

CHAPTER 8

Curve Fitting, Regression, and Correlation

CURVE FITTING

Very often in practice a relationship is found to exist between two (or more) variables, and one wishes to express this relationship in mathematical form by determining an equation connecting the variables.

A first step is the collection of data showing corresponding values of the variables. For example, suppose x and y denote, respectively, the height and weight of an adult male. Then a sample of n individuals would reveal the heights x_1, x_2, \dots, x_n and the corresponding weights y_1, y_2, \dots, y_n .

A next step is to plot the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ on a rectangular coordinate system. The resulting set of points is sometimes called a *scatter diagram*.

From the scatter diagram it is often possible to visualize a smooth curve approximating the data. Such a curve is called an *approximating curve*. In Fig. 8-1, for example, the data appear to be approximated well by a straight line, and we say that a *linear relationship* exists between the variables. In Fig. 8-2, however, although a relationship exists between the variables, it is not a linear relationship and so we call it a *nonlinear relationship*. In Fig. 8-3 there appears to be no relationship between the variables.

The general problem of finding equations of approximating curves that fit given sets of data is called *curve fitting*. In practice the type of equation is often suggested from the scatter diagram. For Fig. 8-1 we could use a straight line

$$y = a + bx \quad (1)$$

while for Fig. 8-2 we could try a *parabola* or *quadratic curve*:

$$y = a + bx + cx^2 \quad (2)$$

Sometimes it helps to plot scatter diagrams in terms of *transformed variables*. For example, if $\log y$ vs. x leads to a straight line, we would try $\log y = a + bx$ as an equation for the approximating curve.

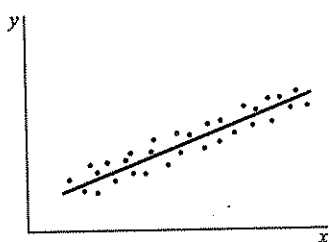


Fig. 8-1

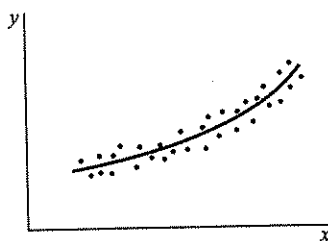


Fig. 8-2

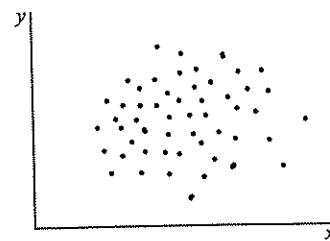


Fig. 8-3

REGRESSION

One of the main purposes of curve fitting is to estimate one of the variables (the *dependent variable*) from the other (the *independent variable*). The process of estimation is often referred to as *regression*. If y is to be estimated from x by means of some equation, we call the equation a *regression equation of y on x* and the corresponding curve a *regression curve of y on x* .

THE METHOD OF LEAST SQUARES

Generally, more than one curve of a given type will appear to fit a set of data. To avoid individual judgment in constructing lines, parabolas, or other approximating curves, it is necessary to agree on a definition of a "best-fitting line," "best-fitting parabola," etc.

To motivate a possible definition, consider Fig. 8-4 in which the data points are $(x_1, y_1), \dots, (x_n, y_n)$. For a given value of x , say, x_1 , there will be a difference between the value y_1 and the corresponding value as determined from the curve C . We denote this difference by d_1 , which is sometimes referred to as a *deviation*, *error*, or *residual* and may be positive, negative, or zero. Similarly, corresponding to the values x_2, \dots, x_n , we obtain the deviations d_2, \dots, d_n .

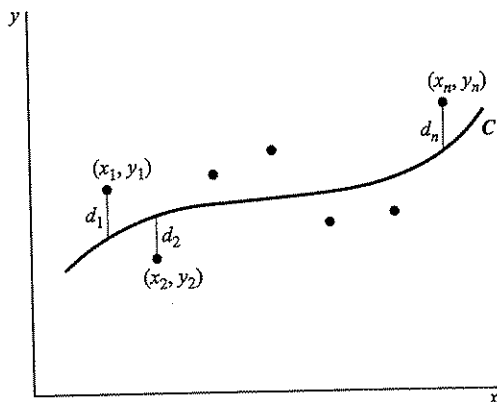


Fig. 8-4

A measure of the goodness of fit of the curve C to the set of data is provided by the quantity $d_1^2 + d_2^2 + \dots + d_n^2$. If this is small, the fit is good, if it is large, the fit is bad. We therefore make the following definition.

Definition. Of all curves in a given family of curves approximating a set of n data points, a curve having the property that

$$d_1^2 + d_2^2 + \dots + d_n^2 = \text{a minimum}$$

is called a *best-fitting curve* in the family.

A curve having this property is said to fit the data in the *least-squares sense* and is called a *least-squares regression curve*, or simply a *least-squares curve*. A line having this property is called a *least-squares line*; a parabola with this property is called a *least-squares parabola*, etc.

It is customary to employ the above definition when x is the independent variable and y is the dependent variable. If x is the dependent variable, the definition is modified by considering horizontal instead of vertical deviations, which amounts to interchanging the x and y axes. These two definitions lead in general to two different least-squares curves. Unless otherwise specified, we shall consider y as the dependent and x as the independent variable.

It is possible to define another least-squares curve by considering perpendicular distances from the data points to the curve instead of either vertical or horizontal distances. However, this is not used very often.

THE LEAST-SQUARES LINE

By using the above definition, we can show (see Problem 8.3) that the least-squares line approximating the set of points $(x_1, y_1), \dots, (x_n, y_n)$ has the equation

$$y = a + bx \quad (3)$$

where the constants a and b are determined by solving simultaneously the equations

$$\begin{aligned} \sum y &= an + b \sum x \\ \sum xy &= a \sum x + b \sum x^2 \end{aligned} \quad (4)$$

which are called the *normal equations* for the least-squares line. Note that we have for brevity used $\sum y$, $\sum xy$ instead of $\sum_{j=1}^n y_j$, $\sum_{j=1}^n x_j y_j$. The normal equations (4) are easily remembered by observing that the first equation can be obtained formally by summing on both sides of (3), while the second equation is obtained formally by first multiplying both sides of (3) by x and then summing. Of course, this is not a derivation of the normal equations but only a means for remembering them.

The values of a and b obtained from (4) are given by

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2} \quad b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \quad (5)$$

The result for b in (5) can also be written

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} \quad (6)$$

Here, as usual, a bar indicates *mean*, e.g., $\bar{x} = (\sum x)/n$. Division of both sides of the first normal equation in (4) by n yields

$$\bar{y} = a + b\bar{x} \quad (7)$$

If desired, we can first find b from (5) or (6) and then use (7) to find $a = \bar{y} - b\bar{x}$. This is equivalent to writing the least-squares line as

$$y - \bar{y} = b(x - \bar{x}) \quad \text{or} \quad y - \bar{y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} (x - \bar{x}) \quad (8)$$

The result (8) shows that the constant b , which is the *slope* of the line (3), is the fundamental constant in determining the line. From (8) it is also seen that the least-squares line passes through the point (\bar{x}, \bar{y}) , which is called the *centroid* or *center of gravity* of the data.

The slope b of the regression line is independent of the origin of coordinates. This means that if we make the transformation (often called a *translation of axes*) given by

$$x = x' + h \quad y = y' + k \quad (9)$$

8.3. Derive the normal equations (4), page 280, for the least-squares line.

Refer to Fig. 8-7. The values of y on the least-squares line corresponding to x_1, x_2, \dots, x_n are

$$a + bx_1, a + bx_2, \dots, a + bx_n$$

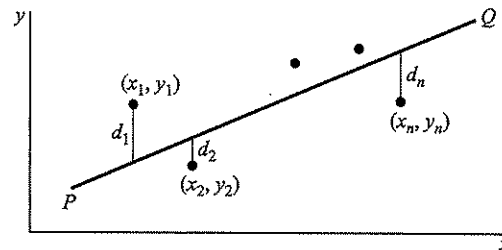


Fig. 8-7

The corresponding vertical deviations are

$$d_1 = a + bx_1 - y_1, \quad d_2 = a + bx_2 - y_2, \quad \dots, \quad d_n = a + bx_n - y_n$$

Then the sum of the squares of the deviations is

$$d_1^2 + d_2^2 + \dots + d_n^2 = (a + bx_1 - y_1)^2 + (a + bx_2 - y_2)^2 + \dots + (a + bx_n - y_n)^2$$

or

$$\sum d^2 = \sum (a + bx - y)^2$$

This is a function of a and b , i.e., $F(a, b) = \sum (a + bx - y)^2$. A necessary condition for this to be a minimum (or a maximum) is that $\partial F/\partial a = 0$, $\partial F/\partial b = 0$. Since

$$\frac{\partial F}{\partial a} = \sum \frac{\partial}{\partial a} (a + bx - y)^2 = \sum 2(a + bx - y)$$

$$\frac{\partial F}{\partial b} = \sum \frac{\partial}{\partial b} (a + bx - y)^2 = \sum 2x(a + bx - y)$$

we obtain

$$\sum (a + bx - y) = 0 \quad \sum x(a + bx - y) = 0$$

i.e.,

$$\sum y = an + b \sum x \quad \sum xy = a \sum x + b \sum x^2$$

as required. It can be shown that these actually yield a minimum.

8.4. Fit a least-squares line to the data of Problem 8.2 using (a) x as independent variable, (b) x as dependent variable. (See Table 8-2 for data)

(a) The equation of the line is $y = a + bx$. The normal equations are

$$\sum y = an + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

The work involved in computing the sums can be arranged as in Table 8-2. Although the last column is not needed for this part of the problem, it has been added to the table for use in part (b).

Since there are 8 pairs of values of x and y , $n = 8$ and the normal equations become

$$8a + 56b = 40$$

$$56a + 524b = 364$$

Solving simultaneously, $a = \frac{6}{11}$ or 0.545, $b = \frac{7}{11}$ or 0.636; and the required least-squares line is $y = \frac{6}{11} + \frac{7}{11}x$ or $y = 0.545 + 0.636x$. Note that this is not the line obtained in Problem 8.2 using the freehand method.

Table 8-2

x	y	x^2	xy	y^2
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
$\sum x = 56$	$\sum y = 40$	$\sum x^2 = 524$	$\sum xy = 364$	$\sum y^2 = 256$

Another method.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum x)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \text{ or } 0.545$$

$$b = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \text{ or } 0.636$$

- (b) If x is considered as the dependent variable and y as the independent variable, the equation of the least-squares line is $x = c + dy$ and the normal equations are

$$\sum x = cn + d \sum y$$

$$\sum xy = c \sum y + d \sum y^2$$

Then using Table 8-2, the normal equations become

$$8c + 40d = 56$$

$$40c + 256d = 364$$

from which $c = -\frac{1}{2}$ or -0.50 , $d = \frac{3}{2}$ or 1.50 .

These values can also be obtained from

$$c = \frac{(\sum x)(\sum y^2) - (\sum y)(\sum xy)}{n \sum y^2 - (\sum y)^2} = \frac{(56)(256) - (40)(364)}{(8)(256) - (40)^2} = -0.50$$

$$d = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum y^2 - (\sum y)^2} = \frac{(8)(364) - (56)(40)}{(8)(256) - (40)^2} = 1.50$$

Therefore, the required equation of the least-squares line is $x = -0.50 + 1.50y$.

Note that by solving this equation for y , we obtain $y = 0.333 + 0.667x$, which is not the same as the line obtained in part (a).

8.5. Graph the two lines obtained in Problem 8.4.

The graphs of the two lines, $y = 0.545 + 0.636x$ and $x = -0.500 + 1.50y$, are shown in Fig. 8-8. Note that the two lines in this case are practically coincident, which is an indication that the data are very well described by a linear relationship.

The line obtained in part (a) is often called the *regression line of y on x* and is used for estimating y for given values of x . The line obtained in part (b) is called the *regression line of x on y* and is used for estimating x for given values of y .

so that

$$\log P + 1.4 \log V = 4.2, \quad \log PV^{1.4} = 4.2, \quad \text{and} \quad PV^{1.4} = 16,000$$

THE LEAST-SQUARES PARABOLA

8.15. Derive the normal equations (19), page 282, for the least-squares parabola.

$$y = a + bx + cx^2$$

Let the sample points be $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Then the values of y on the least-squares parabola corresponding to x_1, x_2, \dots, x_n are

$$a + bx_1 + cx_1^2, \quad a + bx_2 + cx_2^2, \quad \dots, \quad a + bx_n + cx_n^2$$

Therefore, the deviations from y_1, y_2, \dots, y_n are given by

$$d_1 = a + bx_1 + cx_1^2 - y_1, \quad d_2 = a + bx_2 + cx_2^2 - y_2, \quad \dots, \quad d_n = a + bx_n + cx_n^2 - y_n$$

and the sum of the squares of the deviations is given by

$$\sum d^2 = \sum (a + bx + cx^2 - y)^2$$

This is a function of a, b , and c , i.e.,

$$F(a, b, c) = \sum (a + bx + cx^2 - y)^2$$

To minimize this function, we must have

$$\frac{\partial F}{\partial a} = 0, \quad \frac{\partial F}{\partial b} = 0, \quad \frac{\partial F}{\partial c} = 0$$

$$\begin{aligned} \text{Now} \quad \frac{\partial F}{\partial a} &= \sum \frac{\partial}{\partial a} (a + bx + cx^2 - y)^2 = \sum 2(a + bx + cx^2 - y) \\ \frac{\partial F}{\partial b} &= \sum \frac{\partial}{\partial b} (a + bx + cx^2 - y)^2 = \sum 2x(a + bx + cx^2 - y) \\ \frac{\partial F}{\partial c} &= \sum \frac{\partial}{\partial c} (a + bx + cx^2 - y)^2 = \sum 2x^2(a + bx + cx^2 - y) \end{aligned}$$

Simplifying each of these summations and setting them equal to zero yields the equations (19), page 282.

8.16. Fit a least-squares parabola having the form $y = a + bx + cx^2$ to the data in Table 8-8.

Table 8-8

x	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

Then normal equations are

$$\begin{aligned} \sum y &= an + b \sum x + c \sum x^2 \\ (1) \quad \sum xy &= a \sum x + b \sum x^2 + c \sum x^3 \\ \sum x^2 y &= a \sum x^2 + b \sum x^3 + c \sum x^4 \end{aligned}$$

The work involved in computing the sums can be arranged as in Table 8-9.

Table 8-9

x	y	x^2	x^3	x^4	xy	x^2y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1055.58	41.04	233.93
7.1	6.8	50.41	357.91	2541.16	48.28	342.79
8.6	4.5	73.96	636.06	5470.12	38.70	332.82
9.8	2.7	96.04	941.19	9223.66	26.46	259.31
$\sum x =$ 42.2	$\sum y =$ 46.4	$\sum x^2 =$ 291.20	$\sum x^3 =$ 2275.35	$\sum x^4 =$ 18,971.92	$\sum xy =$ 230.42	$\sum x^2y =$ 1449.00

Then the normal equations (1) become, since $n = 8$,

$$\begin{aligned}
 &8a + 42.2b + 291.20c = 46.4 \\
 (2) \quad &42.2a + 291.20b + 2275.35c = 230.42 \\
 &291.20a + 2275.35b + 18971.92c = 1449.00
 \end{aligned}$$

Solving, $a = 2.588$, $b = 2.065$, $c = -0.2110$; hence the required least-squares parabola has the equation

$$y = 2.588 + 2.065x - 0.2110x^2$$

8.17. Use the least-squares parabola of Problem 8.16 to estimate the values of y from the given values of x .

For $x = 1.2$, $y_{\text{est}} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. Similarly, other estimated values are obtained. The results are shown in Table 8-10 together with the actual values of y .

Table 8-10

y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

MULTIPLE REGRESSION

8.18. A variable z is to be estimated from variables x and y by means of a regression equation having the form $z = a + bx + cy$. Show that the least-squares regression equation is obtained by determining a , b , and c so that they satisfy (21), page 282.

Let the sample points be $(x_1, y_1, z_1), \dots, (x_n, y_n, z_n)$. Then the values of z on the least-squares regression plane corresponding to $(x_1, y_1), \dots, (x_n, y_n)$ are, respectively,

$$a + bx_1 + cy_1, \quad \dots, \quad a + bx_n + cy_n$$

Therefore, the deviations from z_1, \dots, z_n are given by

$$d_1 = a + bx_1 + cy_1 - z_1, \quad \dots, \quad d_n = a + bx_n + cy_n - z_n$$

and the sum of the squares of the deviations is given by

$$\sum d^2 = \sum (a + bx + cy - z)^2$$