Linear Correlation

Linear curve-fitting

- If two set of data points have roughly a linear dependency, a least-square-fit line roughly lies between the scatter region.
- If the least-square line has equation y = mx + c, m and c can be found from the data points
- If there are N data points, then
- $$\sum_{i=1}^{N} y_i = m.\sum_{i=1}^{N} x_i + N.c$$
- $$\sum_{i=1}^{N} x_i . y_i = m.\sum_{i=1}^{N} x_i^2 + c.\sum_{i=1}^{N} x_i$$

Linear curve fitting (contd ...)

- Multiplying first equation by $\sum_{i=1}^{N} x_i$,
- second equation by N, we get
- $N.\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i . \sum_{i=1}^{N} y_i = m.(N.\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2)$
- Hence m = $\dfrac{\sum_{i=1}^{N} x_i y_i - N.\overline{x}.\overline{y}}{\overline{x^2} - N\,\overline{x}^2}$
-

Linear Correlation Coefficient

- Let us define
- $$SS_{xx}=\sum_{i=1}^{N}(x_i-\overline{x})^2=\sum_{i=1}^{N}x_i{}^2-N.(\overline{x})^2$$

$$SS_{yy}=\sum_{i=1}^{N}(y_i-\overline{y})^2=\sum_{i=1}^{N}y_i{}^2-N.(\overline{y})^2$$

$$SS_{xy}=\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})=\sum_{i=1}^{N}x_i y_i-N.(\overline{x}\,\overline{y})$$

Linear Correlation Coefficient (contd)

- Now,
- $SS_{xx} = N\, var(x)$
- $SS_{yy} = N\, var(y)$
- $SS_{xy} = N\, cov(x, y)$
-
- and $m = \frac{SS_{xy}}{SS_{xx}}$
- If the axes are altered; i.e x = b y + d, then
- $b = \frac{SS_{xy}}{SS_{yy}}$