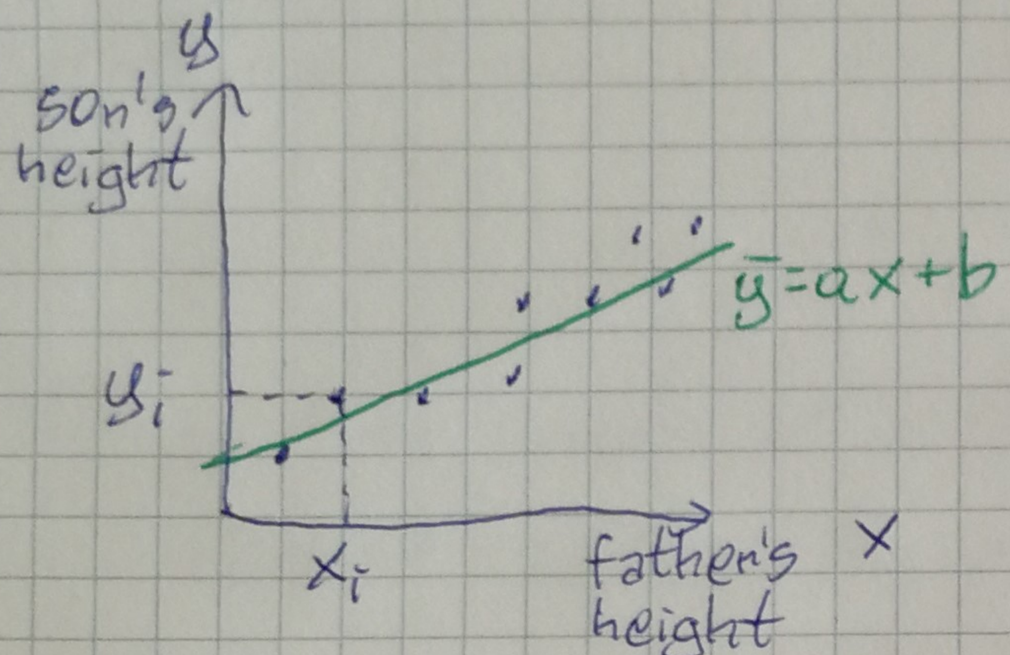
Francis Galton, 1886

„Regression Towards Mediocrity in Hereditary Stature"



Measurement data:

$D=\{(x_i, y_i)\}_{i=1}^{N}$

compose nice-looking cloud of points, seems like we can represent them by a curve, e.g. line. Let's call this curve a model.

Wy do we need the model?

We could replace large amount of data by a simple equation, with the ability to reproduce the data (approximately) or to predict the value of new incoming observations (e.g. predict the value of y for a new x).

The model $\bar{y}=ax+b$ includes two parameters to be determined to make us happy. But when should we feel happy? How to tell whether the model is good or bad?

Let's pick up a single pair (x_i, y_i) out of the dataset D. Then calculate the value of $\bar{y}=ax_i+b$ for some fixed values of parameters a and b (fixed in order the model to be good).
In duggish dialect we say, that the [illegible] input x_i is fed to the model.

{ y_i is measured, $\bar{y}_i$ is calculated } → Compare both: error $e_i=y_i-\bar{y}_i$
Is the value of e_i small?
If yes, we feel happy.

Each measurement pair (x_i, y_i) may be assigned a prediction error:

$$e_1, e_2, \dots, e_N$$

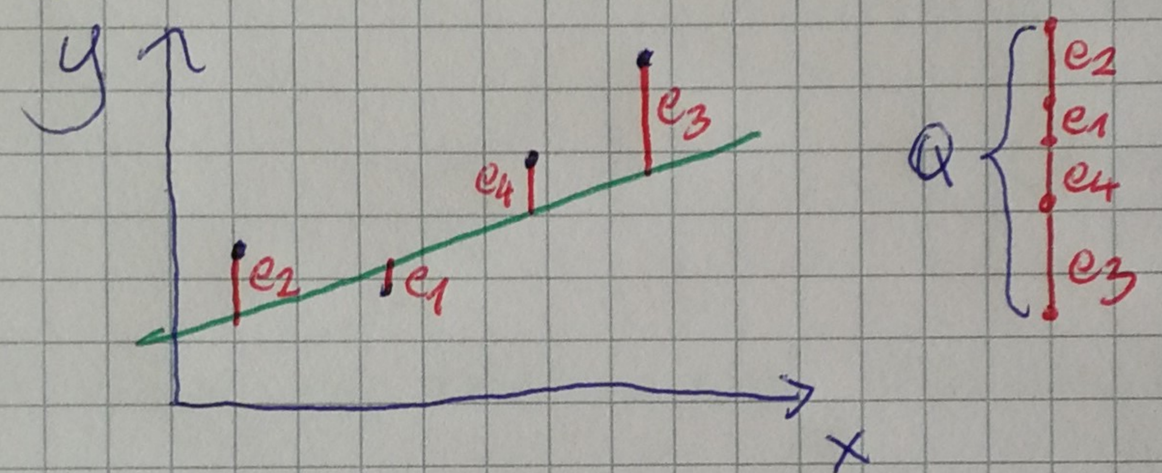
Let's add them up to obtain a single value:

$$\sum_{i=1}^{N} e_i .$$

Notice, that errors of different sign may cancel out, leading to incorrect assessment of the model quality. Better formula:

$$Q_{\text{Quality}} = \sum_{i=1}^{N} |e_i| .$$

It has nice geometrical interpretation:



We can make it even better, averaging out:

$$Q = \frac{1}{N} \sum_{i=1}^{N} |e_i| .$$

Rewrite it, emphasizing that Q is a function dependent on the model parameters:

$$Q(a,b) = \frac{1}{N} \sum_{i=1}^{N} |y_i - \bar{y}_i| =$$

$$= \frac{1}{N} \sum_{i=1}^{N} |y_i - (ax_i + b)| \qquad (1)$$

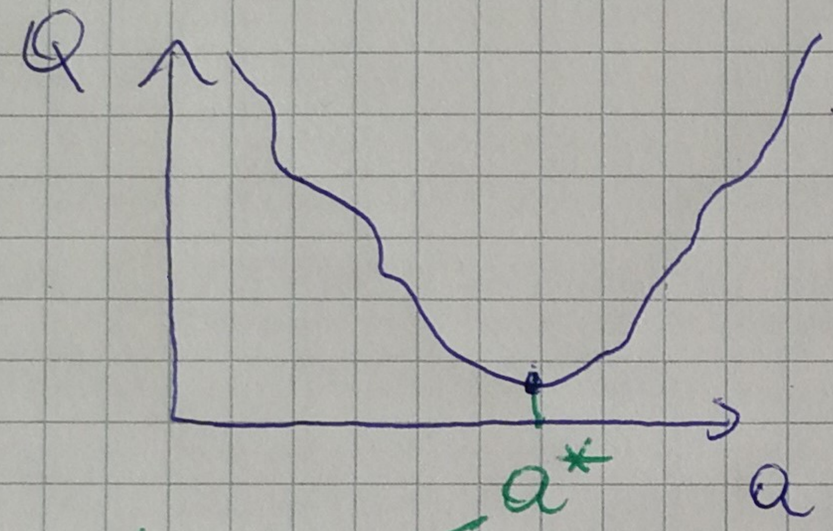
Here, we will show pity for those weakened by the coronavirus, and take - temporarily - simplified version of the model:

$$\bar{y} = ax ,$$

that has only one parameter to be fixed by a user.
The total error Q reduces to:

$$Q(a) = \frac{1}{N} \sum_{i=1}^{N} |y_i - a x_i| , \qquad (2)$$

and can be easily illustrated:



- naive illustration of the shape of Q

(fig I)

the best value of the model parameter, leading to the smallest value of the total error Q

Fitting the model to the data is therefore equivalent to solving an optimization task: $a^* = \arg\min Q(a)$.

You shoul know, that for $a \in \mathbb{R}$ it is sufficient to solve $\frac{dQ}{da} = 0$ for a.

And we run into trouble immediately, because Q is non-differentiable. Let's resume the task we've posed:

Given: $\{(x_i, y_i)\}_{i=1}^{N}$	Find: a that makes
$\bar{y} = ax$	Q minimal
$Q = \frac{1}{N} \sum_{i=1}^{N} \lvert y_i - \bar{y}_i \rvert$	

It turns out that analytical solution cannot be derived.

1) the formula for the solution is direct,
2) the solution is exact.

There are many possible difficulties that may prevent us from deriving analytical solutions. One of them is ill-behaving functions, that can't be differentiated or integrated.
These are situations when a drowning man will clutch at numerical solutions.

1) instead of a direct formula, we are given an iterative / recursive procedure, possibly randomized
2) the solution is approximate.

So, let's bring the function (2) to the labs and give it to a numerical solver (optimization routine) to search for a solution among the space of solutions and resuls in the best approximation a^*.

Let it be your homework to do that. Draw a graph (on a computer screen) similar to that of the page 1. Measurement data may be made up by yourself or acquired from your environment, or even downloaded somewher from internet. Use your imagination, do not expect straight instructions, like machines do.

You will quickly meet the limitations of the model $\bar{y}=ax$ capabilities, e.g.:

it must pass through the origin.

However, you may easily circumvent this problem by replacing measurements. This is for you to try at home.

But we can be fully satisfied only when the model $\bar{y}=ax+b$ is fully operational. Thus, bring the error function (11) to the labs with you and use a numerical solver to find the best values of a and b.

You are probably disappointed with such a quick escape to numerical solver. Is it possible to see an analytical solution of this task? No. But we can take a slightly different task. Let's came up with another error function, that also prevents from cancelling out errors of different signs, but has the nice property of being differentiable:

$$Q=\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y}_i)^2$$ - the averaged square error

Take a look at the fig. I. It illustrates that the change of scale along y axis doe's not affect the value of parameters minimizing a function (to see this just use your imagination). Therefore, we could ged rid of $\frac{1}{N}$ term, since it is unnecessary; the fewer symbols the better.

But be careful: the change of criteria means the change of a task!

A new task:

Given: $\{(x_i, y_i)\}_{i=1}^N$

$y = ax$

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

Find:

$a^* = \arg\min Q(a)$

Let's find out whether an analytical solution can be derived from:

$$\frac{dQ}{da} = 0 \quad .$$

Ready, steady, go!

$$\frac{dQ}{da} = \frac{d}{da} \sum_{i=1}^{N} (y_i - ax_i)^2 = \sum_i \frac{d}{da} (y_i - ax_i)^2 =$$

$$= \sum_i -2(y_i - ax_i) x_i = 0$$

$$\sum_i (y_i x_i - a x_i^2) = 0$$

$$\sum_i x_i y_i - \sum_i a x_i^2 = 0$$

$$a \sum_i x_i^2 = \sum_i x_i y_i$$

$$\boxed{a^* = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}}$$ - direct/explicit formula - exact solution. (3)

I'm sure you can guess what is your homework according to the formula above? It needs to be done in Python, the chart shold look nicely, do not use any solver (because you are equipped with the formula (3)).

Now, recall from mathematical analysis classes the 2nd order conditions of optimality. How do you think, why I was so careless about it when solving the previous task?

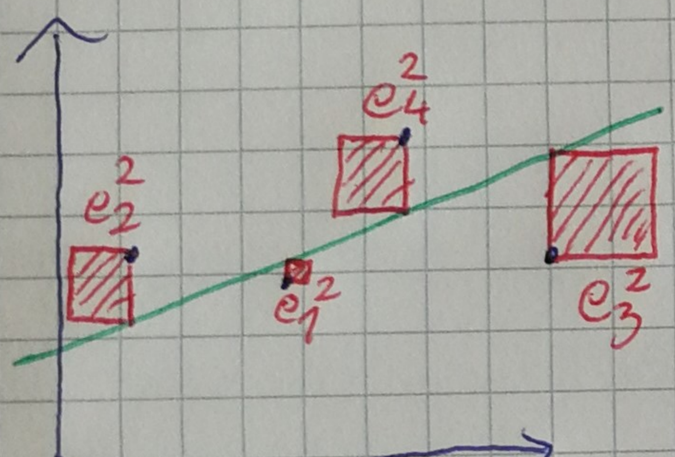
I recommend you to get used to the introduced error function, this is an ordinary 2nd-order polynomial:

$$Q(a) = \sum_{i=1}^{N} (y_i - a x_i)^2 = \alpha a^2 + \beta a + \gamma .$$

Figure out formulas describing coefficients α, β, γ.

Think about the geometrical interpretation behind the error function:

$$Q = \sum_{i=1}^{N} (y_i - \bar{y}_i)^2 .$$



Use your imagination to investigate the effect the outliers exert on:

1) the sum of absolute values of errors,

2) the sum of squared errors.

If your imagination is not enough, support it with computer-based experimentation.

Now you are ready to face the original problem posed and solved by Francis Galton in XIX century.

Given: $\{(x_i, y_i)\}_{i=1}^{N}$

$\bar{y} = ax + b$

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

Find:

$a^*, b^* = \arg\min Q(a,b)$

In duggish dialect, a is called the slope and b is called the intercept.

(strongly) assume that from mathematical analysis classes you know the procedure of optimizing multivariable functions:

$$\begin{cases} \frac{\partial Q}{\partial a} = 0 \\ \frac{\partial Q}{\partial b} = 0 \end{cases}$$

, recall that the vector $\begin{bmatrix} \frac{\partial Q}{\partial a} \\ \frac{\partial Q}{\partial b} \\ \vdots \end{bmatrix}$ is commonly termed „gradient" and denoted ∇Q.

{Perhaps you have noticed that the word „gradient", trendy nowadays, mainly due to rapid development of deep learning field. I recommend gentlemen to use it in order to impress ladies. However, ladies are advised to avoid it in order not to scare out the gentleman.}

We are going to treat gradients later on.

Figure out formulas for a and b, I won't do this for you. You have to tackle the linear system of two equations having two unknowns in total. It would be a shame not to handle it.

Implement the developed formulas in order to reproduce a figure similar to fig. from page 1, including model $\bar{y} = ax + b$.

Try to tame a 2Dimensional sum of squared errors, by groupping its components:

$$Q(a,b)=\sum_{i=1}^{N}(y_i-(a_i x+b))^2=\alpha a^2+\beta b^2+\gamma ab+\varepsilon a+\xi b+\Delta,$$

where $\alpha, \beta, \gamma, \varepsilon, \xi, \Delta$ are coefficients to be determined.

Conclusions and generalizations.

We have dealt with the task of fitting a model $\bar{y}=F(x,a)$ to the dataset $\{(x_i,y_i)\}_{i=1}^{N}$, where F is any given function of variable x, dependent on parameters gathered in a vector a:

$$a=\begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}.$$

The components of a are fixed in such a way, that the error function $Q(a)$ takes on the minimum value.

We have considered two cases of model:

1) $\bar{y}=F(x,a)=a\cdot x$, $a\in\mathbb{R}$

2) $\bar{y}=F(x,a)=a_1\cdot x+a_0$, $a=\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}\in\mathbb{R}^2$,

and two examples of error functions:

1) $Q=\sum_{i=1}^{N}|e_i|$

2) $Q=\sum_{i=1}^{N}e_i^2$

If you will do something cool in the topic, you may send your results by e-mail.