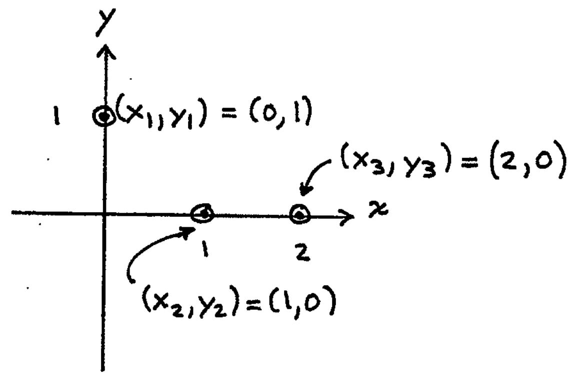
ex: Find the linear regression fit for the following data:



$$f(\vec{a}, x) \equiv a_1 x + a_2$$ Function we use to fit the data.

Define the total squared error:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where $\hat{y}_i \equiv f(\vec{a}, x_i)$ = Estimate produced by function (i.e., line) we chose to fit data.

$$\hat{y}_i = a_1 x_i + a_2$$

We now find a_1 and a_2 to minimize SSE.

The solution, from calculus, is to set derivatives to zero:

$$\frac{\partial}{\partial a_1} SSE = 0 \quad \text{and} \quad \frac{\partial}{\partial a_2} SSE = 0$$

Now we do the math.

$$0 = \frac{\partial}{\partial a_1} SSE = \frac{\partial}{\partial a_1} \sum_{i=1}^{n} [y_i - (a_1 x_i + a_2)]^2$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial a_1} [y_i - (a_1 x_i + a_2)]^2$$

$$= \sum_{i=1}^{n} 2[y_i - (a_1 x_i + a_2)] \frac{\partial}{\partial a_1} [y_i - (a_1 x_i + a_2)]$$

$$= \sum_{i=1}^{n} 2[y_i - (a_1 x_i + a_2)] (-x_i)$$

$$0 = \sum_{i=1}^{n} (-\cancel{2} x_i y_i) + \sum_{i=1}^{n} \cancel{2} a_1 x_i^2 + \sum_{i=1}^{n} \cancel{2} a_2 x_i$$

or

$$a_1 \sum_{i=1}^{n} x_i^2 + a_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$0 = \frac{\partial}{\partial a_2} SSE = \frac{\partial}{\partial a_2} \sum_{i=1}^{n} [y_i - (a_1 x_i + a_2)]^2$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial a_2} [y_i - (a_1 x_i + a_2)]^2$$

$$= \sum_{i=1}^{n} 2[y_i - (a_1 x_i + a_2)] \frac{\partial}{\partial a_2} [y_i - (a_1 x_i + a_2)]$$

$$= \sum_{i=1}^{n} 2[y_i - (a_1 x_i + a_2)] \cdot (-1)$$

$$0 = \sum_{i=1}^{n} (-\cancel{2} y_i) + \sum_{i=1}^{n} \cancel{2} a_1 x_i + \sum_{i=1}^{n} \cancel{2} a_2$$

or

$$a_1 \sum_{i=1}^{n} x_i + a_2 \sum_{i=1}^{n} 1 = \sum_{i=1}^{n} y_i$$

Using the data, (x_i, y_i)'s, we can calculate all the summation terms and think of them as constants.

Solving the resulting simultaneous eq'ns gives:

$$a_1 = \frac{\left(\sum_{i=1}^{n} 1\right)\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left(\sum_{i=1}^{n} 1\right)\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i\right)}$$

$$a_2 = \frac{\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} 1}$$

Note: $\sum_{i=1}^{n} 1 = n$

Another form for the sol'n (from Walpole, et al. <u>Probability and Statistics for Engineers and Scientists</u>, Prentice Hall, 2002) is

$$a_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad a_2 = \bar{y} - a_1 \bar{x}$$

where $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$

$\bar{y} \equiv \frac{1}{n} \sum_{i=1}^{n} y_i$

For the example at hand, we have

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}(0+1+2) = \frac{3}{3} = 1$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{3}(1,0,0) = \frac{1}{3}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = (0-1)(1-\tfrac{1}{3}) + \overset{0}{\cancel{(1-1)}}(0-\tfrac{1}{3})$$

$$+ (2-1)(0-\tfrac{1}{3})$$

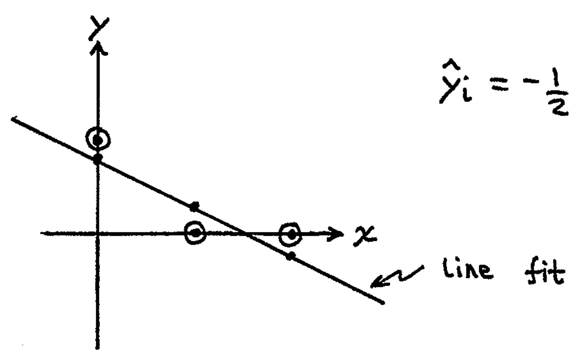
$$= -1\left(\tfrac{2}{3}\right) + 1\left(-\tfrac{1}{3}\right)$$

$$= -1$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = (0-1)^2 + (1-1)^2 + (2-1)^2 = 2$$

$$a_1 = \frac{-1}{2}$$

$$a_2 = \bar{y} - a_1\bar{x} = \frac{1}{3} - \left(-\frac{1}{2}\right)(1) = \frac{1}{3} + \frac{1}{2} = \frac{5}{6}$$



$$\hat{y}_i = -\frac{1}{2}x_i + \frac{5}{6}$$

Our estimated points are found by substituting x_i values into $a_1 x_i + a_2$:

$$(x_1, \hat{y}_1) = \left(0, -\tfrac{1}{2}\cdot 0 + \tfrac{5}{6}\right) = \left(0, \tfrac{5}{6}\right)$$

$$(x_2, \hat{y}_2) = \left(1, -\tfrac{1}{2}\cdot 1 + \tfrac{5}{6}\right) = \left(1, \tfrac{1}{3}\right)$$

$$(x_3, \hat{y}_3) = \left(2, -\tfrac{1}{2}\cdot 2 + \tfrac{5}{6}\right) = \left(1, -\tfrac{1}{6}\right)$$

The coefficient of determination, r, tells us what fraction of the total squared error our straight line fit eliminates compared to just using $y_i = \bar{y}$ for the estimate of every y_i.

$$r \equiv 1 - \frac{SSE}{SST} \qquad \text{where} \quad SSE \equiv \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Here, $$SSE = \left(1 - \tfrac{5}{6}\right)^2 + \left(0 - \tfrac{1}{3}\right)^2 + \left(0 - \tfrac{1}{6}\right)^2$$

$$= \left(\tfrac{1}{6}\right)^2 + \left(\tfrac{2}{6}\right)^2 + \left(\tfrac{1}{6}\right)^2 = \frac{6}{6^2} = \frac{1}{6}$$

$$SST = \left(1 - \tfrac{1}{3}\right)^2 + \left(0 - \tfrac{1}{3}\right)^2 + \left(0 - \tfrac{1}{3}\right)^2$$

$$= \left(\tfrac{2}{3}\right)^2 + \left(\tfrac{1}{3}\right)^2 + \left(\tfrac{1}{3}\right)^2 = \frac{6}{3^2} = \frac{2}{3}$$

Thus, $$r = 1 - \frac{1/6}{2/3} = 1 - 1/4 = \frac{3}{4} \text{ or } 75\%$$