Appendix E. Parameter Estimation[1]

Engineers often need to express experimental data in terms of an equation. They must decide on the equation and then determine the parameters that provide the best fit to the data. The problem is simplest if the equation is linear. This appendix describes simple methods using Excel and MATLAB® for fitting a straight line to data, then for fitting a polynomial to data (polynomial regression), and finally for fitting any set of functions in which the unknown parameters appear linearly (multiple regression). The data used in the examples is only illustrative, but it has scatter included, as you would find in data taken in the laboratory. After mastering the examples in this chapter, you will be ready to fit an equation of your choice to data you measure.

Mathematical formulation

Consider a set of data

$$\{y(x_i)\}, \text{ for } i = 1 \text{ to } n \tag{E.1}$$

and find an equation that models the data. Write the equation in a general form:

$$y(x;a_1,a_2,...,a_M), \tag{E.2}$$

showing that it depends on x, but also on some unknown parameters, $\{a_1,a_2,..a_M\}$. The goal is to find the set of parameters that gives the 'best fit.' The best fit is usually defined by minimizing the sum of the square residuals, where the residual is the difference between the predicted value and the data. Because the data may have errors in it, an exact fit won't be possible in most cases. Thus, you minimize the variance of the residuals (Press, *et al.* 1986, pp. 502-3).

$$\sigma^2 = \sum_{i=1}^{N} \frac{[y_i - y(x_i)]^2}{N}, \quad y \equiv y(x_i,a_1,a_2,...,a_M) \tag{E.3}$$

If the parameters enter the equation linearly, then the minimization problem reduces to a set of linear equations which are solved easily by Excel and MATLAB. The effectiveness of the curve fit is often reported as values of the linear correlation coefficient squared, r^2. The linear correlation coefficient is defined as (Press, *et al.* 1986, p. 484)

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2 \ \sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{E.4}$$

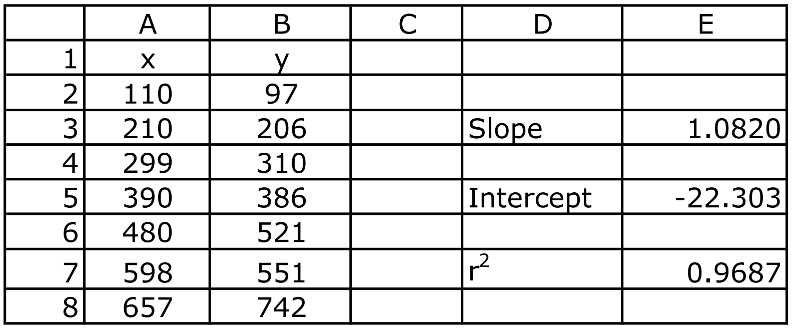
[1] This material was in Appendix E of the 1st edition of *Introduction to Chemical Engineering Computing*. It is copyrighted by Wiley (2006).

Values of r near 1 indicate a positive correlation; r near –1 means a negative correlation and r near zero means no correlation.

Straight line

This section describes how to fit a straight line to tabular data using Excel and MATLAB. The data in Table E-1 represents seven measurements, in columns A and B. The goal is to find an equation, $y = a + bx$, that best represents this data.

Straight line curve fit using Excel. Step One. First insert the data into Excel as shown.



	A	B	C	D	E
1	x	y			
2	110	97			
3	210	206		Slope	1.0820
4	299	310			
5	390	386		Intercept	-22.303
6	480	521			
7	598	551		r^2	0.9687
8	657	742			

Table E-1. Simulated data for two measurements of the same thing (x and y)

Step Two. After inserting the data, select the slope cell, E3, and insert the command:

=SLOPE(B2:B8,A2:A8) (E.5)

The result is 1.082. Notice that the y-locations are entered first, followed by the x-locations.

Step Three. To get the intercept, you do the same thing for the intercept cell, E5:

=INTERCEPT(B2:B8,A2:A8) (E.6)

Step Four. Finally, the r^2 value is found using the following command in cell E7.

=RSQ(B2:B8,A2:A8) (E.7)

The curve fit is then

$$y = -22.303 + 1.082x \tag{E.8}$$

Because the r^2 value is 0.9687, close to 1.0, the curve fit is good. You should plot the data, and this is easily done using the 'trendline' option, as described below.

Plotting the trendline. Step One. Select the data in columns A2:B8. Then choose 'Insert/Chart' and choose the scatter plot with no lines. Follow the instructions, adding titles, etc., and place the chart on the spreadsheet.

Step Two. Put the cursor on any data point and right click (on the Macintosh, use CTRL-click). A menu appears; choose 'Add trendline. ' The trendline is added as shown. Put the cursor on the trendline and right click. Choose 'Format Trendline,' then 'Options' and click 'Display equation on chart;' also choose 'Display R-squared on chart.' The result is Fig. E-1; notice that the equation in the figure agrees with Eq. (E.8).

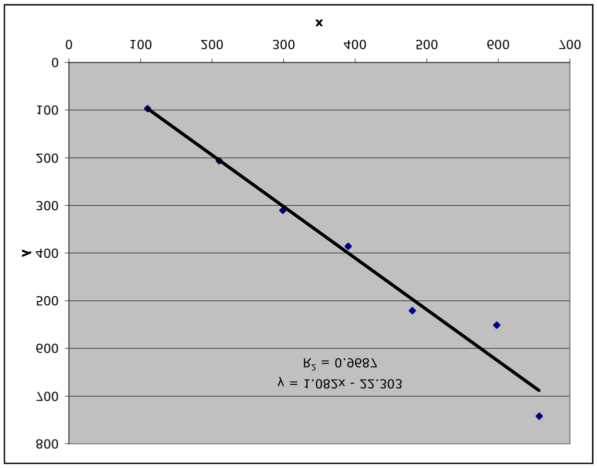


Figure E-1. Linear Curve Fit to data in Table E-1 using Excel

Straight line curve fit using MATLAB. The same problem can be solved using MATLAB, too. **Step One.** Put the *x* values into a vector, *x*, and the *y* values into a vector, *y*. Then issue the command

```
p = polyfit(x,y,1)
```
(E.9)

Step Two. Next evaluate the polynomial at a set of points from x = 100 to 700 for plotting purposes. Because the polynomial is a straight line, only two points are needed.

```
w = [100 700]
v = polyval(p,w)
```
(E.10)

Finally, plot the data and the straight line; use red symbols for the data and a blue line for the straight line.

```
plot(x,y,'ro',w,v,'b-')
```
(E.11)

The result is shown in Fig. E-2. (The figure here appears in black, obtained by changing 'ro' to 'ko' and 'b-' to 'k-'.)

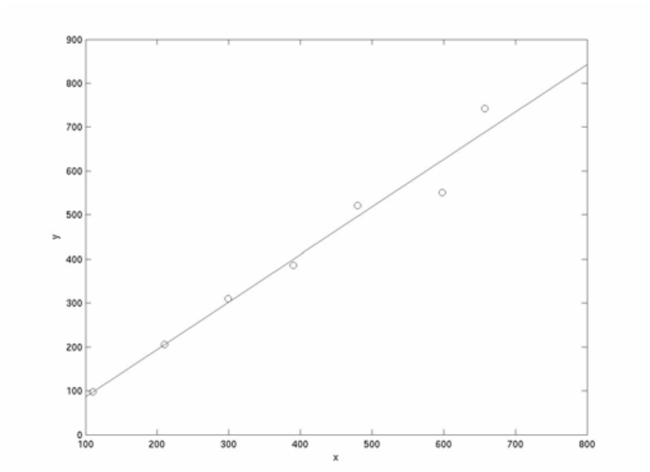


Figure E-2. Linear Curve Fit to data in Table E-1 using MATLAB

Polynomial regression

You needn't be limited to a straight line when trying to fit data. The first extension is to use a polynomial and determine the coefficients that give the best fit. In that case, the formula is

$$y = a + bx + cx^2 + dx^3 + ... \qquad \text{(E.12)}$$

or, more generally,

$$y(x) = \sum_{i=1}^{N} a_i x^{i-1} \qquad \text{(E.13)}$$

You can easily carry out this process using either Excel or MATLAB.

As an example, consider the set of data in Table E-2 giving the measured partial pressure of a chemical at different times in a batch chemical reactor. You wish to fit this data to a linear or quadratic equation.

time (sec)

0.00	1.02	1.82	2.87	3.38	4.96	5.97	6.60	8.11	8.36	9.10
	10.00	10.80	12.40	12.60	14.80	17.50	18.10	18.20	20.20	21.00

partial pressure

7.45	8.98	8.8	10.1	10.9	12.1	12.8	12.8	14.4	13.7	13.9
	14.4	14.6	15.8	15.3	16.9	19.1	18.7	17.9	18.9	18.7

Table E-2. Partial pressure versus time

Polynomial regression using Excel. Steps One and Two. Follow the instructions given above to plot a trendline.

Step Three. Put the cursor on a data point and choose 'Add Trendline'. This time, however, choose the option at the top right labeled 'Polynomial,' and pick an order of polynomial. By choosing 'Polynomial,' then 2, you can fit a quadratic function to the data, while choosing 'Polynomial,' then 3, gives a cubic function. The results are shown in Fig. E-3(a) and E-3(b). The equation is displayed, along with the square of the linear correlation coefficient. The graphs show that the curve fit becomes better as the degree of polynomial is increased, and this is also reflected in the value of r^2. In mathematical notation (*Handbook of Mathematical Functions*, 1964, p. 773), the order of a polynomial is the number of terms (or coefficients) whereas the degree is the highest power. Thus, a 2nd-order polynomial includes two terms – the constant and linear term and is a 1st degree polynomial. A 3rd-order polynomial includes three terms, the constant and the coefficients of x and x^2 and is a 2nd degree polynomial. This nomenclature is not followed in Excel.

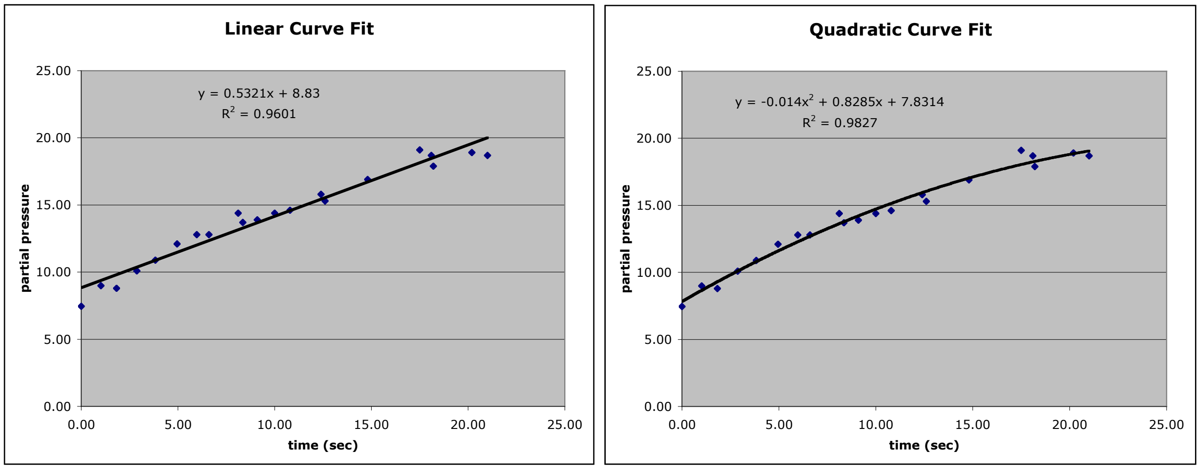


Figure E-3. Linear (a) and quadratic (b) curve fit to the data in Table E-2 using Excel

Polynomial regression using MATLAB. Next consider MATLAB with a polynomial of higher order than a straight line. Put the data into the vectors t and p. The only change is in the 'polyfit' command - just type what order, n, you want.

```
pn = polyfit(t,p,n)
```

(E.14)

This time, though, the curve fit is a curve. Thus, more points are needed to evaluate the curve. You also need to eliminate the previous solution (or at least the size of the vectors and matrices), because they may not be the same in this example. The complete set of commands to generate Fig. E-4(a) is:

```
clear w v          % this eliminates the previous curve fit
```

```
p1 = polyfit(t,p,1) % for a 1-st order polynomial
w = [0:1:22]
v = polyval(p1,w)
plot(t,p,'ro',w,v,'b-')
```

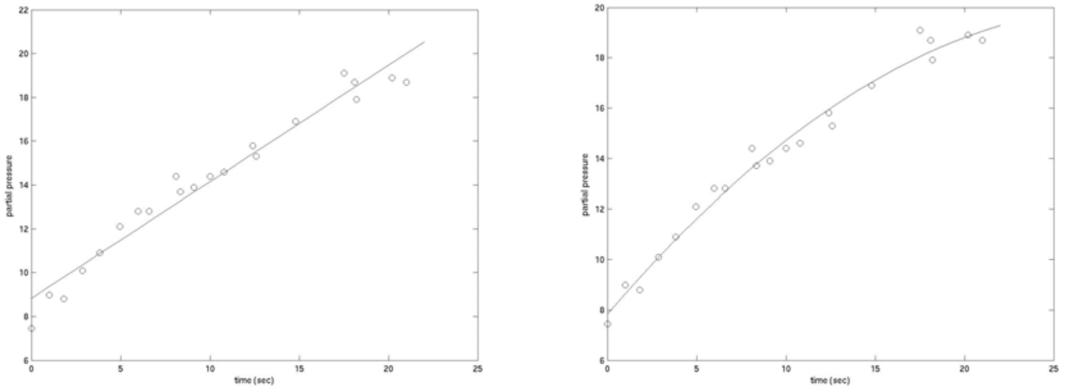


Figure E-4. Linear (a) and quadratic (b) curve fit to the data in Table E-2 using MATLAB

To obtain Fig. E-4(b), use the same commands except with

```
P2 = polyfit(t,p,2)  % for a 2nd order polynomial                (E.15)
```

Multiple regression using Excel

The last example shows how to fit a polynomial to data. The same thing can be done when the functions are not simple powers, but are more complicated functions. However, to keep the problem linear, the unknown coefficients must be coefficients of those functions, i.e., the functions are completely specified. Multiple regression simply determines how much of each one is needed. Thus, the form of the equation is

$$y(x) = \sum_{i=1}^{M} a_i f_i(x) \tag{E.16}$$

The goal is to find the best M values of $\{a_i\}$, given the M functions $f_i(x)$ and data $y_i = y(x_i)$, $i = 1,...,N$.

In Excel, you put the x values in a column and create additional columns, with each column being a function, evaluated for the x value in that row. The example used here is to find the constants in a reaction rate formula. The expected expression is

$$rate = kp_A^n p_B^m \tag{E.17}$$

and the goal is to find the values of *k*, *n*, and *m* that give the best fit of the *rate* for various partial pressures of substances A and B. This form is not linear, which is a requirement of multiple regression, but a transformation can make it linear. Take the logarithm of both sides of the equation.

$$\ln(rate) = \ln k + \quad n \ln p_A \quad + m \ln p_B \tag{E.18}$$

This equation has the following form,

$$y = a + bx_1 + cx_2 \tag{E.19}$$

where the dependence upon two or more variables is clear. The data is entered into the spreadsheet, and the various terms are transformed as shown in Table E-3. Columns A and B are the partial pressures of the two chemicals for which the rate is measured, as indicated in column C.

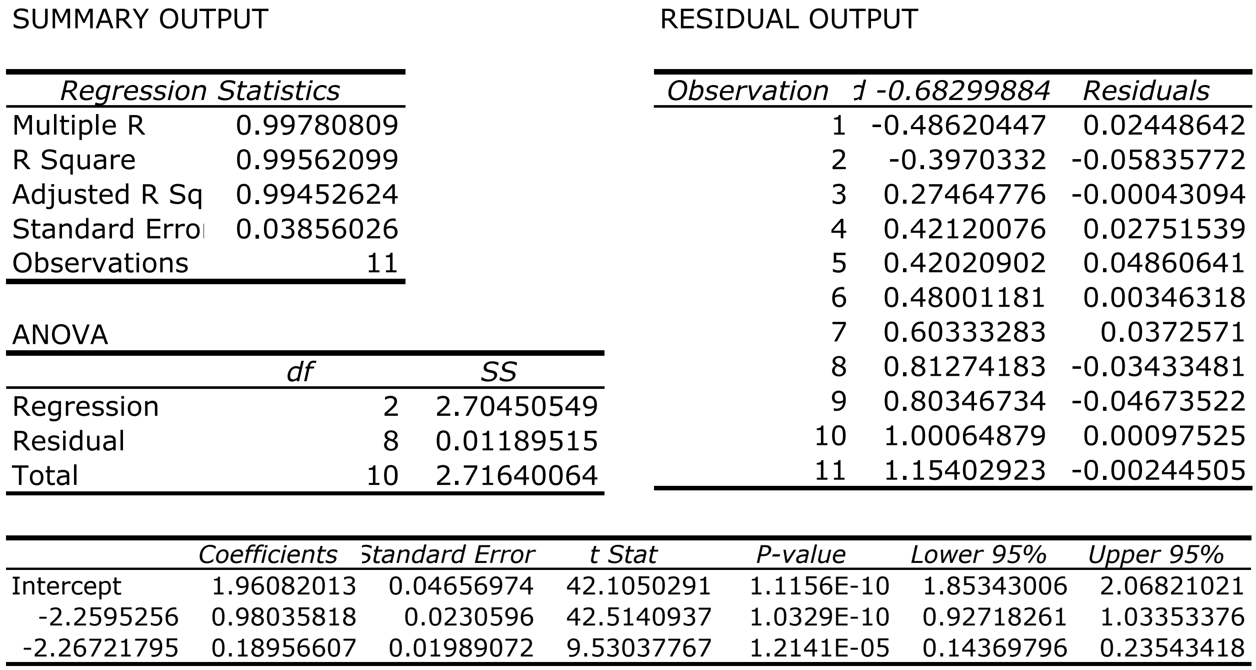
	A	B	C	D	E	F
1	pa	pb	rate	ln(pa)	ln(pb)	ln(rate)
2	0.1044	0.1036	0.5051	-2.2595	-2.2672	-0.6830
3	0.1049	0.2871	0.6302	-2.2547	-1.2479	-0.4617
4	0.1030	0.5051	0.6342	-2.2730	-0.6830	-0.4554
5	0.2582	0.1507	1.3155	-1.3540	-1.8925	0.2742
6	0.2608	0.3100	1.5663	-1.3440	-1.1712	0.4487
7	0.2407	0.4669	1.5981	-1.4242	-0.7616	0.4688
8	0.3501	0.0922	1.6217	-1.0495	-2.3838	0.4835
9	0.3437	0.1944	1.8976	-1.0680	-1.6378	0.6406
10	0.3494	0.5389	2.1780	-1.0515	-0.6182	0.7784
11	0.4778	0.1017	2.1313	-0.7386	-2.2857	0.7567
12	0.4880	0.2580	2.7227	-0.7174	-1.3548	1.0016
13	0.5014	0.5037	3.1632	-0.6904	-0.6858	1.1516

Table E-3. Reaction rate data as a function of partial pressures

Step One. Obtain columns D, E, and F by taking the logarithm of columns A, B, and C, respectively. Do this in the first cell (D2), copy it across the E and F rows, and copy down the three rows (D2, E2, and F2).

Step Two. Proceed with parameter estimation by choosing 'Tools/Data Analysis,' and then choosing 'Regression.' If 'Data Analysis' does not appear in your menu under 'Tools,' you may have to install it from the original Excel CD. Enter F2:F13 for the *y* values and D2:E13 for the *x* values. This tells the computer that you want the best line representing ln(*rate*) depending linearly upon ln *pa* and ln *pb*.

Step Three. There are several other options; choose residuals, residual plots, and line fit plots. These are all useful for evaluating the results. You can place the results on another sheet in the same Workbook or on the same sheet by specifying a location. Table E-4 shows part of the output.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.99780809
R Square	0.99562099
Adjusted R Sq	0.99452624
Standard Erro	0.03856026
Observations	11

ANOVA

	df	SS
Regression	2	2.70450549
Residual	8	0.01189515
Total	10	2.71640064

RESIDUAL OUTPUT

Observation	d -0.68299884	Residuals
1	-0.48620447	0.02448642
2	-0.3970332	-0.05835772
3	0.27464776	-0.00043094
4	0.42120076	0.02751539
5	0.42020902	0.04860641
6	0.48001181	0.00346318
7	0.60333283	0.0372571
8	0.81274183	-0.03433481
9	0.80346734	-0.04673522
10	1.00064879	0.00097525
11	1.15402923	-0.00244505

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	1.96082013	0.04656974	42.1050291	1.1156E-10	1.85343006	2.06821021
-2.2595256	0.98035818	0.0230596	42.5140937	1.0329E-10	0.92718261	1.03353376
-2.26721795	0.18956607	0.01989072	9.53037767	1.2141E-05	0.14369796	0.23543418

Table E-4. Multiple regression of reaction rate data

The best fit is for

$$a = 1.9608,\, b = 0.9804,\, c = 0.1896,\, k = e^{a} = 7.105 \qquad \text{(E.20)}$$

The curve fit is then

$$rate = 7.105\, p_A^{0.9804}\, p_B^{0.1896} \qquad \text{(E.21)}$$

The ‘Standard Error’ gives an idea of how accurately the parameter is determined. If this value is a significant fraction of the parameter itself, the data is probably too scattered to be correlated in the way you have chosen. Also note the residuals, which are the errors in predicting the data. ‘Residuals’ should be both positive and negative with no trends. If the first five residuals were negative and the last six residuals were positive, it would indicate some systematic trend that is not accounted for by the formula used, such as Eq. (E.19). Fig. E-5 shows one residual plot, indicating that the errors are scattered and both positive and negative with no trends. The r^2 value is 0.9956, which indicates a good correlation.

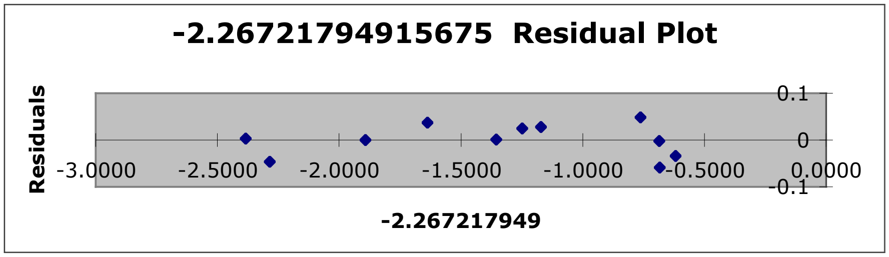


Figure E-5. Residual plot for reaction rate correlation

The rate can be calculated for each data point and compared with the data to obtain a residual. Those residuals are summed to obtain a least squares value for the correlation, as shown in Table E-5.

$$\sigma^2 = \sum_{i=1}^{N} \frac{[rate_i \; - \; rate(pa_i, pb_i)]^2}{N} \qquad \text{(E.22)}$$

F	G	H	I	J
ln(rate)	predicted ln(rate)	predicted rate using ln form	residual using ln form	squares of residuals
-0.6830	-0.6841	0.5045	-0.0006	3.1571E-07
-0.4617	-0.4862	0.6150	-0.0152	2.3238E-04
-0.4554	-0.3970	0.6723	0.0381	1.4525E-03
0.2742	0.2746	1.3161	0.0006	3.2152E-07
0.4487	0.4212	1.5238	-0.0425	1.8071E-03
0.4688	0.4202	1.5223	-0.0758	5.7487E-03
0.4835	0.4800	1.6161	-0.0056	3.1433E-05
0.6406	0.6033	1.8282	-0.0694	4.8161E-03
0.7784	0.8127	2.2541	0.0761	5.7881E-03
0.7567	0.8035	2.2333	0.1020	1.0398E-02
1.0016	1.0006	2.7200	-0.0027	7.0438E-06
1.1516	1.1540	3.1709	0.0077	5.9964E-05
			(sum of squares)/N	2.5285E-03

Table E-5. Least squares calculation for reaction rate correlation using logarithmic form

You might think at this point the correlation is complete. It isn't, though, because the data was transformed to make the parameter estimation problem linear. Thus, the statistics are in terms of the transformed problem. It is always a good idea to calculate the curve fit using the original variables. You can do this most conveniently by duplicating some columns so they are adjacent for plotting purposes, as shown in Table E-6.

L	M	N	O
rate data	predicted rate	rate data	rate data
0.5051	0.5045	0.5051	0.5051
0.6302	0.6150	0.6302	0.6302
0.6342	0.6723	0.6342	0.6342
1.3155	1.3161	1.3155	1.3155
1.5663	1.5238	1.5663	1.5663
1.5981	1.5223	1.5981	1.5981
1.6217	1.6161	1.6217	1.6217
1.8976	1.8282	1.8976	1.8976
2.1780	2.2541	2.1780	2.1780
2.1313	2.2333	2.1313	2.1313
2.7227	2.7200	2.7227	2.7227
3.1632	3.1709	3.1632	3.1632

Table E-6. Collection of data for plotting

Column L is the rate data, duplicated. Column M is the prediction using Eq. (E.21). Columns N and O are duplicated to permit a straight line, as shown in Fig. E-6. This figure plots the predicted value, column M, versus the experimental one, column L. **Step One.** You can make the figure by choosing ‘Insert/Chart’ as before, plotting column L and M using only data points (no line). **Step Two.** If the correlation is perfect, all of the predicted values will lie on a straight line going through the origin with a slope of 1.0. To get that line, right-click (CTRL-click on the Mac) on the diagram, choose ‘Source Data/Series, ‘add a series,’ and add the second curve (columns N and O). **Step Three.** Then right-click on one of those data points, choose ‘Format data series,’ and choose ‘line’ and unselect ‘data points.’ The correlation is reasonably good, and the example is finished.

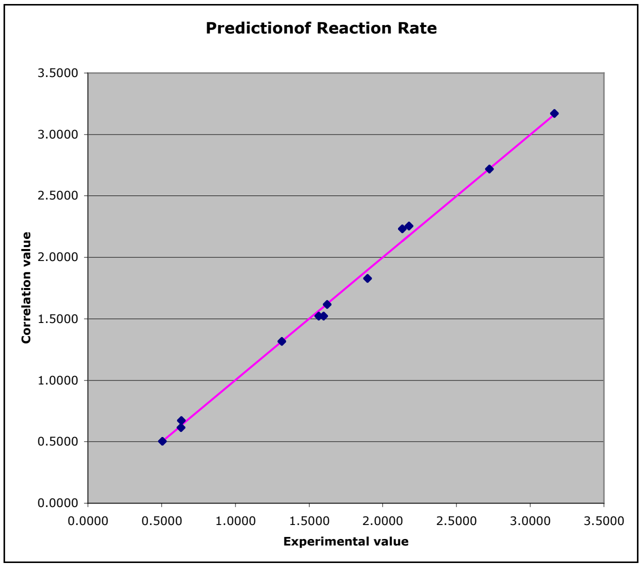


Figure E-6. Comparison of predicted reaction rate versus experimental value

Nonlinear regression

Nonlinear regression is a curve fit in which the unknown parameters enter into the problem in a nonlinear way. Nonlinear regression is much more difficult (for the computer), so it is best to always try to manipulate your model into a form that is linear. Sometimes that isn't possible, and then nonlinear regression must be used. You need to be aware, though, that the methods described here don't always work. Nonlinear regression uses techniques borrowed from the field of optimization, and it is difficult to construct a method that works every single time for every problem.

To use nonlinear regression, you minimize Eq. (E.3) with respect to the unknown parameters. Polynomial and multiple regression do this too (behind the scenes), but for nonlinear curve fits it is necessary to use functions such as 'Solver' in Excel and 'fminsearch' in MATLAB. This is demonstrated using the same example given above for multiple regression.

Nonlinear regression using Excel. Step One. Place the data on a new sheet in the Workbook, columns A, B, and C from Table E-3, as reproduced in Table E-7.

Step Two. Select some cells for the parameters *k*, *n*, and *m*. Arbitrary values are inserted – use your best guess, because that will possibly mean the difference between success and failure.

	A	B	C	D	E	F
1	pa	pb	measured rate	calculated rate	residual	residual squared
2	0.1044	0.1036	0.5051	0.0108	0.4943	2.4432E-01
3	0.1049	0.2871	0.6302	0.0301	0.6001	3.6010E-01
4	0.1030	0.5051	0.6342	0.0520	0.5822	3.3893E-01
5	0.2582	0.1507	1.3155	0.0389	1.2766	1.6297E+00
6	0.2608	0.3100	1.5663	0.0808	1.4855	2.2066E+00
7	0.2407	0.4669	1.5981	0.1124	1.4857	2.2074E+00
8	0.3501	0.0922	1.6217	0.0323	1.5894	2.5263E+00
9	0.3437	0.1944	1.8976	0.0668	1.8308	3.3518E+00
10	0.3494	0.5389	2.1780	0.1883	1.9897	3.9589E+00
11	0.4778	0.1017	2.1313	0.0486	2.0827	4.3377E+00
12	0.4880	0.2580	2.7227	0.1259	2.5968	6.7433E+00
13	0.5014	0.5037	3.1632	0.2526	2.9106	8.4719E+00
14						
15						
16		k	1		(sum of	
17		n	1		squares/N)	3.03139933
18		m	1			

Table E-7. Correlation of rate expression using 'Solver' in Excel; initial guess

Step Three. In column D, calculate the value of rate using the parameters in C16:C18, the data in columns A and B, and the formula, Eq. (E.17). **Step Four.** Make column E the difference between columns C and D, and then square the result and put it in column F. **Step Five.** Sum Column F, divide by the number entries [COUNT(F2:F13)] to obtain the (sum of squares)/N, Eq. (E.3).

Step Six. The goal is to minimize F17 by choosing values C16:C18. To do that, choose 'Tools/Solver.' You might have to add it to the Excel program if that wasn't done when the program was installed. A screen appears in which you insert F17 as the quantity to be affected, and choose 'Min' as the option. Then insert C16:C18 as the cells to be changed, and click 'Solve.' The results are shown in Table E-8.

	A	B	C	D	E	F
			measured	calculated		residual
1	pa	pb	rate	rate	residual	squared
2	0.1044	0.1036	0.5051	0.5162	-0.0111	1.2379E-04
3	0.1049	0.2871	0.6302	0.6331	-0.0029	8.2401E-06
4	0.1030	0.5051	0.6342	0.6948	-0.0606	3.6753E-03
5	0.2582	0.1507	1.3155	1.3203	-0.0048	2.2770E-05
6	0.2608	0.3100	1.5663	1.5351	0.0312	9.7617E-04
7	0.2407	0.4669	1.5981	1.5404	0.0577	3.3307E-03
8	0.3501	0.0922	1.6217	1.6045	0.0172	2.9704E-04
9	0.3437	0.1944	1.8976	1.8242	0.0734	5.3903E-03
10	0.3494	0.5389	2.1780	2.2623	-0.0843	7.1090E-03
11	0.4778	0.1017	2.1313	2.2018	-0.0705	4.9745E-03
12	0.4880	0.2580	2.7227	2.6957	0.0270	7.2844E-04
13	0.5014	0.5037	3.1632	3.1534	0.0098	9.5898E-05
14						
15						
16		k	6.97755309		(sum of	
17		n	0.95605386		squares/N)	0.00222768
18		m	0.195694			

Table E-8. Correlation of rate expression using Solver in Excel

The best correlation is

$$rate = 6.978 p_A^{0.9561} p_B^{0.1957} \tag{E.23}$$

These numbers are slightly different than those obtained using multiple regression. Multiple regression and nonlinear regression obtained the solution by minimizing two different objective functions. Notice that the (sum of squares/N) is smaller when using nonlinear regression, because nonlinear regression minimizes that exact quantity. If nonlinear regression doesn't work, though, then multiple linear regression is your only option.

Nonlinear regression using MATLAB. Step One. Construct a function which calculates the (sum of the squares)/N, using these commands.

```
function value=para(parameters)
k = parameters(1);
n = parameters(2);
m = parameters(3);
pa = [0.1044 0.1049 0.1030 ...];
pb = [0.1036   0.2871   0.5051   ...];
```

```
rate1 = [0.5051    0.6302    0.6342    ...];
value = 0.;
for i=1:12
   rate2(i)=k*(pa(i)^n)*(pb(i)^m);
   value = value + (rate1(i)-rate2(i))^2;
end
value = value/12;
```

Step Two. Test this function by removing the semi-colons. Then issue the following command

```
feval(@para,[2 3 4])
```

and calculate a few terms to check. In this case, Table E-7 can be used to provide values for checking.

Step Three. The minimum of value is found using the fminsearch function.

```
fminsearch(@para,[1 1 1])
ans = 6.9776    0.9561    0.1957
```

These numbers are the same as obtained using Solver in Excel, as expected. The best correlation is Eq. (E.23).

References

Perry, R. H., Green, D. W. (ed.), *Perry's Chemical Engneers' Handbook*, 8th ed. McGraw-Hill: New York, 2008.

Press, W. H., Flannery, E. P., Teukolsky, S. A., Vettering, W. T., *Numerical Recipes,* Cambridge University Press: Cambridge, 1986.