts):

$$(mx_2 + b - y_2)^2 = m^2{x_2}^2 + b^2 + {y_2}^2 + 2mbx_2 - 2mx_2y_2 - 2by_2$$

$$(mx_3 + b - y_3)^2 = m^2{x_3}^2 + b^2 + {y_3}^2 + 2mbx_3 - 2mx_3y_3 - 2by_3$$

And now we can take advantage of the pattern shown in these expressions. The right side of these expressions have the same six terms, and our equation for "s" requires that we add everything in these expressions. We can simplify this addition by grouping together "like terms", i.e.:

$$s = m^2({x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots) + (b^2 + b^2 + b^2 + \cdots) + ({y_1}^2 + {y_2}^2 + {y_3}^2 + \cdots)$$

$$+ 2mb\,(x_1 + x_2 + x_3 + \cdots) - 2m(x_1y_1 + x_2y_2 + x_3y_3 + \cdots) - 2b(y_1 + y_2 + y_3 + \cdots)$$

We can further simplify this expression by replacing each set of parentheses with one letter. To do this, we make the definitions:

$A = (x_1 + x_2 + x_3 + \cdots)$ i.e. "A" is the sum of the "x" data

$B = (y_1 + y_2 + y_3 + \cdots)$ i.e. "B" is the sum of the "y" data

$C = (x_1y_1 + x_2y_2 + x_3y_3 + \cdots)$ i.e. "C" is the sum of the product "xy" for each data point

$D = ({x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots)$ i.e. "D" is the sum of the square of each "x" measurement

$E = ({y_1}^2 + {y_2}^2 + {y_3}^2 + \cdots)$ i.e. "E" is the sum of the square of each "y" measurement

We can now use A B C D E, along with "N" to represent the number of data points, to simply the expression for "s":

$$s = m^2D + Nb^2 + E + 2mbA - 2mC - 2bB$$

This equation may look cryptic, but it's important to recognize that for any given set of data, all of the uppercase letters… *A B C D E N*… represent known values. Which means only *s m b* are unknowns. We want to find *m* and *b* for our *best-fit line*. And we can use this equation to do so!

How? We acknowledge that *m* and *b* are the slope and y-intercept of any "candidate" line, but we want to find the best candidate… the line with the smallest (or minimum) value of *s* (i.e. the line with essentially the smallest gaps.) Which presents us with a classic *optimization problem* from basic calculus.

Since "*s*" is a *function* of *m* and *b*, we can find the parameters that give us the minimum value of this function by taking the derivative of the function and setting the result equal to zero. This problem presents a slight challenge, because "*s*" is a function of two variables, *m* and *b*, but we can handle this.

How do you take a derivative of a function of two variables? Take the derivative twice… once with the first variable and once with the second variable. So we can take the derivative of "*s*" with respect to *m* (while we treat *b* as a constant…) and then we take the derivative of "*s*" with respect to *b* (while we treat *m* as a constant.)

$$\frac{ds}{dm} = 2Dm + 0 + 0 + 2Ab - 2C - 0 = 2\,(Dm + Ab - C)$$

$$\frac{ds}{db} = 0 + 2Nb + 0 + 2Am - 0 - 2B = 2\,(Nb + Am - B)$$

We can now set each of these equal to zero, and divide both sides by two:

$$2\,(Dm + Ab - C) = 0 \qquad or \qquad Dm + Ab - C = 0$$

$$2\,(Nb + Am - B) = 0 \qquad or \qquad Nb + Am - B = 0$$

We now have two equations and two unknowns: *m* and *b*… the slope and y-intercept of the line which gives us the minimum value of "*s*"… that is, the slope and y-intercept of, by definition, *the best-fit line*.

A bit of simple algebra allows to solve first for *m*, and then for *b*. We rearrange each of the two equations above:

$$b = \frac{C-Dm}{A} \qquad and \qquad b = \frac{B-Am}{N}$$

Set these equal:

$$\frac{C-Dm}{A} = \frac{B-Am}{N} \qquad or \qquad N(C - Dm) = A(B - Am)$$

Then:

$$NC - NDm \;=\; AB - \; A^2m$$

$$NC - AB \;=\; NDm - \; A^2m$$

Or:
$$m = \frac{NC - AB}{ND - A^2}$$

Following the same process to solve for b, we start with the two original equations and again rearrange:

$$m = \frac{C - Ab}{D} \qquad and \qquad m = \frac{B - Nb}{A}$$

Set these equal:

$$\frac{C - Ab}{D} = \frac{B - Nb}{A} \qquad or \qquad A(C - Ab) = D(B - \; Nb)$$

Then:

$$AC - A^2b \;=\; DB - \; DNb$$

$$NDb - \; A^2b \;= BD - AC$$

Or:
$$b = \frac{BD - AC}{ND - A^2}$$