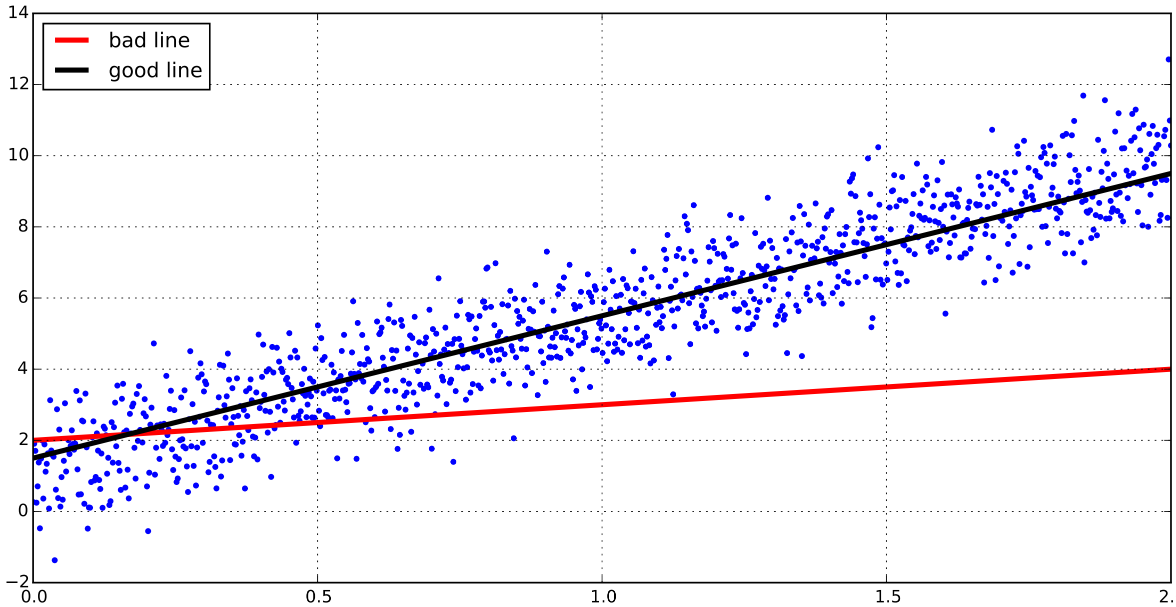
Least Square Fitting

Least Square Fitting

Consider the following data points shown in blue dots.

$$\{(x_i, y_i)|i = 1, 2, 3, 4, \dots, n\}$$

You can think about the x-axis as number of hour a student study for midterm and the y axis would be teh score the student get on the midterm.



If you look at the the blue dots, your gut feeling will tell that this data behaves like a linear function. That means the guess/prediction from the line will be calculated using

$$\hat{y} = ax + b \tag{1}$$

where the symbol $\hat{y}$ indicate that this is the guess value. In particular our guess for the i-th data point is given by

$$\hat{y}_i = ax_i + b \tag{2}$$

Our job here is to find a and b that makes the “best” line.

We can see that the red line on the plot doesn't really represent the data and the black looks like a much better fit. We can this goodness of fit. If we look at the two lines, the reason we way that the red line is worse than the black line is because the line seems to be so far away from the points.

That means we need a quantity that tells us how far our guess using the line is from the the actual point. We can calculate this quantity point-wise. The distance of our guess for x_i from the actual value is given by

$$d_i = \hat{y}_i - y_i \tag{3}$$

But the goodness of fit has to be a global value not just a point-wise one. The most natural thing to do is to add up all the distance

$$\text{Bad Measure} = \sum_{i=1}^{n} d_i \tag{4}$$

This, however, doesn't work as the negative value from one data point and the positive value from another data point will cancel out. We can fix this by squaring the point-wise distance before adding them up[1]. So we define

$$\chi^2 = \sum_{i=1}^{n} d_i^2 \tag{5}$$

$$= \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 . \tag{6}$$

The χ symbol reads chi. If χ^2 is large that means a lot of point are further away from the line indicating a bad fit. If χ^2 is small that means most points are close to the line indicating a good fit.This symbol is also quite meaningful in statistics.

Let us continue on elaborating χ^2. The most important thing about this quantity is that it is a function of a and b. Your data(x_i, y_i) are fixed. The things that changes from line to line are a and b. These two parameters dictate what your line looks like. In particular,

$$\chi^2(a,b) = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{7}$$

$$= \sum_{i=1}^{n} (ax_i + b - y_i)^2 \tag{8}$$

Now that we have a quantitiy that measure the goodness of fit as a function of our parameters. There is only one thing left to do is to find a and b such that χ^2 is minimize. We can do that by just simple differentiation and set it to zero.

[1] Absolute function or the fourth power would do the same job but these do not posess the statistical meaing or the differentiability like the square one

$$\frac{\partial}{\partial a}\chi^2(a,b) = 0 \tag{9}$$
$$\frac{\partial}{\partial b}\chi^2(a,b) = 0 \tag{10}$$

Equation 9 gives

$$\frac{\partial}{\partial a}\chi^2(a,b) = \sum_{i=1}^{n} 2\left(ax_i + b - y_i\right)x_i \tag{11}$$
$$= 2\left(a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i\right) = 0 \tag{12}$$

Therefore we have

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i = 0 \tag{13}$$
$$a\mathbb{E}[x^2] + b\mathbb{E}[x] - \mathbb{E}[xy] = 0 \tag{14}$$

where in the last line we divide through by the number of data points n on both sides and define

$$\mathbb{E}[x] = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{15}$$
$$\mathbb{E}[x^2] = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{16}$$
$$\mathbb{E}[xy] = \frac{1}{n}\sum_{i=1}^{n} x_i y_i \tag{17}$$

Equation 10 gives

$$\frac{\partial}{\partial b}\chi^2(a,b) = \sum_{i=1}^{n} 2\left(ax_i + b - y_i\right) \tag{18}$$
$$= 2\left(a\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i\right) = 0 \tag{19}$$

Therefore we have

$$a\sum_{i=1}^{n}x_i + bn - \sum_{i=1}^{n}y_i = 0 \tag{20}$$

What's left for us to do is to solve Equation 14 and 20 for a and b. From Equation 20 we have

$$b = \frac{\sum_{i=1}^{n}y_i - a\sum_{i=1}^{n}x_i}{n} = \mathbb{E}[y] - a\mathbb{E}[x] \tag{21}$$

where $\mathbb{E}[y] = \frac{1}{n}\sum_{i=1}^{n}y_i$.

Plugging this into Equation 14 we have

$$\begin{align} 0 =& a\mathbb{E}[x^2] + (\mathbb{E}[y] - a\mathbb{E}[x])\,\mathbb{E}[x] - \mathbb{E}[xy] \tag{22} \\ & \tag{23} \end{align}$$

Simplfying the above gives

$$\begin{align} 0 =& a\left(\mathbb{E}[x^2] - Ex^2\right) + \mathbb{E}[y]\mathbb{E}[x] - \mathbb{E}[xy] \tag{24} \\ a =& \frac{\mathbb{E}[xy] - \mathbb{E}[y]\mathbb{E}[x]}{\mathbb{E}[x^2] - \mathbb{E}[x]^2} = \frac{\mathrm{Cov}[x,y]}{\mathrm{Var}[x]} \tag{25} \end{align}$$

remember those from Discrete Math? Then b can then be found by plugging this back into Equation 20.

$$\begin{align} b =& \mathbb{E}[y] - a\mathbb{E}[x] \tag{26} \\ =& \mathbb{E}[y] - \frac{\mathbb{E}[xy] - \mathbb{E}[y]\mathbb{E}[x]}{\mathbb{E}[x^2] - \mathbb{E}[x]^2}\mathbb{E}[x] \tag{27} \\ =& \frac{\mathbb{E}[y]\mathbb{E}[x^2] - \cancel{\mathbb{E}[y]\mathbb{E}[x]^2} - \mathbb{E}[xy]\mathbb{E}[x] + \cancel{\mathbb{E}[y]\mathbb{E}[x]^2}}{\mathbb{E}[x^2] - \mathbb{E}[x]^2} \tag{28} \\ =& \frac{\mathbb{E}[y]\mathbb{E}[x^2] - \mathbb{E}[xy]\mathbb{E}[x]}{\mathbb{E}[x^2] - \mathbb{E}[x]^2} \tag{29} \end{align}$$

This process can be generalized to a much more complicate plot. You will do that in the homework.

Error on Slope(Bonus)

Derivation of this requires quite a bit of understanding in Statistics. You can find the derivation on the internet[2]. Long story short, the error on slope is given by

$$\sigma_a = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i - \mathbb{E}[x])^2}} \tag{30}$$

There is actually another way to find the error on slope called bootstrapping. You will get to do that on the homework.

[2] For example, `http://stats.stackexchange.com/questions/88461`