Least Squares Regression

Note Title 7/2/2010

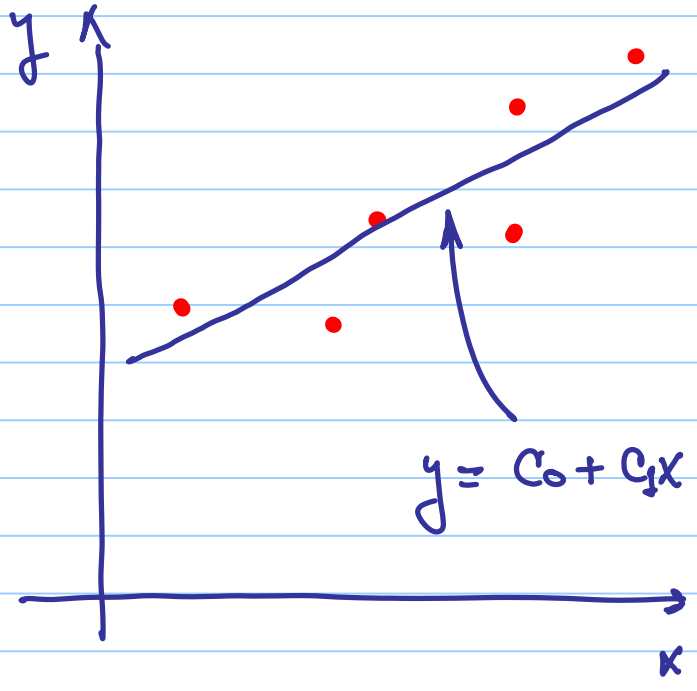
Regression is the method of obtaining the best fit to a given set of data.

The procedure of fitting the best straight line (linear equation) to the data is known as

"Linear Regression"

Data Points:

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_6	y_k
⋮	⋮



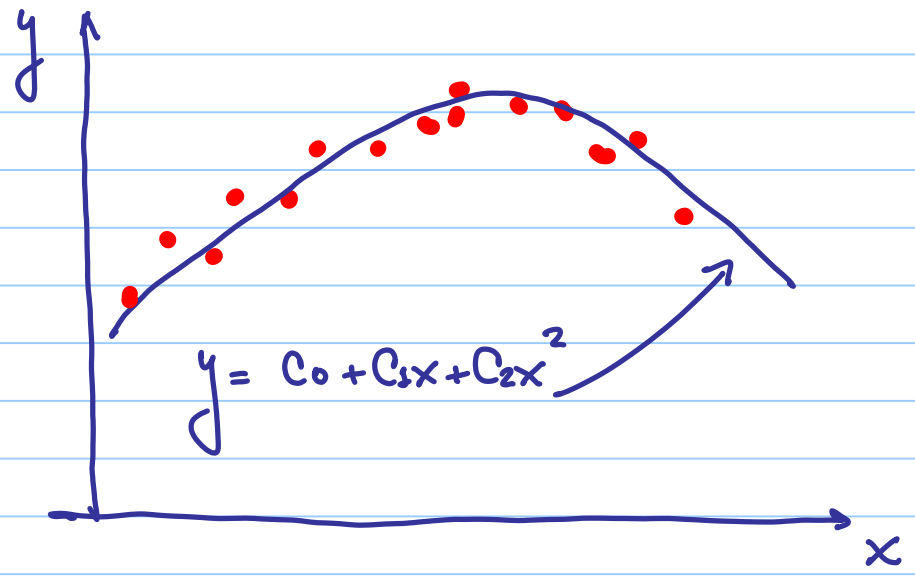
$$y = C_0 + C_1 x$$

We need to find C_0 and C_1.

In many engineering applications the variables may be related nonlinearly.

For data points exhibiting non-linear behavior we can use

'Polynomial Regression'



Polynomial proposed:

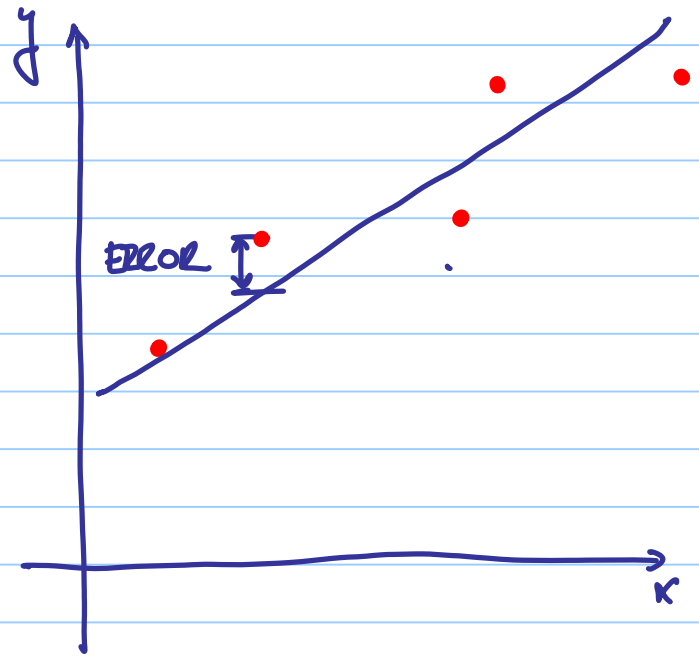
$$y = C_0 + C_1x + C_2x^2$$

We need to find coefficients C_0, C_1 and C_2... How??

WELL, THE BEST COEFFICIENTS WILL BE THOSE THAT MINIMIZE THE ERROR AT EVERY DATA POINT.

ASSUME WE HAVE THE FOLLOWING DATA:

x	y
x_1	y_1
x_2	y_2
x_3	y_3
x_4	y_4
x_5	y_5



WE PROPOSE THE LINE

$$y = C_0 + C_1 x$$

SO THAT MEANS THAT THE FOLLOWING EQUATIONS MUST HOLD:

$$y_1 = C_0 + C_1 x_1$$

$$y_2 = C_0 + C_1 x_2$$

$$y_3 = C_0 + C_1 x_3$$

$$y_4 = C_0 + C_1 x_4$$

$$y_5 = C_0 + C_1 x_5$$

WE COULD USE THIS SYSTEM OF LINEAR EQUATIONS TO FIND C_0 AND C_1,

HOWEVER, THIS SYSTEM HAS NOT AN EXACT SOLUTION BECAUSE IT IS OVERDETERMINE

IT HAS 5 EQUATIONS AND ONLY TWO UNKNOWNS !!

SINCE NO SOLUTION SATISFIES ALL THE EQUATIONS EXACTLY, WE CAN DEFINE A RESIDUAL ERROR VECTOR AND TRY TO FIND THE COEFFICIENTS C_0 AND C_1 THAT MINIMIZE THIS VECTOR.

USE MATRIX NOTATION TO WRITE THE SYSTEM OF EQUATIONS:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \end{bmatrix}$$

$$\vec{y} = [A]\vec{c}$$

RESIDUAL ERROR VECTOR:

$$\vec{e} = [A]\vec{c} - \vec{y}$$

A COMMONLY USED PROCEDURE TO MINIMIZE THE ERROR VECTOR INVOLVES MINIMIZING THE SUM OF SQUARES OF THE COMPONENTS OF THE ERROR VECTOR, KNOW AS 'LEAST SQUARE REGRESSION'

THE SUM OF SQUARES OF THE COMPONENTS OF $\vec{e}$ CAN BE EXPRESSED AS

$$S(c_0, c_1) = \vec{e}^T \cdot \vec{e} = e_1^2 + e_2^2 + e_3^2 \cdots$$

$$S(c_0, c_1) = \left(\vec{c}^T [A]^T - \vec{y}^T\right) \cdot \left([A]\vec{c} - \vec{y}\right)$$

FOR MINIMIZING S, WE SET THE PARTIAL DERIVATIVES OF S WITH RESPECT TO EACH OF THE COMPONENTS OF $\vec{c}$ EQUAL TO ZERO:

$$\frac{\partial S}{\partial c_0} = 0$$

$$\frac{\partial S}{\partial c_1} = 0$$

THIS LEADS TO THE EQUATION

$$[A]^T [A] \vec{c} = [A]^T \vec{y}$$

$$[K] \vec{c} = \vec{b}$$

$(n \times n)$ $n \times 1$ $n \times 1$

THIS IS A SYSTEM OF n EQNS. AND n UNKNOWNS ... HENCE,

IT CAN BE SOLVED EASILY! 😊

ACCURACY OF THE LINEAR REGRESSION:

1.- COMPUTE FIRST THE SUM OF SQUARES OF THE DEVIATION OF THE DATA AROUND THE MEAN

$$S_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$\bar{y}$ DENOTES THE MEAN VALUE OF y_i:

$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_i$$

2.- COMPUTE THE SUME OF SQUARES OF THE DEVIATION OF THE DATA ABOUT THE STRAIGHT LINE (ERRORS)

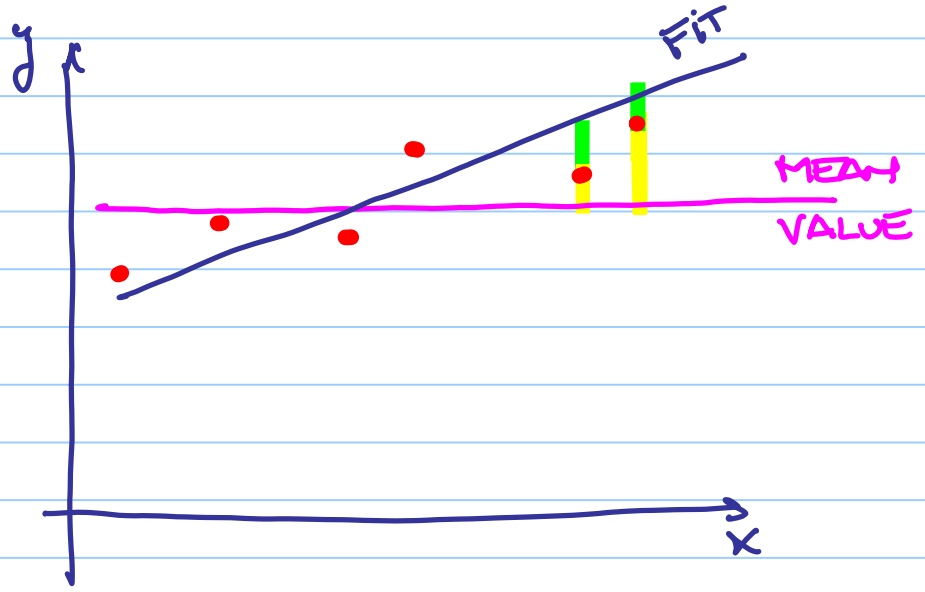
$$S = \sum_{i=1}^{n} (y_i - (C_0 + C_1 x_i))^2$$

THE DIFFERENCE BETWEEN S_0 AND S PROVIDES A MEASSURE OF ACCURACY OF THE REGRESSION

$$\boxed{R^2 = \frac{S_0 - S}{S_0}}$$

FOR A PERFECT FIT S WILL BE ZERO AND R^2 WILL BE 1.

FOR A POOR FIT $S = S_0$ AND $R^2 = 0$.



IN PRACTICE, A GOOD LEAST SQUARES FIT, IS INDICATED BY A VALUE OF R^2 CLOSE TO 1.

THE SAME APPROACH EXPLAINED HERE CAN BE USED FOR POLYNOMIAL REGRESSION, EVEN FOR OTHER TYPE OF NON-LINEAR REGRESSIONS.

Example 1:

The number of graduate students enrolled in the Civil Engineering Department of FIU over a period of 10 years is tabulated as follows:

Year (x_i)	Number (y_i)
1	204
2	216
3	208
4	220
5	240
6	257
7	271
8	270
9	285
10	291

Fit a linear equation to the data

$$y = C_0 + C_1 x$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ \vdots & \vdots \\ 1 & x_{10} \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \end{bmatrix}$$

$\vec{y}$ $\quad$ $[A]$ $\quad$ $\vec{C}$

EXAMPLE 2 :

THE TEMPERATURE MEASURED AT VARIOUS POINTS ALONG THE THICKNESS OF A WALL SUBJECTED TO A SUDDEN CHANGE IN THE INNER TEMPERATURE YIELDED THE FOLLOWING RESULTS :

x →	di (% WALL THICKNESS)	0	0.25	0.5	0.75	1
y →	Ti (°C)	100	70	45	25	15

DEVELOP A SECOND ORDER POLYNOMIAL TO SHOW THE VARIATION OF TEMPERATURE ALONG THE THICKNESS OF THE WALL.

$$y = C_0 + C_1 x + C_2 x^2$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \end{bmatrix} \begin{bmatrix} C_0 \\ C_1 \\ C_2 \end{bmatrix}$$

$\Downarrow$ $\vec{y}$ $\quad$ $\Downarrow$ $[A]$ $\quad$ $\Downarrow$ $\vec{C}$