# Regression Analysis Tutorial

## INTRODUCTION

Regression analysis can be used to identify the line or curve which provides the best fit through a set of data points. This curve can be useful to identify a trend in the data, whether it is linear, parabolic, or of some other form. Regression analysis can be performed using different methods; this tutorial will explore the use of Excel and MATLAB for regression analysis. In addition to fitting a curve to given data, regression analysis can be used in combination with statistical techniques to determine the validity of data points within a data set. For example, the standard deviation for a data set can easily be determined, and any data points existing outside of the $3\sigma$ range can be reviewed to determine if they are valid points.

## REGRESSION ANALYSIS USING EXCEL

### *Exercise A-1*

In Excel, generate a plot of the seven points given in Table 1. If you are unfamiliar with Excel, detailed instructions on how to do this are given in Appendix A.

Table 1. Data points to be used for Excel examples.

| x-value | y-value |
|---|---|
| 2 | 3 |
| 4 | 5 |
| 8 | 7 |
| 11 | 7.5 |
| 14 | 8 |
| 18 | 9 |
| 21 | 12 |

### *Exercise A-2*

Using all data points in the set, use Excel tools to perform a linear regression on the data. To do this, select the graph containing the data set, then select:

Chart
- Add Trendline
  - Type
    - Trend/Regression type → Linear
  - Options
    - Select Display equation on chart
    - Select Display R-squared value on chart
- OK

The graph will resemble Fig. 1. This plot shows the original data points along with the line providing the best fit through the points. The equation for the line is also given.

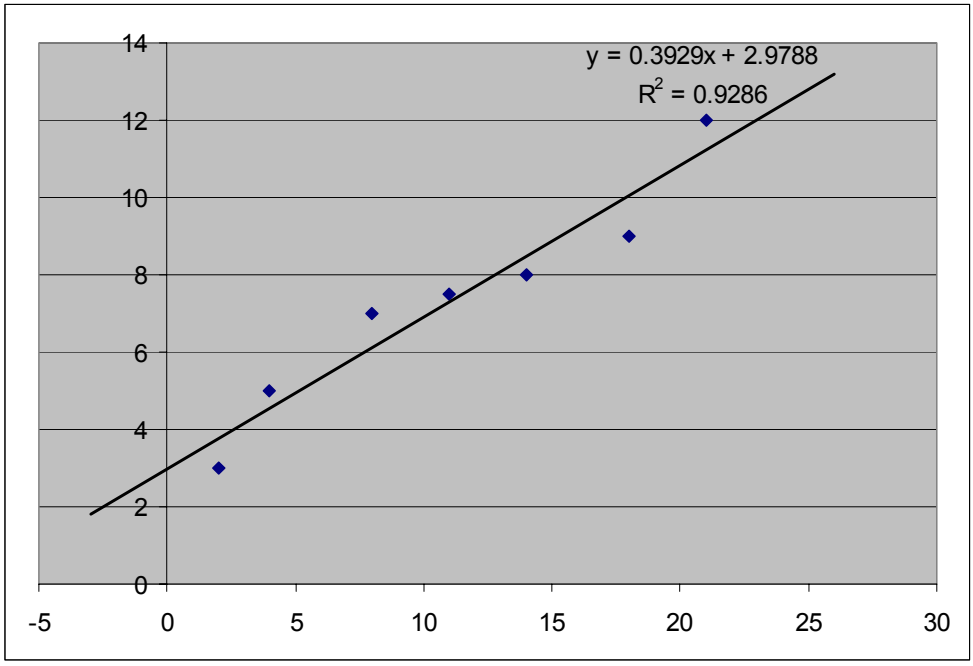


Fig. 1. Plot of data set with line of best fit.

The $R^2$ value shown on the graph indicates the goodness of fit for the line through the given points. An $R^2$ value of 1 would indicate a perfect fit, meaning that all points lie exactly on the line.

Next, under Chart\Add Trendline\Options, select Set intercept = 0. Note the change in the value of $R^2$. This feature is useful, as there are times when the data set being considered has a known initial starting point or y-intercept.

***Exercise A-3***

Now create the same plot as in Exercise A-2, except remove the leading and ending data points, and observe the changes to the linear regression, equation of the line, and $R^2$ value.

***Exercise A-4***

It is apparent from the fit of the line to the original data set that a linear regression may not be the most accurate description of the trend existing in the data. The same Excel tools can be used to perform regressions of higher order. For this exercise, a second order regression will be performed over the full data set. To perform a second order regression, select:

Chart
- Add Trendline
  - Type tab
    - Trend/Regression type → Polynomial, Order 2
  - Options tab
    - Select Display equation on chart
    - Select Display R-squared value on chart
  - OK

Note the shape of the curve and the goodness of fit.

The second-order equation can also be forced to have a y-intercept of zero, as was done for the linear example. After setting a zero y-intercept, note the shape of the curve, the equation of the line, and the $R^2$ value.

***Exercise A-5***

Now perform a second-order curve fit, but without including the first point of the data set. Note how this compares to the original second-order curve.

Try another second-order curve fit, but without the last point of the data set. Note the significant effect this has on the shape of the curve.

When performing a curve fit, especially with small numbers of data points, it must be noted that a single point can have enormous effect on the result obtained.

***Exercise A-6***

Next, perform a third-order regression. To do this, follow the same sequence of commands as given in Exercise A-4, but select Polynomial, Order 3 as the Trend/Regression type. Note the shape of the curve, the equation of the line, and the goodness of fit.

***Exercise A-7***

From previous exercises, it has been seen that, as the order of the regression increases, the $R^2$ value approaches 1. Now, continue to increase the value of the order of the polynomial, as done in exercises A-4 and A-5. At what point does the $R^2$ value seem to reach 1?

Consider that in some cases the $R^2$ value displayed on the chart may appear to be 1, but in reality this is only because the number is being rounded off when it is displayed. The number of displayed decimal places can be changed to fix this. To increase the number of decimal places, right click on the region containing the equation and the $R^2$ value, then select

Format data labels

Number

Category → Number

Decimal places → Enter the desired number of decimal places

Even if the $R^2$ value equals 1, it must also be considered whether the line fit makes physical sense. For example, and object in free-fall should have a position plot which is parabolic. Therefore a second-order line fit is desired, even though a higher-order line might fit the points more closely.

**REGRESSION ANALYSIS USING MATLAB**

***Exercise B-1***

Plot the data set identified in Exercise A-1 in MATLAB. An example of how to do this is given in Appendix B.

***Exercise B-2***

Least squares regression is used to determine the line of best fit through the data points. The mathematical procedure for this method will now be reviewed.

Any curve which can be fit over a data set can be shown to be a function y where

$$y = f(x, a_j), \text{ where } j = 1, 2, \ldots m, \tag{1}$$

with j representing the number of coefficients required to create the curve of the specified order. For example, the 3rd order equation can be expressed in the general form

$$y_i = a_1 + a_2 x + a_3 x^2 + a_4 x^3 . \tag{2}$$

In equation 2, i = 1, 2, … n, which represents the number of points to which this curve will be fit (for this exercise n = 7), and $a_1$ through $a_4$ are the unknown $a_j$ coefficients. These coefficients can be found using the least squares regression method and matrix algebra.

The general formula for least squares regression is

$$\sum_{i=1}^{n} (y_i - f(x_i, a_1 \ldots a_m)) \frac{\partial}{\partial a_j} f(x_i, a_1 \ldots a_m) = 0 . \tag{3}$$

The second half of (3) can be simplified by taking the partial derivative of the terms, producing

$$\frac{\partial f}{\partial a_j}(x_i, a_1 \ldots a_m) = \frac{\partial}{\partial a_j}[a_1 g_1 + a_2 g_2 + \ldots . a_m g_m] = g_i(x_i) . \tag{4}$$

After this partial differentiation, the general equation for least squares regression becomes

$$\sum_{i=1}^{n} [y_i - a_j g_i(x_i)] g_i(x_i) = 0 . \tag{5}$$

From the general equation in (5), the general form matrix can be built,

$$\begin{bmatrix} \sum_{i=1}^{n} g_1(x_i) g_1(x_i) & \ldots & \sum_{i=1}^{n} g_1(x_i) g_m(x_i) \\ .. & .. & .. \\ .. & .. & .. \\ \sum_{i=1}^{n} g_m(x_i) g_1(x_i) & \ldots & \sum_{i-1}^{n} g_m(x_i) g_m(x_i) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \ldots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i g_1(x_i) \\ .. \\ .. \\ \sum_{i=1}^{n} y_i g_m(x_i) \end{bmatrix} \tag{6}$$

Finally, by taking the 3rd order equation identified in (2) and defining the values of $g_i(x_i)$ as shown in (7), the general form of the matrix can be populated and solved using linear algebra, so that

$$f(x_i, a_1 \ldots a_4) = a_1 g_1(x) + a_2 g_2(x) + a_3 g_3(x) + a_4 g_4(x), \text{ where} \tag{7}$$

$$g_1(x) = 1,$$

$$g_2(x) = x,$$

$$g_3(x) = x^2, \text{ and}$$

$$g_4(x) = x^3 .$$

For the third-order regression being performed in this case, the matrix equation to be solved is

$$\begin{bmatrix} \sum_{i=1}^{n} 1 & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 & \sum_{i-1}^{n} x_i^6 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \\ \sum_{i=1}^{n} y_i x_i^2 \\ \sum_{i=1}^{n} y_i x_i^3 \end{bmatrix}. \quad (8)$$

MATLAB can be used to solve for the unknown coefficients in (8), and to compare the resulting coefficient values achieved from the MATLAB solution to those found using Excel. Also, plot the solution for the line over the previously plotted data set in MATLAB. An example of a program which can be used to do this is given in Appendix C. The resulting third-order regression is shown in Fig. 2.

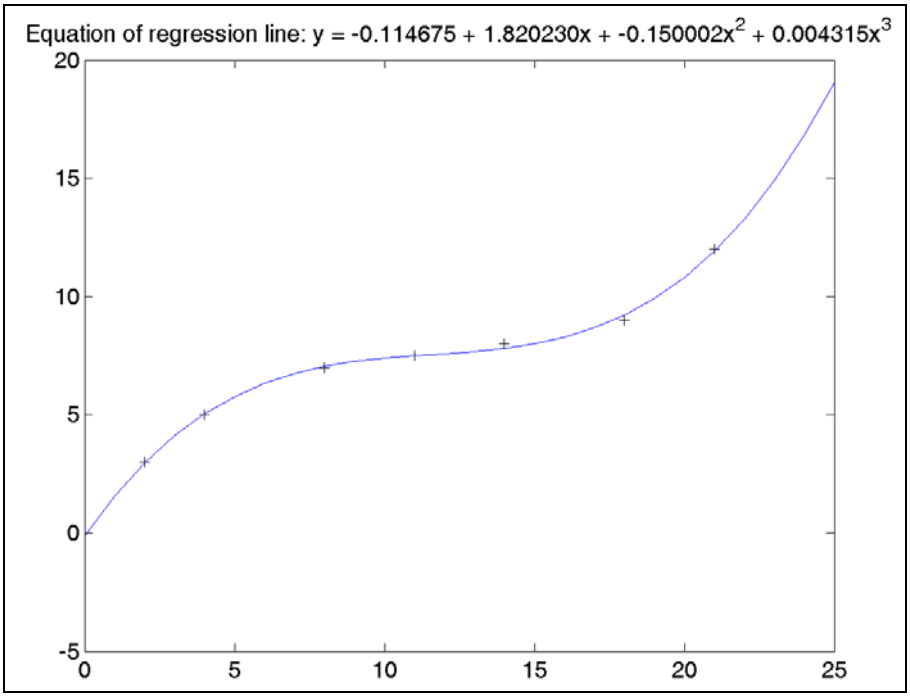


Fig. 2. Result of MATLAB third-order regression.

Regression can also be performed in MATLAB using the built-in commands, as discussed in Appendix D.

## STANDARD DEVIATION OF DATA POINTS

Standard deviation is a measure of how spread-out normally-distributed data is—how far individual points are from the mean value. The usual equation for standard deviation is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}, \tag{9}$$

where

n = the total number of data points,

$y_i$ = an individual data point, and

$\bar{y}$ = the average of all of the data points.

If a set of data has a small standard deviation, that means the data points are closely clustered around the mean value. Fig. 3 shows the standard plot for a normal distribution and indicates how much of the data is contained within 1, 2, and 3 standard deviations.

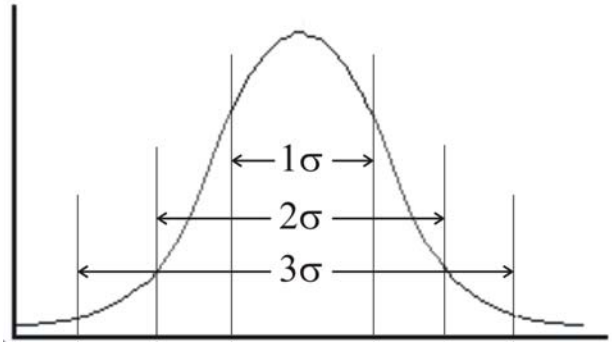


Fig. 3. Standard deviation.

Approximately 99% of the data points are within 3 standard deviations of the mean. This 3σ standard is frequently used to determine whether a given result is valid, and the same concept can be applied to a regression analysis.

When collecting data on a phenomenon which has a linear trend, it is expected that the majority of the points will lie close to the line which was found by linear regression. It can be assumed that the distance of the points from this line will have a normal distribution, like that shown in Fig. 3. The standard deviation of this distribution can be found, using

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - y_{Li})^2}{n-1}}, \tag{10}$$

where $y_{Li}$ is the y-value of the regression line at the x-value corresponding to $y_i$. Once this standard deviation has been found, lines can be added to the plot 3σ above the regression line and 3σ below. This is shown in Fig. 4.

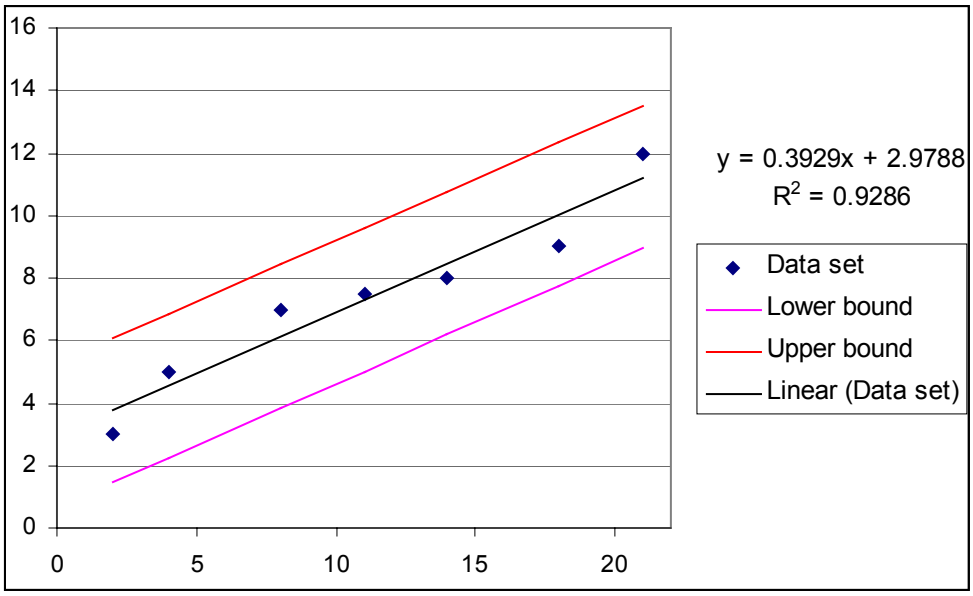


Fig. 4. Use of standard deviation to determine validity of regression line or data points.

A detailed explanation of how this plot was produced can be found in Appendix E.

Note that, with this method, it is being assumed that the regression line is the "correct answer" and the distribution of points around the line is found. Therefore, if a data point lies outside this 3σ range, it could mean one of two things. First, the regression line could be valid, and therefore there is a 99% chance that data point itself invalid. Second, the regression line itself could be incorrect, and the data point is fine.

# APPENDIX A
# Plotting Data Points With Excel

This Appendix details how to plot the data points, as required by Exercise A-1.

First, enter the data points, given in Table 1, in two consecutive columns as shown in Fig. A-1.

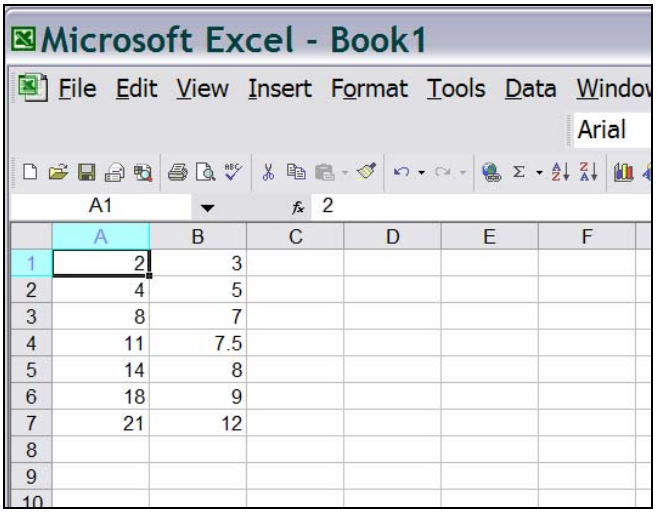


Fig. A-1. Entered data points.

Select Insert, Chart, or select the Chart Wizard icon on the toolbar. The Chart Wizard window will appear. Select XY (scatter) as the Chart type, and select the Chart sub-type which consists of data points only, as shown in Fig. A-2.

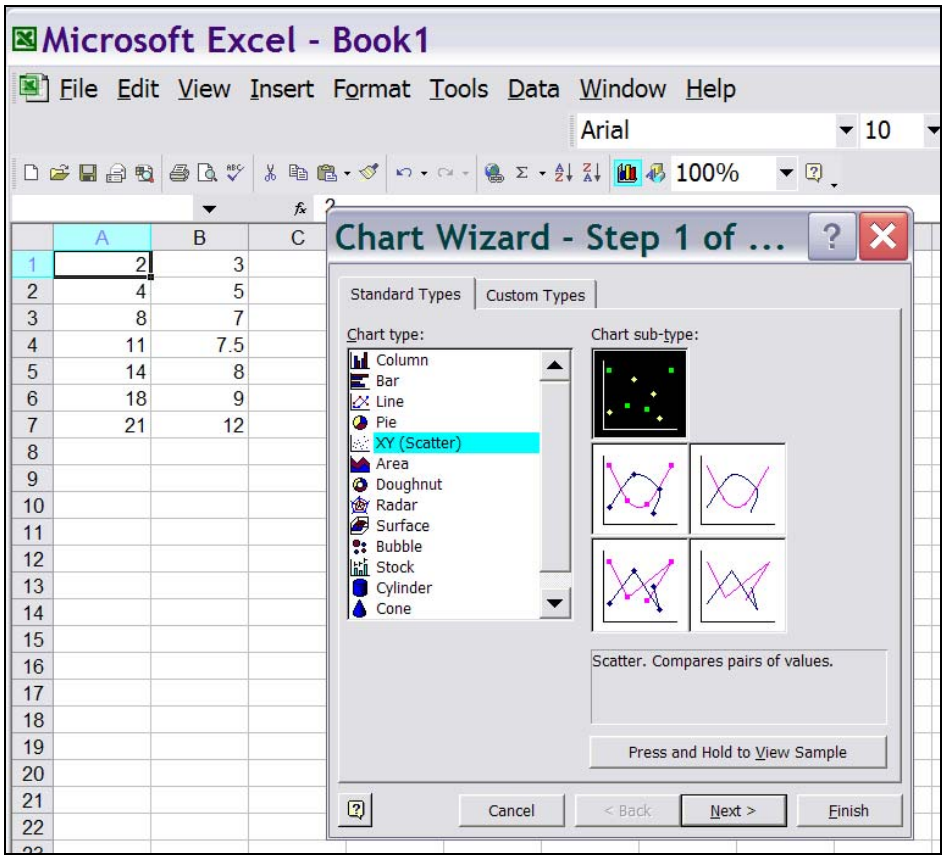


Fig. A-2. Selection of chart type and sub-type.

Click Next, and the Chart Wizard should automatically plot the points, creating a plot like that shown in Fig. A-3.

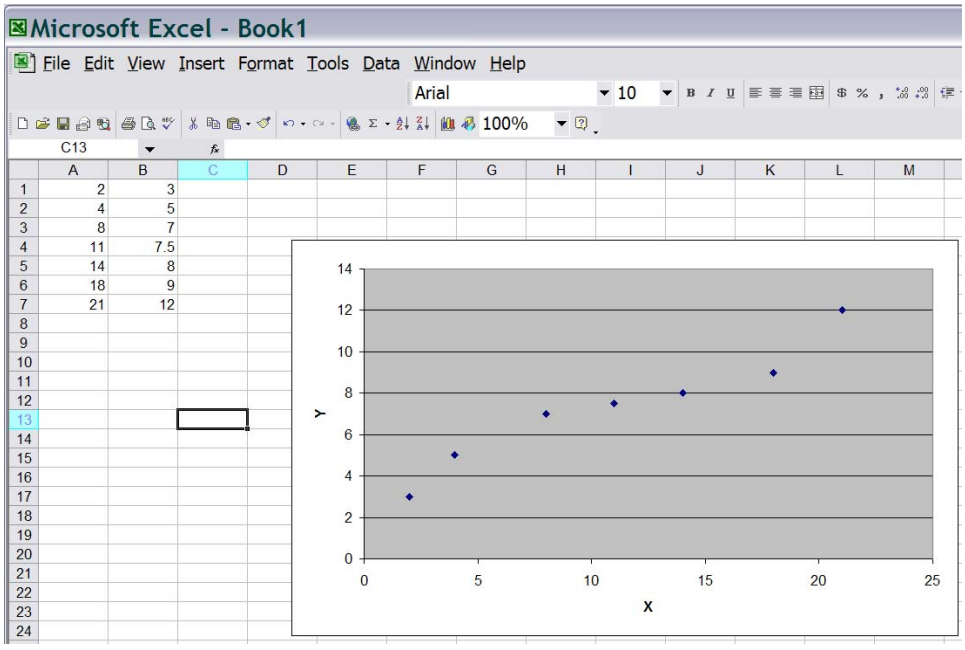


Fig. A-3. Completed chart.

# APPENDIX B
# Plotting Data Points With MATLAB

```
% Define vectors for the x and y values of the data points:

x = [2;4;8;11;14;18;21];
y = [3;5;7;7.5;8;9;12];

% Plot the data points.
% This plots the points as black '+' symbols.

plot(x,y,'k+')

% Set the x and y axes limits so that all of the data points
% can be clearly seen.

xlim([0 22])
ylim([0 13])

% Display a title.

title('Data points')
```

# APPENDIX C
# Regression Analysis with MATLAB, method 1

The MATLAB program below can be used to perform a third-order regression on a set of 7 data points. This program implements the least squares regression method, without using any of the MATLAB built-in regression tools. This is not the most straightforward way to perform regression in MATLAB, but it is helpful in better understanding the theory behind the technique.

```
% Performs a third-order regression on 7 data points

% Define data points being considered:
x = [2;4;8;11;14;18;21];
y = [3;5;7;7.5;8;9;12];

% The matrix which multiplies the coefficient matrix is A.
% The terms needed for the A matrix are:
sumx=sum(x);
sumx2=sum(x.^2);
sumx3=sum(x.^3);
sumx4=sum(x.^4);
sumx5=sum(x.^5);
sumx6=sum(x.^6);

% The matrix on the right side of the equation is B.
% The terms needed for the B matrix are:
sumy=sum(y);
sumyx=sum(y.*x);
sumyx2=sum(y.*x.^2);
sumyx3=sum(y.*x.^3);

% Define A and B.
% Note that the term in the top left corner of the A matrix is equal to the
% number of data points being used, 7 in this case.
A = [7,sumx,sumx2,sumx3;sumx,sumx2,sumx3,sumx4;sumx2,sumx3,sumx4,sumx5;...
        sumx3,sumx4,sumx5,sumx6];
B = [sumy; sumyx; sumyx2; sumyx3];

% The coefficient vector is the inverse of A multiplied by B:
Coeff = inv(A)*B;

% Plug the found values for the coefficients into the form for the fitted
% curve (a cubic equation, in this case):
curvex=linspace(0,25,26);
for i = 1:26;
    curvey(i)=Coeff(1)+Coeff(2)*curvex(i)+Coeff(3)*curvex(i)^2+Coeff(4)*...
        curvex(i)^3;
end

% Create a string variable of the equation, to be used as the title for the
% plot:
equation = sprintf...
    ('Equation of regression line: y = %f + %fx + %fx^2 + %fx^3',...
```

```
    Coeff(1),Coeff(2),Coeff(3),Coeff(4));

% Plot the original data points along with the fitted curve:
plot(x,y,'k+',curvex,curvey)
title(equation)
```

# APPENDIX D
# Regression Analysis with MATLAB, method 2

This Appendix provides instruction on performing regression in MATLAB, using the built-in regression tools. The method for determining the $R^2$ value will also be covered.

The command used to perform curve fitting in MATLAB is:

```
p = polyfit(x,y,n)
```

This function finds the coefficients of a polynomial, p(x), which provides a least-squares best fit to the data provided. The inputs to the function are x, y and n, where

x is a vector containing the x-values of the data points,
y is a vector containing the y-values of the data points, and
n is the order of the polynomial.

The output of `polyfit`, p in this case, is a vector containing the coefficients of the polynomial, starting with the highest order term. For example, the vector

```
p = [5 12 3 1]
```

represents the polynomial

$$5x^3 + 12x^2 + 3x + 1$$

The command `polyval` can be used to plot the resulting polynomial. The syntax for `polyval` is

```
f = polyval(p,x)
```

where p is the array containing the polynomial's coefficients, and x is the original vector of x-values. The polynomial is therefore being evaluated at these x-values, and the result (f) is a vector of the y-values. To plot the original data points along with the regression line, simply enter

```
plot(x,y,'o',x,f,'-')
```

This will plot the original data points with small circles, and the polynomial curve fit as a line.

To calculate the $R^2$ value, the mean, J value and S value must first be found. The mean is simply found using

```
mu = mean(y)
```

The J value is

```
J = sum((f-y).^2)
```

Recall that f is the polynomial evaluated at the x-values, and y contains the original y-values of the data points. The S value is

```
S = sum((y-mu).^2)
```

Using these, the $R^2$ value may then be calculated using

```
r2 = 1-J/S
```

APPENDIX E

# Standard Deviation of Data Points Around Regression Line

This Appendix details how the plot in Fig. 4 was produced. The Excel spreadsheet used is shown in Fig. B-1.

Microsoft Excel - Regression Tutorial Excel.xls

File Edit View Insert Format Tools Data Window Help

C5 = =0.3929*A5+2.9788

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | | | | | | | |
| 2 | | | | | | | |
| 3 | | | y-value of | | | lower | upper |
| 4 | x-value | y-value | regression line | difference | | bound | bound |
| 5 | 2 | 3 | 3.76 | -0.76 | | 1.47 | 6.06 |
| 6 | 4 | 5 | 4.55 | 0.45 | | 2.25 | 6.85 |
| 7 | 8 | 7 | 6.12 | 0.88 | | 3.82 | 8.42 |
| 8 | 11 | 7.5 | 7.30 | 0.20 | | 5.00 | 9.60 |
| 9 | 14 | 8 | 8.48 | -0.48 | | 6.18 | 10.78 |
| 10 | 18 | 9 | 10.05 | -1.05 | | 7.75 | 12.35 |
| 11 | 21 | 12 | 11.23 | 0.77 | | 8.93 | 13.53 |
| 12 | | | | | | | |
| 13 | | | Standard deviation | 0.77 | | | |
| 14 | | | 3 Standard deviations | 2.30 | | | |
| 15 | | | | | | | |

Fig. B-1. Excel spreadsheet used to produce plot in Fig. 4.

To begin, plot the 7 data points (columns A and B). Then perform a linear regression, electing to show the equation on the chart. Using this equation, calculate the y-values of the regression line at each of the x-values in column A. These calculated values are those shown in column C. For this example, the equation

=0.3929*A5+2.9788

was entered in the first cell of column C (cell C5 above) and copied and pasted into the remaining cells in the column.

The difference, in column D, is the difference between the y-value of the data point and the y-value of the regression line. This was calculated by entering

=B5-C5

in cell D5 and copying to the other cells in the column.

The standard deviation of the difference, the values in column D, must then be found. Excel has a built-in function to calculate standard deviation, so the value in cell D13 was calculated by entering

=STDEV(D5:D11)

which finds the standard deviation of the values in cells D5 through D11. In this case, the value of $3\sigma$ is desired, which is simply 3 times the value in D13. This was entered in D14.

The upper and lower bounds (the red and pink lines in Fig. 4) were found by adding and subtracting the $3\sigma$ value from the y-value of the regression line. To find the lower bound, enter

=0.3929*A5+2.9788-$D$14

in F5, copying to the other cells in the column. To find the upper bound, enter

=0.3929*A5+2.9788+$D$14

in G5, copying down to the other cells. Finally, plot the data points and their regression line, along with the upper and lower bounds vs. the x-values.