```
% Adequacy of Regression Models
% 2007 Jamie Trahan, Autar Kaw
% University of South Florida
% United States of America
% kaw@eng.usf.edu
% http://numericalmethods.eng.usf.edu

% This worksheet allows you to determine whether a straight-line regression model adequately
% describes n data points (x1,y1), (x2, y2), (x3,y3),....,(xn,yn).
% Four different checks are used to find if the model is adequate.
% Our discussion, although limited to straight line models, is applicable to any regression model.
%
% Adequacy Check #1: Plot of straight-line regression model versus data to visually inspect
%                    how well the data fits the line.
% Adequacy Check #2: Calculation of the coefficient of determination, r2.
%                    This value quantifies the percentage of the original uncertainty
%                    in the data that is explained by the straight line model.
% Adequacy Check #3: Determine if the residuals as a function of x show nonlinearity.
%                    This is an indication that the model is not adequate.
% Adequacy Check #4: Determine if 95% of the values of scaled residuals are within [-2,2].
%                    If so, this is an indication that the model may be adequate.
% Please note that the above checks are not a complete test of the adequacy of regression models.
% Other tests include testing if indeed y is dependent on x, confidence intervals of the constants
% of the model, etc.
%
% To learn more about the quality of a fitted linear regression model,
% see the worksheet on the Adequacy of Regression models.

%***************************************************************************
% INPUT PARAMETERS
% The (x,y) data pairs
clc
disp('2007 Jamie Trahan, Autar Kaw')
disp('University of South Florida')
disp('United States of America')
disp('kaw@eng.usf.edu')
disp('http://numericalmethods.eng.usf.edu')
% To enter the data, you may enter arrays as given below
% The x and y arrays of data are given below
x=[1  1.5    3  5.1    7   8   13   19   20    25];
y=[1  2.25   9  26.01 49  64 169  381  400   625];
% ***********************************************************

disp(' ')
disp('This worksheet allows you to determine whether a straight-line regression model')
disp ('adequately describes n data points (x1,y1), (x2, y2), (x3,y3),....,(xn,yn)')
disp('Four different checks are used to find if the model is adequate. ')
disp('Our discussion, although limited to straight line models, is applicable to any regression model')
disp(' ')
disp('Adequacy Check #1: Plot of straight-line regression model versus data to visually')
disp('                   inspect how well the data fits the line.')
disp('Adequacy Check #2: Calculation of the coefficient of determination, r2. ')
disp('                   This value quantifies the percentage of the original')
disp('                   uncertainty in the data that is explained by the straight line model.')
disp('Adequacy Check #3: Determine if the residuals as a function of x show nonlinearity.')
disp('                   This is an indication that the model is not adequate.')
disp('Adequacy Check #4: Determine if 95 percent of the values of scaled residuals are within [-2,2]')
disp('                           If so, this is an indication that the model may be adequate.')
disp(' ')
disp('Please note that the above checks are not a complete test of the adequacy of regression models. ')
disp('Other tests include testing if indeed y is dependent on x, confidence intervals')
disp('of the constants of the model, etc.')
disp('To learn more about the quality of a fitted linear regression model,')
disp('see the Adequacy of Regression Models Texbook Notes')


% Finding the straight line regression model
n=length(x);
fprintf('\nNumber of data points =%g\n',n)
disp('The x and y arrays of data are given below')
x
y
a = polyfit(x,y,1);
fprintf ('The straight line regression curve y=%g + (%g) x\n',a(2),a(1))

% ADEQUACY CHECK#1
% The linear regression model is plotted versus the data points in Figure 1.
% See if the straight-line regression model visually explains the data.
disp(' ')
disp(' ADEQUACY CHECK#1')
disp(' The linear regression model is plotted versus the data points in Figure 1.')
disp(' See if the straight-line regression model visually explains the data.')
figure (1)
xx=x(1):(x(n)-x(1))/20:x(n);
yy=a(2)+a(1)*xx;
plot(xx,yy)
hold on;
plot(x,y,'o')
hold off;
legend('Predicted straight line','Data Points')
xlabel ('x')
ylabel('y')


% ADEQUACY CHECK#2
% In this section we will calculate the coefficient of determination, r2.
% r2=(St-Sr)/St
% where
% Sr = the sum of the squares of the residuals
%       (a value that quantifies the spread around the regression line)
% and
% St = the sum of the squares of deviation from the mean
%       (a value that measures the spread between the data and its mean)
%
% This value describes the proportion of variation in the response data
% that is explained by the regression model. When all the points in a data set lie
% on the regression model, the largest possible value of r2 = 1 is obtained,
% while a minimum possible value of r2 = 0 is obtained when there is only one data point
% or if the straight line regression model is a constant line.

disp(' ')
disp(' ADEQUACY CHECK#2')
disp(' In this section we will calculate the coefficient of determination, r2.')
disp('    r2=(St-Sr)/St')
disp(' where ')
disp('    Sr = the sum of the squares of the residuals ')
disp('             (a value that quantifies the spread around the regression line)')
disp('    St = the sum of the squares of deviation from the mean ')
disp('             (a value that measures the spread between the data and its mean)')
disp('This value describes the proportion of variation in the response data ')
disp('that is explained by the regression model. When all the points in a data set lie ')
disp('on the regression model, the largest possible value of r2 = 1 is obtained, ')
disp('while a minimum possible value of r2 = 0 is obtained when there is only one data point ')
disp('or if the straight line regression model is a constant line. ')
St=0;
for i=1:n
   St=St+(y(i)-mean(y))^2;
end

Sr=0;
for i=1:n
   Sr=Sr+(y(i)-a(2)-a(1)*x(i))^2;
end
r2=(St-Sr)/St;
fprintf ('The coefficient of determination is = %g\n\n',r2)


% ADEQUACY CHECK#3
% In this section, the residuals, which are the differences between the observed values
% and predicted values (y(i) - a2 - a1*x(i)), are found and then plotted as a function of x to
% check for increasing variance, outliers, or nonlinearity.

disp('ADEQUACY CHECK#3')
disp('In this section, the residuals, which are the differences between the observed values ')
disp('and predicted values (y(i) - a2 - a1*x(i)), are found and then plotted in Figure 2 as a')
disp('function of x to check for increasing variance, outliers, or nonlinearity.')

residuals=y-a(2)-a(1)*x;
figure(2)
plot(x,residuals,'o')
hold off;
legend('Residuals vs x')
xlabel ('x')
ylabel('Residual')

% ADEQUACY CHECK#4
% In this section, the scaled residuals SR (ratio between residual and the standard error of estimate)
% are calculated. SR at a point is given by
% SR=(residual at the point)/(standard error of estimate)
% Standard error of estimate = sqrt(Sr/(n-constants of model))
% For a straight line, the constants of the model are 2
% You want to check if 95% of the scaled residuals fall in the [-2,2] range
disp(' ')
disp('ADEQUACY CHECK#4')
disp('In this section, the scaled residuals SR (ratio between residual and the standard error of estimate) ')
disp('are calculated.  SR at a point is given by ')
disp('     SR=(residual at the point)/(standard error of estimate)')
disp('where')
disp('     Standard error of estimate = sqrt(Sr/(n-constants of model))')
disp('For a straight line, the constants of the model are 2')
disp('You want to check if 95% of the scaled residuals fall in the [-2,2] range')

standard_error=sqrt(Sr/(n-2));
fprintf('The standard error of estimate is =%g\n',standard_error)
SR=residuals/standard_error;

% Counting how many scaled residuals are in [-2,2] range
number_inside=0;
for i=1:n
  if SR(i)>=-2 & SR(i)<=2
      number_inside=number_inside+1;
  end
end

percent_inside=number_inside/n*100;
if percent_inside>=95
    disp('95 percent of the scaled residuals are between -2 and 2.  See Figure 3')
    fprintf('Actual percentage of scaled residuals between -2 and 2 is =%g',percent_inside)
else
    disp('95 percent of the scaled residuals are NOT between -2 and 2. See Figure 3')
    fprintf('Actual percentage of scaled residuals between -2 and 2 is =%g',percent_inside)
end

figure(3)
plot(x,SR,'o')
hold on;
xx=x(1):(x(n)-x(1))/10:x(n);
yy=xx-xx-2;
plot(xx,yy,'-','LineWidth',3)
hold on;
xx=x(1):(x(n)-x(1))/10:x(n);
yy=xx-xx+2;
plot(xx,yy,'-','LineWidth',3)
hold off;
legend('Scaled residuals vs x')
xlabel ('x')
ylabel('Scaled Residual')
```