Module 6

Simple linear regression

So, now we will finally going to apply all this multivariate calculus to help us fit functions to data

This is very handy, it allows us to make sense of our data, to start doing statistics testing, to start to really apply all the stuff we been doing into real world.

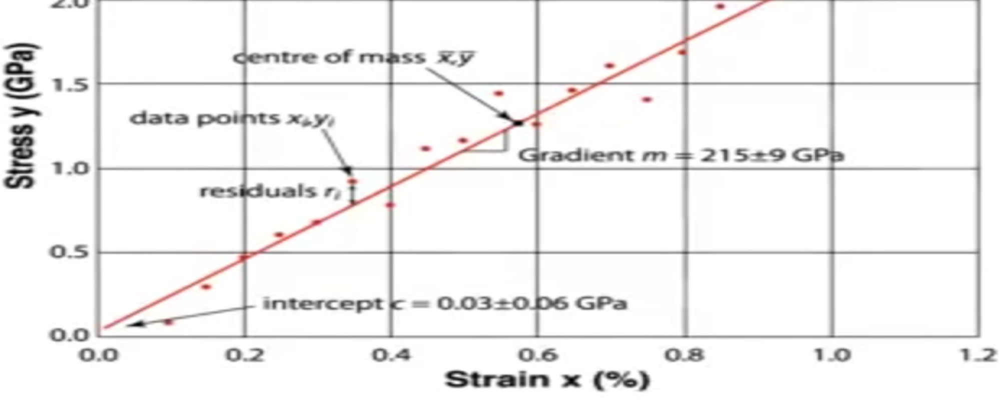
If you have some big bundle of data, the first thing you need to do is to clean it up
there are other course on this..., to get it into shape
That means figuring out what the sensible thing to do is, with things like partially empty data entry, 0's, grouping, binning?

data together, eliminating duplicates, etc
easiest thing to do is to tag the data,
so it can be re-ordered, and sort them,
look for funny values, or replace them
with something sensible!

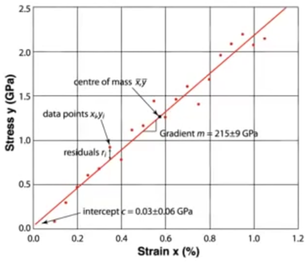
So once cleaned up, now, you can start
graphing it, take averages, look at std dev, etc

It really helps if you interested in data, need
to ~~treat~~ like a person, you want to figure out
and understand, until it becomes an old friend.
If over time you collect more data, and
it starts to change, you need to notice those changes

Once you have graphed it in sensible way,
you often get a plot like this graph: (Oxford)



where's a sample x,y plot with some data (pic)



this data seem to plot like a straight line

If you know something physically about the processes involved in generating the data, or if you have some hypothesis on how the variables are related, then you can try fit that model to the data

Alternatively, you can just try fitting something sensible based on how it looks, like a straight line to the data here. for example.

Now, I can model my straight line,

y here as been a function of the i observation x_i, and vector a of the fitting parameters

$$y = y(x, a_i) = m x_i + c$$

$$a = \begin{bmatrix} m \\ c \end{bmatrix}$$

In case of straight line $y = mx + c$, then the parameters in vector a, would be gradient m and intercept $c \Rightarrow a = \begin{bmatrix} m \\ c \end{bmatrix}$ of straight line

So here I plotted, the optimal straight line, it happens to have gradient 215 GPA, and intercept of 0.03 GPA for C.

I can also find the mean of x, $\bar{x}$, and mean y, $\bar{y}$ which are at the geometric centre of mass of that data set

Now in order to find the optimal value of m and c, lets first define a residual r which we define as the difference between the data items y_i and predicted location for those on line of y, which is $mx + c$

$$r_i = y_i - mx_i - c$$

$$\chi^2 = \sum_i r_i^2 = \sum_i (y_i - mx_i - c)^2$$

If r is $y_i - mx_i - c$

Kai

then I can take a measure of the overall quality of the fit, been quantity I call, χ^2 which is the sum of squares of residuals r

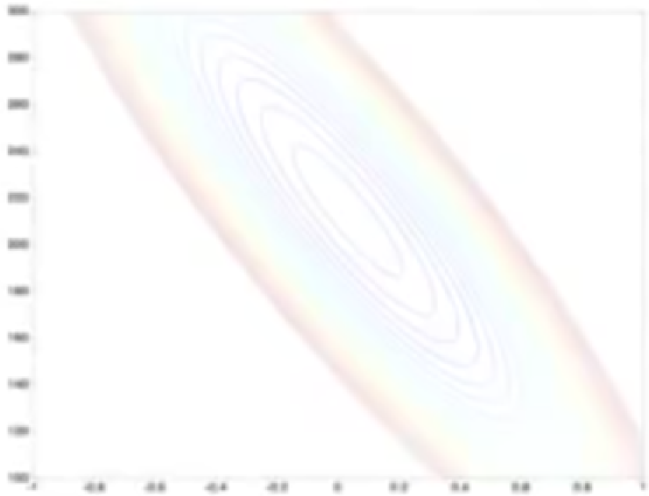
I do this so that I penalize both data above and data below the line, I dont want the +'s and -'s to net off against eachother

I also want to badly penalize data, a long way away from line

(kai)

and try to find the best χ^2 possible, the one that is lowest, I am doing a minimization.

So it's worth plotting what χ^2 is going to look like for lots of different possible values too, made see contour plot pe:



In the middle of about 215 and mean intercept of 0, I found my minimum, and the further I get away from there, the worse it is.

Note, that the χ^2 it's slanted, the bigger the optimum ~~approximate~~ value of gradient m gets, the lower and vice versa.

And if I look ~~at this~~ on this plot, ~~it's got~~, it's quite drawn, if I make the line steeper on the original fit, then in order for it to fit, while the intercept is going to have to get smaller.

actually I am pivoting about the centre of mass

Also, the shallow throught, here in the x value valley, its going to be quite shallow, this is going to be a tricky problem for steepest decent, it going to be easy to get down sides, but difficult to get down bottom valley to find actual minimum

But nevertheless it looks like its going to be an OK problem to solve,

It has 1 easy to spot minimum, therefore we can find it.

Note, to do this with any precision, simply by doing lots of computations here, like I have done, for different m's and c's and finding the minimum that way, and plotting them all out, and finding minimum on the graph,

To do this I will have to do a lot of maths

In matlab, the contour plot, took about 200k computations to make

So even for a simple problem, like this we would really do want to find an algorithm that will let us get there a bit more efficiently.

Note, the minimum is going to be found when the gradient of χ^2 is 0, so

So if we take the grad of χ^2, wrt to fitting parameters and set it to 0, that's going to be our solution

$$\nabla \chi^2 = \begin{bmatrix} \frac{\partial \chi^2}{\partial m} \\ \frac{\partial \chi^2}{\partial c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Now, the neat thing is, that in this particular case, we can actually solve this problem explicitly.

so that's what we going to do in this lesson.