```
% Linear Regression
% 2007 Fabian Farelo, Autar Kaw
% University of South Florida
% United States of America
% kaw@eng.usf.edu
%
% This worksheet demonstrates the use of Matlab to illustrate
% the procedure to make a linear regression on a given data set

% INPUTS
% This is the input data that can be modified by the user
% X and Y are arrays of data points
X=[1,2,3,4,5];
Y=[1,4,9,16,125];

% REST OF THE WORKSHEET
clc
disp(sprintf('Linear Regression'))
disp(sprintf('©2007 Fabian Farelo, Autar Kaw'))
disp(sprintf('University of South Florida'))
disp(sprintf('United States of America'))
disp(sprintf('kaw@eng.usf.edu'))

disp(sprintf('\n\nNOTE: This worksheet demonstrates the use of Matlab to illustrate'))
disp(sprintf('the procedure to make a Linear Regression on a given data set'))
%------------------------------------------------------------------------
disp(sprintf('\n\n***************************** Introduction *******************************'))
disp(sprintf('\nLinear Regression is the most popular regression model. In this model'))
disp(sprintf('we wish to predict response points to n data points (x1,y1), (x2,y2),'))
disp(sprintf('.....,(xn,yn) data by a regression model given by:'))
disp(sprintf('\n                    y=a0 +a1*x                              (1)'))
disp(sprintf('\nwhere a0 and a1 are the constants of the regression model.'))
disp(sprintf('A measure of goodness of fit, that is, how a0 + a1*x predicts the '))
disp(sprintf('response variable y is the magnitude of the residual,ei at each of the'))
disp(sprintf('n data points'))
disp(sprintf('\n ei= yi - (a0 +a1*xi) = (observed value at xi - predicted value at xi)     (2)'))
disp(sprintf('\nIdeally, if all the residuals ei are zero, one may find an equation'))
disp(sprintf('in which all the points lie on the model. Thus, minimization of the '))
disp(sprintf('residual is an objective of obtaining regression coefficients.'))
disp(sprintf('The most popular method to minimize the residual is the least squares'))
disp(sprintf('method, where the estimates of the constants of the models are chosen'))
disp(sprintf('such that the sum of the squared residuals, Sr, is minimized, that is '))
disp(sprintf('minimize'))
disp(sprintf('\n               Sr = sum(ei^2, i=1:n)                         (3)'))
disp(sprintf('\nwhere'))
disp(sprintf('\n      sum(ei^2, i=1:n)=sum((yi - (a0 +a1*xi))^2, i=1:n          (4)'))
disp(sprintf('\nLet us use the least squares criterion where we minimize the sum of the'))
disp(sprintf('squared residuals Sr:'))
disp(sprintf('\n             d/d(a0) (Sr) = 0                                 (5)'))
disp(sprintf('\n             d/d(a1) (Sr) = 0                                 (6)'))
disp(sprintf('\nOnce Sr is minimized with respect to the regression coefficients a0 and a1,'))
disp(sprintf('the coefficients can be solved for:'))
disp(sprintf('\n             a0 = yave - a1*xave                             (7)'))
disp(sprintf('\n             a1 = Sxy/Sxx                                     (8)'))
disp(sprintf('\nwhere Sxy and Sxx can be defined as:'))
disp(sprintf('\n            Sxy=sum(xi*yi , i=1:n)-n*xave*yave                (9) '))
disp(sprintf('\n            Sxx=sum(xi^2 , i=1:n)-n*xave^2                        (10) '))
disp(sprintf('\nand the average values can be defined as'))
disp(sprintf('\n            xav=sum(xi, i=1:n)/n                                 (11) '))
disp(sprintf('\n            yav=sum(yi, i=1:n)/n                                 (12) '))

disp(sprintf('\n\n**************************** Input data *******************************'))
disp(sprintf('\nThese are the simulation parameters that can be modified by the user in'))
disp(sprintf('\the beginning of the M-file. This is the only section that requires user input.'))


%n is the number of elements in the arrays
n=length(X);
%-----------------------------------------------------------------------
disp ('Xarray')
disp (sprintf('    %g',X))
disp ('Yarray')
disp (sprintf('    %g',Y))
disp (sprintf('Number of data points = %g ',n))
xav=sum(X)/n;
yav=sum(Y)/n;
Sxy=0;
Sxx=0;
for i=1:n
   Sxy=Sxy +X(i)*Y(i)-xav*yav;
   Sxx=Sxx + (X(i))^2-xav^2;
end
disp(sprintf('\n\n**************************** Results ********************************'))
disp(sprintf('Using equations (9) and (10) to calculate Sxx and Sxy:'))
disp (sprintf('Sxx = %g ',Sxx))
disp (sprintf('Sxy = %g ',Sxy))

disp(sprintf('\nNow Sxx, Sxy, xave, and yave can be used to calculate the regression  '))
disp(sprintf('coefficients a0 and a1 using equations (7) and (8)'))
a1=Sxy/Sxx;
a0=yav-a1*xav;
Yp=zeros(1,floor(max(X)+1));
for i=0:floor(max(X)+1)
   Yp(i+1)=a0+a1*i;
end
Xp=(0:floor(max(X)+1));
Yp;
disp(sprintf('\nThe Linear model is described as            y=%d',a0))
disp(sprintf('\b + %d',a1))
disp(sprintf('\b*x'))
figure
plot(Xp,Yp)
hold on
plot(X,Y,'bo','MarkerFaceColor','b')

Sr=0;
St=0;
for i=1:n
   Sr=Sr+(Y(i)-a0-a1*X(i))^2;
   St=St+(Y(i)-yav)^2;
end
disp(sprintf('\n\n******************** Coefficient of determination ***********************'))
disp(sprintf('\nOne of the major indicators of how well least squares characterizes or predicts'))
disp(sprintf('the whole data is a quantity called the coefficient of determinaton, r2'))
disp(sprintf('\n                              r2 = (St - Sr)/St                              (13)'))
disp(sprintf('\nwhere \nSr = the sum of the squares of the residuals (a value that quantifies the spread'))
disp(sprintf('around the regression line)'))
disp(sprintf('and \nSt = the sum of the squares of deviation from the mean (a value that measures'))
disp(sprintf('the spread between the data and its mean)'))
disp(sprintf('\nThe objective of least squares method is to obtain a compact equation that best'))
disp(sprintf('describes all data points. The mean can also be used to describe only data points. The'))
disp(sprintf('magnitude of the sum of squares of deviation from the mean or from the least squares'))
disp(sprintf('line is therefore a good indicator of how well the mean or least squares characterizes'))
disp(sprintf('the whole data.'))
disp(sprintf('\nThe difference between these two parameters measures the error due to describing '))
disp(sprintf('or characterizing the data in one from instead the other. A relative comparison of this '))
disp(sprintf('difference (St-Sr), with the sum of squares deviation associated with the mean (i.e. r2), '))
disp(sprintf('describes a proportion of variation in the response data that is explained by the regression'))
disp(sprintf('model. When all the points in a data set lie on the regression model, the largest possible'))
disp(sprintf('value of r2 = 1 is obtained, while a minimum possible value of r2 = 0 is obtained when'))
disp(sprintf('there is only one data point or if the straight line model is a constant line.\n'))

disp(sprintf('     St = %g', St))
disp(sprintf('     Sr = %g', Sr))
r2=(St-Sr)/St;
disp(sprintf('     r2 = %g', r2))
disp(sprintf('\nAs r2 gets closer to 1, the model more accurately describes the data.'))
disp(sprintf('The figure shows the data points as well as the least squares regression\n line.'))
```