Situation: Line of Best Fit

Prepared at the University of Georgia in Dr. Wilson's EMAT 6500 Class
July 10, 2013 - Sarah Major

Prompt:

A teacher in a high school Algebra class has just explained how to find the equation of a line given two data points. A student then asks, "What if you're given more than two data points but there is no line that can be drawn through all of them? How do you find the best fit line for the points?"

Commentary:

First of all, though not explicitly stated in the prompt, we assume that the student is referring to the "line of best fit". This is the terminology most associated with the process the student is asking about. Though there may be other interpretations for his/her question, they all can be connected to the process of finding the line of best fit.

Secondly, the terminology "line of best fit" can be interpreted many different ways. Some may believe the line of best fit may be was visually seems to fit a set of points while others believe there must be a mathematical basis behind it that shows it is the line of best fit. **Focus 1** attempts to sort through these discrepancies and show that though there are different interpretations associated with the terminology, there is a distinct process at play that branch out to functions of higher degrees.

There are multiple methods of finding the line of best fit. This situation focuses specific on methods involved from algebra, statistics, and linear algebra. Each of these perspectives will be analyzed within the foci.

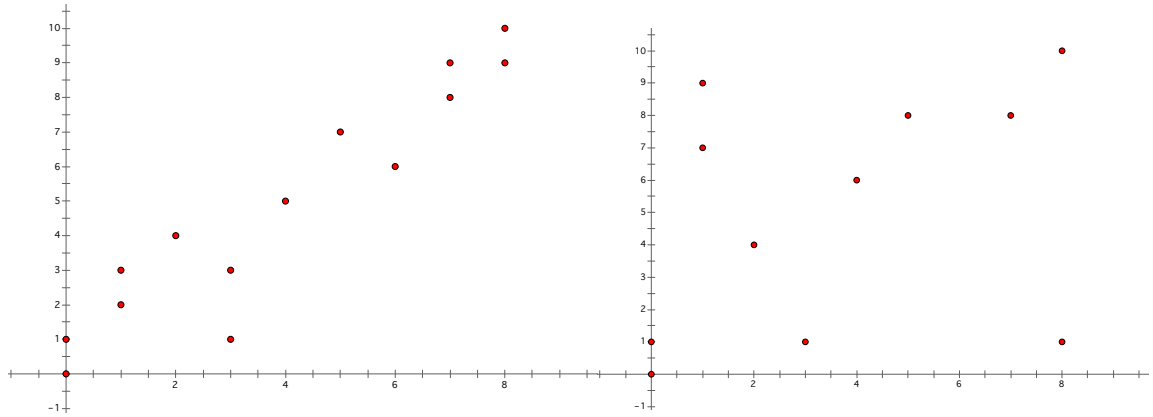
Mathematical Foci:

Mathematical Focus 1

The terminology "line of best fit" may have different implications or definitions depending on the interpretation.

In simplest terms, the line of best fit is a line that "best" represents a set of points. This line may pass through all, some, or none of the points but attempts to minimize the displacement from the points to the line of best fit. However, a line may not be the best function to use to minimize the

displacements. In fact, if the points do not appear linear at all, a line is not the function that would best represent the trend of the points. This is why the line of best fit is often called the “trend line”. The graphs below compare a set of points that appear to have a linear trend to a set of points who do not appear to have a linear trend.



In a broader sense, finding the line of best fit is a form of curve fitting, which is the process of finding a mathematical function that fits a set of data points. In this process, a curve will be an exact fit for a set of points if the degree of the equation given by the curve is exactly one less than the number of points. For instance, a line is an exact fit through any two points, a parabola is an exact fit through a set of three points, and so forth. However, though larger degree polynomials may provide an exact fit for a set of points, this does not mean they are the right fit for the process involved. If a set of points does appear to have a linear trend, the approximation of the line of best fit will provide easier calculations and be more visually simple than higher degree functions.

Along with curve fitting comes the choice of using algebraic fitting or geometric fitting. Algebraic fitting attempts to minimize the displacements of the points from the curve, and geometric fitting tries to provide the best visual fit by minimizing orthogonal distances from the points to the curve. While algebraic fitting is usually the preferred method because of its ease compared to the possible complications associated with the geometric fitting calculations, either method can be used for a line because of the simplicity of a line’s function.

Mathematical Focus 2

Simple algebraic techniques can be utilized to approximate the equation of a line the passes through at least two of the data points and near the other points.

The simplest way to find the line of best fit is to simply draw a line that seems to fit the data well that passes through two points. Once two points are chosen, this method becomes the normal calculation of a line:

1. Find the slope m. $m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$ *or* $\frac{y_1 - y_2}{x_1 - x_2}$. Either calculation can be used as long as the same order is used in the numerator and the denominator (i.e. the order of the coordinates used from each point is the same).
2. The calculation of m and the coordinates of one of the points can be used to find the y-intercept. Simply plus the latter values into the equation $y = mx + b$ (slope-intercept form) where x and y are the coordinates of the point, m is the slope, and b is the y-intercept.
3. Plug in the calculations for m and b into the equation $y = mx + b$ to find the line of best fit.

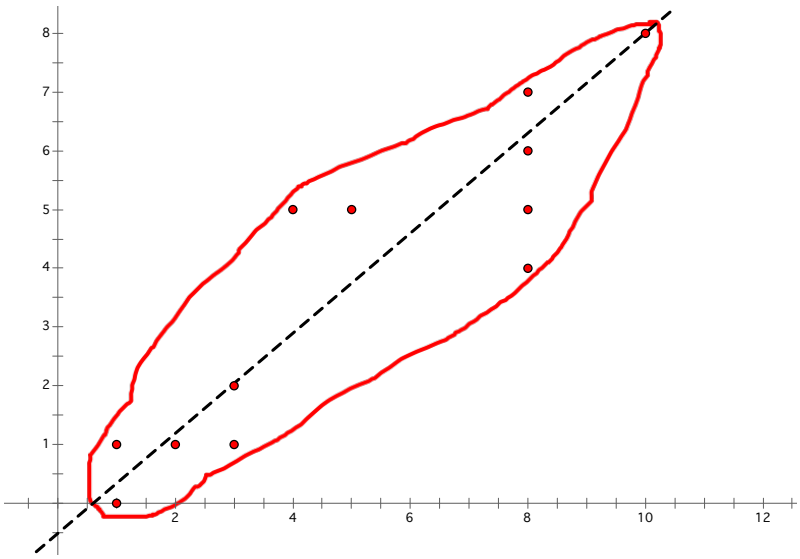
The equation of the line can also be found by plugging in the value of m into the equation $y - y_1 = m(x - x_z)$, called point-slope form), which eliminates the need for the calculation of the y-intercept.

Though this method is ambiguous, it produces the line of best fit to the eye of the beholder.

Mathematical Focus 3

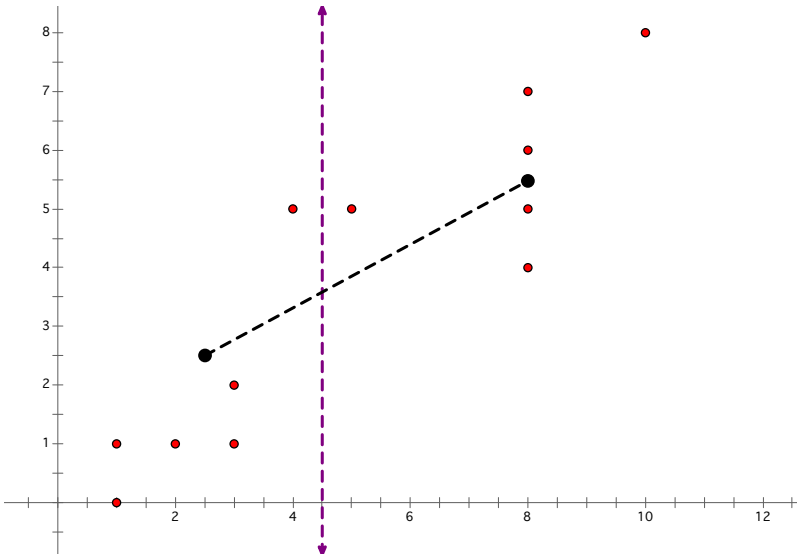
Other methods of simply eyeing the line can be used to approximate a reasonable equation for the line of best fit.

The first method manipulates the area around the points to create the line of best fit and is commonly called the “area method”. First, draw a smooth curve that encloses all of the points making sure the draw it as close to the data points as possible. Next, draw a line that evenly divides the area of the figure created. In other words, “bisect” the area of the figure making sure to go in the direction that mimics the linear trend. In this method, the line does not actually have to go through any of the points but may depending on how the figure is drawn. Below is an example of this process:



One of the methods from **Focus 1** can then be utilized to find the equation of the line of best fit that has been drawn. In the above figure, the line is $y = {}^{6}/_{7}x + {}^{1}/_{2}$.

Another method involves dividing the points into two groups, each with the same number of points, by drawing a vertical line between them. If the scatter plot has an odd number of points, attempt the draw the vertical line through the point that appears in the center of the two groups. Then, somehow mark; either with a point, an x, etc.; where you think the centers of the two groups should be. Draw a line through these two centers, and this will be the line of best fit. Below is a figure that demonstrates this method using the points from the previous method:



Like the previous example, the methods from **Focus 1** can be used to find the equation. In this example, the line of best fit is $y = {}^{27}/_{50}\,x + {}^{23}/_{20}$.

There are many variations of these methods in existence. For instance, some teachers may require students to draw an oval around all of the points instead of any smooth shape so that a line dividing the figure in half can be more easily drawn.

Once again, these methods are a bit ambiguous, especially since two different equations for the line of best fit were found for the same set of points. Though they are easy methods, teachers may run into complications explaining their adequacy, especially when dealing with actual data points and prediction.

Mathematical Focus 4

Statisticians find the line of best fit for a set of data points by finding the least squares regression line.

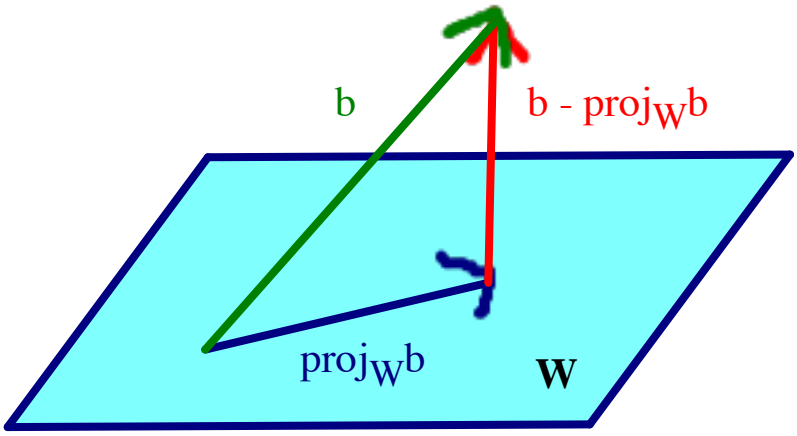
Typically, before calculations can be done to find the best fit line, a correlation coefficient is found to see if the data appears to be approximately linear.

Mathematical Focus 5

Linear algebra can also be used to find the least squares regression line given a set of data points.

In this method, we assume that the matrix equation $Ax = b$ has no solution. If this were not true, the vector b would already lie in the column space of A, so there would be no need to project it into the column space because it would produce the same vector.

However, if indeed the matrix equation above has no solution, the closest solution can be found by projecting b into the column space of A, which we will call W. Visually, this is what we're trying to do:



Since $b - proj_W b$ is orthogonal to W, it is in the null space of A^T. Since we are trying to find x, which is the closest vector and thus the projection, then we have:

$$A^T(b - Ax) = 0 \Rightarrow A^T Ax = A^T b$$

We can then find the least-squares solution, which is also the projection of b onto W by plugging into $A^T Ax = A^T b$ and row reducing. In this case, we have:

$$A = [x^0, x^1] \text{ and } b = y,$$

or:

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \text{ and } b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Finding the transpose of A simply involves converting the columns to rows, or making the first column the first row and the second column the second row:

$$A^T = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_4 \end{bmatrix}$$

We then compute the matrices $A^T A$, which will result in a 2×2 matrix, and $A^T b$, which will result in a 2×1 matrix. We align these two together in an augmented matrix and row reduce until we get something of the form:

$$\left[\begin{array}{cc|c} 1 & 0 & a \\ 0 & 1 & b \end{array}\right]$$

We then plug the corresponding values for a and b to find the equation of the best fit line, which is $y = a + bx$, or the projection of b onto W.

Mathematical Focus 6

Simple forms of technology, including an average graphing calculator or Excel, can calculate the line of best fit for a set of data points.