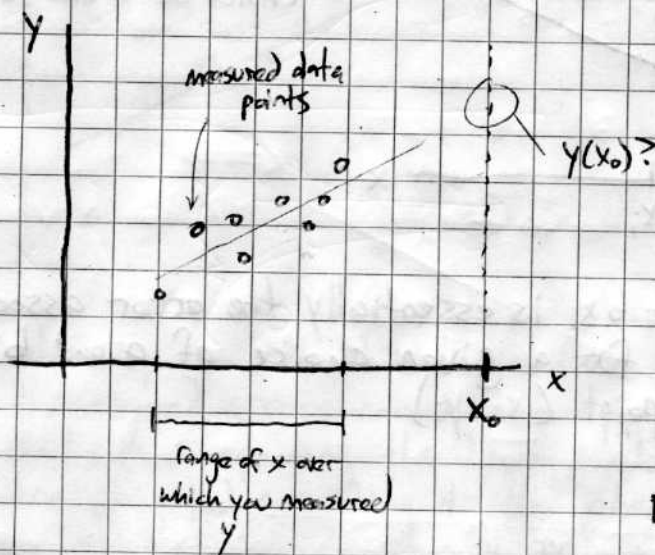


Tangential Topic - Linear Regression (OAE in different context) or 'Least Squares'

12/11/09

- suppose we have an independent variable x and we measure the dependent variable y . In doing so, we find the following relationship:



but say you want to try to determine y for some value x_0 outside of your tested range. How to do it?

1 notice $y(x)$ appears to be fairly linear

2 IDEA: if we can determine a line of best fit to the data, we can interpolate what $y(x_0)$ might be

3 for simplicity @ first, assume there is a small and uniform uncertainty for all our meas. points

Technical Details (fast-forwarded)

- we assume $y(x) = a + bx$ [b-slope, a-y intercept] and we want to determine the 'best' values of a and b such that we get the best 'fit' of this curve to the data
[once we have the best a and b values, it becomes very straight-forward to interpolate what $y(x_0)$ might be]

- we define an 'error' (χ^2 , or chi-squared) as follows:

$$\chi^2 \equiv \sum_{i=1}^N [y_i - y(x_i)]^2 = \sum_{i=1}^N [y_i - a - bx_i]^2$$

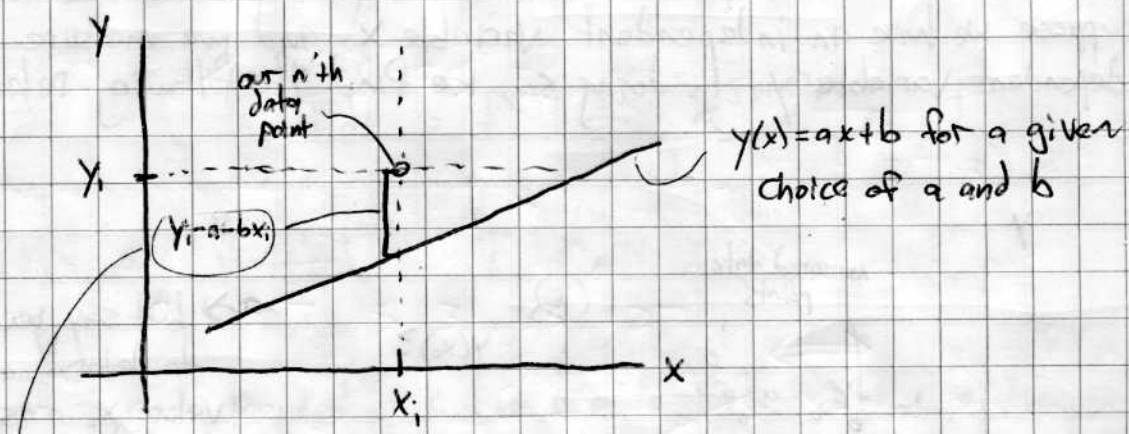
where N is the # of meas. you made, (x_i, y_i) is the i th meas. pair and a and b are your 'guess values' for the eqn. parameters

12/1/09

(cont.)

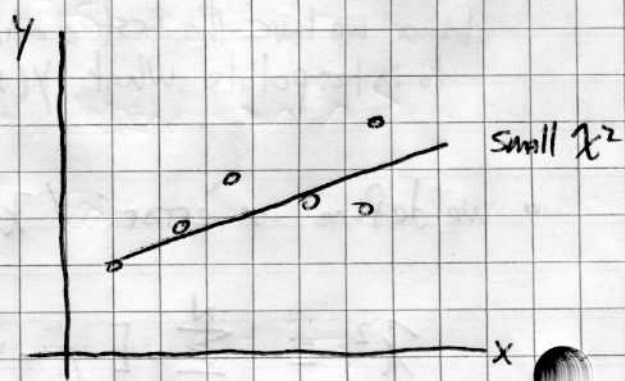
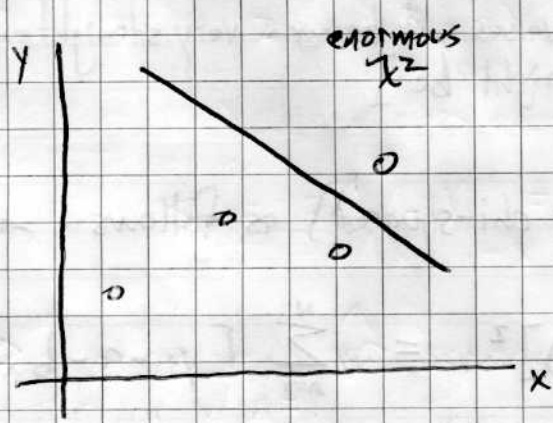
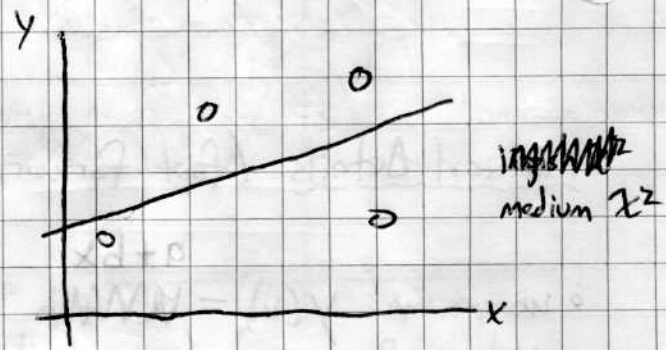
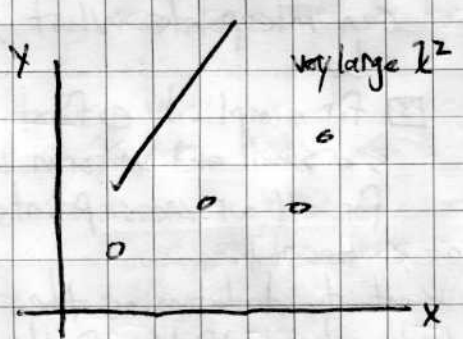
ASIDE

why do we call χ^2 the 'error' and why is it squared?

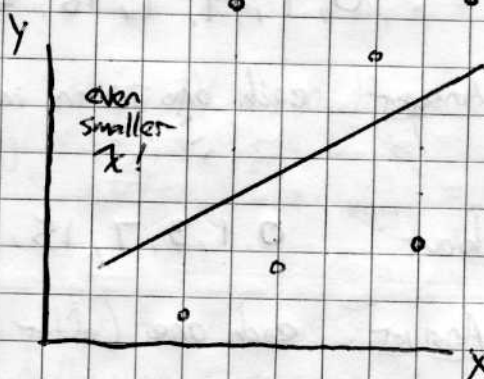


this term $y_i - a - bx_i$ is essentially the error associated with the curve $y(x)$ for a given choice of a and b re. the particular data point (x_i, y_i)

Some Further Examples



12/1/09



(or at least, the total amount of deviation is smaller)

→ obviously the left fit is better than the right, but for the right the signed errors cancelled out perfectly, leading to a wrong conclusion; squaring the 'error' by-passes this and all deviations are weighted according to the absolute value (squared)

ASIDE: well, why not use the absolute value instead of the square?
We'll see in a bit. It basically boils down to calculus: we want things to stay differentiable

IDEA: If we can minimize χ^2 w/ respect to the best values of a and b , we can determine the best fit params.!!

take a ^{partial} derivative and set equal to zero:

$$\frac{\partial}{\partial a} \chi^2 = \frac{\partial}{\partial a} \sum (y_i - a - bx_i)^2 = -2 \sum (y_i - a - bx_i) = 0$$

$$\frac{\partial}{\partial b} L^2 = \frac{\partial}{\partial b} [\quad " \quad] = -2 \sum x_i (y_i - a - bx_i) = 0$$

⇒ so while we have two ODEs, since we are looking for 'optimized' values (i.e. where the derivative is zero), all we need is to solve two equations for two unknowns!!

[cont.]

Ensayo de la Ley de la Gravedad

12/1/09 (cont. from several pgs. back)

Linear Regression (cont.)

So we have two eqns. we need to solve:

$$-2 \sum_{i=1}^N (y_i - a - bx_i) = 0$$

$$-2 \sum_{i=1}^N x_i (y_i - a - bx_i) = 0$$

where N is the total # of measurement pairs (x_i, y_i) made and a and b are the parameter values we are trying to solve for.

$\Rightarrow$ we can rearrange things a bit, using the rule that $\sum \alpha_i + \beta_i = \sum \alpha_i + \sum \beta_i$

$$\sum y_i = a + b \sum x_i$$

$$\sum x_i y_i = a \sum x_i + b \sum x_i^2$$

which is just two eqns. w/ two unknowns (a and b). Using some basic notation from linear algebra, we can solve for the two unknowns:

$$a = \frac{1}{\Delta} \begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix}$$

$\Rightarrow$ thus we have very straight-forward formulae (computationally-wise) for determining the best fit parameters!

$$b = \frac{1}{\Delta} \begin{vmatrix} 1 & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}$$

Question (as food for thought):
what if each data point had some associated uncertainty, i.e. $y_i \sim x_i \pm \sigma_i$?
 $\Rightarrow$ how would our strategy change?

$$\text{and } \Delta \equiv \begin{vmatrix} 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}$$

NOTATION-WISE

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} \equiv ad - bc \quad (\text{this is called the determinant})$$