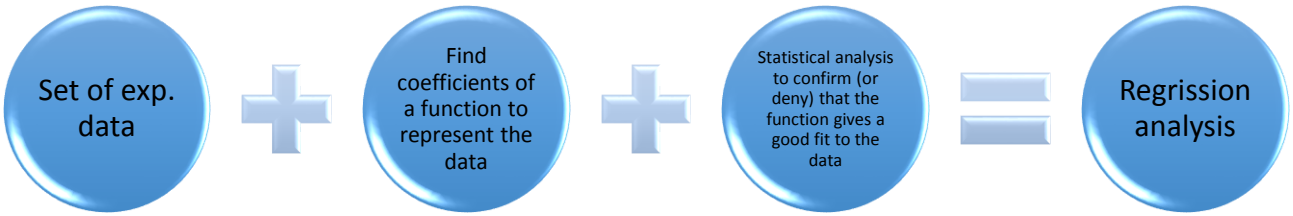
Regression Analysis and Parameter Estimation

What is regression analysis?

The problem of regression analysis is to find coefficients (parameters) of a function that are believed to properly represent a set of experimental data and to perform statistical analyses to confirm (or deny) that the function gives a good fit to the data.



Without the additional statistical analysis, just finding the parameters of some candidate function is called curve fitting. Curve fitting, while useful in certain circumstances, is not as powerful as regression analysis.

Consider data as represented in figure 1:

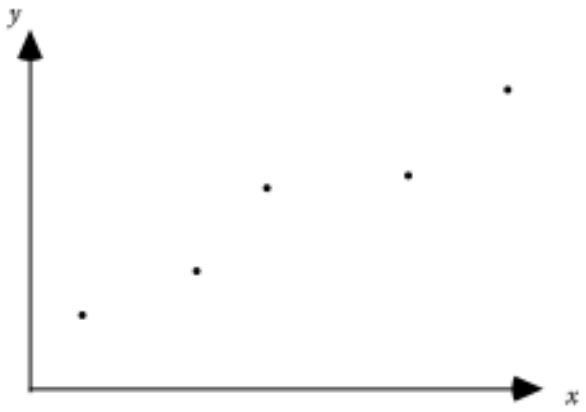


Fig. 1, data with experimental error in the dependent variable

It is assumed that x, the independent variable, is error free (this might be time or temperature, for example), and y, the dependent variable, contains experimental error.

In chemical and biomolecular engineering applications, theoretical knowledge often exists of the function that should "fit" the data.

If the deviations (errors) between the data and the fitting function are statistically distributed with a normal distribution with zero mean and constant variance, then it can be shown that a proper way to find the unknown coefficients of the function is to minimize the sum of squares of the errors (The principle of the least square method)

It is common nomenclature to call the **errors** "residuals," which are defined as follows:

$$r_i = y_i^{calc} - y_i^{data}; \qquad i = 1,2,3 \dots \dots, n_d$$

Where:

r_i is the residual or error	n_d is the number of data point
y_i^{calc} is the value of y calculated from the fitting	y_i^{data} is the associated data value

Probability and Statistics Review

We give a quick introduction to the basic elements of probability and statistics which we need for the Method of Least Squares

Given a sequence of data $x_1, \dots, x_N$ we define the mean (or the expected value) to be:

$$Mean = (x_1 + x_2 + .. + x_N)/N$$

$$Mean = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

The mean is the average value of the data.

Consider the following two sequences of data: {10, 20, 30, 40, 50} and {30, 30, 30, 30, 30}.

Both sets have the same mean; however, the first data set has greater variation about the mean.

This leads to the concept of variance (regression coefficient), which is a useful tool to quantify how much a set of data fluctuates about its mean.

Variance (Regression coefficient) σ_x^2:

The variance of $\{x_1, \dots, x_N\}$, denoted by σ_x^2 is:

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Standard deviation σ_x:

The σ_x standard deviation is the square root of the variance:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Note that if the x's have units of meters then the variance σ_x^2has units of m^2, and the standard deviation σ_x and the mean x have units of meters.

What is standard deviation physical meaning?

It gives a good measure of the deviations of the x's around their mean.

Standard error: s_e

$$s_e = \frac{\sigma_x}{\sqrt{N}}$$

What is the difference between SD and SE?

While the actual calculations for Standard Deviation and Standard Error look very similar, they represent two very different, but complementary, measures. SD tells us about the shape of our distribution, how close the individual data values are from the mean value. SE tells us how close our sample mean is to the true mean of the overall population. Together, they help to provide a more complete picture than the mean alone can tell us.

- The SD (standard deviation) quantifies scatter — how much the values vary from one another.
- The SEM (standard error of the mean) quantifies how precisely you know the true mean of the population. It takes into account both the value of the SD and the sample size.
- Both SD and SEM are in the same units -- the units of the data.
- The SEM, by definition, is always smaller than the SD.

- The SEM gets smaller as your samples get larger. This makes sense, because the mean of a large sample is likely to be closer to the true population mean than is the mean of a small sample. With a huge sample, you'll know the value of the mean with a lot of precision even if the data are very scattered.
- The SD does not change predictably as you acquire more data. The SD you compute from a sample is the best possible estimate of the SD of the overall population. As you collect more data, you'll assess the SD of the population with more precision. But you can't predict whether the SD from a larger sample will be bigger or smaller than the SD from a small sample. (This is a simplification, not quite true. See comments below.)

Note that standard errors can be computed for almost any parameter you compute from data, not just the mean. The phrase "the standard error" is a bit ambiguous. The points above refer only to the standard error of the mean.

The Method of Least Squares

Straight line regression

Given data $\{(x_1, y_1), \dots\dots, (x_N, y_N)\}$ and it is required to fit the data to a straight line equation $y = ax + b$

The method of least squares depends on finding the coefficient a and b which give the minimum value of the summation of squares of errors.

How do we know this is the right line?

What makes it best?

The Method of Least Squares is a procedure, requiring just some calculus and linear algebra, to determine what the “best fit” line is to the data.

We may define the error (residual) by the equation:

$$Error = y_{data} - y_{calculated\ from\ the\ equation}$$

The summation of the squares of the error:

$$E = \sum_{i=1}^{N} \left(y_i - (ax_i + b)\right)^2$$

The above equation shows that the error is a function of two variables a and b.

In order to minimize the error we have to find the values of a and b corresponding to the minimum value of E by taking the derivatives such that $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{N} 2\left(y_i - (ax_i + b)\right) \cdot (-x_i)$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{N} 2\left(y_i - (ax_i + b)\right) \cdot (-1)$$

Setting $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = 0$ and dividing by 2 yields:

$$\sum_{i=1}^{N} \left(y_i - (ax_i + b)\right) \cdot (-x_i) = 0$$

$$\sum_{i=1}^{N} \left(y_i - (ax_i + b)\right) \cdot (-1) = 0$$

These two equations may rearranged as:

$$\left(\sum_{i=1}^{N} x_i^2\right) a + \left(\sum_{i=1}^{N} x_i\right) b = \sum_{i=1}^{N} x_i y_i$$

$$\left(\sum_{i=1}^{N} x_i\right) a + \left(\sum_{i=1}^{N} 1\right) b = \sum_{i=1}^{N} y_i$$

Solving the two equations by any technique (e.g. matrix operation) we can get a and b.

$$\begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix}$$

How good is the fit from a statistical perspective?

There are several tools to help visualize if the chosen function is a good one. The tools will be discussed in the following section:

1. **Residual plots**

   When the residuals $(y_{calc} - y_{data})$versus x are plotted, these values should distribute themselves somewhat evenly about zero, and their magnitude should be approximately constant (these were the basic assumptions for the least-squares method).

2. **Coefficient of determination R^2**

   A quantitative measure of the "goodness of fit" is provided by the coefficient of determination.

   If $\bar{y}$ is the mean of the observed data:

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Then the variability of the data set can be measured using three sums of squares formulas:

$$SS = \sum_{i=1}^{N} (y_i - \bar{y})^2$$

The residual r or the error is defined as:

$$r_i = \sum_{i=1}^{N} (y_{calc} - y_{data})$$

The coefficient of determination is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} r_i^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}$$

An R^2 value near 1 indicates that all residuals are small and that the fit is "good."

3. Parameter t-ratio

When choosing an arbitrary function to fit a set of data (such as the quadratic function, it might be asked if all three terms in the equation are needed or if it can be simplified by dropping perhaps the linear term (the one involving bx). To answer this question, statistics gives the answer. It can be shown that the ratio of the optimal parameter values divided by their standard deviations.

$$t - ratio = \frac{parameter}{its\ standard\ error}$$

A parameter's contribution is insignificant (i.e., its value is zero) if the calculated value $|t_i| \leq 2$.

Curvelinear fitting

Generalize the method of least squares to find the best fit quadratic to $y = ax^2 + ax + b$

We will do the same procedure as in the above case.

Given data $\{(x_1, y_1), \dots\dots, (x_N, y_N)\}$ and it is required to fit the data a quadratic equation $y = ax^2 + bx + c$

We may define the error (residual) by the equation:

$$Error = y_{data} - y_{calculated\ from\ the\ equation}$$

The summation of the squares of the error:

$$E = \sum_{i=1}^{N} \left(y_i - (ax_i^2 + bx_i + c)\right)^2$$

The above equation shows that the error is a function of two variables a and b.

In order to minimize the error we have to find the values of a and b corresponding to the minimum value of E by taking the derivatives such that $\frac{\partial E}{\partial a} = 0$ and $\frac{\partial E}{\partial b} = 0$

$$\frac{\partial E}{\partial a} = \sum_{i=1}^{N} 2\left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-x_i^2)$$

$$\frac{\partial E}{\partial b} = \sum_{i=1}^{N} 2\left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-x_i)$$

$$\frac{\partial E}{\partial c} = \sum_{i=1}^{N} 2\left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-1)$$

Setting $\frac{\partial E}{\partial a} = \frac{\partial E}{\partial b} = \frac{\partial E}{\partial c} = 0$ and dividing by 2 yields:

$$\sum_{i=1}^{N} \left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-x_i^2) = 0$$

$$\sum_{i=1}^{N} \left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-x_i) = 0$$

$$\sum_{i=1}^{N} \left(y_i - (ax_i^2 + bx_i + c)\right) \cdot (-1) = 0$$

These three equations may rearranged as:

$$\left(\sum_{i=1}^{N} x_i^4\right) a + \left(\sum_{i=1}^{N} x_i^3\right) b + \left(\sum_{i=1}^{N} x_i^2\right) c = \sum_{i=1}^{N} x_i^2 y_i$$

$$\left(\sum_{i=1}^{N} x_i^3\right) a + \left(\sum_{i=1}^{N} x_i^2\right) b + \left(\sum_{i=1}^{N} x_i\right) c = \sum_{i=1}^{N} x_i y_i$$

$$\left(\sum_{i=1}^{N} x_i^2\right) a + \left(\sum_{i=1}^{N} x_i\right) b + \left(\sum_{i=1}^{N} 1\right) c = \sum_{i=1}^{N} y_i$$

Solving the three equations by any technique (e.g. matrix operation) we can get a , b and c.

$$\begin{bmatrix} \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 \\ \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} x_i^2 y_i \\ \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{bmatrix}$$

Regression using Excel's (regression add-In)

Excel will do almost everything that has been discussed as long as the problem is one of linear regression. To invoke this package, go to Data/Data Analysis/Regression.

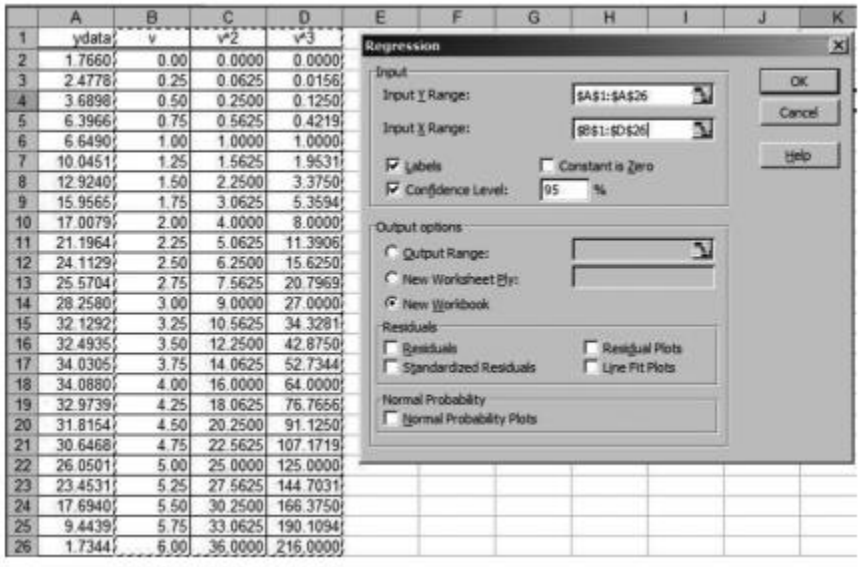
We will it for the following data:

v	*y* Data
0.00	1.766
0.25	2.478
0.50	3.690
0.75	6.397
1.00	6.649
1.25	10.045
1.50	12.924
1.75	15.957
2.00	17.008
2.25	21.196
2.50	24.113
2.75	25.570
3.00	28.258
3.25	32.129
3.50	32.494
3.75	34.031
4.00	34.088
4.25	32.974
4.50	31.815
4.75	30.647
5.00	26.050
5.25	23.453
5.50	17.694
5.75	9.444
6.00	1.734

Assume the data required to fit a third order quadratic equation:

In this case we have to make a spread sheet containing y_{data}, v, v^2, v^3.

The following spreadsheet displays the original data and columns for other terms to be included in the fitting function. Also shown is the window that is presented by the Regression Add-In.



	A	B	C	D
1	ydata	v	v^2	v^3
2	1.7660	0.00	0.0000	0.0000
3	2.4778	0.25	0.0625	0.0156
4	3.6898	0.50	0.2500	0.1250
5	6.3966	0.75	0.5625	0.4219
6	6.6490	1.00	1.0000	1.0000
7	10.0451	1.25	1.5625	1.9531
8	12.9240	1.50	2.2500	3.3750
9	15.9565	1.75	3.0625	5.3594
10	17.0079	2.00	4.0000	8.0000
11	21.1964	2.25	5.0625	11.3906
12	24.1129	2.50	6.2500	15.6250
13	25.5704	2.75	7.5625	20.7969
14	28.2580	3.00	9.0000	27.0000
15	32.1292	3.25	10.5625	34.3281
16	32.4935	3.50	12.2500	42.8750
17	34.0305	3.75	14.0625	52.7344
18	34.0880	4.00	16.0000	64.0000
19	32.9739	4.25	18.0625	76.7656
20	31.8154	4.50	20.2500	91.1250
21	30.6468	4.75	22.5625	107.1719
22	26.0501	5.00	25.0000	125.0000
23	23.4531	5.25	27.5625	144.7031
24	17.6940	5.50	30.2500	166.3750
25	9.4439	5.75	33.0625	190.1094
26	1.7344	6.00	36.0000	216.0000

For the Input Y Range, the first column was selected, and for the Input X Range, the next three columns were identified. Since the first row contains labels, the "Labels" box was checked. Also checked was the Confidence Level box and 95% for the confidence level.

For the Output options, the New Workbook radio button was checked, and the following is the output from the Regression Add-In:

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.998277703					
5	R Square	0.996558373					
6	Adjusted R Square	0.996066712					
7	Standard Error	0.72049303					
8	Observations	25					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	3156.587363	1052.196	2026.922	5.16918E-26	
13	Residual	21	10.90131434	0.51911			
14	Total	24	3167.488678				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	2.224111248	0.499705233	4.450846	0.000221	1.184917331	3.2633052
18	v	0.182815879	0.736261921	0.248303	0.806312	-1.348324599	1.7139564
19	v^2	5.874718908	0.28858153	20.35722	2.63E-15	5.274580765	6.4748571
20	v^3	-0.985488667	0.031585686	-31.2005	4.45E-19	-1.051174697	-0.919803

Note that the t Stat (same as t-ratio) for v is less than 2, so the associated term (involving v) was deleted and the analysis was repeated. To do this, the second data column was not included in the X-Range in the Regression Add-In window. Results when doing that are as follows:

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.998272642					
5	R Square	0.996548269					
6	Adjusted R Square	0.996234475					
7	Standard Error	0.704960337					
8	Observations	25					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*ignificance F*	
12	Regression	2	3156.555358	1578.278	3175.807	8.29E-28	
13	Residual	22	10.9333197	0.496969			
14	Total	24	3167.488678				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	2.32685819	0.274098785	8.489123	2.17E-08	1.758412	2.895304
18	v^2	5.943797784	0.07503901	79.20944	1.6E-28	5.788176	6.099419
19	v^3	-0.992608943	0.012956692	-76.6098	3.32E-28	-1.01948	-0.96574

Important note:

The Regression Add-In works only for linear regression, where the Zs in Equation 7.9 do not involve the unknown parameters. For nonlinear regression, the calculations must be done "by hand."

EXCEL REGRESSION ANALYSIS OUTPUT PART ONE: REGRESSION STATISTICS

These are the "Goodness of Fit" measures. They tell you how well the calculated linear regression equation fits your data.

1. **Multiple R.** This is the correlation coefficient. It tells you how strong the linear relationship is. For example, a value of 1 means a perfect positive relationship and a value of zero means no relationship at all. It is the square root of r squared (see #2).
2. **R squared**. This is R^2, the Coefficient of Determination. It tells you how many points fall on the regression line. For example, 80% means that 80% of the variation of y-values around the mean are explained by the x-values. In other words, 80% of the values fit the model.
3. **Adjusted R square.** The adjusted R-square adjusts for the number of terms in a model. You'll want to use this instead of #2 if you have more than one x variable.
4. **Standard Error of the regression:** An estimate of the standard deviation of the error μ. This is *not* the same as the standard error in descriptive statistics. The standard error of the regression is the precision that the regression coefficient is measured; if the coefficient is large compared to the standard error, then the coefficient is probably different from 0.
5. **Observations**. Number of observations in the sample.

Some notes about the results shown in the above table:

What is the difference between adjusted R^2 , multiple R and R^2 in regression analysis?

Simply put, R is the correlation between the predicted values and the observed values of Y.

R square is the square of this coefficient and indicates the percentage of variation explained by your regression line out of the total variation.

This value tends to increase as you include additional predictors in the model. Thus, one can artificially get higher R square by increasing the number of Xs in the model. To penalize this effect, adjusted R square is used.

When you compare models with their complexity, you should then rely on Adj R square. Predicted R square is another measure which addresses the issue of overfitting the data and explain the prediction power for future observations.