

1. correlation coefficient — detect a data trend

data points $\{x_i, y_i\}$ $i=1, \dots, n$

correlation coefficient:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$

$$r_{xy} \in [-1, 1]$$

$r_{xy} = +1$: perfectly linear relationship with a positive slope

-1 :  negative

0 : no linear correlation

Determine whether a given correlation is significant:

A. compute r_{xy}

B. with given size of data points n , and required confidence level

(then the level of significance α), determine $r_{\alpha} = r_{\alpha}(n, \alpha)$

— the chance for $r_{xy} > r_{\alpha}$ due to pure chance is α

— if $r_{xy} > r_{\alpha}$, the experimental value shows a real correlation with confidence level of $1 - \alpha$.

Least-square linear fit

Give measured data

$$\{x_i, y_i\} \quad i=1, \dots, n$$

seek to fit a straight line

$$z = ax + b$$

such that the line best representative to the data points

Solution: for each value of x_i , the fitted value has an error

$$e_i = z_i - y_i$$

the square of the error is

$$e_i^2 = (z_i - y_i)^2 = (ax_i + b - y_i)^2$$

the sum of the squared error is

$$E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (ax_i + b - y_i)^2$$

Minimizing E w.r.t a and b requires $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^n 2(ax_i + b - y_i) \cdot x_i = \sum_{i=1}^n 2x_i^2 a + 2x_i b - 2x_i y_i = 0$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^n 2(ax_i + b - y_i) \cdot 1 = \sum_{i=1}^n 2x_i a + 2b - 2y_i = 0$$

$$\frac{\partial E}{\partial a} = \left(\sum_{i=1}^n x_i^2 \right) a + \left(\sum_{i=1}^n x_i \right) b - \sum_{i=1}^n x_i y_i = 0 \quad \times n$$

$$\frac{\partial E}{\partial b} = \left(\sum_{i=1}^n x_i \right) a + n \cdot b - \sum_{i=1}^n y_i = 0 \quad \times \sum_{i=1}^n x_i$$

$$\Rightarrow n \left(\sum_{i=1}^n x_i^2 \right) a + n \left(\sum_{i=1}^n x_i \right) b - n \sum_{i=1}^n x_i y_i = 0$$

$$\left(\sum_{i=1}^n x_i \right)^2 a + n \left(\sum_{i=1}^n x_i \right) b - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right) = 0$$

$$\left[n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2 \right] a = n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

$$\Rightarrow a = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

$$b = \frac{\left(\sum_{i=1}^n x_i^2 \right) \left(\sum_{i=1}^n y_i \right) - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right)}{n \left(\sum_{i=1}^n x_i^2 \right) - \left(\sum_{i=1}^n x_i \right)^2}$$

Force the linear fitting goes through origin:

$$y = ax$$

special case for $b=0$

$$\frac{\partial E}{\partial a} = 0 \Rightarrow \left(\sum_{i=1}^n x_i^2 \right) a - \sum_{i=1}^n x_i y_i = 0$$

$$\Rightarrow a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

Fitting by Data Transformation

(1) logarithmic dependence $z = a e^{bx}$

$$\ln z = \ln a + bx$$

(2) polynomial regression

$$z = a_0 + a_1 x + a_2 x^2 + \dots + a_k x^k$$

Define residual squared

$$e_i^2 = (z_i - y_i)^2 = (a_0 + a_1 x_i + a_2 x_i^2 + \dots + a_k x_i^k - y_i)^2$$

$$E = \sum_{i=1}^n e_i^2$$

Determine the fitting coefficients, a_0, \dots, a_k by

$$\partial E / \partial a_i = 0 \quad i = 0, \dots, k$$

Many software can do this.

How well of the fitting?

Adequency of the fitting can be measured by "coefficient of determination"

$$r^2 = 1 - \frac{\sum (z_i - y_i)^2}{\sum (y_i - \bar{y}_i)^2} \quad \text{Good: } r \rightarrow 1$$

Another measure of How well of the fitting: "standard error of estimate"

$$S_{y,x} = \sqrt{\frac{\sum y_i^2 - b \sum z_i y_i - a \sum x_i y_i}{n-2}}$$

Data rejection criterion: modified Thompson τ technique

$$\bar{x} = x_1, x_2, \dots, x_n$$

determine: $\max \{ \delta_i \}$ $\delta_i = |x_i - \bar{x}|$

compare δ_i and τs (s : standard deviation
 τ : from the table)

if $\delta > \tau s$ discard

$\delta < \tau s$ keep

Only one data point is rejected at a time, repeatedly following the procedure to reject more points

Note: Thompson τ technique is applicable for rejecting bad data from a number of measurements of a single variable

Outliers in x - y Data sets

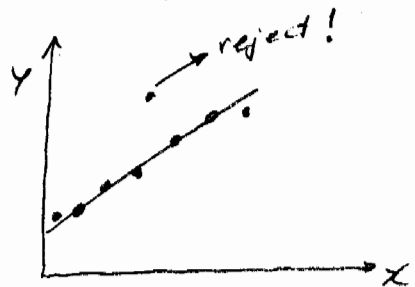
y is measured as a function of an independent variable x

— insufficient data to determine s at each x , thus

Thompson τ technique is not applicable

Data rejection criteria:

- (1) plot the data and the best-fit curve, reject data that have large deviation from the curve



(2) standard residual criterion

$$e_i = z_i - y_i = (ax_i + b) - y_i$$

$$S_r^i = e_i / S_{y,x}$$

Assume: normal distribution
of the residuals

Expect 95% of the standard
residual fall in $[-2\sigma, +2\sigma]$

reject any data fall beyond
the region

