Fit – By Least Squares: Details about the Relationships between Variables.

The relationships between variables can be measured and quantified in many manners.

One can assess,

- The monotonicity between variables – via correlation
- The degree of congruence – via the magnitude of the correlation.
- The direction/sign of the relationship – via correlation.

The relationships between variables can be more thoroughly – that is comprehensively and succinctly – by equation; this is known as modelling.

Construeing that model is difficult and often subject to factors like bias, simplification or improper characterization, or measurement error.

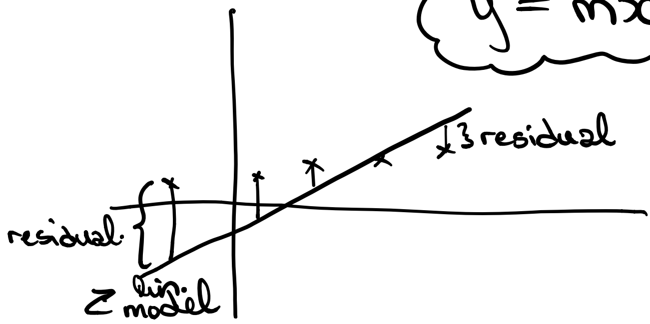
The result of this activity – that is modelling is an equation; and the correctness of that equation is called <u>fit</u>.

The incorrectness – is known as deviation or <u>residual</u>.

"that is an estimation tech?"

Currently a popular modelling technique – is Linear Least Squares.

__Linear Least Squares Fit__ is the most common way of estimating slope. __Linear Fit__ is a line meant to model the relationship between variables. __Least Squares__ is the process of minimizing the Mean Squared Error (MSE) between the line and the data. Can be represented as $\varphi(x)$; the alternative is to use interpolation.



$y = mx + b$, a linear model expressing a proportional offsetted relationship between the variables y and x.

r – the residual is the amount unaccounted for by the model

$$r = y - (mx + b)$$

This residual is the byproduct of inaccurate modelling or data quality issues, or weak correlation.

There are numerous ways to estimate a models parameters.

I. By absolute values
II. By total sum
III. By squares
IV. By cubes

The sum of the squared residuals has these benefits,

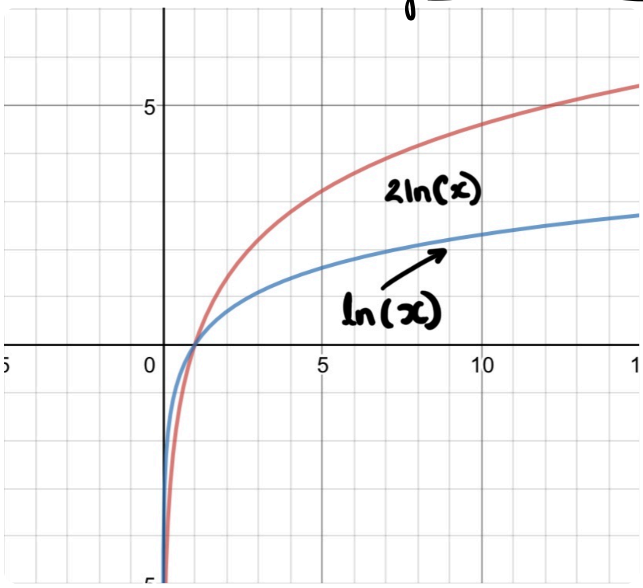
I. Squaring treats positive & negative deviations the same. They become constructive.
II. Stable weighting – expresses large variation and minimizes smaller disturbances.
III. Inexpensive computation of the linear – model calculating the slope or intercept.
IV. If residuals are,
- uncorrelated
- normally distributed
- mean "μ" of service
- constant variance

then the least squares fit is MLE.

If the fit is less important than the nature of the inaccuracy then a cost function would be more apt, than the least squares.

$$\sum \text{cost}(x) \text{ or } \sum x^2.$$

Numerical Method for Linear Least Squares

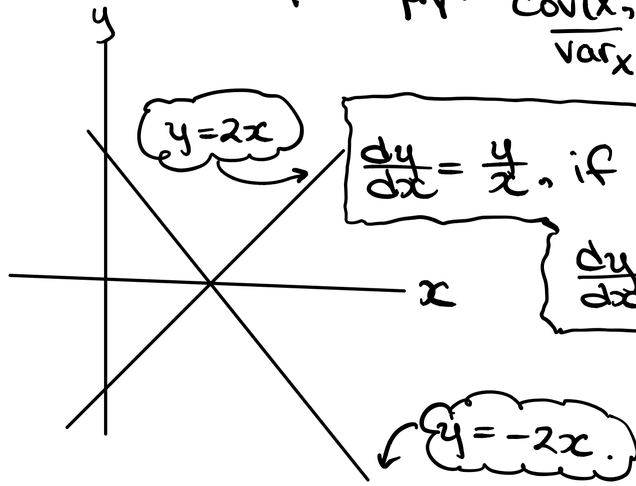


① Get samples of both vars X and Y.

② Calculate μ_X, μ_Y, var_X.

③ Use μ_X, μ_Y, var_X to determine $\dfrac{Cov(X,Y)}{var_X}$ "the slope"

④ "the intercept" $= \mu_Y - \dfrac{Cov(X,Y)}{var_X} \cdot \mu_X$.



$\dfrac{dy}{dx} = \dfrac{y}{x}$, if $Y = 2\ln x$, $X = \ln x$

$\dfrac{dy}{dx} = 2$.

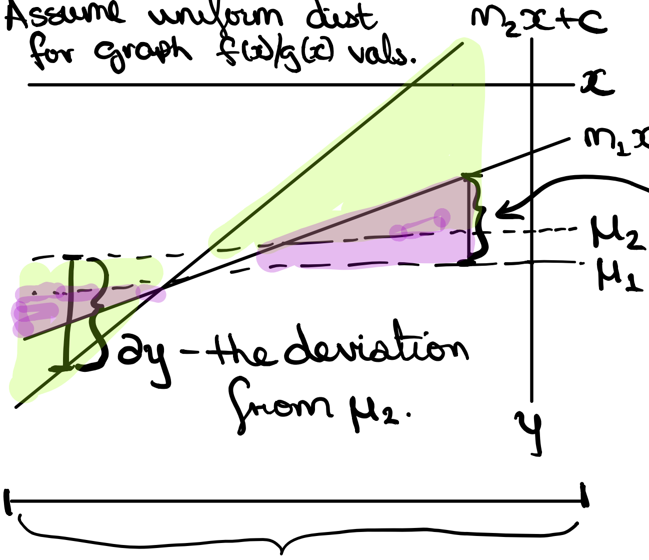
Graph#2: Linear relationships between variables

— suppose the random variables X and Y could be represented by the distributions $2\ln x$ and $\ln x$ respectively. Their relationship would be linear — with the ratio of 2:1.

Graph#3: Find the orthogonal projection

— if there exists a linear relationship between variables the orthogonal projection of X will possess the magnitude equivalent parameter.

* Assume uniform dist for graph $f(x)/g(x)$ vals.



$\|x\| = N$, the sample size.

- The deviation calculations for the variables X and Y will be deviation additive or subtractive (or both) for the range(s) of their sample.

from μ_1 ∂x

As their equations trend towards larger absolute values their means less impact the deviation calculations. At some point the Covariance is the products of X_i and Y_i; their sum standardized in terms of the sample size (N).