Numerical Methods for Civil Engineers

Lecture Notes
CE 311K
Daene C. McKinney
Introduction to Computer Methods
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin

Regression

Introduction

Consider the nature of most experimental data. Typically such data include noise due to many different effects. The noisy data from an experiment might appear as shown in the following Table and Figure. We assume that the *x* values are accurate. Visual inspection of the data suggests a positive relationship between *x* and $y = f(x)$, i.e., higher values of *y* are associated with higher values of *x*. One strategy for deriving an approximating function for this data might be to try to fit the general trend of the data without necessarily matching the individual points. A straight line could be used to generally characterize the trend in the data without passing through any particular point. The line in Figure 1 has been sketched through the points. Although this approach may work well in many cases, it does not provide us with any quantitative measure of how good the fit of the line is to the data. We need a criterion with which to measure the *goodness of fit* of the line to the data. One way to do this is to derive a curve that minimizes the discrepancy between the data points and the curve. The technique for accomplishing this is called *least-squares regression*.

Often data are available at discrete points and we require estimates at points between the discrete values. In this section we will discuss techniques to fit curves to data in order to estimate intermediate, or fitted, values. Two methods of curve fitting are generally considered, depending on the amount of error in the data. When the data are known to be precise, the method of *interpolation* is used. The primary purpose of interpolation is to provide information between tabular data, and, as accurately as possible, to force the approximating function to assume exactly the value provided at each of the points where the data is supplied. For significantly "noisy" data, a single curve representing the general trend of the data is derived by the method of *least-squares regression*.

Table 1. Noisy Data from an Experiment.

i	1	2	3	4	5
x	2.10	6.22	7.17	10.52	13.68
$f(x)$	2.90	3.83	5.98	5.71	7.74

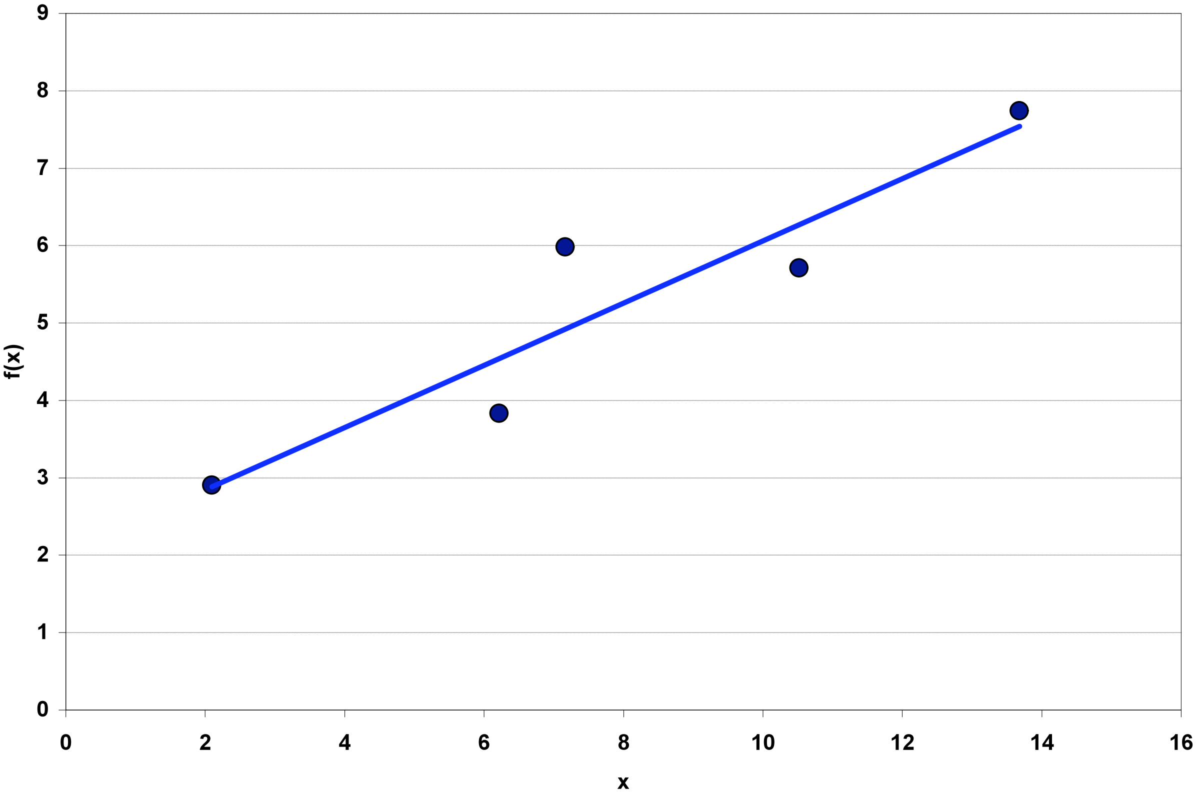


Figure 1. Noisy data from an experiment.

Linear Least Squares Regression

Consider fitting a straight line to a set of data such as those shown the previous Table. Let the data be represented by the set of *n* data points:

$$(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)$$

The equation of a straight line through the data is

$$y = a_0 + a_1 x$$

where a_0 is the vertical intercept and a_1 is the slope of the line. If the relationship between *x* and *y* were indeed truly linear and there was no noise in the data, then the slope and intercept could be estimated such that the line passed through all of the data points. However, as can be seen from Figure 1, this is not the case. These is a discrepancy, or *residual*, between the true value of *y* and the linear approximation $y = a_0 + a_1 x$. This residual is denoted by *e* and is defined by

$$e = y - a_0 - a_1 x$$

The approximating function, the straight line, must now be chosen such that, in some sense, the discrepancy *e* is minimized over the entire range of *x* where the approximation is to be applied.

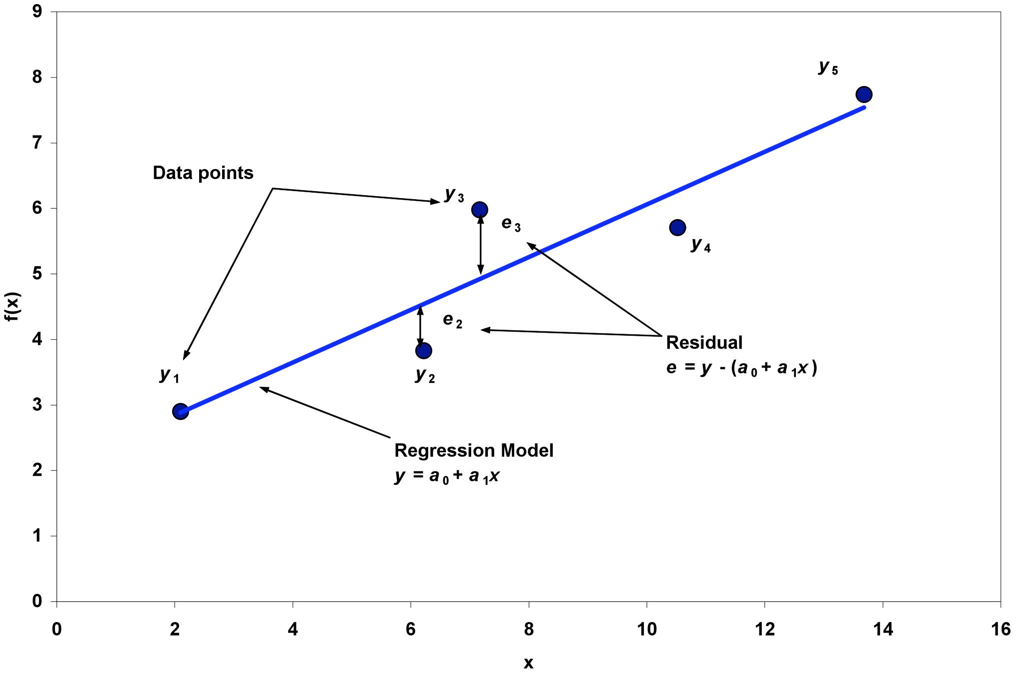


Figure 2. Noisy data from an experiment with residuals.

The sense in which *e* is minimized is clearly a vital factor in determining the character of the approximation. We could minimize the maximum value of *e* for all data points. However, this criterion is usually not an effective one to use in selecting a continuous functional approximation of noisy data, simply because it permits individual points -- which may be badly in error -- to exert overpowering influence on the approximating function. That is, a single point can force the approximating function to shift drastically toward it in order to minimize the maximum error which would tend to occur at that point. A much more favorable condition to apply to minimize *e* for this type of approximation is the *least-squares criterion*.

If we denote the *x* coordinates at which data are available as x_i, then the *i*-th residual e_i is

$$e_i = y_i - a_0 - a_1 x_i$$

and if there are *n* such coordinates, then the sum of squared residuals over all the data points is

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

In order to determine the values of the coefficients a_0 and a_1, we can minimize S_r. The minimization is accomplished by setting the partial derivatives of S_r with respect to each coefficient equal to zero:

$$\begin{aligned}
\frac{\partial S_r}{\partial a_0} &= \frac{\partial}{\partial a_0}\left[\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i\right)^2\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial a_0}[\cdots]^2 = \sum_{i=1}^{n} 2[\cdots]\left\{\frac{\partial}{\partial a_0}[\cdots]\right\} \\
&= \sum_{i=1}^{n} 2\left[y_i - a_0 - a_1 x_i\right](-1) \\
&= 0
\end{aligned}$$

or, dividing by -2 and summing term by term, we have

$$\boxed{n a_0 + \left[\sum_{i=1}^{n} x_i\right] a_1 = \sum_{i=1}^{n} y_i \qquad (1)}$$

Similarly, the second equation is

$$\begin{aligned}
\frac{\partial S_r}{\partial a_1} &= \frac{\partial}{\partial a_1}\left[\sum_{i=1}^{n}\left(y_i - a_0 - a_1 x_i\right)^2\right] \\
&= \sum_{i=1}^{n}\frac{\partial}{\partial a_1}[\cdots]^2 = \sum_{i=1}^{n} 2[\cdots]\left\{\frac{\partial}{\partial a_1}[\cdots]\right\} \\
&= \sum_{i=1}^{n} 2\left[y_i - a_0 - a_1 x_i\right](-x_i) \\
&= 0
\end{aligned}$$

or, dividing by -2 and summing term by term, we have

$$\left[\sum_{i=1}^{n} x_i\right]a_0 + \left[\sum_{i=1}^{n} x_i^2\right]a_1 = \sum_{i=1}^{n} x_i y_i \tag{2}$$

Now, Equations (1) and (2) represent of two simultaneous linear equations in two unknowns (a_0 and a_1):

$$na_0 + \left[\sum_{i=1}^{n} x_i\right]a_1 = \sum_{i=1}^{n} y_i$$

$$\left[\sum_{i=1}^{n} x_i\right]a_0 + \left[\sum_{i=1}^{n} x_i^2\right]a_1 = \sum_{i=1}^{n} x_i y_i$$

These are called the *normal equations*. The solution to these equations is

$$a_0 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2} \qquad a_1 = \frac{\sum_{i=1}^{n} x_i\, y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]^2}$$

Example: Given the following noisy data, fit a straight line to this data by using least squares.

Table 2. Noisy Data from an Experiment.

i	1	2	3	4	5
x	2.10	6.22	7.17	10.52	13.68
f(x)	2.90	3.83	5.98	5.71	7.74

Each element of these equations can now be computed

$$\sum_{i=1}^{5} x_i = 39.69$$

$$\sum_{i=1}^{5} x_i^2 = 392.3201$$

$$\sum_{i=1}^{5} y_i = 26.16$$

$$\sum_{i=1}^{5} x_i y_i = 238.7416$$

The solution of the normal equations is

$$a_0 = \frac{\frac{1}{5}(26.16)(392.3) - \frac{1}{5}(39.69)(238.7)}{392.3 - \frac{1}{5}\left[39.69\right]^2} = 2.038 \qquad a_1 = \frac{238.7 - \frac{1}{5}(39.69)(26.16)}{392.3 - \frac{1}{5}\left[39.69\right]^2} = 0.4023$$

The required straight line is thus

$$y = 2.038 + 0.4023x$$

Polynomial Regression

Previously, we fit a straight line to noisy data

$$(x_1,y_1),(x_2,y_2),\cdots(x_n,y_n)$$

using the least-squares criterion. As we have seen, some data are poorly represented by a straight line and for these cases a curve is better suited to fit the data. The most commonly used function for this purpose is the polynomial such as a parabola

$$y = a_0 + a_1 x + a_2 x^2$$

or a cubic

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

or in general an *m*th degree polynomial:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m$$

where $a_0, a_1, a_2, \cdots, a_m$ are the constant coefficients of the polynomial.

If the relationship between *x* and *y* were indeed truly *m*-th degree polynomial and there was no noise in the data, then the coefficients could be estimated such that the polynomial passed through all of the data points. However, this is hardly ever the case. As in the linear case, the discrepancy (residual) between the true value of *y* and the polynomial approximation is

$$e_i = y_i - (a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \cdots + a_m x_i^m)$$

and if there are *n* such pairs of points (x_i, y_i), then the sum of squared residuals over all the data points is

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y - (a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_m x^m) \right]^2$$

In order to determine the values of the coefficients $a_0, a_1, a_2, \cdots, a_m$, we can minimize S_r. The minimization is accomplished by setting the partial derivatives of S_r with respect to each coefficient equal to zero:

$$\begin{aligned}\frac{\partial S_r}{\partial a_0} &= \frac{\partial}{\partial a_0}\left[\sum_{i=1}^{n}\left(y - a_0 - a_1x - a_2x^2 - a_3x^3 - \cdots - a_mx^m\right)^2\right]\\ &= \sum_{i=1}^{n}\frac{\partial}{\partial a_0}[\cdots]^2 = \sum_{i=1}^{n}2[\cdots]\left\{\frac{\partial}{\partial a_0}[\cdots]\right\}\\ &= \sum_{i=1}^{n}2\left[y - a_0 - a_1x - a_2x^2 - a_3x^3 - \cdots - a_mx^m\right](-1)\\ &= 0\end{aligned}$$

Now, dividing by -2 and summing term by term, we have

$$na_0 + \left[\sum_{i=1}^{n}x_i\right]a_1 + \left[\sum_{i=1}^{n}x_i^2\right]a_2 + \cdots \left[\sum_{i=1}^{n}x_i^m\right]a_m = \sum_{i=1}^{n}y_i$$

Similarly, the second equation is

$$\begin{aligned}\frac{\partial S_r}{\partial a_1} &= \frac{\partial}{\partial a_1}\left[\sum_{i=1}^{n}\left(y - a_0 - a_1x - a_2x^2 - a_3x^3 - \cdots - a_mx^m\right)^2\right]\\ &= \sum_{i=1}^{n}\frac{\partial}{\partial a_1}[\cdots]^2 = \sum_{i=1}^{n}2[\cdots]\left\{\frac{\partial}{\partial a_1}[\cdots]\right\}\\ &= \sum_{i=1}^{n}2\left[y - a_0 - a_1x - a_2x^2 - a_3x^3 - \cdots - a_mx^m\right](-x_i)\\ &= 0\end{aligned}$$

Dividing by -2 and summing term by term, we have

$$\left[\sum_{i=1}^{n}x_i\right]a_0 + \left[\sum_{i=1}^{n}x_i^2\right]a_1 + \left[\sum_{i=1}^{n}x_i^3\right]a_2 + \cdots + \left[\sum_{i=1}^{n}x_i^{m+1}\right]a_m = \sum_{i=1}^{n}x_iy_i$$

It can now be inferred that the complete set of simultaneous linear equations in the coefficients (the normal equations) of the polynomial is

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \cdots & \sum_{i=1}^{n} x_i^m \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \cdots & \sum_{i=1}^{n} x_i^{m+1} \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \cdots & \sum_{i=1}^{n} x_i^{m+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_i^m & \sum_{i=1}^{n} x_i^{m+1} & \sum_{i=1}^{n} x_i^{m+2} & \cdots & \sum_{i=1}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \\ \vdots \\ \sum_{i=1}^{n} x_i^m y_i \end{bmatrix}$$

Example: Given the following data, choose the most suitable low order polynomial and fit it to this data using the least-squares criterion.

Table 3. Data for polynomial fitting example.

x	0	1.0	1.5	2.3	2.5	4.0	5.1	6.0	6.5	7.0	8.1	9.0
y	0.2	0.8	2.5	2.5	3.5	4.3	3.0	5.0	3.5	2.4	1.3	2.0
x	9.3	11.0	11.3	12.1	13.1	14.0	15.5	16.0	17.5	17.8	19.0	20.0
y	-0.3	-1.3	-3.0	-4.0	-4.9	-4.0	-5.2	-3.0	-3.5	-1.6	-1.4	-0.1

The data are plotted in the following Figure. The data appear to have a maximum near $x = 5$ and a minimum near $x = 15$. The lowest order polynomial which can reproduce such behavior is a cubic. The least-squares equations (normal equations) for this set of data ($n = 24$) and for $m = 3$ are

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^5 & \sum_{i=1}^{n} x_i^6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i^3 y_i \end{bmatrix}$$

$$\begin{bmatrix} 24 & 229.6 & 3060.2 & 46342.8 \\ 229.6 & 3060.2 & 46342.8 & 752835.2 \\ 3060.2 & 46342.8 & 752835.2 & 12780147.7 \\ 46342.8 & 752835.2 & 12780147.7 & 223518116.8 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} -1.30 \\ -316.9 \\ -6037.2 \\ -9943.36 \end{bmatrix}$$

Gauss elimination yields

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} -0.3593 \\ 2.3051 \\ -0.3532 \\ 0.0121 \end{bmatrix}$$

Thus the equations for the interpolating polynomial is

$$y = 0.0121x^3 - 0.3532x^2 + 2.3051x - 0.3593$$

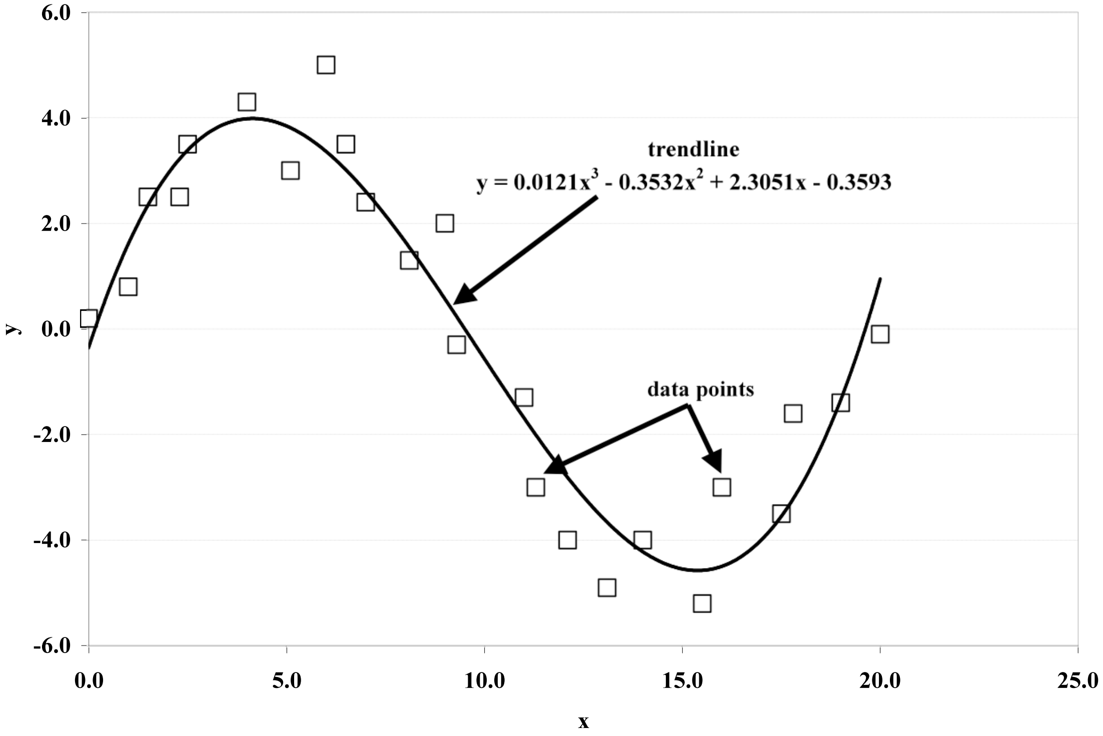


Figure 3. Plot of data for polynomial fitting example

Linearization of Nonlinear Relationships

In order to apply the techniques of linear least-squares regression, the function whose coefficients are being approximated must be linear in the coefficients. Many relationships among independent and dependent variables in engineering are not linear. However, in many cases a transformation can be applied to the relationships to render them linear in the coefficients. Consider an exponential relationship,

$$y = ae^{bx}$$

where the base is the number e, and a and b are constants. If we take the natural logarithm of both sides of the equation, we have

$$\ln(y) = \ln(a) + bx$$

which is a linear relationship between ln(y) and x. The coefficients to be determined in this expression are ln(a) and b. A power law relationship be written as

$$y = ax^{b}$$

If we take the natural logarithm of both sides of this equation, we have

$$\ln(y) = \ln(a) + b\ln(x)$$

which is a linear relationship between ln(y) and ln(x). Again, the coefficients to be determined in this expression are ln(a) and b.

Example. Given the data in the following table, use the least-squares criterion to fit a function of the form Ax^{B} to these data.:

i	1	2	3	4	5	6

x	1.2	2.8	4.3	5.4	6.8	7.9
y	2.1	11.5	28.1	41.9	72.3	91.4

The power law relationship is

$$y = Ax^B$$

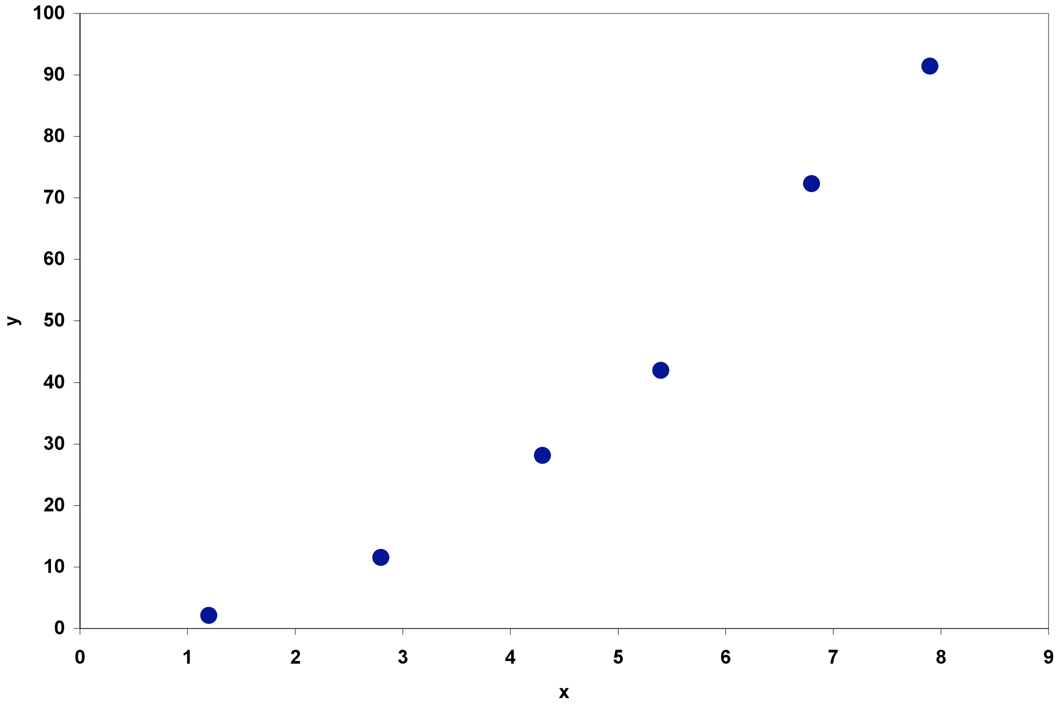
Take the natural logarithm of both sides

$$\ln(y) = \ln(A) + B\ln(x)$$

which is a linear relationship between ln(y) and ln(x). The coefficients to be determined are ln(A) and B. Another way to look at this is

$$Y = a + BX$$

which is a linear relationship between Y=ln(y) and X=ln(x). The coefficients to be determined are a=ln(A) and B



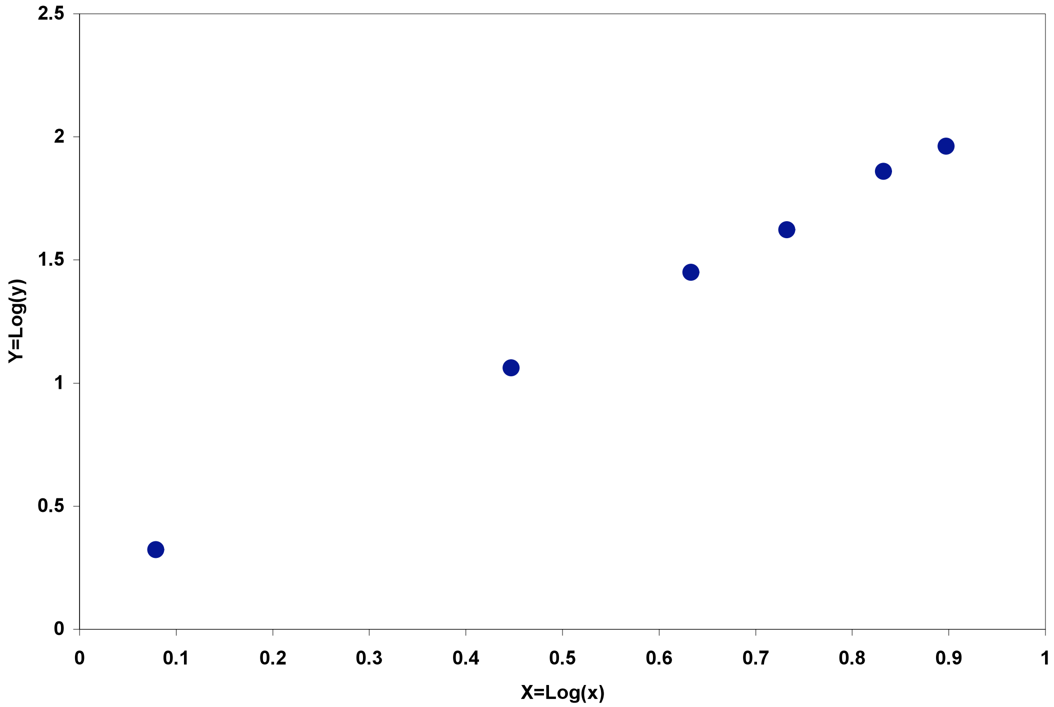


Figure. Plot of data on arithmetic and log-log axes.

The normal equations are

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix} \begin{bmatrix} a \\ B \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i Y_i \end{bmatrix}$$

x_i	$X_i = \ln(x_i)$	X_i^2	y_i	$Y_i = \ln(y_i)$	X_iY_i
1.2	0.18	0.03	2.1	0.74	0.14
2.8	1.03	1.06	11.5	2.44	2.51
4.3	1.46	2.13	28.1	3.34	4.87
5.4	1.69	2.84	41.9	3.74	6.3
6.8	1.92	3.67	72.3	4.28	8.21
7.9	2.07	4.27	91.4	4.52	9.33

$$\sum_{i=1}^{5} X_i = \sum_{i=1}^{5} \ln(x_i) = 8.34$$

$$\sum_{i=1}^{5} X_i^2 = \sum_{i=1}^{5} \ln(x_i)^2 = 14.0$$

$$\sum_{i=1}^{5} Y_i = \sum_{i=1}^{5} \ln(y_i) = 19.1$$

$$\sum_{i=1}^{5} X_i Y_i = \sum_{i=1}^{5} \ln(x_i)\ln(y_i) = 31.4$$

Plugging in the numerical values from the data table, the normal equations are

$$\begin{bmatrix} 6 & 8.34 \\ 8.34 & 14.0 \end{bmatrix} \begin{bmatrix} a \\ B \end{bmatrix} = \begin{bmatrix} 19.1 \\ 31.4 \end{bmatrix}$$

Solution yields

$$a = ? \qquad so \qquad A = \exp(a) = ?$$
$$B = ?$$

Example - Carbon Adsorption

Adsorption involves the accumulation of dissolved substances at interfaces of and between material phases. Adsorption may occur as the result of the attraction of a surface or interface for a chemical species, such as the adsorption of substances from water by activated carbon (Weber and DiGiano, 1996) as commonly used in home water filters. Carbon is well known for its adsorptive properties. Activated carbon is regularly used to remove taste and odors from drinking water since carbon has a unique ability to remove synthetic organic chemicals from water supplies.

Adsorption is the process where molecules of a liquid or gas are attached to and then held at the surface of a solid. Physical adsorption is the process whereby surface tension causes molecules to be held at the surface of a solid. Chemical adsorption occurs when a chemical reaction occurs to cause molecules to be held at the surface by chemical bonding. Physical adsorption occurs on activated carbon. The large surface area of the carbon makes it an excellent adsorbent material. Macropores in the surface of the activated carbon granules provide an entrance into the interior of the granual. Adsorption requires three processes: (1) diffusion through a liquid phase to reach the carbon granule, (2) diffusion of molecules through macropores in the carbon granule to an adsorption site, and (3) adsorption of the molecule to the surface. These processes occur at different rates for different molecules of different substances.

Sorption studies are conducted by equilibrating known quantities of sorbent (say, carbon) with solutions of solute (the pollutant). Plots of the resulting data relating the variation of solid-phase concentration, or amount of the solute (pollutant) sorbed per unit mass of solid (carbon), to the variation of the solution-phase concentration are termed sorption isotherms (Weber and DiGianno, 1996). They are referred to as isotherms because the data are collected at constant temperature.

To evaluate the effectiveness of using activated carbon to remove pollutants from water, the first step is to perform a liquid-phase adsorption isotherm test. Data are generated by adding known weights of carbon to water containing a known concentration of pollutant. The carbon-water mixture is mixed at constant temperature, then the carbon is removed by filtration. The residual pollutant concentration in the water is measured and the amount of pollutant adsorbed on to the carbon is calculated. This value if divided by the weight of carbon to determine the carbon loading (q).

Several models have been developed to represent sorption isotherms mathematically. These include:

(1) Linear isotherm model

$$q = Kc$$

where q is the mass of pollutant sorbed per unit mass of carbon at equilibrium with a solution of pollutant concentration c, and K is called the distribution coefficient. The distribution coefficient can be determined by fitting a straight line through the origin to the data.

(2) Langmuir isotherm model

$$q = \frac{Qbc}{1 + bc}$$

where Q is the maximum adsorption capacity, and b is a rate constant, and

(3) Freundlich isotherm model

$$q = K(c)^n$$

where, K is called the specific capacity, an indicator of sorption capacity at a specific pollutant concentration; n is a measure of the energy of the sorption reaction. Both of the parameters can be determined fitting a straight line to the logarithmic transformation

$$\ln q = \ln K + (n)\ln c$$

or

$$\log_{10} q = \log_{10} K + n \log_{10} c$$

Table. Adsorption data for a pollutant (phenol).

c	**q**	**logc**	**logq**
2.8	77.8	0.45	1.89
3.1	90.9	0.49	1.96
12.1	132.7	1.08	2.12
18	153.6	1.26	2.19
30.4	171.4	1.48	2.23
36.2	185.4	1.56	2.27
48.5	196.2	1.69	2.29
46.4	187.2	1.67	2.27
63	193.4	1.80	2.29
71.4	232.6	1.85	2.37
78.1	204.4	1.89	2.31
87.7	206.2	1.94	2.31
102	210.8	2.01	2.32
109	218	2.04	2.34
102	230.5	2.01	2.36
180	259.2	2.26	2.41
273	271.4	2.44	2.43
353	285.2	2.55	2.46
434	294.3	2.64	2.47
526	279.9	2.72	2.45
600	268	2.78	2.43

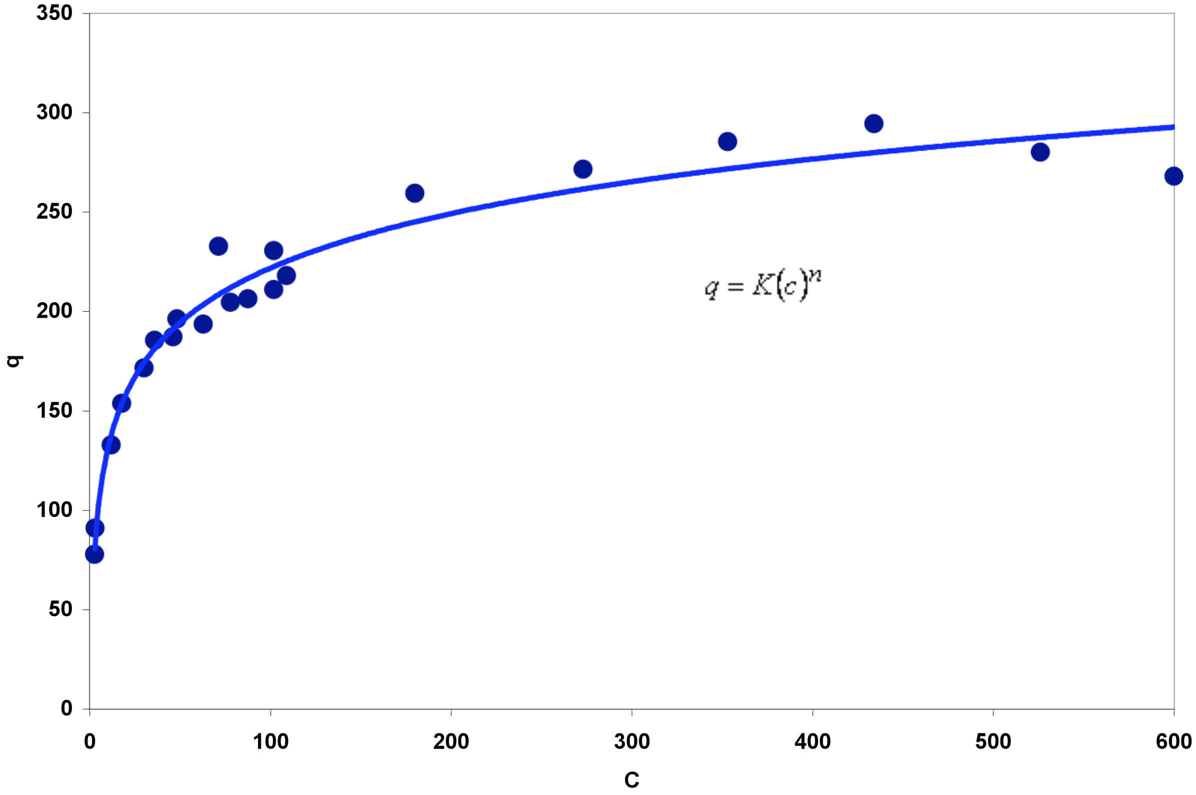


Figure. Phenol isotherm (arithmetic scales on axes).

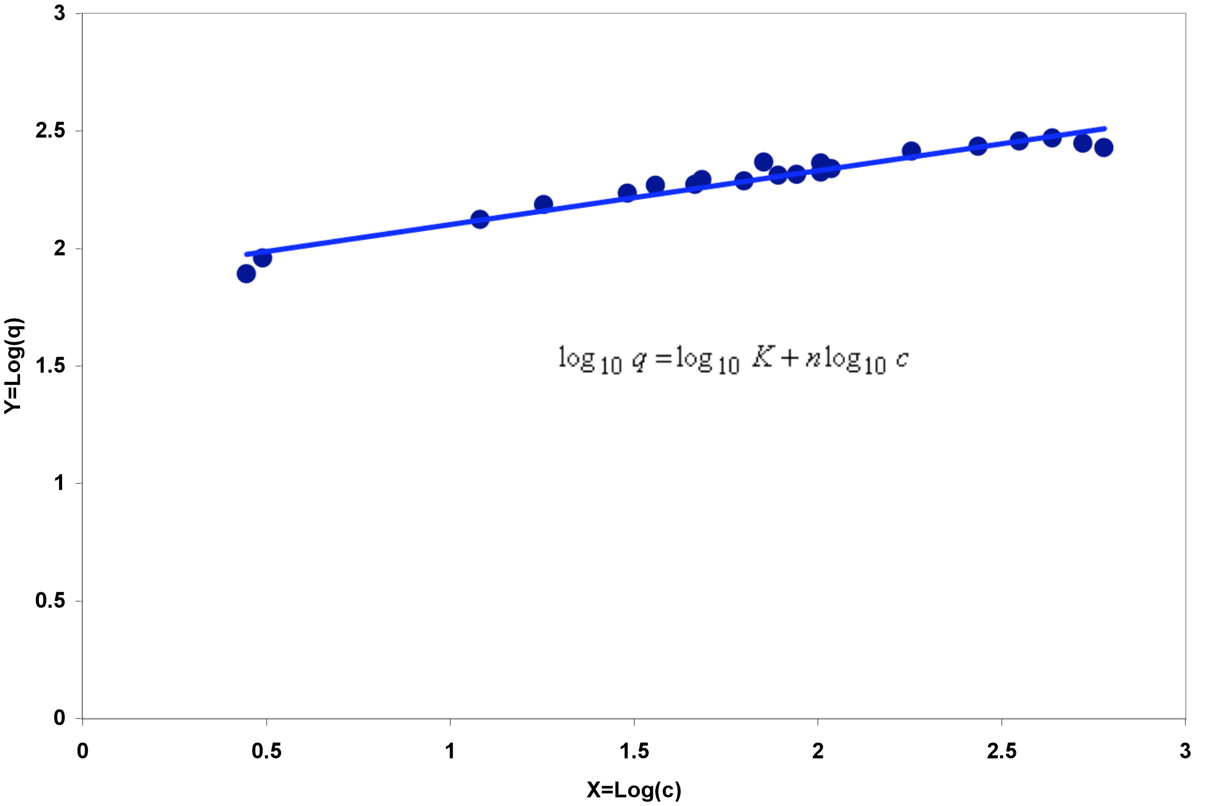


Figure. Phenol isotherm (logarithmic scales on axes).

Using the Arithmetic axes:

$$q = K(c)^n$$

so

$$K = 74.702\text{, and } n = 0.2289$$

Using the Logarithmic axes:

$$\log_{10} q = \log_{10} K + n \log_{10} c$$

so

$$logK = 1.8733$$

or

$$K = 10^{1.6733} = 74.696$$

and

$$n = 0.2289$$

Exercises

1. Use least-squares regression to fir a straight line to the following data:

X	1	3	5	6	10	12	13	16	18	20
Y	4	5	6	5	8	7	6	9	12	11

2. An example of a nonlinear model that is sometimes fitted to data is the saturation-growth-rate equation

$$y = a\frac{x}{b + x}$$

where *a* and *b* are constant coefficients. This model is often used for population growth rate models under limiting conditions where the population *y* levels off (saturates) as *x* increases. This model can be linearized by inverting it to give

$$\frac{1}{y} = \frac{b}{a}\frac{1}{x} + \frac{1}{a}$$

$$Y = mX + s$$

Thus a plot of $Y = 1/y$ versus $X = 1/x$ is linear, with a slope of $m = b/a$ and an intercept of $s = 1/a$.

Fit a saturation-growth-rate model to

x	0.75	2	2.5	4	6	8	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7

Show your work. Plot the data and the resulting equation.

3. Given the data

X	5	10	15	20	25	30	35	40	45	50
y	16	25	32	33	38	36	39	40	42	42

Use least-squares regression to fit

(a) a straight line;

(b) a power equation;

(c) a saturation-growth-rate equation; and

(d) a parabola.

Plot the data along with all the curves. Is any one of the curves superior to the others?

4. Using least-squares regression, fit a parabola (second-order polynomial) to the data

X	1	2	2.5	4	6	8	8.5
Y	0.4	0.7	0.8	1.0	1.2	1.3	1.4

(a) Write the equation for your parabola?

(b) What are the unknown coefficients in the equation from part (a)?
(c) Write the 3 x 3 set of “normal” equations needed to compute the unknown coefficients using the least-squares method.
(d) Solve for the coefficients using Gauss Elimination.

5. Using least-squares regression, fit a power function to the carbon adsorption isotherm data given in the following table. Be sure to:

a. Write the equation for your function?
b. Write the set of “normal” equations needed to compute the unknown coefficients using the least-squares method.

Table. Adsorption data for phenol. Ce is the equilibrium liquid phase concentration of phenol and qe is the GAC loading (mg/g) of phenol on carbon at equilibrium.

Ce	Qe				
(mg/L)	(mg/g)				
2.8	77.8				
12.1	132.7				
30.4	171.4				
48.5	196.2				
63.0	193.4				
78.1	204.4				
102	210.8				
102	230.5				
273	271.4				
434	294.3				

600	268.0				
	Sums =				

6. Risk assessment studies subject animals to a lifetime exposure to a fixed amount of a toxic chemical. The number of animals of a population of *n(d)* showing a response *r(d)* at dose *d* are counted. Response means that the animal dies or develops a cancerous tumor. The probability of a response is estimated as a frequency, $\theta_j = {}^{r_j}\!/_{n_j}$, at chemical dose *d*.

The following table shows the liver cancer responses of animals fed a daily dose of DDT.

Group j	Dose d_j	Response r_j	Number of animals n_j	Lifetime risk probability estimate $\theta_j = {}^{r_j}\!/_{n_j}$
1	0	4	111	0.036
2	2	4	105	0.038
3	10	11	124	0.089
4	50	13	104	0.125
5	250	60	90	0.667

Using linear regression and these data, estimate the parameters of the Weibull model of the form

$$\theta = 1 - \exp\left(-\beta d^{\gamma}\right) \qquad (1)$$

a. Show any transformation(s) that you must apply to the equation (1) to make it linear in the (transformed) parameters. What is the final equation that you will use in the least squares regression?
b. What are the transformed values of the data that you must use in the linear regression?

Group
j
1
2
3
4
5

c. What are the normal equations that you must solve to find the values of the parameters of the linearized equation?

d. What are the numerical values which satisfy the normal equations?

e. What is the final form of the Weibull equation (1) using numerical values for the parameters.

7. Fit a saturation-growth-rate model $y = a\dfrac{x}{b + x}$ to the data

I	*x*	*y*				
1	0.75	0.80				
2	2.00	1.30				
3	2.50	1.20				
4	4.00	1.60				
5	6.00	1.70				
6	8.00	1.80				
7	8.50	1.70				
Sum						

8. Regression – Power Function. An example of a nonlinear model that is sometimes fitted to data is the power function equation

$$y = ax^b$$

where a and b are constant coefficients. Fit a power function model to the data in the following table, that is, find the values of a and b. Show your work.

9. Given the following data, fit a straight line to the data using the least-squares criterion:

x	1.1	2.9	4.3	6.2
y	50	43	28	25

Show ALL steps in the computation.

10. Given the following data:

x	1.2	2.8	4.3	5.4	6.8	7.9
y	2.1	11.5	28.1	41.9	72.3	91.4

Using the least-squares criterion, fit a power function of the form to this data.

Interpolation

Introduction

Linear Interpolation

If we assume that the graph between two points is a straight line, then we can use linear interpolation to find approximate values for a function between known pairs of points. The familiar formula is

$$f(x) = f(x_1) + \frac{(x - x_1)}{(x_2 - x_1)}[f(x_2) - f(x_1)]$$

where $x_1 \le x \le x_2$.

Example. (from Etter and Ingber, 2000).

Table. Data from an Experiment.

time (s)	Temp (deg F)
0	0
1	20
2	60
3	68
4	77
5	110

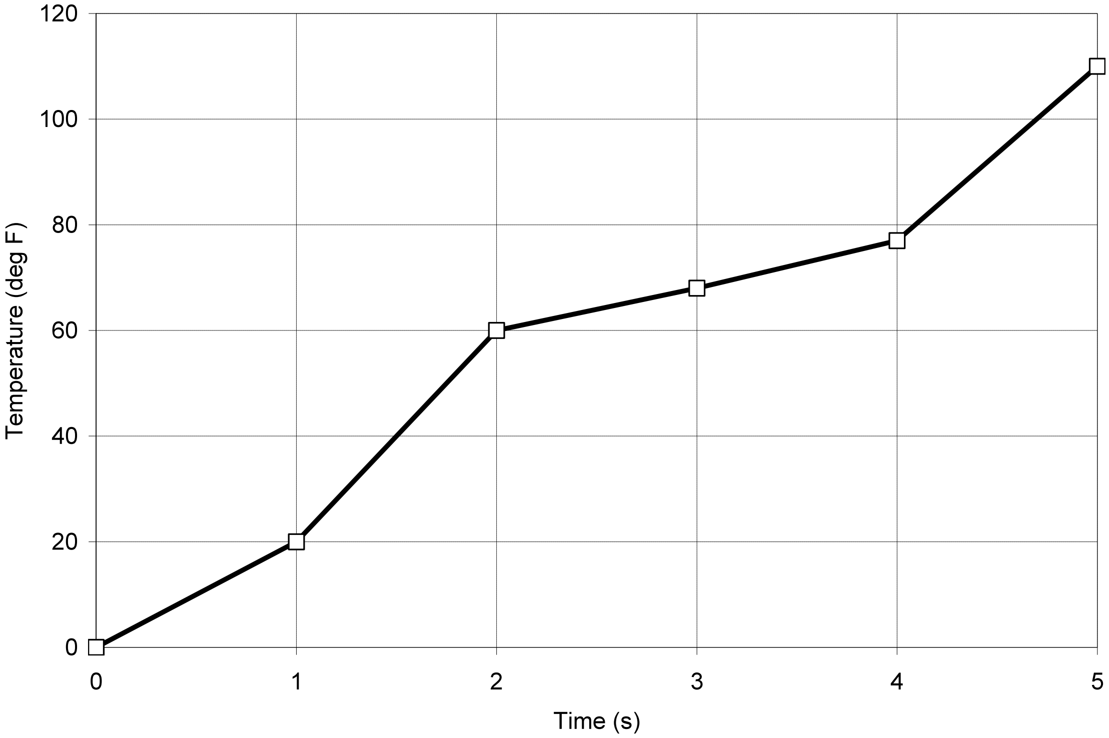


Figure. Data from an experiment.

If we interpolate the value for 2.6 seconds we have

$$f(2.6) = 60 + \frac{0.6}{1} = 64.8$$

Lagrange Interpolating Polynomials

Consider a series of points $f(x_i)$ where the x_i are unevenly spaced, and *i* can take on all integer values from 0 to *n* (there are *n*+1 points). The Lagrange interpolating polynomial can be represented as

$$f_n(x) = \sum_{i=0}^{n} L_i(x) f(x_i)$$

where

$$L_i(x) = \prod_{\substack{j=0 \\ j\neq i}}^{n} \left[\frac{x - x_j}{x_i - x_j} \right] = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

where $\prod$ represents the “product of.”

Note that

$$f_n(x_j) = \sum_{i=0}^{n} L_i(x_j) f(x_i) = L_j(x_j) f(x_j) = f(x_j)$$

so that at the nodes (data points) the interpolation function is identical to the data points, that is, the function passes through each data point.

The first-order Lagrange polynomial is

$$f_1(x) = \sum_{i=0}^{1} L_i(x) f(x_i) = L_0(x) f(x_0) + L_1(x) f(x_1)$$

where

$$L_0(x) = \prod_{\substack{j=0 \\ j\neq i=0}}^{1} \left[\frac{x - x_j}{x_0 - x_j} \right] = \frac{x - x_1}{x_0 - x_1}$$

$$L_1(x) = \prod_{\substack{j=0 \\ j\neq i=1}}^{1} \left[\frac{x - x_j}{x_1 - x_j} \right] = \frac{x - x_0}{x_1 - x_0}$$

$$f_1(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)$$

which is the same as the formula for linear interpolation discussed above.

The second order Lagrange polynomial is

$$f_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}f(x_2)$$

Example: Consider the following set of data:

i	0	1	2	3
x_i	1	2	4	8
$f(x_i)$	1	3	7	11

Suppose we wish to interpolate for $f_3(7)$ using a third order Lagrange polynomial. The third order Lagrange polynomial is

$$f_3(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)}f(x_0) + \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)}f(x_1) + \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)}f(x_2) + \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)}f(x_3)$$

Substituting in the data values, we have

$$f_3(x) = \frac{(x-2)(x-4)(x-8)}{(1-2)(1-4)(1-8)}1 + \frac{(x-1)(x-4)(x-8)}{(2-1)(2-4)(2-8)}3 + \frac{(x-1)(x-2)(x-8)}{(4-1)(4-2)(4-8)}7 + \frac{(x-1)(x-2)(x-4)}{(8-1)(8-2)(8-4)}11$$

Now, if we substitute in the value at which we want the interpolated value (7) we obtain

$$\begin{aligned} f_3(7) &= \frac{(7-2)(7-4)(7-8)}{(1-2)(1-4)(1-8)}1+\frac{(7-1)(7-4)(7-8)}{(2-1)(2-4)(2-8)}3 \\ &+\frac{(7-1)(7-2)(7-8)}{(4-1)(4-2)(4-8)}7+\frac{(7-1)(7-2)(7-4)}{(8-1)(8-2)(8-4)}11 \\ &= 0.71429(1)+(-1.5)(3)+1.25(7)+0.53571(11) \\ &= 10.85710 \end{aligned}$$

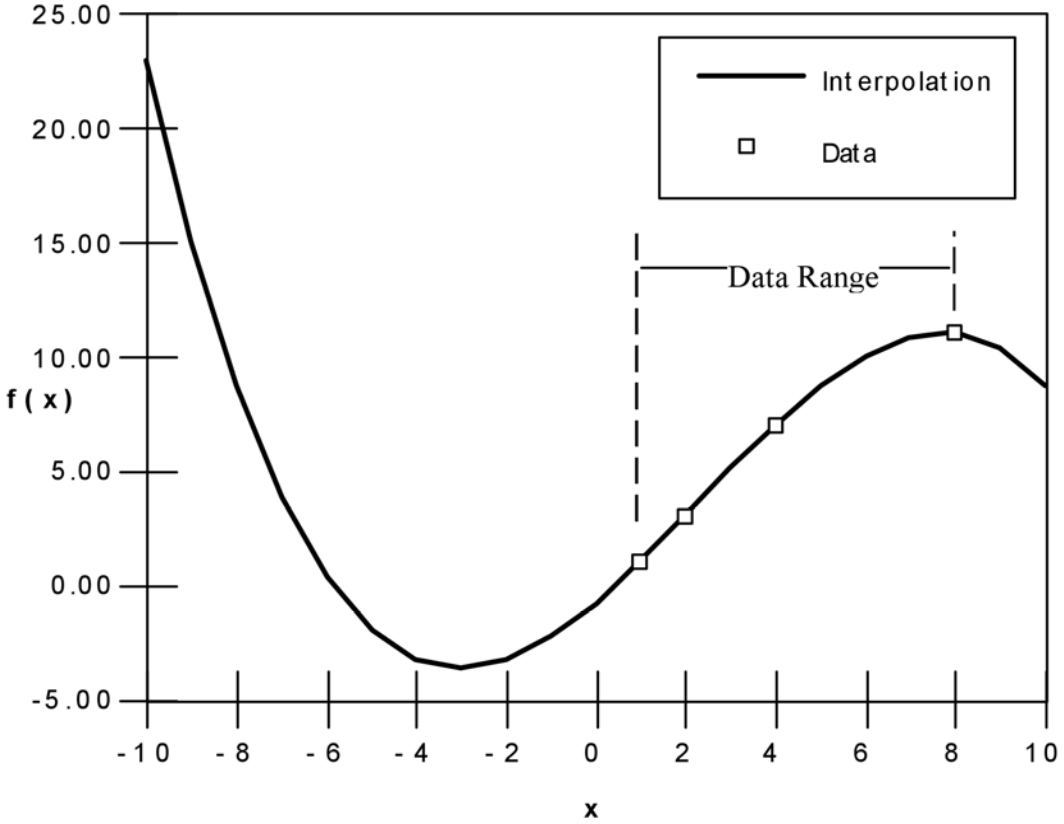


Figure. Plot of Lagrange interpolation polynomial.

Example. Use Lagrange interpolation to find *f(2.9)* using:

i	0	1	2	3	4
x	0	1	2	3.8	5
y=f(x)	0	0.569	0.791	0.224	-0.185

We have 5 data points, so use a fourth order Lagrange polynomial. The fourth order Lagrange polynomial is

$$f_4(x) = \sum_{i=0}^{4} L_i(x) f(x_i)$$

$$= L_0(x)f(x_0) + L_1(x)f(x_1) + L_2(x)f(x_2) + L_3(x)f(x_3) + L_4(x)f(x_4)$$

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)(x-x_4)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)(x_0-x_4)} = ?$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)(x-x_4)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)(x_1-x_4)} = \frac{4.9329}{-11.20} = -0.440438$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)(x-x_4)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)(x_2-x_4)} = \frac{10.4139}{10.8} = 0.96425$$

$$L_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_4)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)(x_3-x_4)} = \frac{10.4139}{22.9824} = 0.453125$$

$$L_4(x) = \frac{(x-x_0)(x-x_1)(x-x_2)(x-x_3)}{(x_4-x_0)(x_4-x_1)(x_4-x_2)(x_4-x_3)} = \frac{-4.4631}{72} = -0.061987$$

$$f_4(2.9) = 0.62508$$

Exercises

1. Given the following data (which were generated using a polynomial function):

X	1	2	3	5	6
F(X)	4.75	4	5.25	19.75	36

(a) Calculate *F*(4) using Lagrange interpolating polynomials of order 1 through 4.

(b) Plot your results using Excel.

(c) What do your results indicate regarding the order of the polynomial used to generate the data in the table.

2. Estimate the logarithm of 5 to the base 10 (log 5) using linear interpolation:

$$f(x) = f(x_1) + \frac{(x - x_1)}{(x_2 - x_1)}[f(x_2) - f(x_1)]$$

(a) Interpolate between log 4 = 0.60206 and log 6 = 0.7781513

(b) Interpolate between log 4.5 = 0.6532125 and log 5.5 = 0.7403627

3. Densities of sodium at three temperatures are given as follows:

	Temperature	Density
I	T_i (^{0}C)	ρ_i (kg/m^3)
0	94	929
1	205	902
2	371	860

a. Write the second-order Lagrange interpolation formula that fits these three data points.

b. Find the density for T = 251 ^{0}C by using the Lagrange interpolation formula from part (a).

4. The data describing the storage volume to surface area relationship for Toktogul Reservoir on the Naryn River in the Central Asian Kyrgyz Republic (see Climbing Magazine, No 199, December, 2000) for an interesting story about these mountains) is given in the following table. Find the second-order Lagrange interpolation polynomial which agrees with the data in the table. Use it to estimate the value of surface area when the storage volume is 11.0 km^3. Be sure to show your work.

StorageVolume	Surface Area
km^3	10^6 m^2
19.5	284
14.19	241.6
9.71	207.4
5.92	168.8
3	124.5

5. The density of sodium at three temperatures is given in the following table:

Temperature	Density
T_i (^{0}C)	ρ_i (kg/m^3)
94	929
205	902
371	860

Find the density for T = 251 ^{0}C by using a second-order Lagrange interpolation formula.

6. Write the third-order Lagrange interpolation polynomial using the values in the table:

X	0	2	3	4
F(x)	7	11	28	63