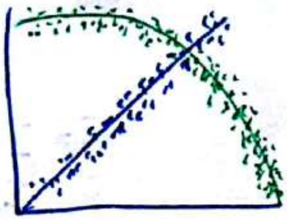
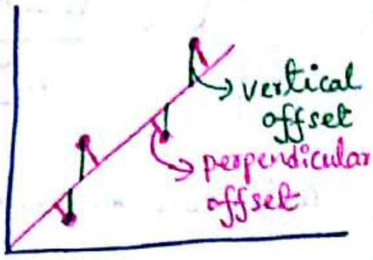
Least square fitting



To find best fitting curve to a set of point is by minimizing the sum of the squares of the offsets (residuals) of the points from the curve. As it is squared, function is continuous differentiable.

Vertical offsets from a line, surface is computed and not perpendicular offset, as former provides a fitting function for the independent variable x to estimate $y = f(x)$, easy to implement along x or y axis, & also provides simpler analytic form for fit parameters.



linear least square fit / linear regression, formulated by Gauss & Legendre solves for a straight line best fit. This also works good for simple non-linear function like log, exp, power law as one can transform to linear, e.g. $T = 2\pi\sqrt{\frac{l}{g}}$ for simple pendulum, fit T vs. $\sqrt{l}$ which is a straight line.

Vertical least square fit of n data point $R^2 = \sum_{i=1}^{n}[y_i - f(x_i, a_1, a_2, \ldots, a_n)]^2$

R^2 to minimum, $\frac{\partial R^2}{\partial a_i} = 0 \in i = 1, \ldots, n.$

for linear fit $f(a, b) = a + bx_i$, So $R^2(a,b) = \sum_{i=1}^{n}[y_i - (a+bx_i)]^2$

$$\frac{\partial R^2}{\partial a} = -2\sum_{i=1}^{n}[y_i - (a+bx_i)] = 0 \text{ or, } na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\frac{\partial R^2}{\partial b} = -2\sum_{i=1}^{n}[y_i - (a+bx_i)]x_i = 0 \text{ or, } a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

In matrix form, $$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

so $$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1}\begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

$$= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i \\ n\sum x_i y_i - \sum x_i \sum y_i \end{pmatrix}$$

So, $\boxed{a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}}$ "Regression coefficients"

$\bar{x} = \sum x_i$
$\bar{y} = \sum y_i$

This can be rewritten in simpler form by defining the sum of squares

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2 = n\sigma_x^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2 = n\sigma_y^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\bar{x}\,\bar{y}. = n\,\mathrm{cov}(x,y)$$

σ_x^2, σ_y^2 = variance, $\mathrm{cov}(x,y)$ = covariance. So,

$$b = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} = \frac{S_{xy}}{S_{xx}}, \quad a = \bar{y} - b\bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}}\bar{x}$$

The quality of the fit is parametrized in terms of correlation coefficient $r^2 = \frac{S_{xy}^2}{S_{xx}\,S_{yy}}$. If $\hat{y}_i$ is the vertical coordinate of the best fit line with coordinate x_i, $\hat{y}_i = a + bx_i$, then error between actual vertical point y_i & fitted point is $e_i = y_i - \hat{y}_i$, so that variance of e_i is defined as

$$s^2 = \sum_{i=1}^{n} \frac{e_i^2}{n-2}, \quad s = \sqrt{\frac{S_{yy} - bS_{xy}}{n-2}} = \sqrt{\frac{S_{yy} - S_{xy}^2/S_{xx}}{n-2}}$$

So that standard error for a and b is

$$a_{SE} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}, \quad b_{SE} = \frac{s}{\sqrt{S_{xx}}}.$$

Goodness of fit is calculated from coefficient of determination $R^2 = 1 - \frac{S_{residue}}{S_{total}}$

$$S_{residue} = \sum_{i=1}^{n} e_i^2, \quad S_{total} = (n-1)\sigma_y^2$$