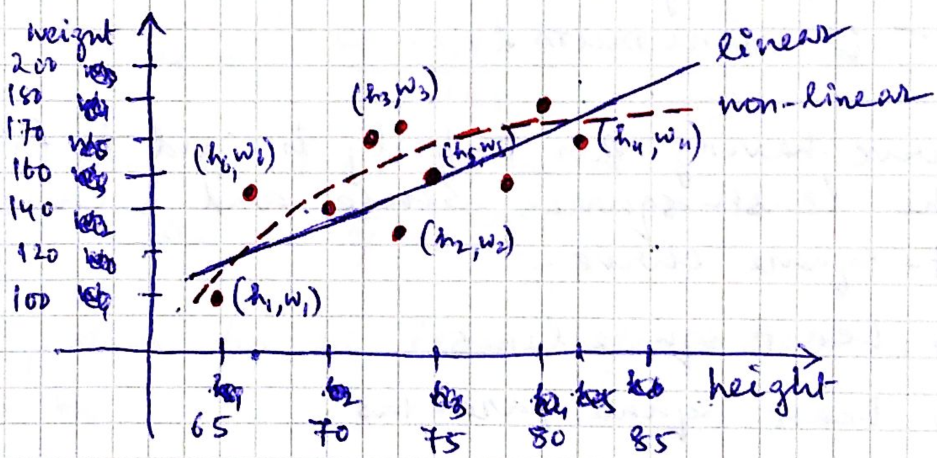
Curve fitting and the Method of Least squares.

The main objective of many statistical investigations is to express this relationships in mathematical form by determining an equation that connects the variables. This leads us to the Concept of curve fitting.

Suppose we have a given data, for example; consider the height and weight of adult males; then a sample of N individuals would have a data of the form (h_i, w_i), $i = 1, 2, \cdots, N$. If we plot those data on a rectangular co-ordinate system, we will get a scattered diagram as given below.



From the diagram it is often possible to visualize a smooth curve that approximates the data. Such a curve is called an approximate curve. This curve may ~~could~~ be linear or non-linear. Our objective here will be to find the best fitted curve.

Common types of approximating curves for a data (X, Y).

(i) Straight line: $Y = a_0 + a_1 X$.

(ii) Parabola: $Y = a_0 + a_1 X + a_2 X^2$

(iii) n-th degree polynomial: $Y = a_0 + a_1 X + \cdots + a_n X^n$

(iv) Exponential: $Y = ab^X$

(v) Logistic: $Y^{-1} = ab^X + g$

etc.,

Method of Least squares.

Suppose we have a given set of data point (x_i, y_i), $i=1,2,\cdots,N$. and which is approximated by a curve $y = f(x)$ where $f(x)$ can take any of the form as discussed earlier. Therefore, $y_i^c = f(x_i)$ are the approximate values of the y_i, obtained from the curve.

Suppose $D_i = (y_i - y_i^c)$, $i=1,2,\cdots,N$. denotes the error / residual or deviation between the actual and approximate data.

* A measure of "goodness of fit" of the curve $y=f(x)$ is the quantity $(D_1^2 + D_2^2 + \cdots D_N^2)$. If this is small, then the fit is good; if it is large, then the fit is bad.

* The best-fitting curve is the curve for which $\sum_{i=1}^{N} D_i^2$ is minimum.

* A curve having this property is said to fit the data in the least-squares sense and is called a least-square curve.

  - Least square lines.
  - Least square parabolas.

Here it is to be noted that

$x \rightarrow$ independent variable.
$y \rightarrow$ dependent variable.

Least square lines

The least square-line approximating the set of points (x_i, y_i), $i=1,2,\cdots,N$, has the equation

$$y = a_0 + a_1 x \quad \text{——} \quad (1)$$

where a_0, a_1 are constants to be determined from the equations

$$\left.\begin{aligned} \sum y_i &= a_0 N + a_1 \sum x_i \\ \sum x_i y_i &= a_0 \sum x_i + a_1 \sum x_i^2 \end{aligned}\right\} \text{—} (2)$$

which are called the normal equations for the least-squares lines.

From ②

$$\frac{1}{N}\sum y_i = a_0 + a_1 \frac{1}{N}\sum x_i$$

$$\frac{1}{N}\sum x_i y_i = a_0 \frac{1}{N}\sum x_i + a_1 \frac{1}{N}\sum x_i^2$$

$$\Rightarrow \quad \bar{y} = a_0 + a_1 \bar{x}$$

$$\frac{1}{N}\sum x_i y_i = a_0 \bar{x} + a_1 \frac{1}{N}\sum x_i^2$$

Multiplying the above eqn. by $\bar{x}$ and substracting from the second gives.

$$\left(\frac{1}{N}\sum x_i y_i - \bar{x}\bar{y}\right) = a_1\left(\frac{1}{N}\sum x_i^2 - \bar{x}^2\right)$$

$$Cov(x, y) = a_1 \, Var(x)$$

$$\Rightarrow \quad a_1 = \frac{Cov(x,y)}{Var(x)} = \frac{Cov(x,y)}{\sigma_x^2}$$

Then from the first equation

$$a_0 = \bar{y} - \frac{Cov(x,y)}{\sigma_x^2}\bar{x}$$

The equation for the least-square line is then given by

$$y = \bar{y} + \frac{Cov(x,y)}{\sigma_x^2}(x - \bar{x})$$

$$\text{or, } (\hat{y} - \bar{y}) = \frac{Cov(x,y)}{\sigma_x^2}(x - \bar{x})$$

$$\text{or, } \boxed{(y - \bar{y}) = r\frac{\sigma_y}{\sigma_x}(x - \bar{x})} \quad \text{— ③}$$

where $\boxed{r = \frac{Cov(x,y)}{\sigma_x \sigma_y}}$ is called the correlation coefficient.

Eqn. ③ is called as linear regression of y on x.

Similarly one can have the linear regression of x on y as

$$(x - \bar{x}) = r\frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad \text{— ④}$$

* Equation ③ and ④ are equal if and only if $r = \pm 1$. In such a case two lines are identical and there is perfect linear correlation between X and Y.
* If $r = 0$, the lines are at right angles and there is no linear correlation between X and Y.
* The value of r lies between

$$-1 \leq r \leq 1$$

* For data grouped as in a bivariate frequency table, then

$$r = \frac{Cov(X,Y)}{\sigma_x \, \sigma_y}$$

where

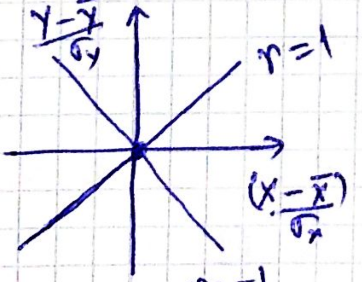
$$Cov(X,Y) = \frac{1}{N} \sum_{i,j} f_{ij} x_i y_j - \bar{X}\bar{Y}$$

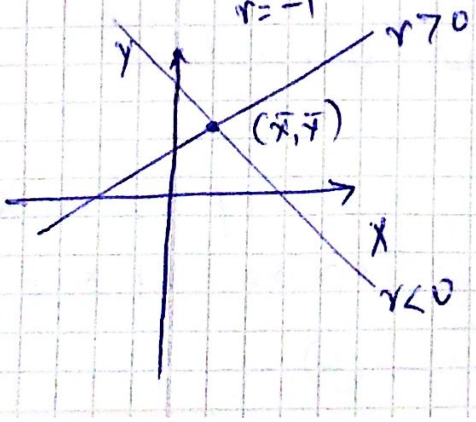
$$\sigma_x^2 = \frac{1}{N} \sum_i f_{i\circ} x_i^2 - \bar{x}^2$$

$$\sigma_y^2 = \frac{1}{N} \sum_j f_{\circ j} y_j^2 - \bar{y}^2$$

where $f_{i\circ}$ and $f_{\circ j}$ are marginal distributions for X and Y respectively.

* When $r > 0$, we call it positive correlation and when $r < 0$, we call it negative correlation.





$\sigma_x > 0$

$\sigma_y > 0$

when $r = \pm 1$

$$\frac{Y - \bar{Y}}{\sigma_y} = \pm \frac{X - \bar{X}}{\sigma_x}.$$

$$\frac{Y - \bar{Y}}{\sigma_y} = r \frac{X - \bar{X}}{\sigma_x}.$$

$$Y = r\frac{\sigma_y}{\sigma_x} X + \left(\bar{y} - r\frac{\sigma_y}{\sigma_x}\bar{x}\right)$$

$$Y = mX + c$$

If $r > 0$, $m > 0$

$r < 0$, $m < 0$