Least square and Curve fitting

7.1 Introduction

The experimental data $(x_1, y_1), (x_2, y_2),......, (x_n, y_n)$ are plotted on a rectangular coordinate system. Such a curve is known as an approximating curve that the data appears to be approximated by a straight line and it clearly exhibits a linear relationship between the two variables. *Curve fitting* is the general problem of finding equations of approximating curves which best fit the given set of data.

7.2 linear least square

We wish to predict response to n data points $(x_1, y_1), (x_2, y_2),......, (x_n, y_n)$ by a straight line given by

$$y = f(x) = a_0 + a_1 x, \tag{7.3}$$

where a_0 and a_1 are the constants of the least square straight line.

A measure of goodness of fit, that is, how well $a_0 + a_1 x$ predicts the response variable y is the magnitude of the residual ε_i at each of the n data points.

$$E_i = y_i - f(x_i) = y_i - (a_0 + a_1 x_i).$$

Ideally, if all the residuals ε_i are zero, one may have found an equation in which all the points lie on the straight line. Thus, minimization of the residual is an objective of obtaining coefficients.

The most popular method to minimize the residual is the least squares methods, where the estimates of the constants of the method are chosen such that the sum of the squared residuals is minimized, that is minimize $\sum_{i=1}^{n} E_i^2$.

To find a_0 and a_1, which minimize *S, i.e.* we want to minimize

$$S_r = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2. \tag{7.4}$$

where S_r is called the sum of the square of the residuals.

Differentiating Equation (7.4) with respect to a_0 and a_1, we get

$$\frac{\partial S_r}{\partial a_0} = 2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i)(-1) = 0, \tag{7.5}$$

$$\frac{\partial S_r}{\partial a_1} = 2\sum_{i=1}^{n}(y_i - a_0 - a_1 x_i)(-x_i) = 0, \tag{7.6}$$

giving

$$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = 0.$$

$$-\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^2 = 0.$$

Noting that

$$n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{7.7}$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{7.8}$$

Solving the above Equations (7.7) and (7.8) gives

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

Example 7.1: Find the least square line $f(x)=ax+b$ which fits the following data:

a)

x	-2	-1	0	1	2
y	1	2	3	3	4

7.3 Polynomial Models

Given n data points $(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_n, y_n)$ use least squares method to regress the data to an m^{th} order polynomial.

$$y = a_0 + a_1 x + a_2 x^2 + \cdots\cdots + a_m x^m, \; m < n.$$

The residual at each data point is given by

$$E_i = y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m .$$

The sum of the square of the residuals is given by

$$S_r = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m \right)^2 .$$

To find the constants of the polynomial regression model, we put the derivatives with respect to a_i to zero, that is,

$$\frac{\partial S_r}{\partial a_0} = \sum_{i=1}^{n} 2\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m \right)(-1) = 0,$$

$$\frac{\partial S_r}{\partial a_1} = \sum_{i=1}^{n} 2\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m \right)(-x_i) = 0,$$

$$\begin{array}{ccccccccccccccccccccc} . & & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \\ . & & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . & . \end{array}$$

$$\frac{\partial S_r}{\partial a_m} = \sum_{i=1}^{n} 2\left(y_i - a_0 - a_1 x_i - \ldots - a_m x_i^m \right)(-x_i^m) = 0.$$

Setting those equations in matrix form gives

$$\begin{bmatrix} n & \left(\sum_{i=1}^{n} x_i\right) & \ldots & \left(\sum_{i=1}^{n} x_i^m\right) \\ \left(\sum_{i=1}^{n} x_i\right) & \left(\sum_{i=1}^{n} x_i^2\right) & \ldots & \left(\sum_{i=1}^{n} x_i^{m+1}\right) \\ . \quad . \quad . & . \quad . \quad . & & . \quad . \quad . \quad . \quad . \\ \left(\sum_{i=1}^{n} x_i^m\right) & \left(\sum_{i=1}^{n} x_i^{m+1}\right) & \ldots & \left(\sum_{i=1}^{n} x_i^{2m}\right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \ldots \\ a_m \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ . \quad . \quad . \\ \sum_{i=1}^{n} x_i^m y_i \end{bmatrix}.$$

The above are solved for $a_0, a_1, \ldots, a_m$

Example 7.3: To find contraction of a steel cylinder, one needs to regress the thermal expansion coefficient data to temperature

S_1	80	40	-40	-120	-200	-280	-340
S_2	$6.47\times S_3$	$6.24\times S_3$	$5.72\times S_3$	$5.09\times S_3$	$4.30\times S_3$	$3.33\times S_3$	$2.45\times S_3$

where S_1=Temperature, T $({}^\circ\mathrm{F})$; S_2= Coefficient of thermal expansion, α (in/in/ $^\circ$F) and $S_3=10^{-6.}$

Fit the above data to $\alpha = a_0 + a_1 T + a_2 T^2$.

Solution: Since $\alpha = a_0 + a_1 T + a_2 T^2$. is the quadratic relationship between the thermal expansion coefficient and the temperature, the coefficients a_0, a_1, a_2 are found as follows

$$\begin{bmatrix} n & \left(\sum_{i=1}^{n} T_i\right) & \left(\sum_{i=1}^{n} T_i^2\right) \\ \left(\sum_{i=1}^{n} T_i\right) & \left(\sum_{i=1}^{n} T_i^2\right) & \left(\sum_{i=1}^{n} T_i^3\right) \\ \left(\sum_{i=1}^{n} T_i^2\right) & \left(\sum_{i=1}^{n} T_i^3\right) & \left(\sum_{i=1}^{n} T_i^4\right) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \alpha_i \\ \sum_{i=1}^{n} T_i \alpha_i \\ \sum_{i=1}^{n} T_i^2 \alpha_i \end{bmatrix}.$$

Summations for calculating constants of model given in the following table

i	$T\ (^\circ F)$	α (in/in/ $^\circ$F)	T^2	T^3
1	80	6.4700×10^{-6}	6.4000×10^{3}	5.1200×10^{5}
2	40	6.2400×10^{-6}	1.6000×10^{3}	6.4000×10^{4}
3	-40	5.7200×10^{-6}	1.6000×10^{3}	-6.4000×10^{4}
4	-120	5.0900×10^{-6}	1.4400×10^{4}	-1.7280×10^{6}
5	-200	4.3000×10^{-6}	4.0000×10^{4}	-8.0000×10^{6}
6	-280	3.3300×10^{-6}	7.8400×10^{4}	-2.1952×10^{7}
7	-340	2.4500×10^{-6}	1.1560×10^{5}	-3.9304×10^{7}
$\sum_{i=1}^{7}$	-8.6000×10^{2}	3.3600×10^{-5}	2.5800×10^{5}	-7.0472×10^{7}

i	T^4	$T \times \alpha$	$T^2 \times \alpha$
1	4.0960×10^{7}	5.1760×10^{-4}	4.1408×10^{-2}
2	2.5600×10^{6}	2.4960×10^{-4}	9.9840×10^{-3}
3	2.5600×10^{6}	-2.2880×10^{-4}	9.1520×10^{-3}
4	2.0736×10^{8}	-6.1080×10^{-4}	7.3296×10^{-2}
5	1.6000×10^{9}	-8.6000×10^{-4}	1.7200×10^{-1}
6	6.1466×10^{9}	-9.3240×10^{-4}	2.6107×10^{-1}
7	1.3363×10^{10}	-8.3300×10^{-4}	2.8322×10^{-1}
$\sum_{i=1}^{7}$	2.1363×10^{10}	-2.6978×10^{-3}	8.5013×10^{-1}

Since $n = 7$,

$$\sum_{i=1}^{7} T_i = -8.6000 \times 10^{-2}, \qquad \sum_{i=1}^{7} T_i^{\,2} = 2.5580 \times 10^{5},$$

$$\sum_{i=1}^{7} T_i^{\,3} = -7.0472 \times 10^{7}, \qquad \sum_{i=1}^{7} T_i^{\,4} = 2.1363 \times 10^{10},$$

$$\sum_{i=1}^{7} \alpha_i = 3.3600 \times 10^{-5}, \qquad \sum_{i=1}^{7} T_i \alpha_i = -2.6978 \times 10^{-3}$$

and $\sum_{i=1}^{7} T_i^{\,2} \alpha_i = 8.5013 \times 10^{-1}$.

We have

$$\begin{bmatrix} 7.0000 & -8.6000 \times 10^{2} & 2.5800 \times 10^{5} \\ -8.600 \times 10^{2} & 2.5800 \times 10^{5} & -7.0472 \times 10^{7} \\ 2.5800 \times 10^{5} & -7.0472 \times 10^{7} & 2.1363 \times 10^{10} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 3.3600 \times 10^{-5} \\ -2.6978 \times 10^{-3} \\ 8.5013 \times 10^{-1} \end{bmatrix}.$$

Solving the above system of simultaneous linear equations, we get

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 6.0217 \times 10^{-6} \\ 6.2782 \times 10^{-9} \\ -1.2218 \times 10^{-11} \end{bmatrix}.$$

The polynomial regression model is

$$\alpha = a_0 + a_1 T + a_2 T^2 = 6.0217 \times 10^{-6} + 6.2782 \times 10^{-9} T - 1.2218 \times 10^{-11} T^2.$$

7.4 Transforming the data to use linear least square formulas

Data for nonlinear models such as exponential, power, and growth can be transformed.

7.4.1 Exponential Model

As given in Example 7.1, many physical and chemical processes are governed by the exponential function.

$$\gamma = a e^{bx}. \tag{7.16}$$

Taking natural log of both sides of Equation (7.16) gives

$$\ln \gamma = \ln a + bx.$$

Let $z = \ln \gamma$, $a_0 = \ln a$ implying $a = e^{a_o}$, $a_1 = b$ then

$$z = a_0 + a_1 x.$$

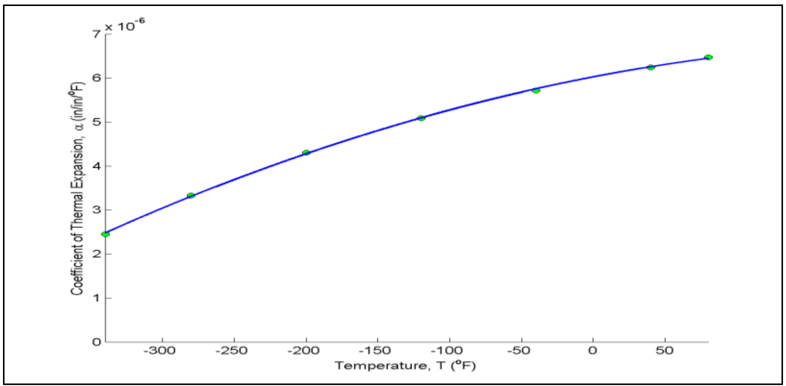

Figure 7.1 Second-order polynomial fitting for coefficient of thermal expansion as a function of temperature.

The data z versus x is now a linear model. The constants a_0 and a_1 can be found using the equation for the linear model as

$$a_1 = \frac{n\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}; \qquad a_0 = \bar{z} - a_1 \bar{x}.$$

Now since a_0 and a_1 are found, the original constants with the model are found as

$$b = a_1; \quad a = e^{a_0}.$$

7.4.2 Logarithmic Functions

The form for the log models is

$$y = \beta_0 + \beta_1 \ln(x).$$

This is a linear function between y and $\ln(x)$ the usual least squares method applies in which y is the response variable and $\ln(x)$ is the regressor.

7.4.3 Power Functions

The power function equation describes many scientific and engineering phenomena. In chemical engineering, the rate of chemical reaction is often written in power function form as

$$y = ax^b.$$

The method of least squares is applied to the power function by first linearizing the data (the assumption is that b is not known). If the only unknown is a, then a linear relation exists between x^b and y. The linearization of the data is as follows.

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

The resulting equation shows a linear relation between $\ln(y)$ and $\ln(x)$.

Let $z = \ln y$, $w = \ln(x)$, $a_0 = \ln a$ implying $a = e^{a_0}$, $a_1 = b$. We get

$$z = a_0 + a_1 w.$$

Hence

$$a_1 = \frac{n\sum_{i=1}^{n} w_i z_i - \sum_{i=1}^{n} w_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2 - \left(\sum_{i=1}^{n} w_i\right)^2}; \qquad a_0 = \frac{\sum_{i=1}^{n} z_i}{n} - a_1\frac{\sum_{i=1}^{n} w_i}{n}. \text{(from Section 7.2)}$$

Since a_a and a_1 can be found, the original constants of the model are

$$b = a_1; \qquad a = e^{a_0}.$$

7.4.4 Growth Model

An example of a growth model in which a measurable quantity y varies with some quantity x is

$$y = \frac{ax}{b + x}.$$

For $x = 0$, $y = 0$ while as $x \to \infty$, $y \to a$. To linearize the data for this method,

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

Let $z=\dfrac{1}{y}$, $w=\dfrac{1}{x}$, $a_0=\dfrac{1}{a}$ implying that $a=\dfrac{1}{a_0}$, $a_1=\dfrac{b}{a}$ implying $b=a_1\times a=\dfrac{a_1}{a_0}$,

then $z=a_0+a_1 w$.

The relationship between z and w is linear with the coefficients a_0 and found as follows:

$$a_1=\frac{n\sum_{i=1}^{n} w_i z_i-\sum_{i=1}^{n} w_i\sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} w_i^2-\left(\sum_{i=1}^{n} w_i\right)^2},\qquad a_0=\left(\frac{\sum_{i=1}^{n} z_i}{n}\right)-a_1\left(\frac{\sum_{i=1}^{n} w_i}{n}\right).$$

Finding a_0 and a_1, then gives the constants of the original growth model as

$$a=\frac{1}{a_0},\quad b=\frac{a_1}{a_0}.$$

EXERCISES

1. Find the normal equation of the curve $y=\dfrac{b}{x(x-a)}$.

2. Find the normal equation of $y=a+bxy$ find *a* and *b*, for the points (-4, 4), (1, 6), (2, 10), (3, 8).

3. Find the normal equations of the curve $y=(1+be^{ax})$ for the following data:

(0, 200), (1, 400), (2, 650), (3, 850) and (4, 950).

4. Find the normal equations of the curve $y=ax+bx^2$ for the following data, and find the maximum errors, (1.1, 5.3), (2, 14.2), (3.2, 30.1), (4, 43.8), (5.5, 77.3), (6.3, 97.6).

5. Drive the normal equations for fit the $y=ax+\dfrac{b}{\sqrt{x}}$, (0.2, 16), (0.3, 14), (0.5, 11), (1, 6), (2, 3), and find the least square error.