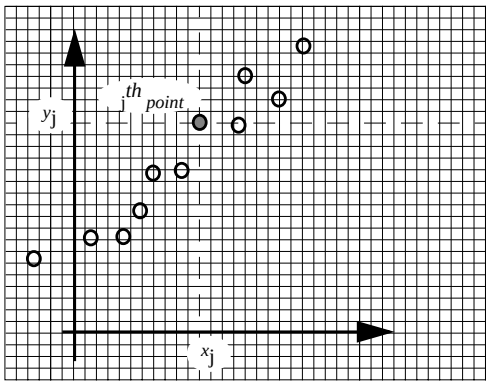
1.105 Solid Mechanics Laboratory

Least Squares Fit of Straight Line to Data

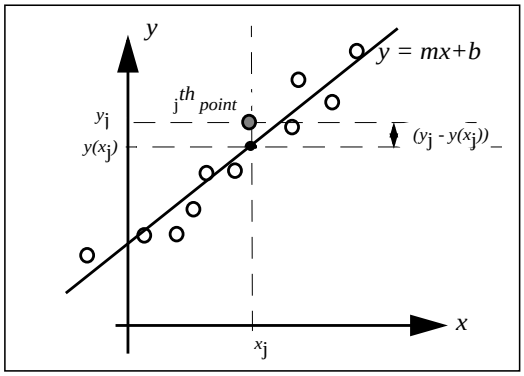
We start with a data set of n points, x_j, y_j, through which we wish to fit a straight line

$$y = mx + b.$$

In the figure, we show 11 pairs of x,y values, x_j,y_j where n is the number of points, $n = 11$. A ***jth*** point is shown as a shaded circle.



Say we try to fit a line by eye; we might draw a line as displayed in the next graph.



The aim now is to set m, the slope of our "best fit" line and b, its intercept with the y axis. The criteria we use is to minimize the "least square error" of the y coordinate. (This is not the only criterion that might be usefully applied. Can you think of another?) That is we seek to choose m and b to minimize the quantity

$$\text{Error} = \sum_{j=1}^{n} [y(x_j) - y_j]^2$$

We can imagine moving the line around to minimize this sum. That's in effect what is going through my mind as I "eyeball" a best fit line. You might, hold the slope, m, constant and slide the line up and down until it looks good. Or you might pin the line at its y intercept, b, at $x = 0$ and rotate the line around until it looks even better. Or, better yet, we can rely upon the differential calculus and set the partial derivatives of this Error sum with respect to *both* m and b to zero - they are independent variables - in order to find their values exactly. This we do now. We have:

$$\frac{\partial}{\partial m}\text{Error} = \sum_{j=1}^{n} 2 \cdot [mx_j + b - y_j] \cdot x_j = 0 \qquad \text{and} \qquad \frac{\partial}{\partial b}\text{Error} = \sum_{j=1}^{n} 2 \cdot [mx_j + b - y_j] \cdot 1 = 0$$

Given the n pairs of points x_j,y_j, these can be taken as two linear equations for determining the slope and intercept. Rewriting, canceling the common factor, 2, we have

$$\left(\sum_{j=1}^{n} x_j^2\right) \cdot m + \left(\sum_{j=1}^{n} x_j\right) \cdot b = \sum_{j=1}^{n} x_j \cdot y_j$$

and

$$\left(\sum_{j=1}^{n} x_j\right) \cdot m + \left(\sum_{j=1}^{n} 1\right) \cdot b = \sum_{j=1}^{n} y_j$$

Dividing all by *n*, the number of points, and noting that the sum of *1*, *n* times is just equal to *n*, we have two linear equations for the unknowns *m* and *b*. The coefficients appearing in these two are set by the values of the data point pairs, x_j, y_j.

$$\frac{1}{n}\sum_{j=1}^{n} x_j^2 \cdot m + \frac{1}{n}\sum_{j=1}^{n} x_j \cdot b = \frac{1}{n}\sum_{j=1}^{n} x_j \cdot y_j$$

and

$$\frac{1}{n}\sum_{j=1}^{n} x_j \cdot m + b = \frac{1}{n}\cdot\sum_{j=1}^{n} y_j$$

The solution is:

$$b = \frac{\left[\frac{1}{n}\sum_{j=1}^{n} y_j\right]\cdot\left[\frac{1}{n}\sum_{j=1}^{n} x_j^2\right] - \left[\frac{1}{n}\sum_{j=1}^{n} x_j\right]\cdot\left[\frac{1}{n}\sum_{j=1}^{n} x_j\cdot y_j\right]}{\left[\frac{1}{n}\sum_{j=1}^{n} x_j^2\right] - \left[\frac{1}{n}\sum_{j=1}^{n} x_j\right]^2}$$

and

$$m = \frac{\left[\frac{1}{n}\sum_{j=1}^{n} x_j\cdot y_j\right] - \left[\frac{1}{n}\sum_{j=1}^{n} x_j\right]\cdot\left[\frac{1}{n}\sum_{j=1}^{n} y_j\right]}{\left[\frac{1}{n}\sum_{j=1}^{n} x_j^2\right] - \left[\frac{1}{n}\sum_{j=1}^{n} x_j\right]^2}$$

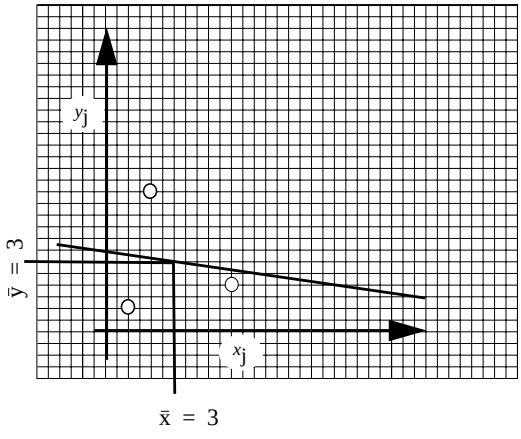
or, letting $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ $\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ $\overline{xy} = \frac{1}{n}\sum_{j=1}^{n} x_j\cdot y_j$ $\overline{x^2} = \frac{1}{n}\sum_{j=1}^{n} x_j^2$ recognizing the first sum is just the mean of x_j, the second, the mean of y_j, etc., we have more simply:

$$b = \frac{\bar{y}\cdot\overline{x^2} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - [\bar{x}]^2} \quad \text{and} \quad m = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - [\bar{x}]^2}$$

Example 1.

Given the three pairs of points:

x	y
1	1
6	2
2	6



$n = 3$
and
$\bar{x} = 3$
$\bar{y} = 3$
$\overline{xy} = 3$
$\overline{x^2} = 3$

The "averages" compute to those shown at the right. With these we find

$b = 3.43$ and $m = -0.143$

The best fit line is shown.

You might have expected another result, e.g., a line through the origin, passing through the point 1,1. But consider the criterion we applied: The error we minimize is proportional to the vertical distance between the data point and the line. Try computing the error of the best fit line and compare it with a line through 0,0 and 1,1.

We note that the best fit line goes through a point defined by the mean of x_j and y_j. You can verify this, in general, checking to see if, when you put $x = \bar{x}$, you obtain $\bar{y}$ from $y = mx + b$.

Try adding another point or two to the set of three given. E.g., the points 0,0 and/or 6,6. What happens to the best fit line?