

Linear Least Squares Regression ©

by

Sidney H. Young and Andrzej Wierzbicki
Department of Chemistry
University of South Alabama
Mobile, AL. 36688

syoung@jaguar1.usouthal.edu

© Copyright 2000 by the Division of Chemical Education, Inc., American Chemical Society. All rights reserved. For classroom use by teachers, one copy per student in the class may be made free of charge. Write to JCE Online, jceonline@chem.wisc.edu, for permission to place a document, free of charge, on a class Intranet.

Introduction

Linear least squares regression is the workhorse of the Physical Chemistry Laboratory. This template demonstrates various implicit and explicit methods for determination of the slope and intercept of the regressed line. It also produces the correlation coefficient, 95% confidence ranges, standard deviation of fit, standard deviation of slope and standard deviation of intercept. Residual analysis is used to demonstrate techniques of removing bad data points from the fit. This template reads data from a file, allowing the template to be used as a general analysis tool.

Occasionally, the data is poorly fit with linear least squares. In this case, a quadratic term, or some other term, can be added to improve the fit. A second set of data, one that requires a quadratic term, will also be analyzed in this document.

Goal

To teach the techniques of linear least squares regression, and to be a practical aid in Physical Chemistry Laboratory.

Prerequisites

1. Moderate skill with Mathcad in performing simple calculations and preparing plots.
2. Familiarity with the basic concepts of linear regression.

Performance Objectives

At the end of the exercise, you will be able to:

1. determine the slope and intercept of a regression line, given x-y data pairs;
2. find the standard deviation of fit, standard deviation of slope, and standard deviation of intercept;
3. evaluate and interpret the correlation coefficient and understand its significance;
4. use residual analysis to look for bad data points;
5. observe the effect of adding a quadratic term and testing to see if the quadratic term is statistically significant.

I. HOW TO INPUT DATA FROM AN ASCII FILE

The input data for the first part of the template is in the file "data1.prn". This ASCII file consists of two columns, containing pairs of x-y data.

The data file, "data1.prn", should be placed in the same directory where this template resides. Use READPRN(data1) function to read the data into the template in Mathcad 6.0. For Mathcad 7 or higher, use READPRN("data1.prn").

$Z := \text{READPRN}(\text{"data1.prn"})$ Z is 2xN vector of x,y data

The "READPRN" function reads the data in the file into a vector Z. This is the standard method to input large data sets into MATHCAD.

Other methods are available. For example, one could use an Excel spreadsheet and generate an ASCII file. Place the x-data in column A and y-data in column B. Then save the spreadsheet as **text (tab delimited)**. The text can then be read in using the READPRN function as described above.

Exercise 1: Take a set of x-y data, place it into a spreadsheet, create a tab delimited file, and then read this file into a Mathcad worksheet.

$Y := Z^{<1>}$ $X := Z^{<0>}$ $N := \text{length}(X)$

Here we used the data in data1.prn after it was put into the Z array by the READPRN function. Notice how we used the array superscript operator to equate a column of the matrix Z with a variable name. To perform this procedure, type Y:Z followed by Ctrl-6. Then type in the column number between the brackets.

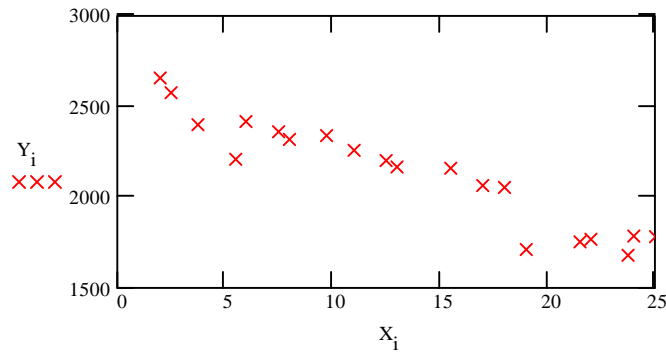
X and Y are vectors of length N containing the x and y data points of the data array Z.

Note that the numbering system for rows and columns of vectors and matrices in Mathcad is such that the first column is designated "0", the second column is designated "1", etc.

The length function is useful to find the number of elements in a vector so that you do not have to count the number of input data points that you are entering.

On the next page is a plot of the x-y data to be analyzed.

$i := 0..(N - 1)$



INPUT DATA

II. HOW TO DETERMINE THE SLOPE AND INTERCEPT OF A FUNCTION

1. Built-in functions

MATHCAD contains built in functions for slope and intercept, as shown below:

$$m := \text{slope}(X, Y) \quad m = -37.154$$

$$b := \text{intercept}(X, Y) \quad b = 2.628 \cdot 10^3$$

We traditionally use m for slope and b for intercept.

2. Explicit functions

The model for least squares is based upon the idea that the standard deviation of fit should be minimized. One determines the parameters which result in this minimization.

Thus:

$$\text{stdev} = \sum_{i=0}^{N-1} (y_i - m \cdot x_i - b)^2 \text{ is minimized by setting } \frac{d}{dm} \text{stdev} = 0 \text{ and } \frac{d}{db} \text{stdev} = 0$$

Exercise 2: Show that the two equations resulting from setting the above derivatives to zero leads directly to the equations given below. Derivations like this can be found in many physical chemistry laboratory texts and general statistics texts. Find such a derivation and compare the slope and intercept definitions to those given here.

The following procedure is from Draper and Smith, "Applied Regression Analysis", New York: Wiley (1981)

$$SXX := \sum_{k=0}^{N-1} (X_k)^2 - \frac{\left(\sum_{k=0}^{N-1} X_k \right)^2}{N} \quad SXX = 1.107 \cdot 10^3$$

$$SXY := \left[\sum_{k=0}^{N-1} (X_k \cdot Y_k) \right] - \frac{\left(\sum_{k=0}^{N-1} X_k \right) \cdot \left(\sum_{k=0}^{N-1} Y_k \right)}{N} \quad SXY = -4.111 \cdot 10^4$$

$$m := \frac{SXY}{SXX} \quad m = -37.154$$

Compare this value of the slope to what you would obtain using the Mathcad built-in function for the slope.

$$ybar := \frac{\left[\sum_{k=0}^{N-1} (Y_k) \right]}{N}$$

$$xbar := \frac{\left[\sum_{k=0}^{N-1} (X_k) \right]}{N}$$

$$b := ybar - m \cdot xbar$$

$$b = 2.628 \cdot 10^3$$

Compare this value of the intercept to what you would obtain using the Mathcad built-in function for the intercept.

3. Matrix Function

Another method to produce the least-squares equations is to use matrix methods. Although more intricate and abstract, the matrix method can easily be extended for quadratic least squares or multiple least squares regression. These formulas are derived in "Draper and Smith" and other texts.

If one defines a Mx2 matrix XX (M=number of data points) and a M vector YY, then the least squares parameters can be automatically determined as given below.

$$XX_{i,0} := 1 \qquad XX_{i,1} := X_i \qquad YY_i := Y_i$$

XX has the form $\begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$ and YY has the form $\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$ for 5 data points.

$$\text{params} := (XX^T \cdot XX)^{-1} \cdot XX^T \cdot YY$$

This is a 2x1 vector in which params_0 is the intercept and params_1 is the slope. Use your knowledge of matrix multiplication to verify that params is a 2x1 vector

$$\text{params} = \begin{bmatrix} 2.628 \cdot 10^3 \\ -37.154 \end{bmatrix}$$

This gives intercept and slope directly and has the added advantage that it can easily be adapted for use in multivariate least squares. **Compare the values in params to the slope and intercept computed on the previous page.**

The following is a proof that the matrix method is equivalent to the previous method.

First we will first develop the matrices that we need for the proof.

Consider an example, consisting of three data points, called (x_1,y_1) , (x_2,y_2) , and (x_3,y_3) .

$$XX^T = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \end{bmatrix} \quad XX = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix}$$

$$XX^T \cdot XX = \begin{bmatrix} 1+1+1 & x_1+x_2+x_3 \\ x_1+x_2+x_3 & (x_1)^2+(x_2)^2+(x_3)^2 \end{bmatrix}$$

Identify the elements in this 2x2 array. Verify each element using the identities given immediately above.

In general,
$$XX^T \cdot XX = \begin{bmatrix} N & \sum_k x_k \\ \sum_k x_k & \sum_k (x_k)^2 \end{bmatrix}$$

N = number of data points (3 in this demonstration). **What is the value of k in the summations in the array shown here?**

The inverse of a matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is given by $\frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{ad - bc}$

$$(XX^T \cdot XX)^{-1} = \frac{\begin{bmatrix} \sum_k (x_k)^2 - \left(\sum_k x_k\right)^2 & \\ -\left(\sum_k x_k\right) & N \end{bmatrix}}{N \cdot \left[\sum_k (x_k)^2 \right] - \left(\sum_k x_k\right)^2}$$

When would the inverse of a matrix be undetermined?

$$YY = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \quad XX^T \cdot YY = \begin{bmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} y_1 + y_2 + y_3 \\ x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 \end{bmatrix} = \begin{bmatrix} \sum_k y_k \\ \sum_k x_k \cdot y_k \end{bmatrix}$$

Finally,

$$(XX^T \cdot XX)^{-1} \cdot (XX^T \cdot YY) = \frac{\begin{bmatrix} \sum_k (x_k)^2 - \left(\sum_k x_k\right)^2 & 0 \\ 0 & N \end{bmatrix} \cdot \begin{bmatrix} \sum_k y_k \\ \sum_k x_k \cdot y_k \end{bmatrix}}{N \cdot \left[\sum_k (x_k)^2 \right] - \left(\sum_k x_k\right)^2} = \begin{bmatrix} b \\ m \end{bmatrix}$$

$$b = \frac{\left[\sum_k (x_k)^2 \right] \cdot \left(\sum_k y_k\right) - \left(\sum_k x_k\right) \cdot \left(\sum_k x_k \cdot y_k\right)}{N \cdot \left[\sum_k (x_k)^2 \right] - \left(\sum_k x_k\right)^2}$$

$$m = \frac{\left[-\left(\sum_k x_k\right) \cdot \left(\sum_k y_k\right) + N \cdot \left(\sum_k x_k \cdot y_k\right) \right] \cdot \sum_k x_k \cdot y_k - \frac{\left(\sum_k x_k\right) \cdot \left(\sum_k y_k\right)}{N}}{N \cdot \left[\sum_k (x_k)^2 \right] - \left(\sum_k x_k\right)^2} = \frac{SXY}{SXX}$$

Exercise 3: If time permits, show that the intercept b , defined above, is equivalent to the expected results using the explicit formulas for slope and intercept derived in exercise 2. (This is a difficult problem; see the answer key for results if necessary.)

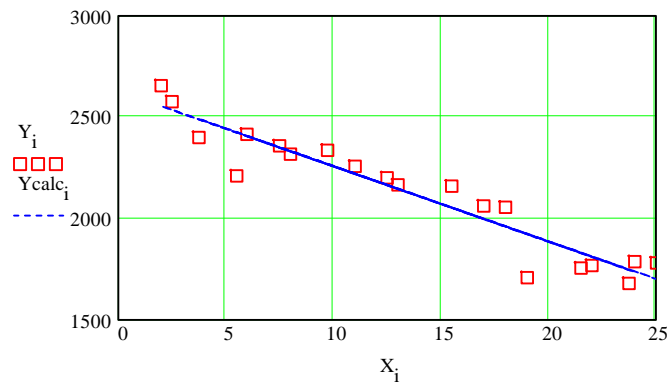
III. HOW TO PRESENT DATA

When presenting data, the preferred method is to include both the least squares line and the raw data as points around the line. To do this, use the appropriate Create a Graph technique from your version of Mathcad. For this exercise you should choose X-Y Plot.

Enter both Y_i and Y_{calc}_i (defined below) as ordinates, with X_i as the abscissa. Then click anywhere inside the graph. Choose **traces** from the pull down menu. For trace 1 (Y) choose the type to be **points** and the symbol (I chose **box**). For trace 2 choose the type to be **line**. This will produce the desired graph.

$$Y_{calc}_i := m \cdot X_i + b$$

Y is the original data and Y_{calc} is a vector of predicted values of Y from the least squares fit. Notice the point style in the figure below.



Exercise 4 - Use the method of least squares to find the slope and intercept of the line $y=mx+b$ that best fits the 4 points $(-3,3)$, $(-1,4)$, $(1,8)$, and $(3,9)$. Try all three methods: built-in functions, explicit formulas, and matrix method. Plot the points and the fitted line on an XY graph.

IV. ANALYSIS OF VARIANCE

Very often we would like to know something about the goodness of fit of an equation to a set of data. This means we need to know about the quality of the parameters used to define the fitted function..

Consider a general set of x-y data. We would like to observe what happens to y as x changes. Consider how y deviates from its mean value as a function of x. Two possibilities are 1) y varies randomly, independent of x and 2) y changes monotonically along with x. The former is due to random fluctuations in the data and the latter is due to the functional relationship between x and y. In the linear case, we can easily determine these two sources of deviations.

In the variables below, SYY is the sum of squares between y and its mean value. SSE is the sum of squares due to random fluctuations and SSR is the sum of squares due to the regression (relationship) between x and y.

For data that is exactly on a straight line, SSE = 0 and SYY=SSR. Alternatively, if the data were randomly situated around the mean, there would be no regression at all, SSR=0 and SSE=SYY. The latter does not usually occur in practice, since random fluctuations sometimes seem to act as if a functional relationship were present.

A. DETERMINATION OF THE CORRELATION COEFFICIENT

For the data1.prn set of data we obtain:

$$SYY := \sum_{k=0}^{N-1} (Y_k)^2 - \frac{\left(\sum_{k=0}^{N-1} Y_k \right)^2}{N} \quad SYY = 1.694 \cdot 10^6 \quad \text{Total sum of squares}$$

$$SSE := \sum_{k=0}^{N-1} (Y_k - m \cdot X_k - b)^2 \quad SSE = 1.663 \cdot 10^5 \quad \text{Residual sum of squares}$$

NOTE: If the regression is perfect, then there is no "residual" error and SSE = 0.

$$SSR := SYY - SSE \quad SSR = 1.527 \cdot 10^6 \quad \text{Regression sum of squares}$$

Notice how SYY, SSE, and SSR are large numbers.

The correlation coefficient R^2 , defined below, gives the fraction of the entire $SY\bar{Y}$ which is due to the regression only. If the regression were perfect, then $R^2 = 1$ or -1 . If the data were randomly placed around the mean value, $R^2=0$. In general, we wish R^2 to be close to 1 or -1 . Typically, we consider a correlation coefficient above 0.95 to indicate a good fit.

WARNING: If the magnitude of y is much smaller than x , then R^2 approaches 0/0, and gives meaningless numbers.

$$R^2 := \left(\frac{SSR}{SY\bar{Y}} \right) \quad R^2 = 0.902$$

Compare this result to that obtained when the Mathcad defined function is used.

Exercise 5 - For the data in Exercise 4, determine the correlation coefficient of the fit.

B. STANDARD DEVIATION OF FIT

The "mean square error", MSE, is defined as the residual sum of squares divided by the number of degrees of freedom, which in the linear case is $N-2$. (One degree of freedom for each data point, minus the two constraints: the slope and intercept.) This term is often used as an estimate for the standard deviation of fit.

$$MSE := \frac{SSE}{N - 2}$$

$$MSE = 9.236 \cdot 10^3$$

SSE is the total residual of all the points, and MSE is the residual per degree of freedom. The residual mean squares is often called 'chi-squared'.

Exercise 6: What is more useful in determining the uncertainty of fit, SSE or MSE? Why?

C. STANDARD DEVIATION OF SLOPE AND INTERCEPT

Since many applications of Physical Chemistry involve linear least squares in which the slope or intercept is related to a physically meaningful parameter, it is useful to determine the standard deviation in the slope or intercept. The following formulas are derived in Seber, "Linear Regression Analysis", New York: Wiley (1977) Note that the use of the matrix method here causes the equations to be more accessible, and these equations generalize to the polynomial or nonlinear case more easily.

$$sb := \sqrt{MSE \cdot \left[(XX^T \cdot XX)^{-1} \right]_{0,0}}$$

$$sb = 44.184$$

Standard deviation of intercept

$$sm := \sqrt{MSE \cdot \left[(XX^T \cdot XX)^{-1} \right]_{1,1}}$$

$$sm = 2.889$$

Standard deviation of slope

Exercise 7: The standard deviation of slope is quite important in the Physical Chemistry lab, since slope is often related to a physical value of interest. For example, suppose that you are determining the activation energy of a reaction using the Arrhenius equation. If a plot of $\ln k$ vs. $1/T$ had a slope of -2.34 K with a standard deviation of slope 0.09 K , determine the value and standard deviation in the activation energy. Explain in words the expression used to compute the standard deviation of the slope.

D. CONFIDENCE RANGES FOR SLOPE AND INTERCEPT

Another useful quantity is the "95% confidence range". This determines a range of values for a parameter such that there is a 95% chance that the parameter will fall within this range due to random errors. The range is defined as **parameter $\pm (t \cdot \text{STDEV})$** , where t is "Student's t " and STDEV is the standard deviation associated with a parameter. The value of t depends upon the number of data points, and is usually near 2.

To obtain the "student's t " value we use a built in Mathcad function. This function is **$qt((1 - (1-\text{range})/2), \{N-2\})$**

Exercise 8: Find a table of Student's t -test values and show that for 10 data points, the value of t from the table for 95% confidence level is equal to Mathcad's $qt(0.975,10)$.

Here we calculate the minimum and maximum values for the slope and intercept using the "student's t " values. these calculations refer to the data in data1.prn.

$m_{\min} := m - qt(0.975, N - 2) \cdot sm$	$m_{\max} := m + qt(0.975, N - 2) \cdot sm$	
$m_{\min} = -43.223$	$m_{\max} = -31.084$	range for the slope
$b_{\min} := b - qt(0.975, N - 2) \cdot sb$	$b_{\max} := b + qt(0.975, N - 2) \cdot sb$	
$b_{\min} = 2.535 \cdot 10^3$	$b_{\max} = 2.721 \cdot 10^3$	range for intercept

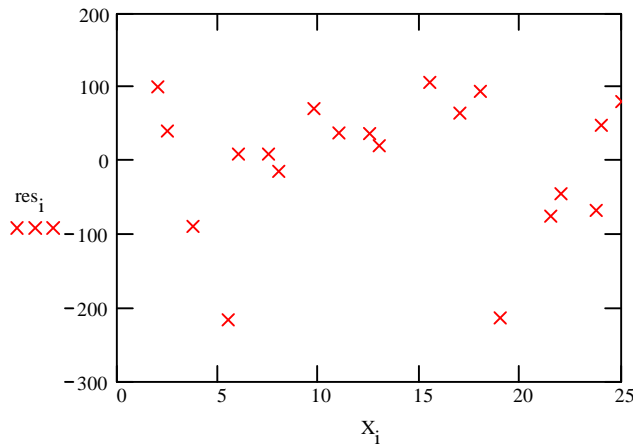
What % of m and b is represented by the confidence limits calculated here?

Exercise 9: A good "rule of thumb" is that qt is nearly 2 for most applications (at 95% confidence interval). Check this by making a table of qt vs. $N-2$ for N going from 4 to 100.

V. RESIDUAL ANALYSIS

The residual of a data point is the difference between the experimental and calculated value of y . A graph of residuals vs. X can be used to determine bad points, or to determine whether weighted least squares is needed.

$$\text{res}_i := (Y_i - m \cdot X_i - b) \quad \text{Definition of residual for point } i.$$



Graph of res_i vs X_i

The points plotted here are plotted in order of increasing value of X_i .

Note the two points below -200 ; they might be candidates for removal. Look at the **res** and **X** vectors to determine the row number of the points to be discarded. The two points we wish to discard are the 4th and 5th in the array of the original data. Be careful to note that the numbering system in Mathcad places the first element of a matrix row or column in row or column zero.

In the following, the suffix REV refers to the revised data set, with the questionable points removed.

$$\begin{aligned} \text{iii} &:= 0..3 & \text{XREV}_{\text{iii}} &:= \text{X}_{\text{iii}} & \text{YREV}_{\text{iii}} &:= \text{Y}_{\text{iii}} & \text{This could also be done by} \\ & & & & & & \text{producing submatrices} \\ \text{iii} &:= 6..N-1 & \text{XREV}_{\text{iii}-2} &:= \text{X}_{\text{iii}} & \text{YREV}_{\text{iii}-2} &:= \text{Y}_{\text{iii}} & \text{and augmenting them} \\ \text{iii} &:= 0..N-3 & & & & & \text{together.} \end{aligned}$$

$$\text{XXREV}_{\text{iii},0} := 1 \quad \text{XXREV}_{\text{iii},1} := \text{XREV}_{\text{iii}} \quad \text{YYREV}_{\text{iii}} := \text{YREV}_{\text{iii}}$$

Exercise 10: Show that the revised matrices above consist of the original matrices with rows 5 and 6 removed.

Here we repeat the process used on page 5. Compare this page to page 5 to be sure you understand the basic concepts.

$$\text{paramsREV} := (\text{XXREV}^T \cdot \text{XXREV})^{-1} \cdot \text{XXREV}^T \cdot \text{YYREV}$$

$$\text{paramsREV} = \begin{bmatrix} 2.659 \cdot 10^3 \\ -37.694 \end{bmatrix}$$

$$\text{params} = \begin{bmatrix} 2.628 \cdot 10^3 \\ -37.154 \end{bmatrix}$$

Remember that
params is equal
to $\begin{bmatrix} \text{intercept} \\ \text{slope} \end{bmatrix}$

$$\text{mREV} := \text{paramsREV}_1$$

$$\text{bREV} := \text{paramsREV}_0$$

Note that removal of two data points does not change the slope and intercept significantly. However, removing two out of twenty data points reduced the MSE by more than a factor of two! This is shown here.

$$\text{MSEREV} := \sum_{k=0}^{N-3} \frac{(\text{YREV}_k - \text{mREV} \cdot \text{XREV}_k - \text{bREV})^2}{N-4}$$

$$\text{MSEREV} = 3.965 \cdot 10^3$$

$$\text{MSE} = 9.236 \cdot 10^3$$

Now consider the uncertainties in slope and intercept

$$\text{sbREV} := \sqrt{\text{MSEREV} \cdot [(\text{XXREV}^T \cdot \text{XXREV})^{-1}]_{0,0}}$$

The revised uncertainty in intercept is $\text{sbREV} = 30.533$, whereas the original uncertainty was $\text{sb} = 44.184$ The uncertainty has dropped by 31%.

$$\text{smREV} := \sqrt{\text{MSEREV} \cdot [(\text{XXREV}^T \cdot \text{XXREV})^{-1}]_{1,1}}$$

$$\text{smREV} = 1.979$$

$$\text{sm} = 2.889$$

The revised uncertainty in slope is $\text{smREV} = 1.979$, whereas the original uncertainty was $\text{sm} = 2.889$ The uncertainty has dropped by 31%.

Therefore, removing the two questionable data points sometimes significantly reduces the standard deviation of slope and standard deviation of intercept.

There is a stronger statistical test that may be used to determine whether a bad data point should be discarded. The method is called ***Chauvenet's criterion***.

Suppose that you have a set of residuals, and one residual is unusually large and a possible candidate for deletion. Determine the standard deviation of the residuals using the mathcad function `stdev(x)`, where `x` is a vector of residuals. Since the mean of the set of residuals is zero, we can calculate the number of standard deviations by which the

suspect residual differs from zero, that is, $t_{\text{suspect}} = \frac{|\text{res}_{\text{suspect}}|}{\text{stdev}(\text{res})}$. The probability of obtaining a point that is expected to deviate as much as the suspect point is given by the formula $\frac{dt(t_{\text{suspect}}, N)}{2}$ where `N` is the total number of points and `dt` is the density

function. We divide by 2 because the probability calculated with `dt` is for both edges (positive and negative) of the probability distribution. For `N` data points, the number of measurements that would be expected to deviate by this amount is $N \cdot \frac{dt(t_{\text{suspect}}, N)}{2}$. If this probability is less than one-half of one data point, then the data point may be rejected.

Note that there are several objections to this method, so that it should be used with caution.

For example, consider a set of residuals:

```

resid :=
[ 0
-2
 2
 4
-3
-1
-6
 1
-1
 1
 1
 2
-1
 3 ]

```

```

mean(resid) = 0
stdev(resid) = 2.507

```

Note: `dt` is the probability density for the student's `t` distribution. It is one of Mathcad's built in functions.

In Mathcad 7 and 8 use `Stdev` instead of `stdev`. `Stdev` is the sample standard deviation while `stdev` is the population standard deviation.

Consider the residual -6. $t_{\text{sus}} := \frac{|-6|}{\text{stdev}(\text{resid})}$ $t_{\text{sus}} = 2.393$

The probability that a point would deviate from the normal distribution by as much as t_{sus} is given by $\frac{dt(t_{\text{sus}}, 14)}{2} = 0.015$ For 14 data points, the number of data points expected to be found in this range is $14 \cdot \frac{dt(t_{\text{sus}}, 14)}{2} = 0.21$ Since this value is less than 0.5, we are justified in throwing out the data point.

The data point may be removed by sorting the residuals and truncating the resulting vector. Then the standard deviation may easily be recalculated.

`residsort := reverse(sort(resid))`

This is the sorted matrix.

`stdev(submatrix(residsort, 12, 0, 0, 0)) = 1.946`

Nearly a 20% drop in stdev by removing one point in 14.

Exercise 11 - Use the data "res" given at the beginning of section V, and use Chauvenet's criterion to determine whether the two points near -200 should be removed, as we have done above.

For further information, see "**Rejection of Data**", by Kevin Lehmann, Princeton University, in the statistics section of the Mathcad in the Chemistry Curriculum website.
<http://www.monmouth.edu/~zielins/mathcad/index.htm>

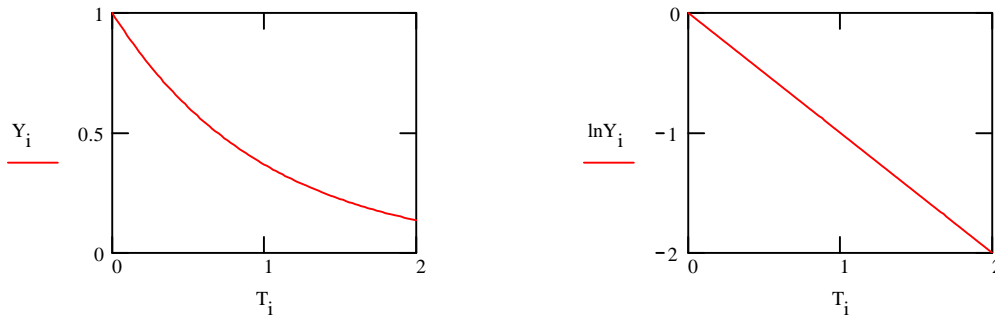
VI. WEIGHTED LEAST SQUARES REGRESSION

A. INTRODUCTION

Consider the following problem. First-order kinetics data are fit using a $\ln(x)$ vs. t plot.

$$i := 0..100 \quad T_i := \frac{i}{50} \quad k := 1 \quad Y_i := e^{-k \cdot T_i} \quad \ln Y_i := \ln(Y_i)$$

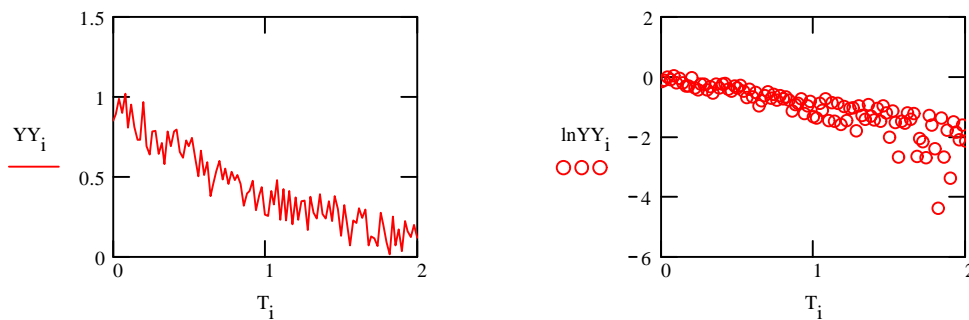
In this example, 50 data points are simulated and fit assuming a first-order decay with unity rate constant.



Now, let's add some random noise. since most instruments are calibrated to produce data proportional to the concentration (absorbance, angles of rotation with optically active species, fluorescence intensities, etc.), then random noise would be uniform with respect to conc. vs. time. Adding random noise, the following graphs are produced.

$$YY_i := Y_i + (\text{rnd}(0.3) - 0.15) \quad \ln YY_i := \ln(YY_i) \quad \text{YY is the vector of Y components with noise added.}$$

Note: Annotation will be reduced in following sections. Methodologies are similar to what has been done so far. Students should master the material on pages 1-16 in order to be able to work through the following sections efficiently.



Note the surprising result: Noise which is linear when plotting conc. vs. time produces nonlinear noise when $\ln(\text{conc})$ is plotted with time! This phenomenon contradicts one of the assumptions of linear regression: that the uncertainty of each data point is constant.

To remove this problem, use weighted least squares or non-linear curve fitting on the original data. When using weighted least squares, each data point is "weighted" with respect to the uncertainty. Often the weights used are $1/\text{uncertainty}$. When non-linear curve fitting is used the uncertainty is the same for each measurement.

For our case, since we are taking the logarithm of the absorbance, a Taylor's series expansion of the log function gives a first term which produces the uncertainty.

$$\ln(a + x) \quad \text{converts to the series} \quad \ln(a) + \frac{1}{a} \cdot x + O(x^2)$$

Highlight x in the expression $\ln(a+x)$ and use "Symbolics", "Variable", "Expand to Series" from the Pull-down menu.

The interpretation is that if x is the noise added to a , then the noise becomes x/a when converted to logarithm. Thus the uncertainty of each point becomes proportional to $1/a$; this is why the uncertainties increase as concentration heads to zero as time proceeds.

Weights may also be determined by multiple measurement of a variable. Once the standard deviation of the measurement is determined, one generally sets the weight as $\frac{1}{\text{stdev}^2}$.

Exercise 12 - Suppose you have three measurements: (2 +/- 0.5), (3 +/- 0.5), and (2 +/- 1). Determine the weights used for each measurement.

B. PERFORMING WEIGHTED LEAST SQUARES

Least squares is most easily performed using the matrix method, by defining a new matrix. The weights are given by the following formula.

$$N := 101 \quad w_i := \left(\left(\frac{1}{\ln Y Y_i} \right) \right)$$

The matrix V is defined as 1/weights.

$$V_{i,i} := \frac{1}{w_i} \quad V \text{ is an } N \times N \text{ matrix with diagonal elements only.}$$

The slope and intercept of the fitted line are given by the following formula from "Draper and Smith". Note that for the logarithmic line, we are plotting $\ln Y Y$ vs. T .

$$X_{i,0} := 1 \quad X_{i,1} := T_i \quad Y_i := \ln Y Y_i$$

$$\text{weighted_matrix} := (X^T \cdot V^{-1} \cdot X)^{-1} \cdot X^T \cdot V^{-1} \cdot Y \quad \text{weighted_matrix} = \begin{bmatrix} 0.04 \\ -1.002 \end{bmatrix}$$

If all of the weights are unity, this equation reverts to the equation used earlier.

If we ignored least squares, we would obtain the following:

$$\text{un}V_{i,i} := 1 \quad \text{All weights are unity.}$$

$$\text{unweighted_matrix} := (X^T \cdot X)^{-1} \cdot X^T \cdot Y \quad \text{unweighted_matrix} = \begin{bmatrix} 0.062 \\ -1.119 \end{bmatrix}$$

There is thus a large difference between weighted and unweighted least squares. However, the residual analysis shown below, indicates that the unweighted least squares analysis, does not even spread out the residuals and is "skewed".

Exercise 13 - Determine the percent change in slope and intercept between unweighted and weighted least squares for the above example.

C. UNCERTAINTIES OF PARAMETERS IN WEIGHTED LEAST SQUARES

In weighted least squares, the value of "weighted sigma-squared" is given by

$$\text{sigma2} := \sum_{i=0}^{100} \frac{w_i \cdot \left[(Y_i) - \text{weighted_matrix}_1 \cdot T_i - \text{weighted_matrix}_0 \right]^2}{101 - 2}$$

$$\text{sigma2} = 0.099$$

This is a "weighted sigma squared" and cannot directly be compared to sigma squares for the unweighted case, unless the weights are normalized.

The standard deviation of slope and intercept are given by

$$\text{stdev_slope} := \left[\left(X^T \cdot V^{-1} \cdot X \right)^{-1} \right]_{1,1} \cdot \text{sigma2} \quad \text{stdev_slope} = 1.481 \cdot 10^{-3}$$

$$\text{stdev_intercept} := \left[\left(X^T \cdot V^{-1} \cdot X \right)^{-1} \right]_{0,0} \cdot \text{sigma2} \quad \text{stdev_intercept} = 4.742 \cdot 10^{-4}$$

The final results of weighted least squares fitting is:

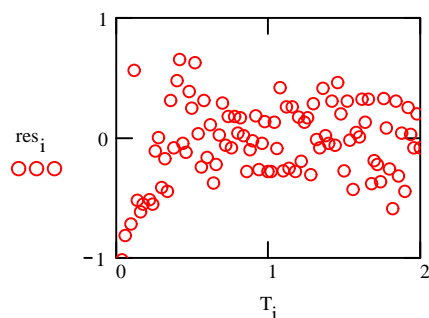
$$\text{weighted_matrix}_0 = 0.04 \quad \text{stdev_intercept} = 4.742 \cdot 10^{-4}$$

$$\text{weighted_matrix}_1 = -1.002 \quad \text{stdev_slope} = 1.481 \cdot 10^{-3}$$

D. RESIDUALS IN WEIGHTED LEAST SQUARES

The residuals are given by the formula

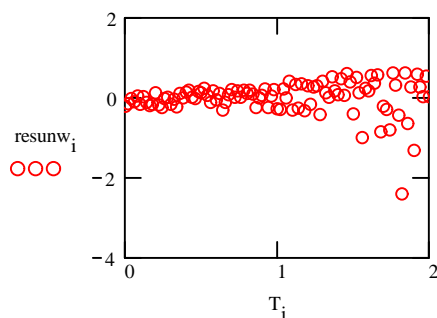
$$\text{res}_i := w_i \cdot [(Y_i) - \text{weighted_matrix}_1 \cdot T_i - \text{weighted_matrix}_0]$$



Except for a few questionable points, note that in general, the residual plot is well-behaved and the residuals remain relatively constant throughout the region.

Exercise 14 - The graph below is the residual graph for unweighted least squares. Compare with the graph for weighted residuals.

$$\text{resunw}_i := [(Y_i) - \text{unweighted_matrix}_1 \cdot T_i - \text{unweighted_matrix}_0]$$



Exercise 15 - What conclusions can be deduced from this graph. Do you think that weighted or unweighted least squares performs better in spreading out the residuals among the data more evenly?

For further information on weighted least squares, see "**Least Squares Fitting of Nonlinear Data in the Undergraduate Laboratory**", by T. Zielinski and R. Allendoerfer, J. Chem. Ed., 1997, v74, n8, 1001.

VII. WHEN AND HOW TO ADD A QUADRATIC TERM

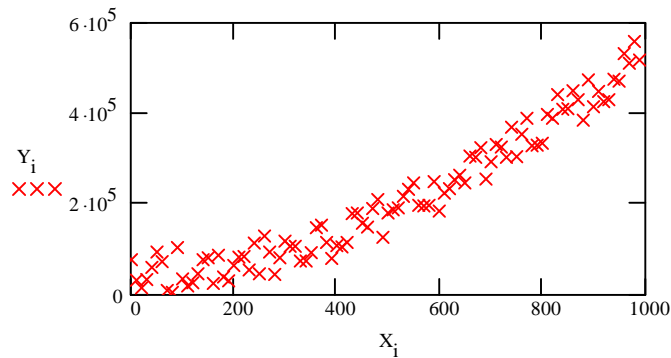
A. USING A RESIDUAL PLOT TO DETERMINE WHEN A QUADRATIC TERM IS NEEDED

Now consider a NEW set of data, with a definite nonlinear response.

```
X := READPRN("linear_x.prn")    x and y can be read from two different files, if
Y := READPRN("linear_y.prn")    necessary. Exercise caution here with the
                                  READPRN function as it is Mathcad version
                                  sensitive.
```

```
N := length(X)
```

```
i := 0..(N - 1)
```



Note the curvature of the plot.

Try linear curve fit

```
XXi,0 := 1    XXi,1 := Xi
```

```
parms := (XXT·XX)-1·XXT·Y    parms = [ -3.351·104 ]
                                         [ 497.081 ]
```

```
R2 := (corr(X, Y))2    R2 = 0.916
```

Built-in correlation function

$$\text{MSE} := \frac{\sum_{k=0}^{\text{length}(X) - 1} (Y_k - \text{parms}_1 \cdot X_k - \text{parms}_0)^2}{N - 2} \quad \text{MSE} = 1.914 \cdot 10^9$$

$$\text{sm} := \sqrt{\text{MSE} \cdot \left[(\text{XX}^T \cdot \text{XX})^{-1} \right]_{1,1}} \quad \text{sm} = 15.158 \quad \text{parms}_1 = 497.081$$

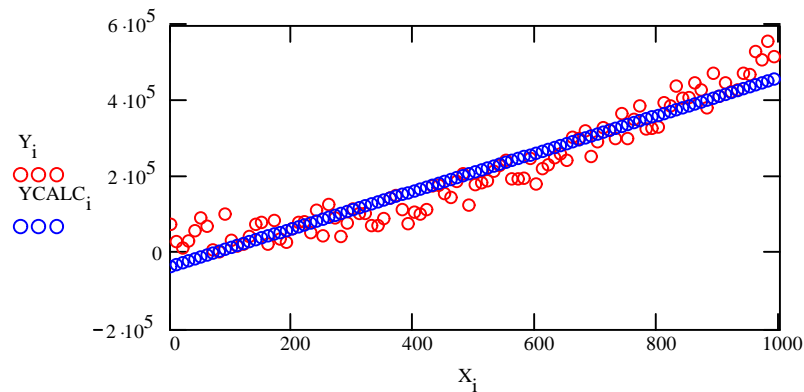
$$\text{sb} := \sqrt{\text{MSE} \cdot \left[(\text{XX}^T \cdot \text{XX})^{-1} \right]_{0,0}} \quad \text{sb} = 8.686 \cdot 10^3 \quad \text{parms}_0 = -3.351 \cdot 10^4$$

slope is $\text{parms}_1 = 497.081$ uncertainty in slope is $\text{sm} = 15.158$

intercept is $\text{parms}_0 = -3.351 \cdot 10^4$ uncertainty in intercept is $\text{sb} = 8.686 \cdot 10^3$

An xy plot gives the following:

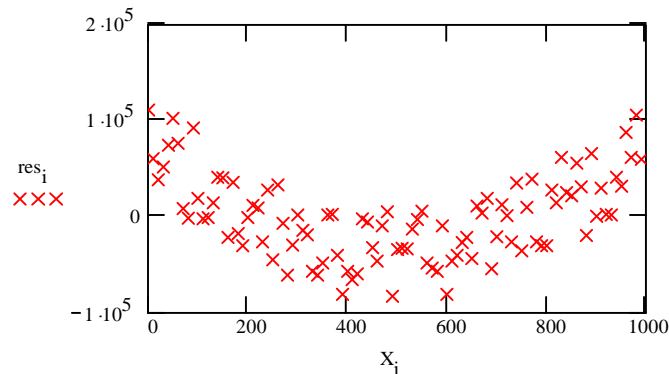
$\text{YCALC} := \text{parms}_0 + \text{parms}_1 \cdot X$ YCALC is linear prediction for Y



Exercise 16 - Would a linear fit be appropriate for this data? Why or why not?

The residuals reveal the true nature of the problem.

$$\text{res}_i := Y_i - \text{YCALC}_i$$



Residuals obtained from a linear fit to the data are seen here in this plot. This is not a random distribution about zero. This residual plot suggests the addition of a quadratic term, because the residuals are parabolic as if following $y = x^2 + c$.

B. HOW TO ADD A QUADRATIC TERM TO THE FIT.

Remember that the linear matrix for regression was given by $\text{XX}_{i,0} := 1$ and $\text{XX}_{i,1} := X_i$

We may add a quadratic term to this matrix, which automatically converts linear least squares to quadratic least squares. This is simply done by adding a third column to the matrix, which contains the square of the values of x .

$$\text{XX}_{i,2} := (X_i)^2$$

Exercise 17 - If you have five data points for x : (1,2,3,4,5), then what would the matrix XX be if a quadratic term were added?

$$\text{parmsq} := (\text{XX}^T \cdot \text{XX})^{-1} \cdot \text{XX}^T \cdot Y \quad \text{parmsq} = \begin{bmatrix} 3.753 \cdot 10^4 \\ 62.153 \\ 0.439 \end{bmatrix}$$

The three elements of parms include a constant term, a linear term, and a quadratic term. If we wished to add a quartic term, we would define $XX_{i,3}$ to be $(X_i)^3$.

$$\text{MSE} := \frac{\sum_{k=0}^{\text{length}(X)-1} \left[Y_k - \text{parmsq}_1 \cdot X_k - \text{parmsq}_0 - \text{parmsq}_2 \cdot (X_k)^2 \right]^2}{N - 3}$$

$$\text{MSE} = 8.294 \cdot 10^8$$

Without the quadratic term, MSE would be

$$\text{MSEold} := \frac{\sum_{k=0}^{\text{length}(X)-1} \left(Y_k - \text{parms}_1 \cdot X_k - \text{parms}_0 \right)^2}{N - 2}$$

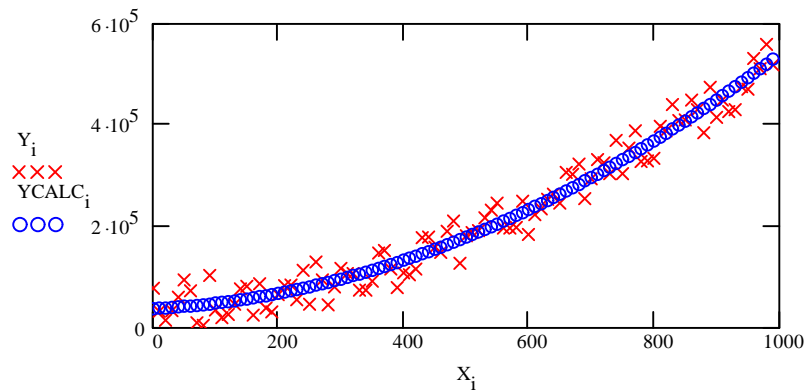
$$\text{MSEold} = 1.914 \cdot 10^9$$

Exercise 18 - Why is it beneficial to have a reduction in MSE?

We can now obtain uncertainties in the three terms:

	Uncertainties	Parameters
$\text{sm} := \sqrt{\text{MSE} \cdot \left[(XX^T \cdot XX)^{-1} \right]_{1,1}}$	sm = 15.158	parmsq ₁ = 62.153
$\text{sb} := \sqrt{\text{MSE} \cdot \left[(XX^T \cdot XX)^{-1} \right]_{0,0}}$	sb = 8.686 · 10 ³	parmsq ₀ = 3.753 · 10 ⁴
$\text{squad} := \sqrt{\text{MSE} \cdot \left[(XX^T \cdot XX)^{-1} \right]_{2,2}}$	squad = 0.039	parmsq ₂ = 0.439

$$Y_{\text{CALC}} := \text{parmsq}_0 + \text{parmsq}_1 \cdot X + \text{parmsq}_2 \cdot X^2$$



Note how smoothly YCALC passes through the data, once the quadratic term is added.

C. HOW TO DETERMINE IF THE QUADRATIC TERMS IS STATISTICALLY SIGNIFICANT

To test for significance of the quadratic term, calculate the value of the parameter divided by its standard deviation, and compare with Student's t (at the 95% significance level). If this ratio is greater than the value of t, then the parameter is significant.

$$\frac{\text{parmsq}_2}{\text{squad}} = 11.367 \quad \text{qt}(0.975, N - 2) = 1.984$$

Thus the quadratic term is significant, since the value of parameter/uncertainty is greater than Student's t.

What about the other terms?

$$\frac{\text{parmsq}_0}{\text{sb}} = 4.431 \quad \text{Constant term}$$

$$\frac{\text{parmsq}_1}{\text{sm}} = 1.572 \quad \text{Linear Term}$$

Exercise 19 - Explain why the linear term is not statistically different from zero.

Now, refit the data using only the constant and the cubic term, as suggested by the above statistical test.

$$\text{XXX}_{i,0} := 1 \quad \text{XXX}_{i,1} := (X_i)^2$$

Use "nom" as suffix for "no m term" in the following

$$\text{parmsnom} := (\text{XXX}^T \cdot \text{XXX})^{-1} \cdot \text{XXX}^T \cdot Y$$

$$\text{parmsnom} = \begin{bmatrix} 4.899 \cdot 10^4 \\ 0.498 \end{bmatrix}$$

$$\text{MSEnom} := \frac{\sum_{k=0}^{\text{length}(X) - 1} [Y_k - \text{parmsnom}_0 - \text{parmsnom}_1 \cdot (X_k)^2]^2}{N - 2}$$

$$\text{MSEnom} = 8.418 \cdot 10^8$$

$$\text{MSE} = 8.294 \cdot 10^8$$

MSE does not change much, but MSEnom has only two parameters, whereas MSE has three.

$$sbnom := \sqrt{\text{MSEnom} \cdot \left[(\text{XXX}^T \cdot \text{XXX})^{-1} \right]_{0,0}}$$

$$sbnom = 4.339 \cdot 10^3$$

$$sb = 8.47 \cdot 10^3$$

Note drop of standard deviation of b

$$squadnom := \sqrt{\text{MSEnom} \cdot \left[(\text{XXX}^T \cdot \text{XXX})^{-1} \right]_{1,1}}$$

$$squadnom = 9.824 \cdot 10^{-3}$$

$$squad = 0.039$$

Same with quadratic term.

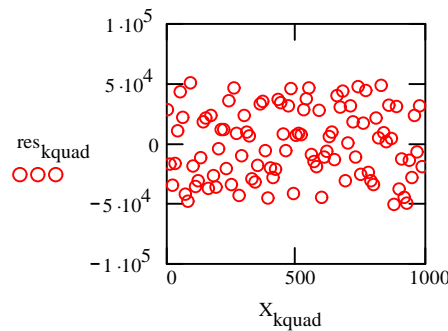
Exercise 20 - Compare the standard deviations of slope and intercept of the original data set with the new data set without the "slope" term. Which of the two fits is preferred?

In both cases, removal of the "slope" term leads to a reduced standard deviation of the remaining parameters. Thus one can conclude that it is advantageous to use the Student t test on each of the parameters, and remove the parameters which are not statistically different from zero.

Finally, look at the residuals for the quadratic fit described above.

$$kquad := 0..N - 1$$

$$res_{kquad} := Y_{kquad} - parmsnom_0 - parmsnom_1 \cdot (X_{kquad})^2$$



Quadratic fit

Exercise 21: Compare this residual plot with the residual plot of the linear fit.

VIII. SUMMARY

This template has demonstrated many statistical concepts, all of which are useful when fitting to a straight line or a line with an added quadratic term. The following problem should be performed on a new spreadsheet, demonstrating many of the principles described above.

Exercise 22: Use the data in "exercise22x.prn" and "exercise22y.prn", and determine the following statistical parameters"

slope
intercept
correlation coefficient
standard deviation of fit
standard deviation of slope
standard deviation of intercept
plots: y vs x
residual plot

Then add a quadratic term, and tell whether this improves the fit and if the quadratic term is statistically different from zero.