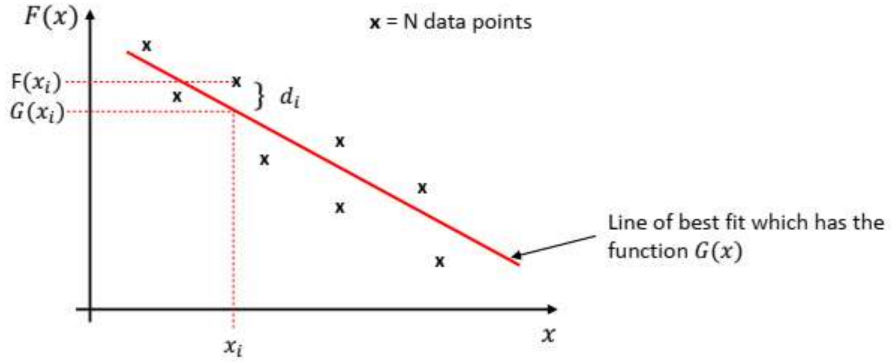
Regression analysis - Least squares regression

- If the data points for a function ($F(x)$) are plotted, but it is known that the data points contain an error, then a line of best fit (with function $G(x)$) can be plotted -



$G(x)$ = function of the line of best fit

$G(x_i)$ = value of function $G(x)$ at x_i

$F(x_i)$ = value of the function $F(x)$ data point at x_i

$d_i = F(x_i) - G(x_i)$

- Assume that the approximate function of the line of best fit ($G(x)$) is:

$$G(x) = a_1 g_1(x) + a_2 g_2(x) + \cdots a_n g_n(x)$$

where $a_1, a_2 \dots a_n$ are constants and $g(x)$ are functions of x (e.g. x, x^2, x^3)

- The deviation function (D) is given by:

$$D = \sum_{i=1}^{N} d_i^2$$

- Therefore to minimise D:

$$\frac{\partial D}{\partial a_1} = 0, \quad \frac{\partial D}{\partial a_2} = 0 \cdots \cdots \frac{\partial D}{\partial a_n} = 0$$

This produces N simultaneous equations for $a_1, a_2 \cdots \cdots a_n$

Linear regression

- If the approximate function of the line of best fit is given by $G(x) = a_1 g_1(x) + a_2 g_2(x) + \cdots a_n g_n(x)$, then the function for a linear regression is given by:

$$G(x) = a_1 + a_2 x \qquad \text{(i.e. } g_1(x) = 1 \; and \; g_2(x) = x\text{)}$$

- Therefore, the deviation function (D) can be written as:

$$D = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} [F(x_i) - a_1 - a_2 x_i]^2$$

- To minimise D we need $\frac{\partial D}{\partial a_1} = 0$ and $\frac{\partial D}{\partial a_2} = 0$, so using partial derivatives:

$$\frac{\partial D}{\partial a_1} = 2 \sum_{i=1}^{N} [F(x_1) - a_1 - a_2 x_i](-1) = 0$$

$$\frac{\partial D}{\partial a_2} = 2 \sum_{i=1}^{N} [F(x_1) - a_1 - a_2 x_i](-x_i) = 0$$

- Using the partial derivatives we can now solved for a_1 and a_2:

$$a_1 = \frac{\sum f_i \sum x_i^2 - \sum f_i x_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}$$

$$a_2 = \frac{N \sum f_i x_i - \sum f_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2}$$

where $f_i = F(x_i)$ and N = number of data points

- This can be used in the equation $G(x) = a_1 + a_2 x$ to plot a straight line through the data points.

Linear regression (example)

Use linear regression to plot the line of best fit through the following data points:

x_i	0	1	2	3	5
f_i	0	1.4	2.2	3.5	4.4

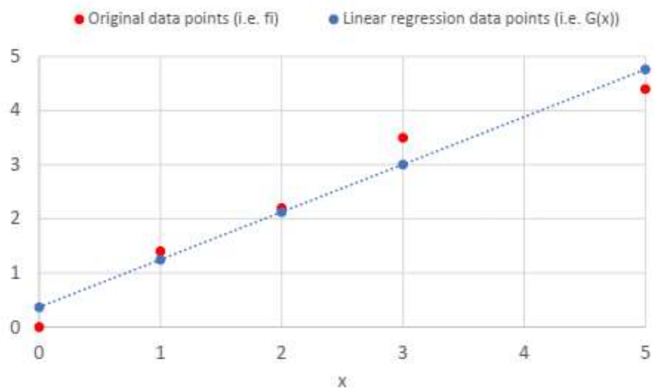
The line of best fit is given by the function $G(x) = a_1 + a_2 x$, therefore we need to calculate the values of a_1 and a_2. To do this we need to find $\sum x_i$, $\sum f_i$, $\sum x_i f_i$ and $\sum x_i^2$:

	i	x_i	f_i	$x_i f_i$	x_i^2
	1	0	0	0	0
	2	1	1.4	1.4	1
	3	2	2.2	4.4	4
	4	3	3.5	10.5	9
	5	5	4.4	22	25
$\sum$		**11**	**11.5**	**38.3**	**39**

$$a_1 = \frac{\sum f_i \sum x_i^2 - \sum f_i x_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2} = \frac{(11.5 \times 39) - (38.3 \times 11)}{(5 \times 39) - 11^2} = 0.368$$

$$a_2 = \frac{N \sum f_i x_i - \sum f_i \sum x_i}{N \sum x_i^2 - (\sum x_i)^2} = \frac{(5 \times 38.3) - (11.5 \times 11)}{(5 \times 39) - 11^2} = 0.878$$

$$\therefore G(x) = 0.368 + 0.878x$$



Plotting the function of $G(x)$ produces a linear regression line through the original data points.

Example 2

The torque T needed to turn the torsional spring of a mousetrap through an angle, θ is given below

Angle, θ Radians	Torque, T $\mathrm{N \cdot m}$
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707

Find the constants k_1 and k_2 of the regression model

$$T = k_1 + k_2\theta$$

Solution

Table below shows the summations needed for the calculation of the constants of the regression model.

Tabulation of data for calculation of needed summations.

i	θ	T	θ^2	$T\theta$
1	0.698132	0.188224	4.87388×10^{-1}	1.31405×10^{-1}
2	0.959931	0.209138	9.21468×10^{-1}	2.00758×10^{-1}
3	1.134464	0.230052	1.2870	2.60986×10^{-1}
4	1.570796	0.250965	2.4674	3.94215×10^{-1}
5	1.919862	0.313707	3.6859	6.02274×10^{-1}
$\sum_{i=1}^{5}$	**6.2831**	**1.1921**	**8.8491**	**1.5896**

$$n = 5$$

$$k_2 = \frac{n\sum_{i=1}^{5}\theta_i T_i - \sum_{i=1}^{5}\theta_i \sum_{i=1}^{5} T_i}{n\sum_{i=1}^{5}\theta_i^2 - \left(\sum_{i=1}^{5}\theta_i\right)^2}$$

$$= \frac{5(1.5896) - (6.2831)(1.1921)}{5(8.8491) - (6.2831)^2}$$

$$= 9.6091\times 10^{-2} \text{ N - m/rad}$$

$$\bar{T} = \frac{\sum_{i=1}^{5} T_i}{n}$$

$$= \frac{1.1921}{5}$$

$$= 2.3842\times 10^{-1} \text{N-m}$$

$$\bar{\theta} = \frac{\sum_{i=1}^{5}\theta_i}{n}$$

$$= \frac{6.2831}{5}$$

$$= 1.2566 \text{radians}$$

$$k_1 = \bar{T} - k_2\,\bar{\theta}$$

$$= 2.3842\times 10^{-1} - (9.6091\times 10^{-2})(1.2566)$$

$$= 1.1767\times 10^{-1} \text{ N - m}$$

Linear regression (MATLAB)

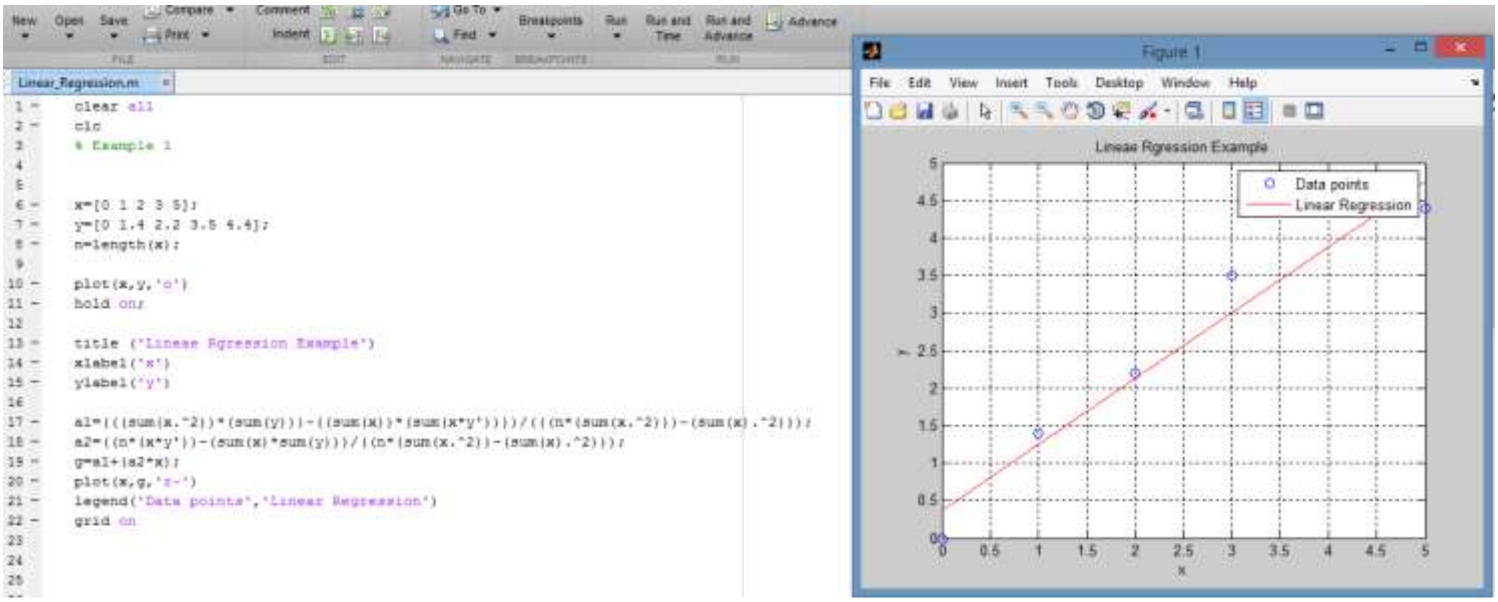
```
% Example 1
clear all
clc


x=[0 1 2 3 5];
y=[0 1.4 2.2 3.5 4.4];
n=length(x);

plot(x,y,'o')
hold on;

title ('Lineae Rgression Example')
xlabel('x')
ylabel('y')

a1=(((sum(x.^2))*(sum(y)))-((sum(x))*(sum(x*y'))))/(((n*(sum(x.^2)))-(sum(x).^2)));
a2=((n*(x*y'))-(sum(x)*sum(y)))/((n*(sum(x.^2))-(sum(x).^2)));
g=a1+(a2*x);
plot(x,g,'r-')
legend('Data points','Linear Regression')
grid on
```



```
% Example 2
clear all
clc

x=[0.698132 0.959931 1.134464 1.570796 1.919862];      %Angle
y=[0.188224 0.209138 0.230052 0.250965 0.313707];      %Torque
n=length(x);
plot(x,y,'o')
hold on;

title ('Lineae Rgression Example')
xlabel('Angle')
ylabel('Torque')

a1=(((sum(x.^2))*(sum(y)))-((sum(x))*(sum(x*y'))))/(((n*(sum(x.^2)))-
(sum(x).^2)));
a2=((n*(x*y'))-(sum(x)*sum(y)))/((n*(sum(x.^2))-(sum(x).^2)));
g=a1+(a2*x);
plot(x,g,'r-')
legend('Data points','Linear Regression')
grid on
```