Linear Regression:

The simplest example of a least squares approximation is fitting a straight line to a set of paired observations:

$$(x_1, y_1), (x_2, y_2) \ldots . (x_n, y_n)$$

The mathematical expression for the straight line is:

$$y = \overset{\text{intercept}}{a_0} + \overset{\text{slop}}{a_1} x + \overset{\text{error}}{e}$$

where a_0 and a_1 are coefficients representing the intercept and the slope respectively. e is the error between the model and observation.

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

Criteria for best fit:

For fitting a best line through the data would be to minimize the sum of the residual errors for all the available data.

Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date :......./......./.........

$\sum a_0 = na_0$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i) \quad \text{--- (1)}$$

Where, n = total no. of points.

$$\sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - a_0 - a_1 x_i|$$

Let,

$$\underbrace{S_r = \sum_{i=1}^{n} e_i^2}_{} = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2 \quad \text{--- (2)}$$

$$\frac{\delta S_r}{\delta a_0} = -2 \sum (y_i - a_0 - a_1 x_i) \quad \text{--- (3)}$$

$$\frac{\delta S_r}{\delta a_1} = -2 \sum (y_i - a_0 - a_1 x_i) x_i \quad \text{--- (4)}$$

derivaties equal to zero will result in a minimum S_r.

(3) $\Rightarrow$ $0 = \sum y_i - \sum a_0 - \sum a_1 x_i$ --- (5)

(4) $\Rightarrow$ $0 = \sum y_i x_i - \sum a_0 x_i - \sum a_1 x_i^2$ --- (6)

From equation (5),

$$n a_0 + (\sum x_i) a_1 = \sum y_i \quad \text{--- (7)}$$

From equation (6),

$$(\sum x_i) a_0 + (\sum x_i^2) a_1 = \sum x_i y_i$$

Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date :......../......./.........

From (6) & (7)

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$\frac{\sum sum}{n} = \overline{sum}$$

From equation (7)

$$na_0 + (\sum x_i)\, a_1 = \sum y_i$$

or, $na_0 = \sum y_i - (\sum x_i) a_1$

or, $a_0 = \frac{\sum y_i}{n} - \frac{(\sum x_i) a_1}{n}$

$$\therefore a_0 = \bar{y} - a_1 \bar{x}$$

▶ Find out the value of a_0 and a_1 in the case of least square regression.

Example 17.1:

Fit a straight line to the x and y value in the first two columns of the table. Computations for an error analysis of the linear fit.

Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date : সাহিত্য করতে হবে।

Given value

x_i	y_i	$(y_i-\bar{y})$	$(y_i-a_0-a_1x_i)^2$
1	0.5	8.5765	0.1687
2	2.5	0.8622	0.5625
3	2.0	2.0408	0.3473
4	4.0	0.3265	0.3265
5	3.5	0.0051	0.5896
6	6.0	6.6122	0.7972
7	5.5	4.2908	0.1993
28	24.0	22.7143	2.9911

Solution:

The following quantities can be computed

$n=7$ $\quad \sum x_i y_i = 119.5$ $\quad \sum x_i^2 = 140$

$\sum x_i = 28$ $\quad \bar{x} = \frac{28}{7} = 4$

$\sum y_i = 24$ $\quad \bar{y} = \frac{24}{7} = 3.428571$

Now, $$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$= \frac{7(119.5) - 28(24)}{7(140) - (28)^2} = 0.8392857$$

$$a_0 = \bar{y} - a_1\bar{x}$$

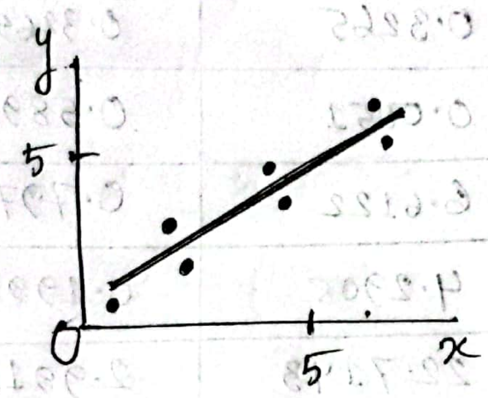
$$= 3.428571 - 0.8392857 \times 4$$

$$= 0.07142857$$

Therefore, the least square fit is :

$$y = a_0 + a_1 x$$

$$\Rightarrow y = 0.07142857 + 0.8392857\, x$$



This line along with the data shown in this figure.

Estimate of ERRORS for the linear least square fit:

Standard deviation:

$$S_y = \sqrt{\frac{22.7143}{7-1}} = 1.9457 \qquad \boxed{S_y = \frac{y_i - \bar{y}}{n-1}}$$

Standard error:

$$S_{y/x} = \sqrt{\frac{2.9911}{7-2}} = 0.7735 \qquad \boxed{S_{y/x} = \sqrt{\frac{S_r}{n-2}}}$$

$$r^2 = \frac{22.7143 - 2.9911}{22.7143} = 0.868 \qquad \boxed{r^2 = \frac{S_t - S_r}{S_t}}$$

$$r = \sqrt{0.868} = 0.93$$

$r^2 \rightarrow$ Coefficient of determination

$r \rightarrow$ Correlation coefficient

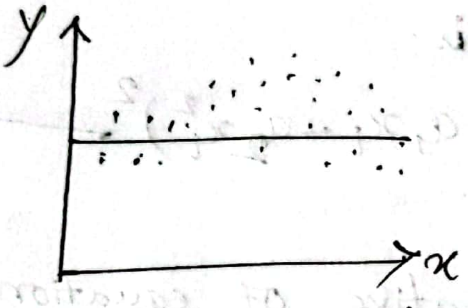
Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date :......../......./.........

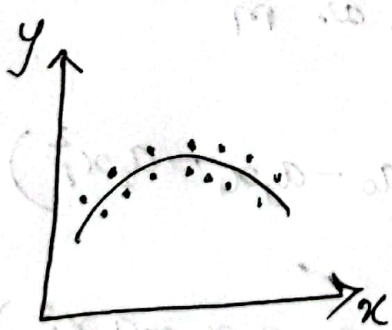
These result indicate that 86.8 percent of the original uncertainity has been explained by the linear model.

polinomial Regression:

The process of going back to an earlier or less advanced from an state.



ill-suited data for linear regression



parabola is preferable

Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date :......../......../.........

The least squares procedure can be readily extended to fit the data to a higher order polinomial.

For example, Lets suppose that we fit a second order polynomial or quadratic.

$$y = a_0 + a_1 x + a_2 x^2 + e$$

For this case the sums of the squares of the residuals is:

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2 \quad \text{--- (1)}$$

Taking the derivative of equation (1) w.r.t. each of the unknown coefficients of the polynomial, as in

$$\frac{\delta S_r}{\delta a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\delta S_r}{\delta a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\delta S_r}{\delta a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

These equations can be set equal to zero and rearranged to developed the following set of normal equations:

$$n a_0 + \left(\Sigma x_i\right) a_1 + \left(\Sigma x_i^2\right) a_2 = \Sigma y_i$$

$$\left(\Sigma x_i\right) a_0 + \left(\Sigma x_i^2\right) a_1 + \left(\Sigma x_i^3\right) a_2 = \Sigma x_i y_i$$

$$\left(\Sigma x_i^2\right) a_0 + \left(\Sigma x_i^3\right) a_1 + \left(\Sigma x_i^4\right) a_2 = \Sigma x_i^2 y_i$$

Standard error is formulated as:

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

[P.T.O]

Example-17.5:

Fit a second order polinomial to the data in the first two columns of the data.

x_i	y_i	$(y_i-\bar{y})^2$	$(y_i-a_0-a_1x_i-a_2x_i^2)^2$
0	2.1	544.44	0.14332
1	7.7	314.47	1.00286
2	13.6	140.03	1.08158
3	27.2	3.12	0.80491
4	40.9	239.22	0.61951
5	61.1	1272.11	0.09439
15	152.6	2513.39	3.74657

Solution:

From the given data,

order, $m = 2$

$n = 6$

$\bar{x} = 2.5$

$\bar{y} = 25.433$

$\Sigma x_i = 15$

$\Sigma y_i = 152.6$

$\Sigma x_i^2 = 55$

$\Sigma x_i^3 = 255$

$\Sigma x_i^4 = 979$

$\Sigma x_i y_i = 585.6$

$\Sigma x_i^2 y_i = 2488.8$

Therefore simultaneous linear equation are-

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 152.6 \\ 585.6 \\ 2488.8 \end{bmatrix}$$

using gauss elimination we get,

$a_0 = 2.47857$

$a_1 = 2.35929$

$a_2 = 1.86071$

The least square quadratic equation for this case is,

$$y = a_0 + a_1 x + a_2 x^2$$

$$\Rightarrow y = 2.47857 + 2.35929\, x + 1.86071\, x^2$$

The standard error of the estimate based on the regression polinomial is

$$S_{y/x} = \sqrt{\frac{3.74657}{6-3}}$$

$$= 1.12$$

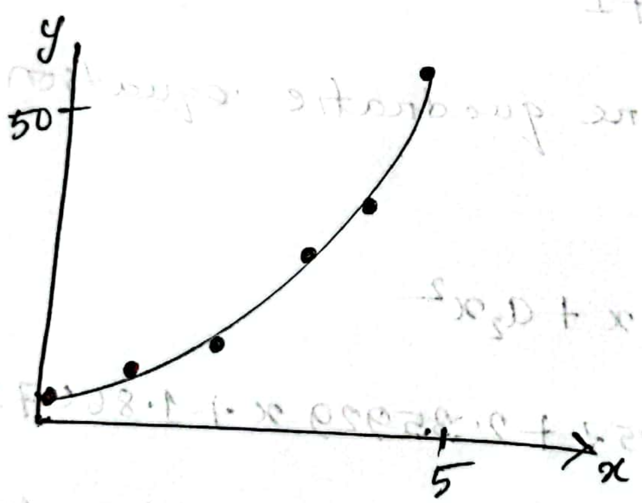
The coefficient of determination

$$r^2 = \frac{2513.39 - 3.74657}{2513.39}$$

$$= 0.99851$$

and the correlation coefficient is, $r = 0.99925$

These result indicate that 99.851 percent of the original uncertainty has been explained by the model.



Fit of a second order polinomial

Runge-Kutta Method

Sat / Sun / Mon / Tue / Wed / Thu / Fri

Date :......../......../.........

Euler's Method:

New value = old value + slop × step size

Euler-cauchy method

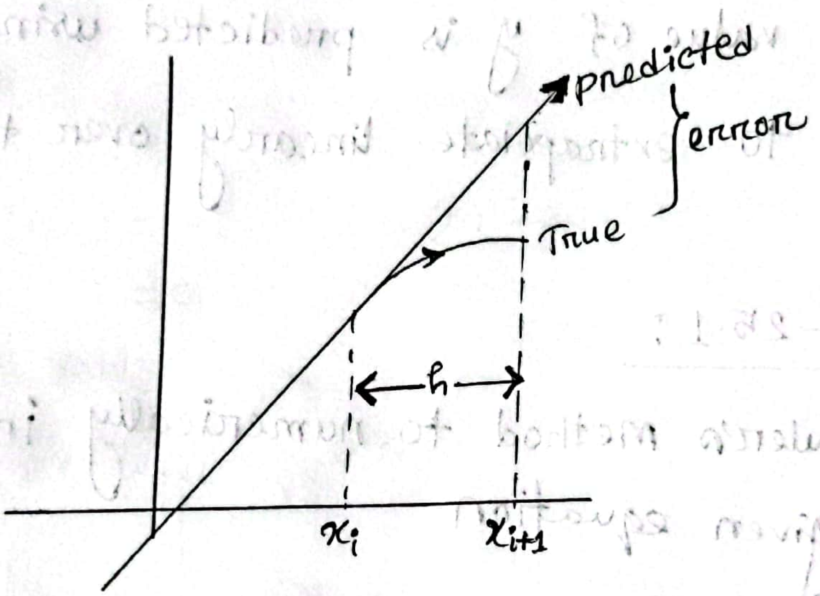
point slop method

$$y_{i+1} = y_i + \phi h$$

$$\boxed{y_{i+1} = y_i + f(x,y)h}$$

slop = $\frac{dy}{dx} = f(x,y)$

▶ This formula referred to as Euler's method



Euler's method

The first derivative provides a direct estimate of the slope at x_i

$$\phi = f(x_i, y_i)$$

where, $f(x_i, y_i)$ is the differential equation evaluated at x_i and y_i.

This estimate can be substitute into the equation,

$$\boxed{y_{i+1} = y_i + f(x_i, y_i)\,h}$$

This formula is referred to as Euler's method.
A new value of y is predicted using the slope to extrapolate linearly over the step size h.

Example - 25.1:

Use Euler's method to numerically integrate the given equation

$$\frac{dy}{dx} = -2x^3 + 12x^2 - 20x + 8.5$$

from $x=0$ to $x=4$ with a step size of 0.5.
The initial condition at $x=0$ is $y=1$.
Recall that the exact equation is given by

$$y = -0.5x^4 + 4x^3 - 10x^2 + 8.5x + 1$$

$x=0 \rightarrow y=1$

Solution:

$y_{i+1} = y_i + f(x_i, y_i)h$ can be ~~use~~ used to implement Euler's method.

$$y(0.5) = y(0) + f(0,1)\,0.5$$

Where,

$y(0)=1$ and the slope estimate at $x=0$ is

$$f(0,1) = -2(0)^3 + 12(0)^2 - 20(0) + 8.5$$
$$= 8.5$$

Therefore,

$$y(0.5) = 1 + 8.5 \times 0.5$$
$$= 5.25$$

The true solution at $x=0.5$ is

$$y = -0.5(0.5)^4 + 4(0.5)^3 - 10(0.5)^2 + 8.5(0.5) + 1$$
$$= 3.21875$$

Thus the error is,

$$E_t = \text{true} - \text{approximate}$$
$$= 3.21875 - 5.25$$
$$= -2.03125$$

~~or, expected~~

or, expressed as percent relative error, $\varepsilon_t = -63.1\%$

$$y(1) = y(0.5) + f(0.5, 5.25)\,0.5$$

$$= 5.25 + \left[-2(0.5)^3 + 12(0.5)^2 - 20(0.5) + 8.5\right]0.5$$

$$= 5.875$$

The true solution at $x=1$ is 3.0

The percent relative error is -95.8%.

The computation is repeated.

x	y_{true}	y_{Euler}	percent Relative ERROR
0	1	1	
0.5	3.21875	5.25	-63.1
1	3	5.875	-95.8
1.5	2.21875	5.125	131
2	2	4.5	-125
2.5	2.71875	4.75	-74.7
3	4	5.875	46.9
3.5	4.71875	7.125	-51.0
4	3	7	-133.3