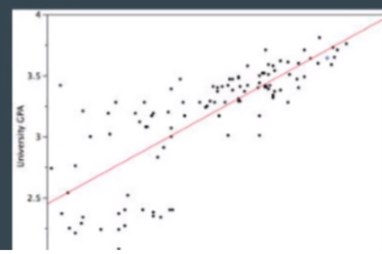
Simple Linear Regression

- So let's re-state our problem in more general terms
- We are given a set of points: { $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$ }
- 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!



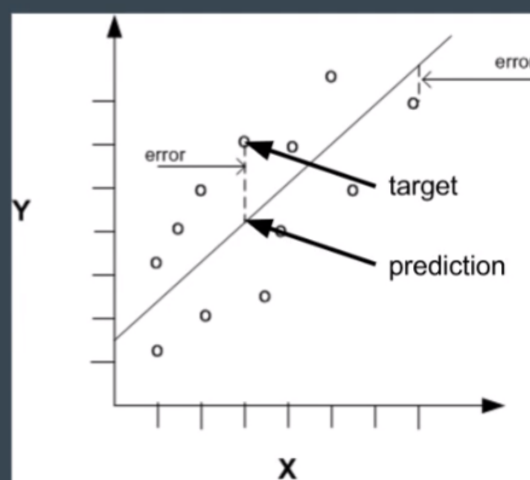
Simple Linear Regression

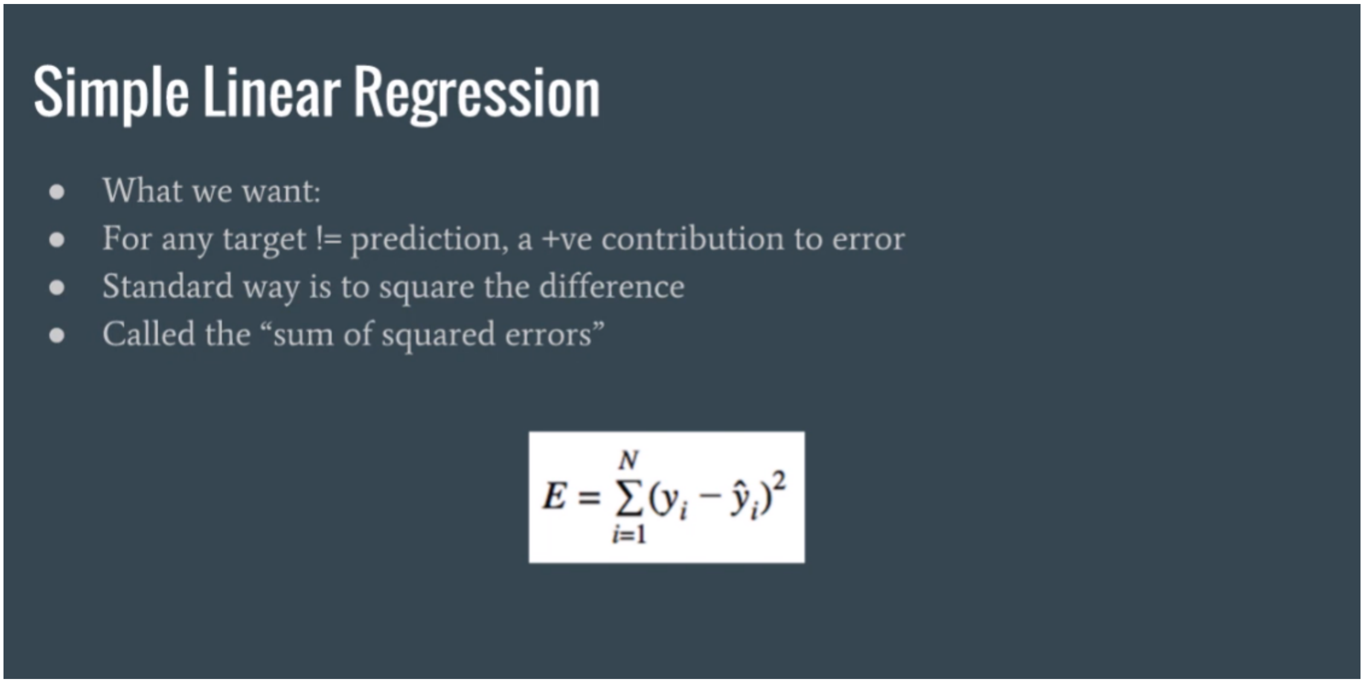
- Our line of best fit is defined as:

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

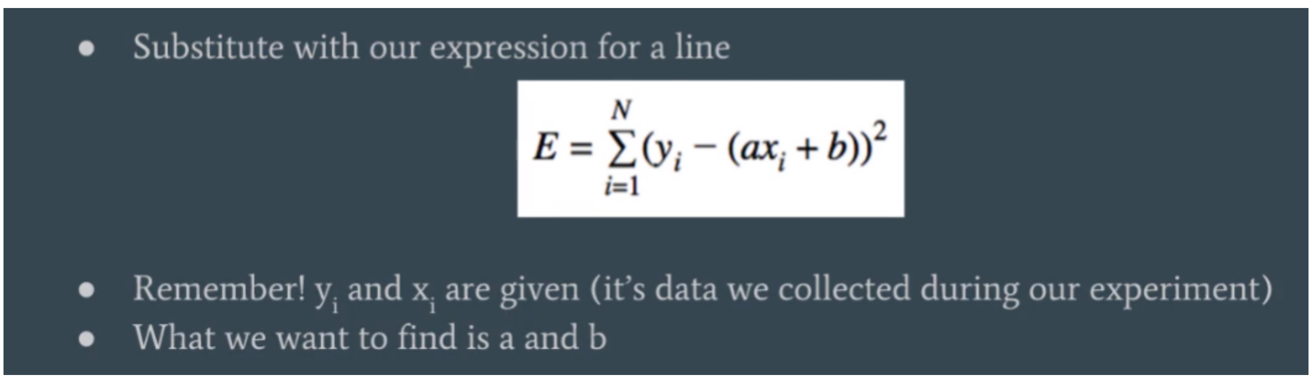
- How can we make sure this "fits" the data well? We would like:

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

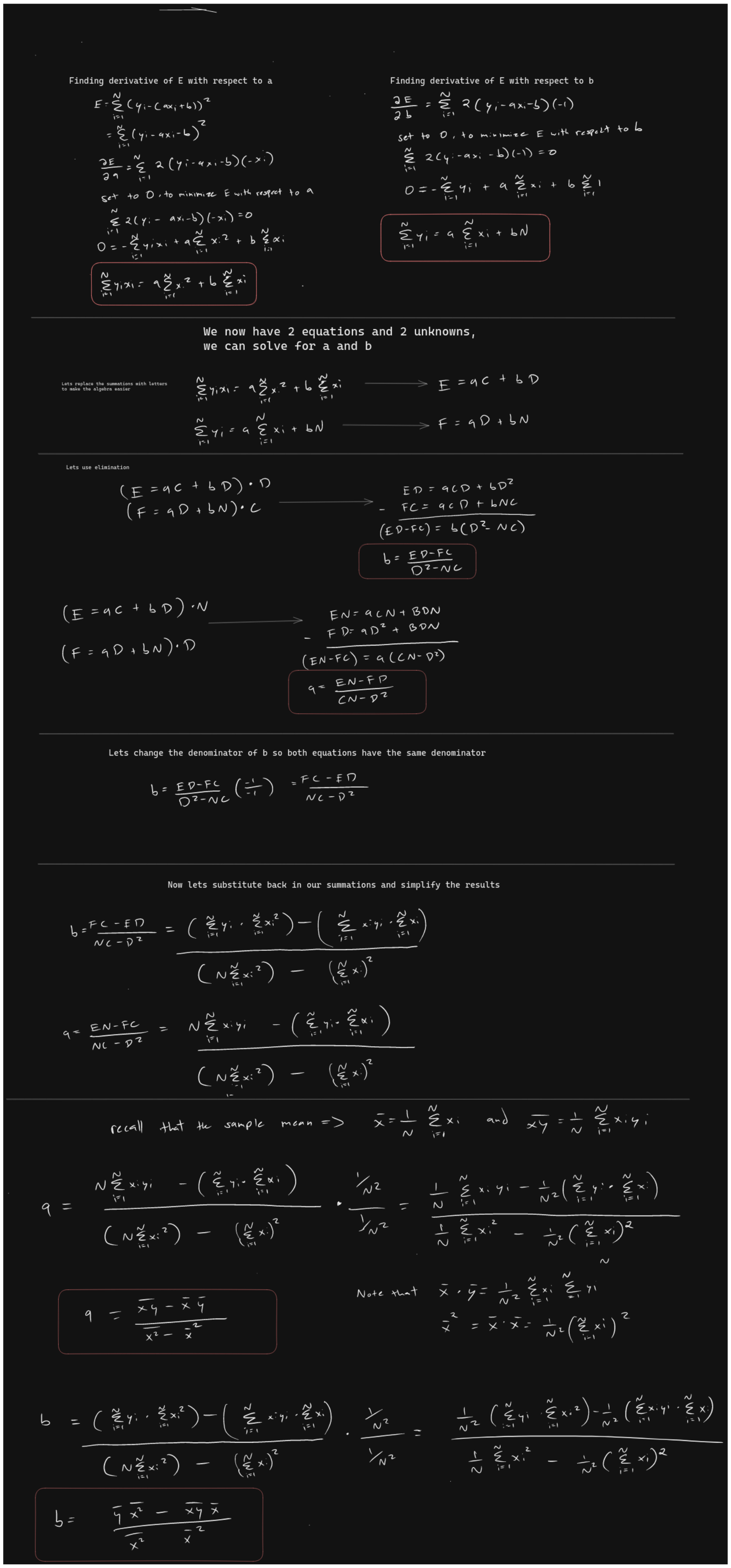




- Note that we use the square of the error to counteract the fact that any given error can be positive or negative!
- We want to find E such that E is the smallest possible value for the sum of squared errors. We are trying to minimize E

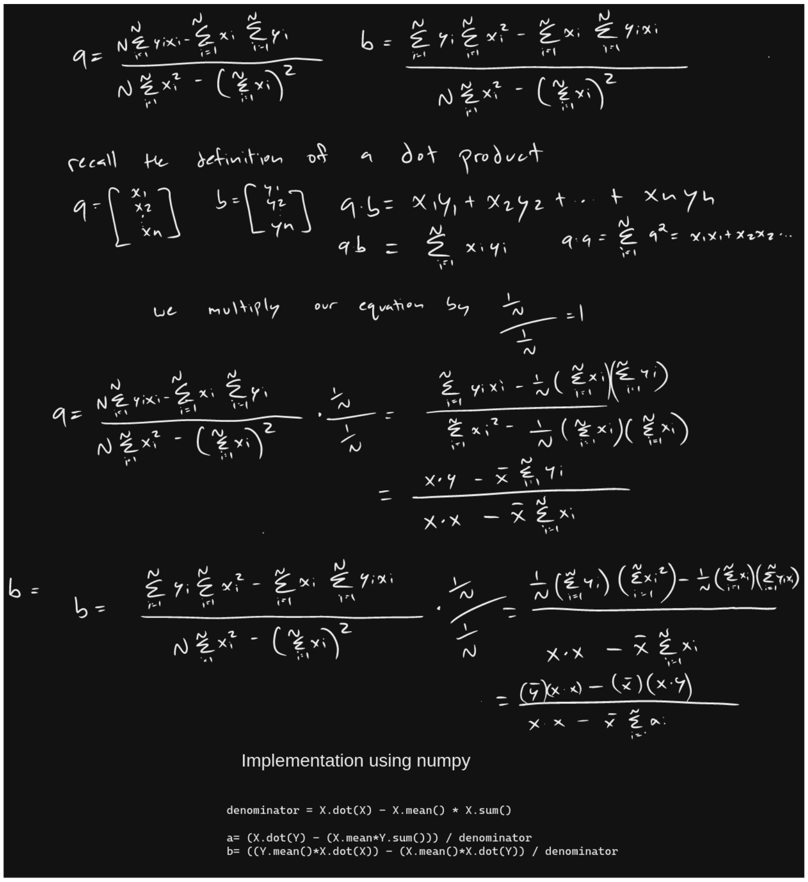


- We want to minimize E with respect to a and b, we can use partial derivatives

- 

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
- 

Determining how good the model is

- We can use the R^2 formula to determine how good the model is, the formula is
  - $R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$
  - $SS_{residual} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
  - $SS_{total} = \sum_{i=1}^{N}(y_i - \bar{y}_i)^2$
- Notice that $SS_{residual}$ is just our error formula
- SS_{total} is the difference between each y against the mean of y.

Cases

- If our error formula is close to 0 then our formula will be $R^2 = 1 - 0 = 1$ which means our model is pretty much perfect, the closer the value is to 1 the better
- If $R^2 = 0$ that means to formula was $R^2 = 1 - 1 = 0$ which means that our $SS_{residual}$ was very close to SS_{total} meaning our predictions were just taking the mean of y to make predictions, which is not good.
  - This can happen if your data does not have a clear trend
- If $R^2 < 0$ that means that $\frac{SS_{residual}}{SS_{total}} > 1$ and this points out that your model is performing worse than just predicting the mean of y