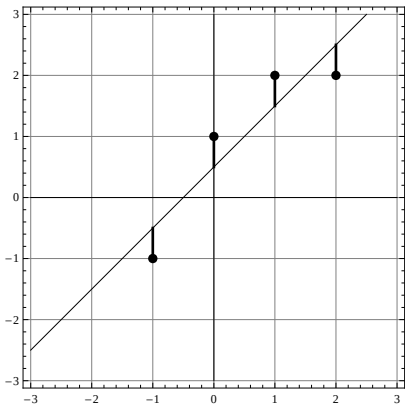
Math 2130 Workshop: Regression Lines

Suppose you have a collection of real world data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. In some circumstances, it is useful to find a line $y = mx + b$ to model these data points. If the points already form a line, figuring out the equation of that line is simple. If the points are completely random, it won't be useful at all to model them with a line. But if the points are close to forming a line, we might be interested in computing the closest fit line (called the regression line or least squares line).

1) The *least squares line* minimizes the sum of the squares of the vertical distances from the points to the line. On the grid below, plot the points $(-1, -1), (0, 1), (1, 2)$ and $(2, 2)$. Draw a straight line going roughly in the same direction as the points ($y = \frac{3}{2}x$, for example) and vertical lines connecting the points to the line. Our goal is to minimize the sum of the squares of the lengths of these vertical lines[1].



2) If our line is given by $y = mx + b$, write out an expression for the sum of the squares of these vertical distances as a function of the m and b we choose. Call this $E(m, b)$.

$$E(m, b) = (-m + b + 1)^2 + (b - 1)^2 + (m + b + 2)^2 + (2m + b - 2)^2.$$

3) Find the critical point of E.

$$E_m(m, b) = 6m + 2b - 7$$
$$E_b(m, b) = 2m + 4b - 4$$

These are both 0 when $m = 1$ and $b = \frac{1}{2}$.

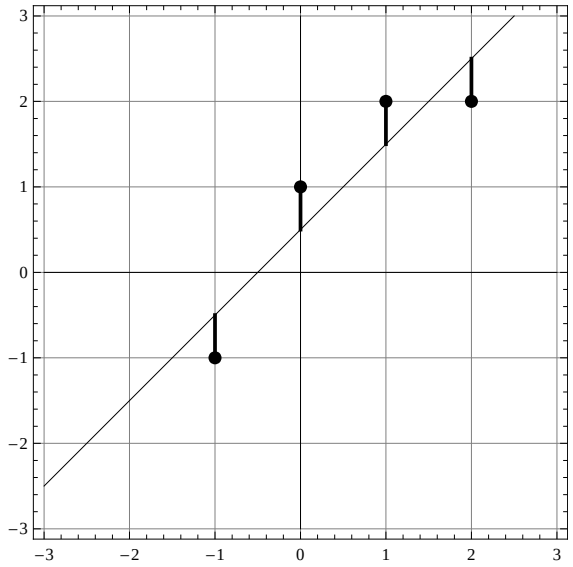
[1]Why the sum of the squares of the vertical distances and not the sum of the vertical distances? The sum of the distances gives multiple solutions, whereas the sum of the squares of the distances has a unique solution, with a relatively nice formula. Meanwhile, the sum of the squares penalizes large distances significantly more than small distances, which is usually desirable. Using the sum of squares formula also allows us to interpret the best fit line as a projection of vectors, as we will see later in this assignment.

4) Use the second derivative test to show that the critical point is a local minimum.

$$E_{mm}(m,b) = 6 \qquad E_{mb}(m,b) = 2 \qquad E_{bb}(m,b) = 4$$

These are constant. Note that $E_{mm} > 0$ and $E_{mm}E_{bb} - E_{mb}^2 = 20 > 0$, so the critical point is a local minimum.

5) Draw in the data points again and the line $y = mx + b$ for the m and b you found.



6) Consider two vectors: the vector of y values of the data points $\vec{v} = (-1, 1, 2, 2)$ and the vector of the corresponding points on the regression line:

$$\vec{w} = (m(-1) + b, m(0) + b, m(1) + b, m(2) + b),$$

For the m and b you computed. Show that $\vec{w}$ is the projection of $\vec{v}$ onto $\vec{w}$.

$$\vec{w} = \left(-\frac{1}{2}, \frac{1}{2}, \frac{3}{2}, \frac{5}{2}\right).$$

The projection of $\vec{v}$ onto $\vec{w}$ is:

$$\left(\frac{\vec{v} \cdot \vec{w}}{||\vec{w}||^2}\right)\vec{w} = \frac{9}{3^2}\vec{w} = \vec{w}.$$

Appreciate that you could, in principle, work out the mathematics behind finding many other curves to model data. For instance, if you wanted to, you could find the parameters of a sinusoid function (amplitude, wavelength, displacement) to fit a collection of data. This technique is used to break sound waves up into their pure sinusoidal components.