

Module 6

Simple linear regression

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

This is very handy, & allows us to make sense of our data, to start doing statistics thing, 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 because other course on this..., to get it into shape

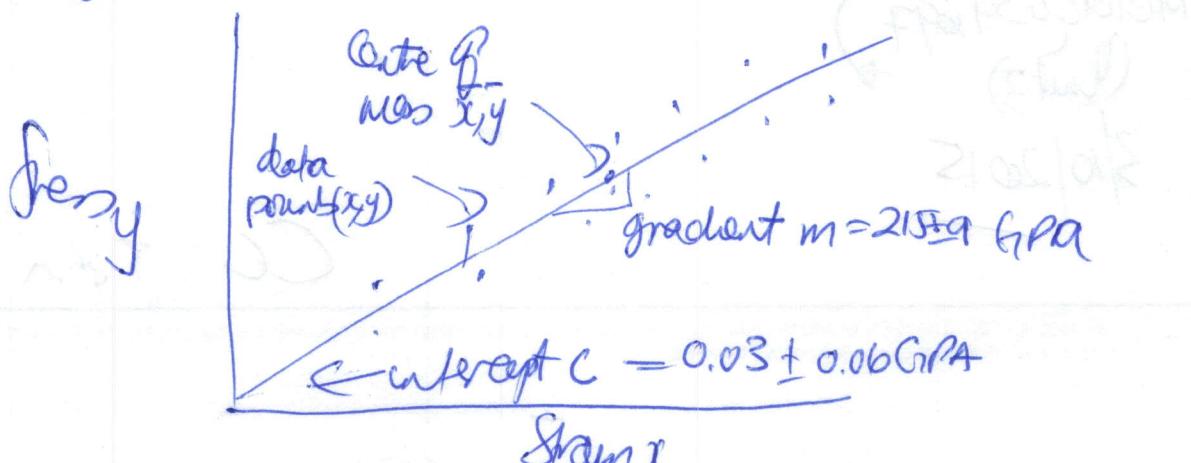
That means figuring out what reasonable things to do's, with things like potentially empty data entry, 0's, grouping, binning?

2

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

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 guy: (Pache)



Makes a sample x,y plot with some data file

The 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 hypotheses on
how the variables are related,
then you can fit that model to the data
Alternatively, you can just try fitting some
straight line to the data file. For example:
Now, I can model my straight line;
 y here has been a function of the ~~the~~ observations
 x_1, x_2 , and we can fit a set of the fitting parameters

$$y = y(x, a_i) = mx_i + c$$

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

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

As here I plotted, the optimal straight line,
 it happens to have gradient 215.6 GPA, and
 intercept of 0.03 GPA for C.

I can also find the mean of x, \bar{x} , and 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 , let's first define a residual r
 which we define as the difference between
 the data items y_i and predicted location
 for those on the line $a(y)$, which is $mx + c$

$$r_i^o = y_i - mx_i^o - c$$

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

$$\text{for } r_i = y_i - mx_i - c$$

then χ^2 can take a measure of the overall quality of the fit, being quantity I call, χ^2 which is

Sum of squares of residuals r

I do this so that I penalize both data above and data below the line, I don't want the +1's and -1's to net off against each other

I also want to badly penalize data, along 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 I want plotting what it's going to look like
for lots of different possible values b_0 , made
to b_0 range to 3000

See scatter plot p:

REGULAR ERROR

so I can see what's happening and what's occurring at the top of the graph. So I can see what's occurring at the top of the graph. So I can see what's occurring at the top of the graph. So I can see what's occurring at the top of the graph.

STANDARD ERROR

In the middle of about 215 and when intercept of 0,
I found my minimum, and the flatter & get
away from those, the worse it's.

Note, that the χ^2 it's stated, the biggest C
gets, the lower optimum value of gradient m
and vice versa.

And if I look at ~~at~~ on the plot, it got, is
quite down, if I make the line steeper
or be a good 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 ⁷

Also, the shallow through here in the ^{value} ~~12~~ Valley,

it's going to be quite shallow, this is going to be a tricky problem for steep descent, it's going to be easy to get down sides, but difficult to get down bottom valley, to find actual minimum

But nevertheless it looks like it's going to be an OR 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 Shoveling, 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 want to find an algorithm that will let us get there a bit more efficiently.

Now, the minimum's going to be found when the gradient of $x^2 \rightarrow 0$, so

so if we take the grad of x^2 , with respect to all the parameters and set it to 0, that's going to be our answer.

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

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

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