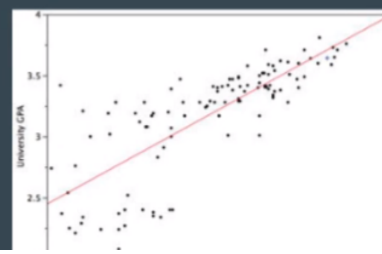
Simple Linear Regression

- So let's re-state our problem in more general terms
- We are given a set of points: { $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ }
- We plot them in a 2-D chart
- We find the line of best fit
- Is there a more systematic way of doing it, other than drawing it using paper and a ruler? Of course!



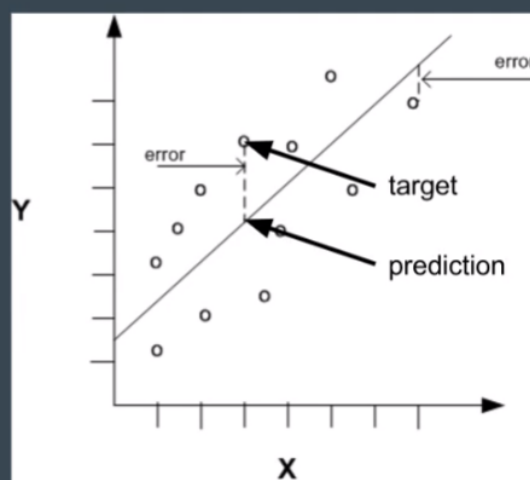
Simple Linear Regression

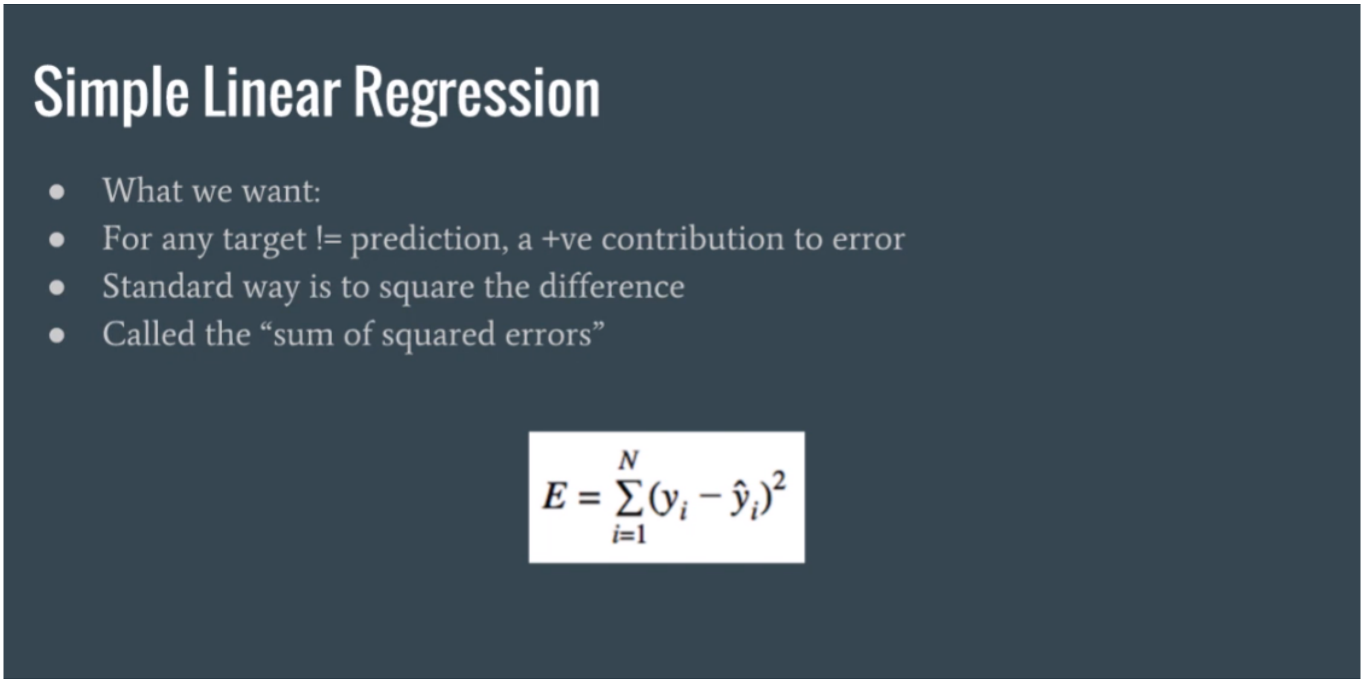
- Our line of best fit is defined as:

$$\hat{y}_i = ax_i + b$$

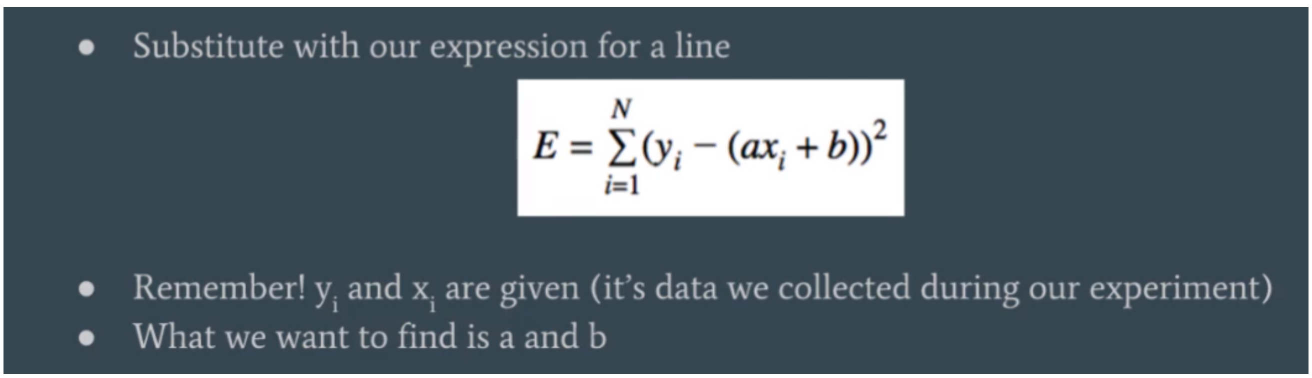
- How can we make sure this "fits" the data well? We would like:

$$y_i \text{ close to } \hat{y}_i,\ i = 1..N$$

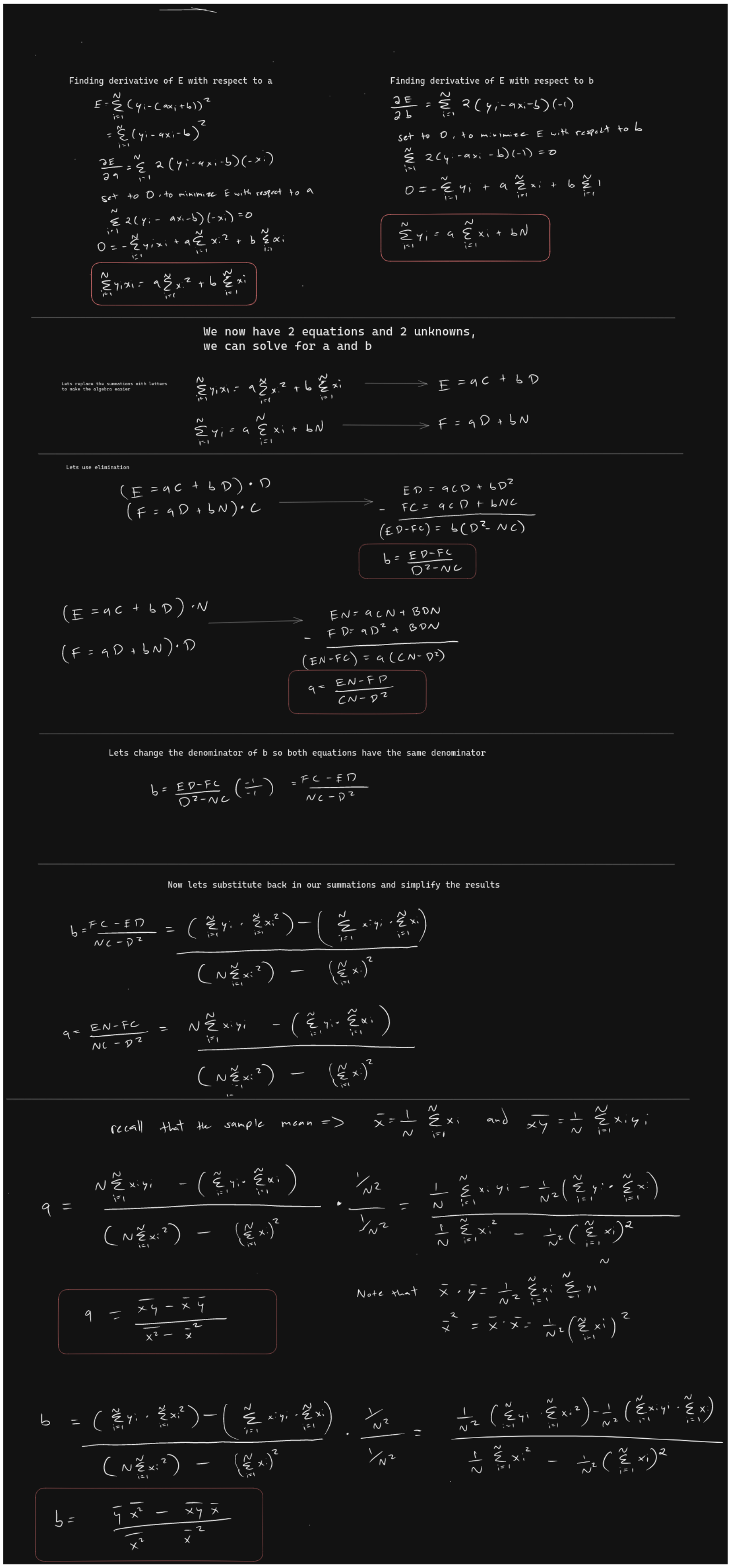




- Note that we use the square of the error to counteract the fact that any given error can be positive or negative!
- We want to find E such that E is the smallest possible value for the sum of squared errors. We are trying to minimize E



- We want to minimize E with respect to a and b, we can use partial derivatives

- 

Finding derivative of E with respect to a

$$E = \sum_{i=1}^{N} (y_i - (ax_i + b))^2 = \sum_{i=1}^{N} (y_i - ax_i - b)^2$$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{N} 2(y_i - ax_i - b)(-x_i)$$

set to 0, to minimize E with respect to a

$$\sum_{i=1}^{N} 2(y_i - ax_i - b)(-x_i) = 0$$

$$0 = -\sum_{i=1}^{N} y_i x_i + a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i$$

$$\boxed{\sum_{i=1}^{N} y_i x_i = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i}$$

Finding derivative of E with respect to b

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{N} 2(y_i - ax_i - b)(-1)$$

set to 0, to minimize E with respect to b

$$\sum_{i=1}^{N} 2(y_i - ax_i - b)(-1) = 0$$

$$0 = -\sum_{i=1}^{N} y_i + a\sum_{i=1}^{N} x_i + b\sum_{i=1}^{N} 1$$

$$\boxed{\sum_{i=1}^{N} y_i = a\sum_{i=1}^{N} x_i + bN}$$

We now have 2 equations and 2 unknowns, we can solve for a and b

Lets replace the summations with letters to make the algebra easier

$$\sum_{i=1}^{N} y_i x_i = a\sum_{i=1}^{N} x_i^2 + b\sum_{i=1}^{N} x_i \longrightarrow E = aC + bD$$

$$\sum_{i=1}^{N} y_i = a\sum_{i=1}^{N} x_i + bN \longrightarrow F = aD + bN$$

Lets use elimination

$$(E = aC + bD)\cdot D \qquad (F = aD + bN)\cdot C \longrightarrow \begin{array}{l} ED = aCD + bD^2 \\ -\ FC = aCD + bNC \\ \hline (ED - FC) = b(D^2 - NC) \end{array}$$

$$\boxed{b = \frac{ED - FC}{D^2 - NC}}$$

$$(E = aC + bD)\cdot N \qquad (F = aD + bN)\cdot D \longrightarrow \begin{array}{l} EN = aCN + BDN \\ -\ FD = aD^2 + BDN \\ \hline (EN - FC) = a(CN - D^2) \end{array}$$

$$\boxed{a = \frac{EN - FD}{CN - D^2}}$$

Lets change the denominator of b so both equations have the same denominator

$$b = \frac{ED - FC}{D^2 - NC}\left(\frac{-1}{-1}\right) = \frac{FC - ED}{NC - D^2}$$

Now lets substitute back in our summations and simplify the results

$$b = \frac{FC - ED}{NC - D^2} = \frac{\left(\sum_{i=1}^{N} y_i \cdot \sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i y_i \cdot \sum_{i=1}^{N} x_i\right)}{\left(N\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

$$a = \frac{EN - FC}{NC - D^2} = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} y_i \cdot \sum_{i=1}^{N} x_i\right)}{\left(N\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

recall that the sample mean => $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$

$$a = \frac{N\sum_{i=1}^{N} x_i y_i - \left(\sum_{i=1}^{N} y_i \cdot \sum_{i=1}^{N} x_i\right)}{\left(N\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2} \cdot \frac{1/N^2}{1/N^2} = \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{1}{N^2}\left(\sum_{i=1}^{N} y_i \cdot \sum_{i=1}^{N} x_i\right)}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2}$$

$$\boxed{a = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}}$$

Note that $\bar{x}\cdot\bar{y} = \frac{1}{N^2}\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i$

$$\bar{x}^2 = \bar{x}\cdot\bar{x} = \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2$$

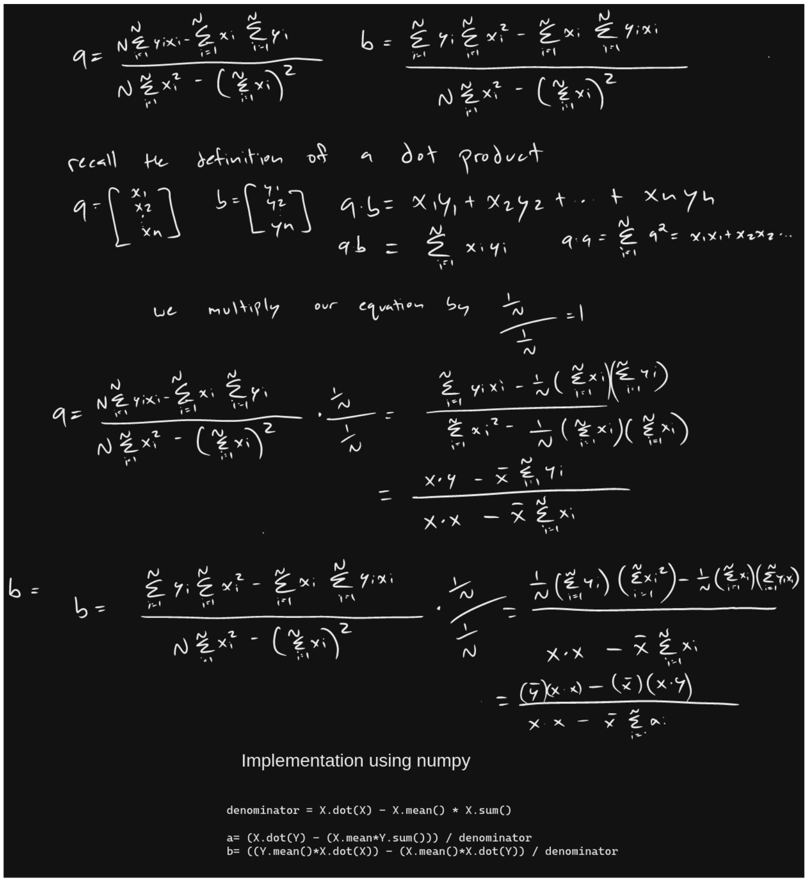
$$b = \frac{\left(\sum_{i=1}^{N} y_i \cdot \sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i y_i \cdot \sum_{i=1}^{N} x_i\right)}{\left(N\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2} \cdot \frac{1/N^2}{1/N^2} = \frac{\frac{1}{N^2}\left(\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2\right) - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i y_i \cdot \sum_{i=1}^{N} x_i\right)}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2}$$

$$\boxed{b = \frac{\bar{y}\,\overline{x^2} - \overline{xy}\,\bar{x}}{\overline{x^2} - \bar{x}^2}}$$

Programming linear regression using numpy

- In order to make the computation of a and b easier we can actually use.

- We can actually re-write the equation we just derived to something simpler that will make it easier to calculations with Numpy



-