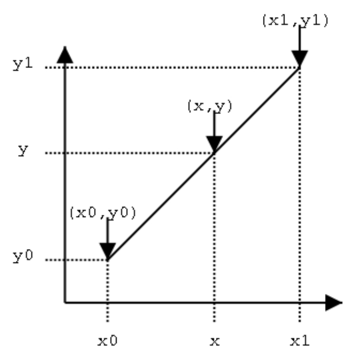
Brief note on interpolation and least square

It often happens in experiments that we know the values of a function $f(x_i)$ at a bunch of points x_i, not necessarily equally spaced, but do not quite know the analytical expression for $f(x)$. Such an analytical form lets us calculate its value $f(x)$ at arbitrary intermediate points $x \neq x_i$. Therefore, the job is to draw a smooth curve or polynomial through the x_i and approximate the form of $f(x)$. If the desired x at which we want to know $f(x)$ is between the largest and smallest of the x_i's, then the problem is called *interpolation*. If outside, then it is *extrapolation* but that is a risky endeavor unless some solid theoretical idea backs that up.

To begin with, let us start with simplest of the cases – to interpolate (x, y) between two points (x_0, y_0) and (x_1, y_1) with a straight line.



The equation of straight line gives,

$$\begin{aligned} \frac{y - y_0}{x - x_0} &= \frac{y_1 - y_0}{x_1 - x_0} \\ y &= y_0 + \frac{x - x_0}{x_1 - x_0}(y_1 - y_0) \end{aligned} \tag{1}$$

Let us generalize this idea by considering a polynomial of degree N to approximate a function from a set of $N + 1$ points, such that

$$P_N(x_i) = f(x_i) = y_i \quad \text{where} \quad i = 0, 1, \ldots, N \tag{2}$$

The general form of the polynomial $P_N(x)$ can be assumed to be

$$P_N(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \cdots + a_N(x-x_0)(x-x_2)\ldots(x-x_{N-1}) \equiv y(x) \tag{3}$$

$$P_1(x_0) = a_0 = y_0 = f(x_0) \tag{4}$$

$$P_2(x_1) = a_0 + a_1(x_1-x_0) = y_1 = f(x_1) \tag{5}$$

$$P_3(x_2) = a_0 + a_1(x_2-x_0) + a_2(x_2-x_0)(x_2-x_1) = y_2 = f(x_2) \quad \text{etc.} \tag{6}$$

If we want linear interpolation, then $n = 1$ and it follows from (4) and (5) that

$$a_0 = y_0 \quad \text{and} \quad a_1 = \frac{y_1 - y_0}{x_1 - x_0} \quad \Rightarrow \quad P_1(x) = y(x) = y_0 + \frac{y_1 - y_0}{x_1 - x_0}\,(x - x_0) \tag{7}$$

which is the same that we obtained in (1). Let us rewrite the $P_1(x)$ in (7) as

$$\begin{aligned} P_1(x) &= y_0 - \frac{x-x_0}{x_1-x_0}\,y_0 + \frac{x-x_0}{x_1-x_0}\,y_1 \\ &= \frac{x-x_1}{x_0-x_1}\,y_0 + \frac{x-x_0}{x_1-x_0}\,y_1 \end{aligned} \tag{8}$$

As an exercise, show

$$P_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}\,y_0 + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}\,y_1 + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}\,y_2 \tag{9}$$

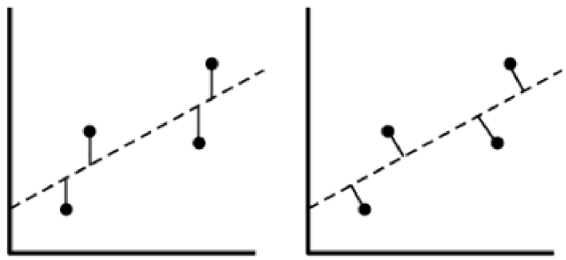
This leads to the classic interpolation formula by Lagrange,

$$P_N(x) = \sum_{i=0}^{N} \prod_{k \neq i} \frac{x - x_k}{x_i - x_k}\,y_i \tag{10}$$

To determine the coefficients of y_i's through an iterative process see any textbooks.

Least Square fitting

Suppose the analytical form of the function $f(x)$ is known, unlike the previous interpolation case, and a set of data points x_i is generated that are supposed to be described by $f(x)$. For instance, in Lab we generate stress-strain data points (x_i, y_i) and we know that within elastic limit the linear relation stress *propto* strain holds. So the task is to calculate the modulus of elasticity using these data points. One of the numerical ways to do this is Least Square fitting and for this we need *sum of squares of the offset (residuals) of the points from the $f(x)$ curve.* Offsets can be either vertical and perpendicular.



The dashed line is the fit or model function which we claim to know. The vertical offsets are most often used as they allow uncertainties of data points along x and y axis to be independent of each other and lead to simpler analytical form for tunable fitting parameters. However, for *good* data, the difference is small.

But why squares, and not absolute values of the offset are used for fitting? One of the reasons being, squares permits offsets or residuals to be treated as continuous differentiable quantities (although unsquared sum of distance seems more appropriate but the absolute values result in discontinuous derivatives). One apparent problem with squares is disproportionate contributions of the outliers to the fitting parameters, which may or may not be desirable.

Suppose we are fitting N data points (x_i, y_i) where $i = 1, \ldots, N$ to a model function $f(x_i, a_1, a_2, \ldots, a_n)$ where a_j's are $nleN$ are adjustable parameters. The model predicts a functional relationship between the measured independent and dependent variables. To determine the a_j's, Least Square suggests to minimize the square of the vertical offset R^2 w.r.t. the a_j's,

$$R^2 = \sum_{i=1}^{N} \Big[y_i - f(x_i, a_1, a_2, \ldots, a_n) \Big]^2 \quad \rightarrow \quad \frac{\partial R^2}{\partial a_j} = 0 \tag{11}$$

Often a weight function w_i is added in the definition of R^2,

$$R^2 = \sum_{i=1}^{N} w_i \Big[y_i - f(x_i, a_1, a_2, \ldots, a_n) \Big]^2 \tag{12}$$

where w_i addresses the problem of large contributions of the outliers. When $w_i = 1/\sigma_i^2$ then we have $R^2 = \chi^2$ fitting. We usually call fitting to be *uncorrelated* when $w_i = 1$ or $1/\sigma_i^2$ and *correlated* when we use *covariance matrix* instead $w_{ij} = \sigma_{ij}^2$.

Suppose, the model function is a straight line $f(x) = a_1 + a_2 x$ where a_1 and a_2 are the parameters we wish to find in order to describe the data points well. Then

$$
\begin{aligned}
R^2 &= \sum_{i=1}^{N} \Big[y_i - (a_1 + a_2 x_i) \Big]^2 \\
\frac{\partial R^2}{\partial a_1} &= -2 \sum_i \Big[y_i - (a_1 + a_2 x_i) \Big] = 0 && (13) \\
\frac{\partial R^2}{\partial a_2} &= -2 \sum_i x_i \Big[y_i - (a_1 + a_2 x_i) \Big] = 0 && (14)
\end{aligned}
$$

Noting that $\sum_i a_1 = N a_1$, from above we obtain

$$
\begin{aligned}
a_1 N + a_2 \sum_i x_i &= \sum_i y_i \\
a_1 \sum_i x_i + a_2 \sum_i x_i^2 &= \sum_i x_i y_i && (15)
\end{aligned}
$$

The solution of the above equations is straight forward,

$$
a_1 = \frac{\overline{y} \sum_i x_i^2 - \overline{x} \sum_i x_i y_i}{\sum_i x_i^2 - N \overline{x}^2} \qquad a_2 = \frac{\sum_i x_i y_i - N \overline{xy}}{\sum_i x_i^2 - N \overline{x}^2} \tag{16}
$$

These solutions can be cast into much simpler form that follows from introducing the following notations,

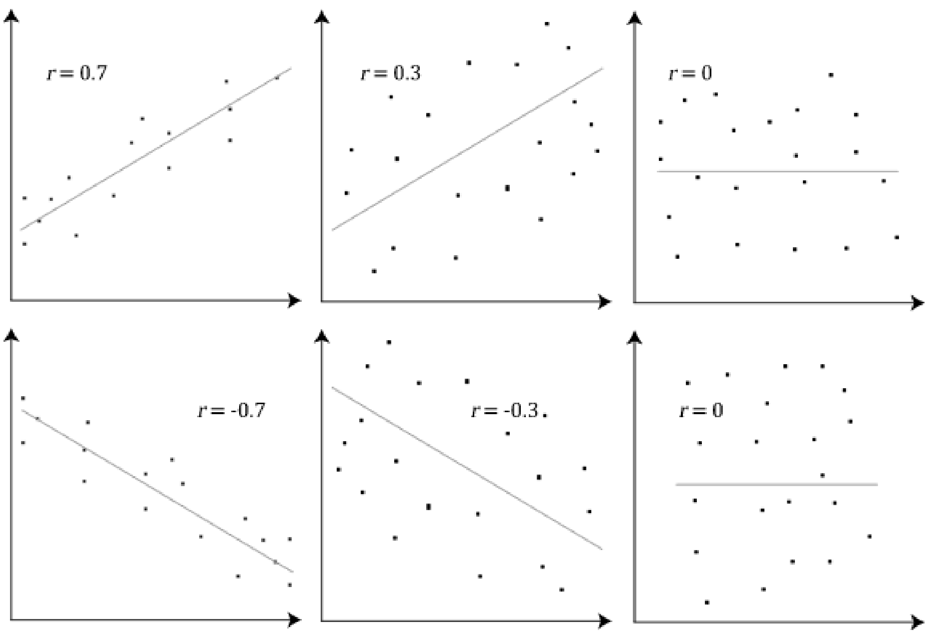
$$
\begin{aligned}
S_{xx} &= \sum_i \Big(x_i - \overline{x} \Big)^2 = \sum_i x_i^2 - N \overline{x}^2 && (17) \\
S_{yy} &= \sum_i \Big(y_i - \overline{y} \Big)^2 = \sum_i y_i^2 - N \overline{y}^2 && (18) \\
S_{xy} &= \sum_i \Big(x_i - \overline{x} \Big) \Big(y_i - \overline{y} \Big) = \sum_i x_i y_i - N \overline{xy} && (19) \\
\sigma_x^2 &= \frac{S_{xx}}{N}, \quad \sigma_y^2 = \frac{S_{yy}}{N} \ \text{ and } \ \mathrm{cov}(x, y) = \frac{S_{xy}}{N} && (20)
\end{aligned}
$$

Hence, a_1 and a_2 become

$$
a_2 = \frac{\mathrm{cov}(x, y)}{\sigma_x^2} = \frac{S_{xy}}{S_{xx}}, \qquad a_2 = \overline{y} - a_2 \overline{x} \tag{21}
$$

Once we know a_1 and a_2, we can fit the data to $a_1 + a_2 x$ but how good is this *linear* description or correlation? This quality is judged by *Pearson correlation coefficient* r where $-1 \leq r \leq 1$. Better the fit, closer $|r|$ is to 1. It defined as,

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{22}$$



What if the (x_i, y_i) is anything but linear? The above method works for any expressions that can be written in straight line form such as *exponential*, *logarithmic* and *power law* as discussed below.

- **exponential** : The fitting function is $f(x)$ and the corresponding straight line form is
$$y = f(x) = a\, e^{bx} \;\;\Rightarrow\;\; \ln y = \ln a + bx \;\;\equiv\;\; Y = A + bx \tag{23}$$

- **logarithmic** : The fitting function is $f(x)$ and the corresponding straight line form is
$$y = f(x) = a + b\, \ln x \;\;\equiv\;\; y = a + bX \tag{24}$$

- **power law** : The fitting function is $f(x)$ and the corresponding straight line form is
$$y = f(x) = a\, x^b \;\;\Rightarrow\;\; \ln y = \ln a + b\, \ln x \;\;\equiv\;\; Y = A + bX \tag{25}$$

Least square polynomial fitting : Polynomial fitting is essentially a generalization of a straight line fitting [see Wolfram maths]. Consider a k-th degree polynomial and the residual to be,

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k \quad \text{and} \quad R^2 = \sum_{i=1}^{n} \Big[y_i - \big(a_0 + a_1 x_i + \cdots + a_k x_i^k \big) \Big]^2 \tag{26}$$

The minimization of R^2 with respect to a_i's leads to the following set of equations,

$$\frac{\partial R^2}{\partial a_0} = -2 \sum \Big[y_i - \big(a_0 + a_1 x_i + \cdots + a_k x_i^k \big) \Big] = 0 \tag{27}$$

$$\frac{\partial R^2}{\partial a_1} = -2 \sum \Big[y_i - (a_0 + a_1 x_i + \cdots + a_k x k_i) \Big] x_i = 0 \tag{28}$$

$$\vdots$$

$$\frac{\partial R^2}{\partial a_k} = -2 \sum \Big[y_i - (a_0 + a_1 x_i + \cdots + a_k x k_i) \Big] x_i^k = 0 \tag{29}$$

The set of equations for a_i's are,

$$a_0 n + a_1 \sum x_i + \cdots + a_k \sum x_i^k = \sum y_i \tag{30}$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + \cdots + a_k \sum x_i^{k+1} = \sum x_i y_i \tag{31}$$

$$\vdots$$

$$a_0 \sum x_i^k + a_1 \sum x_i^{k+1} + \cdots + a_k \sum x_i^{2k} = \sum x_i^k y_i \tag{32}$$

which when written in matrix form appears as,

$$\begin{pmatrix} n & \sum x_i & \cdots & \sum x_i^k \\ \sum x_i & \sum x_i^2 & \cdots & \sum x_i^{k+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_i^k & \sum x_i^{k+1} & \cdots & \sum x_i^{2k} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_k \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^k y_i \end{pmatrix} \quad \Rightarrow \quad \mathbf{X} \cdot \mathbf{a} = \mathbf{Y} \tag{33}$$

Assuming the inverse of the above matrix $\mathbf{X}$ exists, implying that the data can be fitted with a k-degree polynomial, then the coefficients a_i's follow from solving the above matrix equation in (33),

$$\mathbf{a} = \mathbf{X}^{-1} \mathbf{Y} \tag{34}$$

Setting $k = 1$ in the above, the equations reduce to the linear solution.