

Can model the data as being a function of the  $i$  observations  $x_i$  and a vector  $a$  of the fitting parameters,

$$y = y(x; a_i), \text{ given some params } a_i. \\ = mx_i + c$$

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

The overall quality of the fit can be measured using chi-squared.

The residual is the difference between  $y$  and the predicted items.

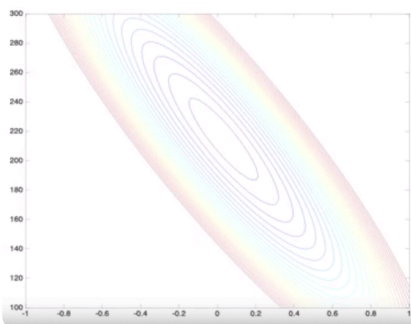
$$r_i = y_i - (mx_i + c)$$

$$\chi^2 = \sum r_i^2$$

$$= \sum (y_i - mx_i - c)^2 \cdot \text{squaring} \begin{matrix} \textcircled{1} \\ \textcircled{2} \end{matrix}$$

ensures equal penalty to positive and negative residuals  
weighted penalties to greater residuals.

The goal is to find the lowest chi-squared possible - that is the best fitting parameters. This is an optimization problem!



This is what the chi-squared looks like for many different values of  $m$  and  $c$ . The min is around  $m=215, c=0$ .

There is some relationship between  $m$  and  $c$ , as  $m$  gets bigger  $c$  must get smaller - and vice versa.

(\*) Also this function has a really shallow gradient/through the min. This makes steepest descent have to perform more computations — and is a slower process.

The minimum of the chi-squared function can be found when the chi-sq. gradient is zero. Solution is to find chi-sq. with respect to the fitting params.

This can be solved explicitly,

- can also be solved by linear descent.
- can also — by non-explicit solution.

$$\nabla x^2 = \begin{bmatrix} \frac{\partial x^2}{\partial m} \\ \frac{\partial x^2}{\partial c} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -2 \sum_i x_i (y_i - mx_i - c) \\ -2 \sum_i (y_i - mx_i - c) \end{bmatrix}$$

In manipulating the second partial derivative,  $-2 \sum_i (y_i - mx_i - c)$ , can find an expression for  $c$ ,

$$\frac{\partial x^2}{\partial c} = -2 \sum_i (y_i - mx_i - c)$$

$$0 = -2 \sum_i (y_i - mx_i) + 2NC, \text{ where } N \text{ is the number of observations.}$$

$$-2NC = -2 \sum_i y_i + 2 \sum_i mx_i$$

$$NC = \sum_i y_i - \sum_i mx_i$$

$$C = \frac{\sum_i y_i}{N} - \frac{\sum_i mx_i}{N}$$

$$C = \bar{y} - m\bar{x}, \text{ where } \bar{y} \text{ is avg } y \text{ and } \bar{x} \text{ is avg } x.$$

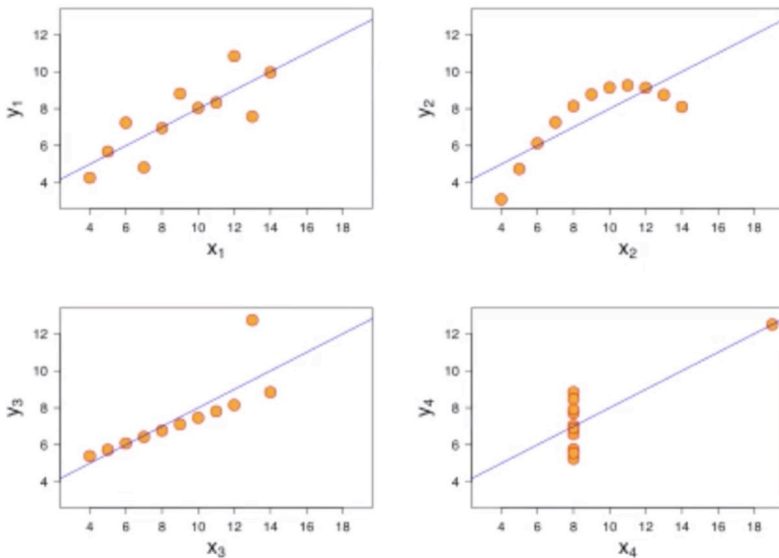
By then solving for  $m$ , can find,

$$m = \frac{\sum (x - \bar{x})y}{\sum (x - \bar{x})^2}$$

Uncertainties are,

$$\sigma_c \approx \sigma_m \sqrt{x^2 + \frac{1}{n} \sum_i (x - \bar{x})^2}$$

$$\sigma_m^2 \approx \frac{x^2}{\sum (x - \bar{x})^2 (n-2)}$$



Anscombe's quartet

It's important to validate that your model correctly expresses the nature of your data—in general that your description via descriptive stat. matches the description

The relationships of your model are important too—and if you introduce unneeded complexity

of your data.

$C = \bar{y} - m\bar{x}$ , is dependent on the  $m$  param.

Can remove this dependency by expressing observations as deviations from the centre of mass (i.e.  $\bar{x}$ ) and the intercept is the centre of mass  $\bar{y}$ . This produces a linear formulation where  $c$  is no longer afflicted by the gradient!

$$y = (m \pm \sigma_m)(x - \bar{x}) + (b \pm \sigma_b)$$

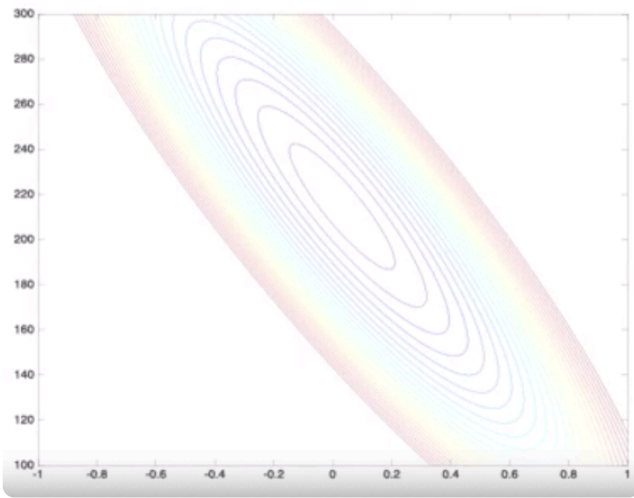
$$m = \frac{\sum (x - \bar{x})y}{\sum (x - \bar{x})^2}$$

$$b = \bar{y}$$

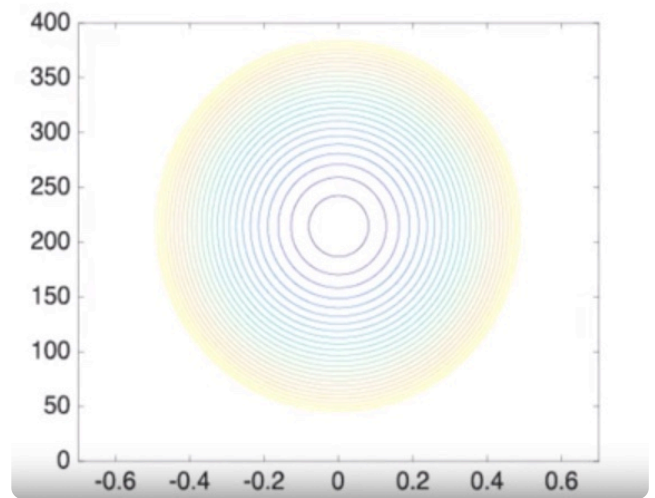
$$\sigma_m^2 \approx \frac{x^2}{\sum (x - \bar{x})^2 (n-2)}$$

$$\sigma_b^2 \approx \frac{x^2}{n(n-2)}$$

By reformulating the linear models to remove the dependencies between parameter definitions, the minimization problem simplifies.



⇒



Contour map of  $\chi^2$  test  
when  $c = \bar{y} - m\bar{x}$ , and  
deviation is based on  
origin  $x=0$ .

Contour map of  $\chi^2$  test  
when  $c = \bar{y}$  and linear  
model is based off deviation  
from centre-mass.

All of this minimization was to answer -

"how to fit a line to some data".

Chi squared measures goodness of fit by squares and deviations of fit from the data.