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 $x_1, x_2, \ldots, x_n$ and the corresponding weights $y_1, y_2, \ldots, y_n$.

A next step is to plot the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ 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

$$y = a + bx \tag{1}$$

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

$$y = a + bx + cx^2 \tag{2}$$

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.

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*.

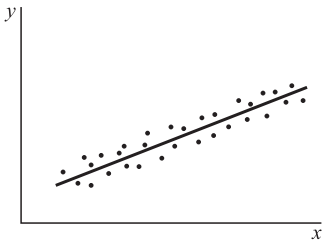

Fig. 8-1

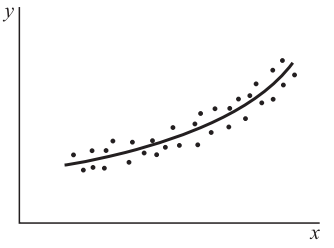

Fig. 8-2

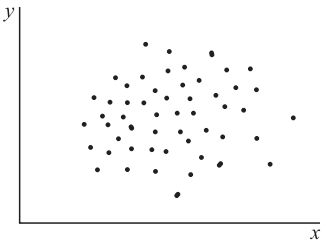

Fig. 8-3

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 $(x_1, y_1), \ldots, (x_n, y_n)$. 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 $x_2, \ldots, x_n$, we obtain the deviations $d_2, \ldots, d_n$.

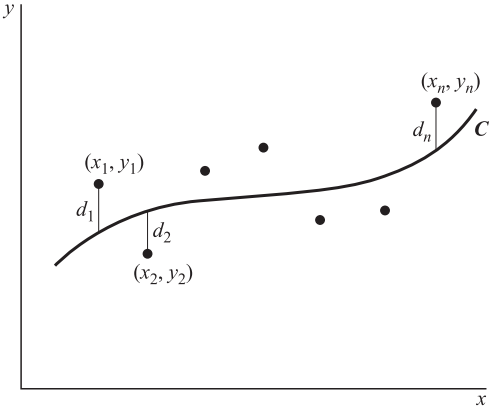

Fig. 8-4

A measure of the goodness of fit of the curve C to the set of data is provided by the quantity $d_1^2 + d_2^2 + \cdots + d_n^2$. 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

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

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 $(x_1, y_1), \ldots, (x_n, y_n)$ has the equation

$$y = a + bx \tag{3}$$

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

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

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

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2} \qquad b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{5}$$

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

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

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

$$\bar{y} = a + b\bar{x} \tag{7}$$

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

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

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

$$x = x' + h \qquad y = y' + k \tag{9}$$

where h and k are any constants, then b is also given by

$$b = \frac{n\sum x'y' - \left(\sum x'\right)\left(\sum y'\right)}{n\sum x'^2 - \left(\sum x'\right)^2} = \frac{\sum (x' - \overline{x'})(y' - \overline{y'})}{\sum (x' - \overline{x'})^2} \tag{10}$$

where x, y have simply been replaced by x', y' [for this reason we say that b is *invariant under the transformation* (9)]. It should be noted, however, that a, which determines the intercept on the x axis, does depend on the origin (and so is not invariant).

In the particular case where $h = \bar{x}$, $k = \bar{y}$, (10) simplifies to

$$b = \frac{\sum x'y'}{\sum x'^2} \tag{11}$$

The results (10) or (11) are often useful in simplifying the labor involved in obtaining the least-squares line.

The above remarks also hold for the regression line of x on y. The results are formally obtained by simply interchanging x and y. For example, the least-squares regression line of x on y is

$$x - \bar{x} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (y - \bar{y})^2}(y - \bar{y}) \tag{12}$$

It should be noted that in general (12) is not the same line as (8).

The Least-Squares Line in Terms of Sample Variances and Covariance

The sample variances and covariance of x and y are given by

$$s_x^2 = \frac{\sum (x - \bar{x})^2}{n}, \qquad s_y^2 = \frac{\sum (y - \bar{y})^2}{n}, \qquad s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n} \tag{13}$$

In terms of these, the least-squares regression lines of y on x and of x on y can be written, respectively, as

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x}) \qquad \text{and} \qquad x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y}) \tag{14}$$

if we formally define the *sample correlation coefficient* by [compare (54), page 82]

$$r = \frac{s_{xy}}{s_x s_y} \tag{15}$$

then (14) can be written

$$\frac{y - \bar{y}}{s_y} = r\left(\frac{x - \bar{x}}{s_x}\right) \qquad \text{and} \qquad \frac{x - \bar{x}}{s_x} = r\left(\frac{y - \bar{y}}{s_y}\right) \tag{16}$$

In view of the fact that $(x - \bar{x})/s_x$ and $(y - \bar{y})/s_y$ are standardized sample values or standard scores, the results in (16) provide a very simple way of remembering the regression lines. It is clear that the two lines in (16) are different unless $r = \pm 1$, in which case all sample points lie on a line [this will be shown in (26)] and there is *perfect linear correlation and regression*.

It is also of interest to note that if the two regression lines (16) are written as $y = a + bx$, $x = c + dy$, respectively, then

$$bd = r^2 \tag{17}$$

Up to now we have not considered the precise significance of the correlation coefficient but have only defined it formally in terms of the variances and covariance. On page 270, the significance will be given.

The Least-Squares Parabola

The above ideas are easily extended. For example, the *least-squares parabola* that fits a set of sample points is given by

$$y = a + bx + cx^2 \tag{18}$$

where a, b, c are determined from the *normal equations*

$$\begin{aligned} \sum y &= na + b\sum x + c\sum x^2 \\ \sum xy &= a\sum x + b\sum x^2 + c\sum x^3 \\ \sum x^2 y &= a\sum x^2 + b\sum x^3 + c\sum x^4 \end{aligned} \tag{19}$$

These are obtained formally by summing both sides of (18) after multiplying successively by 1, x and x^2, respectively.

Multiple Regression

The above ideas can also be generalized to more variables. For example, if we feel that there is a linear relationship between a dependent variable z and two independent variables x and y, then we would seek an equation connecting the variables that has the form

$$z = a + bx + cy \tag{20}$$

This is called a *regression equation of z on x and y*. If x is the dependent variable, a similar equation would be called a *regression equation of x on y and z*.

Because (20) represents a plane in a three-dimensional rectangular coordinate system, it is often called a *regression plane*. To find the least-squares regression plane, we determine a, b, c in (20) so that

$$\begin{aligned} \sum z &= na + b\sum x + c\sum y \\ \sum xz &= a\sum x + b\sum x^2 + c\sum xy \\ \sum yz &= a\sum y + b\sum xy + c\sum y^2 \end{aligned} \tag{21}$$

These equations, called the *normal equations* corresponding to (20), are obtained as a result of applying a definition similar to that on page 266. Note that they can be obtained formally from (20) on multiplying by 1, x, y, respectively, and summing.

Generalizations to more variables, involving linear or nonlinear equations leading to *regression surfaces* in three- or higher-dimensional spaces, are easily made.

Standard Error of Estimate

If we let y_{est} denote the estimated value of y for a given value of x, as obtained from the regression curve of y on x, then a measure of the scatter about the regression curve is supplied by the quantity

$$s_{y.x} = \sqrt{\frac{\sum (y - y_{\text{est}})^2}{n}} \tag{22}$$

which is called the *standard error of estimate of y on x*. Since $\sum (y - y_{\text{est}})^2 = \sum d^2$, as used in the Definition on page 266, we see that out of all possible regression curves the least-squares curve has the smallest standard error of estimate.

In the case of a regression line $y_{\text{est}} = a + bx$, with a and b given by (4), we have

$$s_{y.x}^2 = \frac{\sum y^2 - a\sum y - b\sum xy}{n} \tag{23}$$

or

$$s_{y.x}^2 = \frac{\sum (y - \bar{y})^2 - b\sum (x - \bar{x})(y - \bar{y})}{n} \tag{24}$$

We can also express $s_{y.x}^2$ for the least-squares line in terms of the variance and correlation coefficient as

$$s_{y.x}^2 = s_y^2(1 - r^2) \tag{25}$$

from which it incidentally follows as a corollary that $r^2 \leq 1$, i.e., $-1 \leq r \leq 1$.

The standard error of estimate has properties analogous to those of standard deviation. For example, if we construct pairs of lines parallel to the regression line of y on x at respective vertical distances $s_{y.x}$, and $2s_{y.x}$, and $3s_{y.x}$ from it, we should find if n is large enough that there would be included between these pairs of lines about 68%, 95%, and 99.7% of the sample points, respectively. See Problem 8.23.

Just as there is an unbiased estimate of population variance given by $\hat{s}^2 = ns^2/(n-1)$, so there is an unbiased estimate of the square of the standard error of estimate. This is given by $\hat{s}_{y.x}^2 = ns_{y.x}^2/(n-2)$. For this reason some statisticians prefer to give (22) with $n-2$ instead of n in the denominator.

The above remarks are easily modified for the regression line of x on y (in which case the standard error of estimate is denoted by $s_{x.y}$) or for nonlinear or multiple regression.

The Linear Correlation Coefficient

Up to now we have defined the correlation coefficient formally by (15) but have not examined its significance. In attempting to do this, let us note that from (25) and the definitions of $s_{y.x}$ and s_y, we have

$$r^2 = 1 - \frac{\sum(y - y_{\text{est}})^2}{\sum(y - \bar{y})^2} \tag{26}$$

Now we can show that (see Problem 8.24)

$$\sum(y - \bar{y})^2 = \sum(y - y_{\text{est}})^2 + \sum(y_{\text{est}} - \bar{y})^2 \tag{27}$$

The quantity on the left of (27) is called the *total variation*. The first sum on the right of (27) is then called the *unexplained variation*, while the second sum is called the *explained variation*. This terminology arises because the deviations $y - y_{\text{est}}$ behave in a random or unpredictable manner while the deviations $y_{\text{est}} - \bar{y}$ are explained by the least-squares regression line and so tend to follow a definite pattern. It follows from (26) and (27) that

$$r^2 = \frac{\sum(y_{\text{est}} - \bar{y})^2}{\sum(y - \bar{y})^2} = \frac{\text{explained variation}}{\text{total variation}} \tag{28}$$

Therefore, r^2 can be interpreted as the fraction of the total variation that is explained by the least-squares regression line. In other words, r measures *how well* the least-squares regression line fits the sample data. If the total variation is *all* explained by the regression line, i.e., if $r^2 = 1$ or $r = \pm 1$, we say that there is *perfect linear correlation* (and in such case also *perfect linear regression*). On the other hand, if the total variation is all unexplained, then the explained variation is zero and so $r = 0$. In practice the quantity r^2, sometimes called the *coefficient of determination*, lies between 0 and 1.

The correlation coefficient can be computed from either of the results

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2}\sqrt{\sum(y - \bar{y})^2}} \tag{29}$$

or

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum(y_{\text{est}} - \bar{y})^2}{\sum(y - \bar{y})^2} \tag{30}$$

which for linear regression are equivalent. The formula (29) is often referred to as the *product-moment formula* for linear correlation.

Formulas equivalent to those above, which are often used in practice, are

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2 - \left(\sum x\right)^2\right]\left[n\sum y^2 - \left(\sum y\right)^2\right]}} \tag{31}$$

and

$$r = \frac{\overline{xy} - \bar{x}\bar{y}}{\sqrt{(\bar{x}^2 - \bar{x}^2)(\bar{y}^2 - \bar{y}^2)}} \tag{32}$$

If we use the transformation (9), page 267, we find

$$r = \frac{n\sum x'y' - \left(\sum x'\right)\left(\sum y'\right)}{\sqrt{\left[n\sum x'^2 - \left(\sum x'\right)^2\right]\left[n\sum y'^2 - \left(\sum y'\right)^2\right]}} \tag{33}$$

which shows that r is invariant under a translation of axes. In particular, if $h = \bar{x}$, $k = \bar{y}$, (33) becomes

$$r = \frac{\sum x'y'}{\sqrt{\left(\sum x'^2\right)\left(\sum y'^2\right)}} \tag{34}$$

which is often useful in computation.

The linear correlation coefficient may be positive or negative. If r is positive, y tends to *increase* with x (the slope of the least-squares line is positive) while if r is negative, y tends to *decrease* with x (the slope is negative). The sign is *automatically* taken into account if we use the result (29), (31), (32), (33), or (34). However, if we use (30) to obtain r, we must apply the proper sign.

Generalized Correlation Coefficient

The definition (29) [or any of the equivalent forms (31) through (34)] for the correlation coefficient involves only sample values x, y. Consequently, it yields the same number for all forms of regression curves and is useless as a measure of fit, except in the case of linear regression, where it happens to coincide with (30). However, the latter definition, i.e.,

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum(y_{\text{est}} - \bar{y})^2}{\sum(y - \bar{y})^2} \tag{35}$$

does reflect the form of the regression curve (via the y_{est}) and so is suitable as the definition of a *generalized correlation coefficient* r. We use (35) to obtain nonlinear correlation coefficients (which measure how well a *nonlinear regression curve* fits the data) or, by appropriate generalization, *multiple correlation coefficients*. The connection (25) between the correlation coefficient and the standard error of estimate holds as well for nonlinear correlation.

Since a correlation coefficient merely measures how well a given regression curve (or surface) fits sample data, it is clearly senseless to use a linear correlation coefficient where the data are nonlinear. Suppose, however, that one does apply (29) to nonlinear data and obtains a value that is numerically considerably less than 1. Then the conclusion to be drawn is not that there is *little correlation* (a conclusion sometimes reached by those unfamiliar with the fundamentals of correlation theory) but that there is *little linear* correlation. There may in fact be a *large* nonlinear correlation.

Rank Correlation

Instead of using precise sample values, or when precision is unattainable, the data may be ranked in order of size, importance, etc., using the numbers $1, 2, \ldots, n$. If two corresponding sets of values x and y are ranked in such manner, the *coefficient of rank correlation*, denoted by r_{rank}, or briefly r, is given by (see Problem 8.36)

$$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} \tag{36}$$

where d = differences between ranks of corresponding x and y

n = number of pairs of values (x, y) in the data

The quantity r_{rank} in (36) is known as *Spearman's rank correlation coefficient*.

SOLVED PROBLEMS

The least-squares line

8.1. A straight line passes through the points (x_1, y_1) and (x_2, y_2). Show that the equation of the line is

$$y - y_1 = \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x - x_1)$$

The equation of a line is $y = a + bx$. Then since (x_1, y_1) and (x_2, y_2) are points on the line, we have

$$y_1 = a + bx_1, \qquad y_2 = a + bx_2$$

Therefore,

(1) $$y - y_1 = (a + bx) - (a + bx_1) = b(x - x_1)$$

(2) $$y_2 - y_1 = (a + bx_2) - (a + bx_1) = b(x_2 - x_1)$$

Obtaining $b = (y_2 - y_1)/(x_2 - x_1)$ from (2) and substituting in (1), the required result follows.

The graph of the line PQ is shown in Fig. 8-5. The constant $b = (y_2 - y_1)/(x_2 - x_1)$ is the slope of the line.

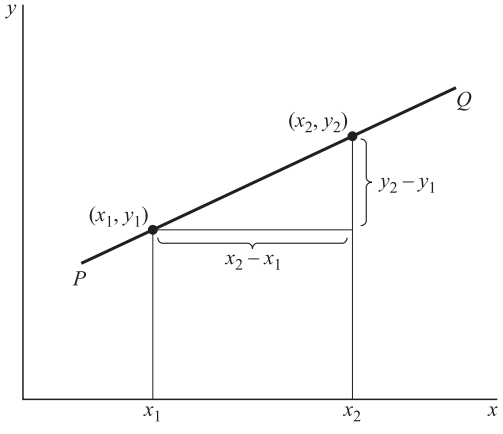

Fig. 8-5

8.2. (a) Construct a straight line that approximates the data of Table 8-1. (b) Find an equation for this line.

Table 8-1

x	1	3	4	6	8	9	11	14
y	1	2	4	4	5	7	8	9

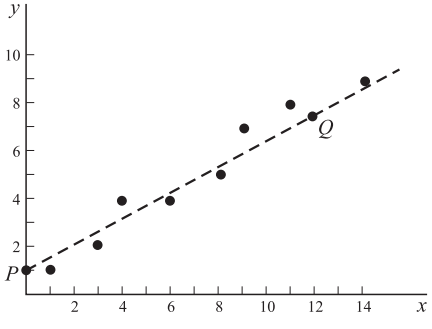

Fig. 8-6

(a) Plot the points (1, 1), (3, 2), (4, 4), (6, 4), (8, 5), (9, 7), (11, 8), and (14, 9) on a rectangular coordinate system as shown in Fig. 8-6.

A straight line approximating the data is drawn *freehand* in the figure. For a method eliminating the need for individual judgment, see Problem 8.4, which uses the method of least squares.

(b) To obtain the equation of the line constructed in (a), choose any two points on the line, such as P and Q. The coordinates of these points as read from the graph are approximately (0, 1) and (12, 7.5). Then from Problem 8.1,

$$y - 1 = \frac{7.5 - 1}{12 - 0}(x - 0)$$

or $y - 1 = 0.542x$ or $y = 1 + 0.542x$.

8.3. Derive the normal equations (4), page 267, for the least-squares line.

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

$$a + bx_1, \qquad a + bx_2, \qquad \ldots, \qquad a + bx_n$$

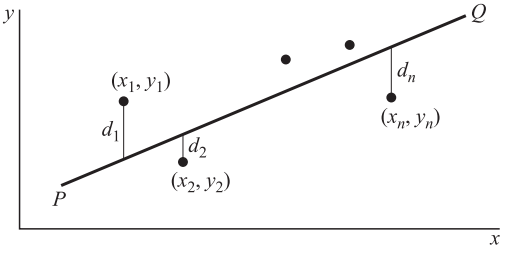

Fig. 8-7

The corresponding vertical deviations are

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

Then the sum of the squares of the deviations is

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

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

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

i.e.,

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

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.

(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

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

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

(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

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

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

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.

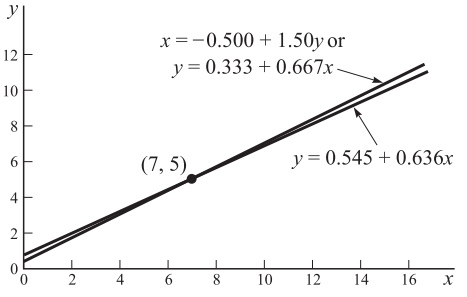

Fig. 8-8

8.6. (a) Show that the two least-squares lines obtained in Problem 8.4 intersect at point $(\bar{x}, \bar{y})$. (b) Estimate the value of y when $x = 12$. (c) Estimate the value of x when $y = 3$.

$$\bar{x} = \frac{\sum x}{n} = \frac{56}{8} = 7, \qquad \bar{y} = \frac{\sum y}{n} = \frac{40}{8} = 5$$

Then point $(\bar{x}, \bar{y})$, called the *centroid*, is (7, 5).

(a) Point (7, 5) lies on line $y = 0.545 + 0.636x$ or, more exactly, $y = \frac{6}{11} + \frac{7}{11}x$, since $5 = \frac{6}{11} + \frac{7}{11}(7)$. Point (7, 5) lies on line $x = -\frac{1}{2} + \frac{3}{2}y$, since $7 = -\frac{1}{2} + \frac{3}{2}(5)$.

Another method

The equations of the two lines are $y = \frac{6}{11} + \frac{7}{11}x$ and $x = -\frac{1}{2} + \frac{3}{2}y$. Solving simultaneously, we find $x = 7$, $y = 5$. Therefore, the lines intersect in point (7, 5).

(b) Putting $x = 12$ into the regression line of y on x, $y = 0.545 + 0.636(12) = 8.2$.

(c) Putting $y = 3$ into the regression line of x on y, $x = -0.50 + 1.50(3) = 4.0$.

8.7. Prove that a least-squares line always passes through the point $(\bar{x}, \bar{y})$.

Case 1

x is the independent variable.

The equation of the least-squares line is	(1)	$y = a + bx$
A normal equation for the least-squares line is	(2)	$\sum y = an + b\sum x$
Dividing both sides of (2) by n gives	(3)	$\bar{y} = a + b\bar{x}$

Subtracting (3) from (1), the least-squares line can be written

$$y - \bar{y} = b(x - \bar{x}) \tag{4}$$

which shows that the line passes through the point $(\bar{x}, \bar{y})$.

Case 2

y is the independent variable.

Proceeding as in Case 1 with x and y interchanged and the constants a, b, replaced by c, d, respectively, we find that the least-squares line can be written

$$x - \bar{x} = d(y - \bar{y}) \tag{5}$$

which indicates that the line passes through the point $(\bar{x}, \bar{y})$.

Note that, in general, lines (4) and (5) are not coincident, but they intersect in $(\bar{x}, \bar{y})$.

8.8. Prove that the least-squares regression line of y on x can be written in the form (8), page 267.

We have from (4) of Problem 8.7, $y - \bar{y} = b(x - \bar{x})$. From the second equation in (5), page 267, we have

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \tag{1}$$

Now
$$\begin{aligned}\sum(x - \bar{x})^2 &= \sum(x^2 - 2\bar{x}x + \bar{x}^2)\\ &= \sum x^2 - 2\bar{x}\sum x + \sum \bar{x}^2\\ &= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2\\ &= \sum x^2 - n\bar{x}^2\\ &= \sum x^2 - \frac{1}{n}\left(\sum x\right)^2\\ &= \frac{1}{n}\left[n\sum x^2 - \left(\sum x\right)^2\right]\end{aligned}$$

Also
$$\begin{aligned}\sum(x - \bar{x})(y - \bar{y}) &= \sum(xy - \bar{x}y - \bar{y}x + \bar{x}\bar{y})\\ &= \sum xy - \bar{x}\sum y - \bar{y}\sum x + \sum \bar{x}\bar{y}\\ &= \sum xy - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}\\ &= \sum xy - n\bar{x}\bar{y}\\ &= \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}\\ &= \frac{1}{n}\left[n\sum xy - \left(\sum x\right)\left(\sum y\right)\right]\end{aligned}$$

Therefore, (1) becomes

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

from which the result (8) is obtained. Proof of (12), page 268, follows on interchanging x and y.

8.9. Let $x = x' + h$, $y = y' + k$, where h and k are any constants. Prove that

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{n\sum x'y' - \left(\sum x'\right)\left(\sum y'\right)}{n\sum x'^2 - \left(\sum x'\right)^2}$$

From Problem 8.8 we have

$$b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

Now if $x = x' + h$, $y = y' + k$, we have

$$\bar{x} = \bar{x}' + h, \quad \bar{y} = \bar{x}' + k$$

Thus

$$\frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = \frac{\sum(x' - \overline{x'})(y' - \overline{y'})}{\sum(x' - \overline{x'})^2}$$

$$= \frac{n\sum x'y' - \left(\sum x'\right)\left(\sum y'\right)}{n\sum x'^2 - \left(\sum x'\right)^2}$$

The result is useful in developing a shortcut for obtaining least-squares lines by subtracting suitable constants from the given values of x and y (see Problem 8.12).

8.10. If, in particular, $h = \bar{x}$, $k = \bar{y}$ in Problem 8.9, show that

$$b = \frac{\sum x'y'}{\sum x'^2}$$

This follows at once from Problem 8.9 since

$$\sum x' = \sum(x - \bar{x}) = \sum x - n\bar{x} = 0$$

and similarly $\sum y' = 0$.

8.11. Table 8-3 shows the respective heights x and y of a sample of 12 fathers and their oldest sons. (a) Construct a scatter diagram. (b) Find the least-squares regression line of y on x. (c) Find the least-squares regression line of x on y.

Table 8-3

Height x of Father (inches)	65	63	67	64	68	62	70	66	68	67	69	71
Height y of Son (inches)	68	66	68	65	69	66	68	65	71	67	68	70

(a) The scatter diagram is obtained by plotting the points (x, y) on a rectangular coordinate system as shown in Fig. 8-9.

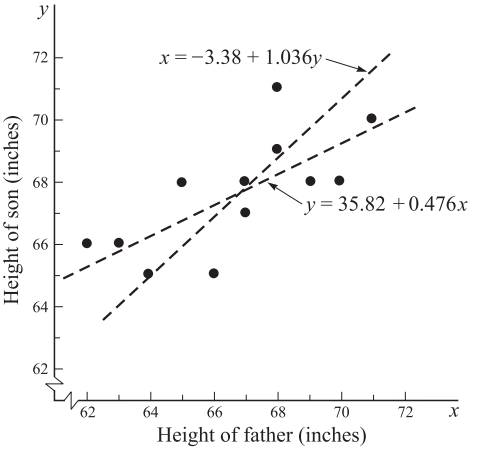

Fig. 8-9

(b) The regression line of y on x is given by $y = a + bx$, where a and b are obtained by solving the normal equations

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

The sums are shown in Table 8-4, and so the normal equations become

$$12a + 800b = 811$$
$$800a + 53{,}418b = 54{,}107$$

from which we find $a = 35.82$ and $b = 0.476$, so that $y = 35.82 + 0.476x$. The graph of this equation is shown in Fig. 8-9.

Another method

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\sum x^2 - \left(\sum x\right)^2} = 35.82, \qquad b = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} = 0.476$$

Table 8-4

x	y	x^2	xy	y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum x = 800$	$\sum y = 811$	$\sum x^2 = 53{,}418$	$\sum = 54{,}107$	$\sum y^2 = 54{,}849$

(c) The regression line of x on y is given by $x = c + dy$, where c and d are obtained by solving the normal equations

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

Using the sums in Table 8-4, these become

$$12c + 811d = 800$$
$$811c + 54{,}849d = 54{,}107$$

from which we find $c = -3.38$ and $d = 1.036$, so that $x = -3.38 + 1.036y$. The graph of this equation is shown in Fig. 8-9.

Another method

$$c = \frac{\left(\sum x\right)\left(\sum y^2\right) - \left(\sum y\right)\left(\sum xy\right)}{n\sum y^2 - \left(\sum y\right)^2} = -3.38, \qquad d = \frac{n\sum xy - \left(\sum y\right)\left(\sum x\right)}{n\sum y^2 - \left(\sum y\right)^2} = 1.036$$

Then

(1) $$r_{\text{rank}} = \frac{s_x^2 + s_y^2 - s_d^2}{2s_x s_y}$$

Since $\bar{d} = 0, s_d^2 = (\sum d^2)/n$ and (1) becomes

(2) $$r_{\text{rank}} = \frac{(n^2 - 1)/12 + (n^2 - 1)/12 - \left(\sum d^2\right)/n}{(n^2 - 1)/6} = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

8.37. Table 8-17 shows how 10 students were ranked according to their achievements in both the laboratory and lecture portions of a biology course. Find the coefficient of rank correlation.

Table 8-17

Laboratory	8	3	9	2	7	10	4	6	1	5
Lecture	9	5	10	1	8	7	3	4	2	6

The difference of ranks d in laboratory and lecture for each student is given in Table 8-18. Also given in the table are d^2 and $\sum d^2$.

Table 8-18

Difference of Ranks (d)	−1	−2	−1	1	−1	3	1	2	−1	−1	
d^2	1	4	1	1	1	9	1	4	1	1	$\sum d^2 = 24$

Then $$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(24)}{10(10^2 - 1)} = 0.8545$$

indicating that there is a marked relationship between achievements in laboratory and lecture.

8.38. Calculate the coefficient of rank correlation for the data of Problem 8.11, and compare your result with the correlation coefficient obtained by other methods.

Arranged in ascending order of magnitude, the fathers' heights are

(1) 62, 63, 64, 65, 66, 67, 67, 68, 68, 69, 70, 71

Since the 6th and 7th places in this array represent the same height (67 inches), we assign a *mean rank* 6.5 to both these places. Similarly, the 8th and 9th places are assigned the rank 8.5. Therefore, the fathers' heights are assigned the ranks

(2) 1, 2, 3, 4, 5, 6.5, 6.5, 8.5, 8.5, 10, 11, 12

Similarly, the sons' heights arranged in ascending order of magnitude are

(3) 65, 65, 66, 66, 67, 68, 68, 68, 68, 69, 70, 71

and since the 6th, 7th, 8th, and 9th places represent the same height (68 inches), we assign the *mean rank* 7.5 $(6 + 7 + 8 + 9)/4$ to these places. Therefore, the sons' heights are assigned the ranks

(4) 1.5, 1.5, 3.5, 3.5, 5, 7.5, 7.5, 7.5, 7.5, 10, 11, 12

Using the correspondences (1) and (2), (3) and (4), Table 8-3 becomes Table 8-19.

Table 8-19

Rank of Father	4	2	6.5	3	8.5	1	11	5	8.5	6.5	10	12
Rank of Son	7.5	3.5	7.5	1.5	10	3.5	7.5	1.5	12	5	7.5	11

The differences in ranks d, and the computations of d^2 and $\sum d^2$ are shown in Table 8-20.

Table 8-20

d	-3.5	-1.5	-1.0	1.5	-1.5	-2.5	3.5	3.5	-3.5	1.5	2.5	1.0	
d^2	12.25	2.25	1.00	2.25	2.25	6.25	12.25	12.25	12.25	2.25	6.25	1.00	$\sum d^2 = 72.50$

Then
$$r_{\text{rank}} = 1 - \frac{6\sum d^2}{n(n^2 - 1)} = 1 - \frac{6(72.50)}{12(12^2 - 1)} = 0.7465$$

which agrees well with the value $r = 0.7027$ obtained in Problem 8.26(b).