Curve Fitting and
The Method of Least Squares

RELATIONSHIP between VARIABLES

Very often in practice a relationship is found to exist between two (or more) variables. For example: weights of adult males depend to some degree on their heights; circumferences of circles depend on their radii; and the pressure of a given mass of gas depends on its temperature and volume.

It is frequently desirable to express this relationship in mathematical form by determining an equation connecting the variables.

CURVE FITTING

To aid in determining an equation connecting variables, a first step is the collection of data showing corresponding values of the variables under consideration.

For example, suppose X and Y denote respectively the height and weight of adult males. 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. In Fig. 1, for example, the data appears to be approximated well by a straight line and we say that a linear relationship exists between the variables. In Fig. 2, however, although a relationship exists between the variables it is not a *linear relationship* and so we call it a *non-linear* relationship.

The general problem of finding equations of approximating curves which fit given sets of data is called *curve fitting*.

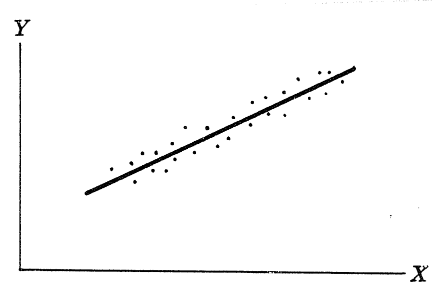


Fig. 1

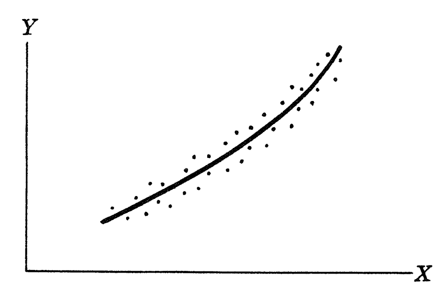


Fig. 2

EQUATIONS of APPROXIMATING CURVES

For purposes of reference we have listed below several common types of approximating curves and their equations. All letters other than X and Y represent constants. The variables X and Y are often referred to as *independent* and *dependent variables* respectively, although these roles can be interchanged.

$Y = mX + b$	Straight line
$Y = m_1 X + m_2 X^2 + b$	Parabola or Quadratic curve

$Y = m_1 X + m_2 X^2 + m_3 X^3 + b$ Cubic curve

$Y = m_1 X + m_2 X^2 + ... + m_n X^n + b$ *n*th degree curve

The right sides of the above equations are called polynomials of the first, second, third, fourth and nth degrees respectively. The functions defined by the first four of these equations are sometimes called *linear, quadratic, cubic* and *quartic* functions respectively.

As other possible equations (among many) used in practice we mention the following.

$Y = \frac{1}{mX + b}$ or $\frac{1}{Y} = mX + b$ Hyperbola

$Y = \alpha\beta^X$ or $\log Y = \log\alpha + X\log\beta = b + mX$ Exponential curve
(where $b = \log\alpha$ and $m = \log\beta$)

$Y = \alpha X^\beta$ or $\log Y = \log\alpha + \beta\log X$ Geometric curve

$Y = \alpha\beta^X + g$ Modified exponential curve

$Y = \alpha X^\beta + g$ Modified geometric curve

$Y = pq^{\beta^X}$ or $\log Y = \log p + \beta^X \log q = \alpha\beta^X + g$ Gompertz curve

$Y = pq^{\beta^X} + h$ Modified Gompertz curve

$Y = \frac{1}{\alpha\beta^X + g}$ or $\frac{1}{Y} = \alpha\beta^X + g$ Logistic curve

$Y = m_1(\log X) + m_2(\log X)^2 + b$ Logrithmic curve

To decide which curve should be used, it is helpful to obtain scatter diagrams of transformed variables. For example, if a scatter diagram of log Y vs. X shows a linear relationship the equation has the form $Y = \alpha\beta^X$, while if log Y vs. log X shows a linear relationship the equation has the form $Y = \alpha X^\beta$. To facilitate this we frequently employ special graph paper for which one or both scales are calibrated logarithmically. These are referred to as *semi-log* or *log-log graph paper* respectively.

THE STRAIGHT LINE

The simplest type of approximating curve is a straight line, whose equation can be written $Y = mX + b$. Given any two points (X_1, Y_1) and (X_2, Y_2) on the line, the constants m and b can be determined. The resulting equation of the line can be written

$$Y - Y_1 = \left(\frac{Y_2 - Y_1}{X_2 - X_1}\right)(X - X_1) \qquad \text{or} \qquad Y - Y_1 = m(X - X_1)$$

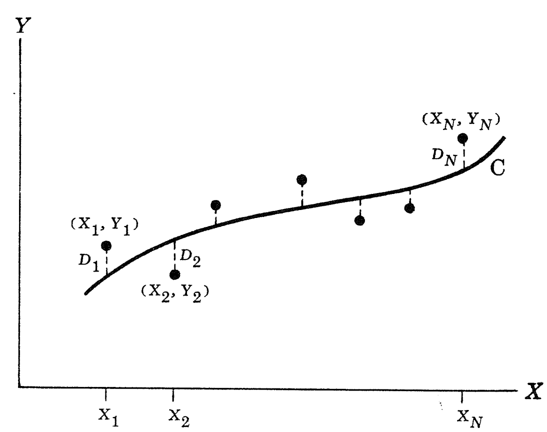
Where $m = \left(\dfrac{Y_2 - Y_1}{X_2 - X_1}\right)$ is called the slope of the line and represents the change in Y divided by the corresponding change in X.

The constant b, which is the value of Y when $X = 0$, is called the *Y intercept.*

THE METHOD of LEAST SQUARES

To avoid individual judgment in constructing lines, parabolas or other approximating curves to fit sets of data, it is necessary to agree on a definition of a "best fitting line", "best fitting parabola", etc.

To motivate a possible definition, consider Fig. 3 in which the data points are given by (X_1, Y_1,), (X_2, Y_2),..., (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. As indicated in the figure 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,..., X_n we obtain the deviations D_2,..., D_n.



~~Fig. 13.3~~ Fig. 3

A measure of the "goodness of fit" of the curve C to the given data is provided by the quantity $D_1^2 + D_2^2 + ... + D_n^2$. If this is small the fit is good, if it is large the fit is bad. We therefore make the following

Definition. Of all curves approximating a given set of data points, the curve having the property that minimizes the *square* of the distance between two points, i.e., where

$$D_1^2 + D_2^2 + ... + D_n^2 \qquad \text{is a minimum}$$

is called a *best fitting curve.*

A curve having this property is said to fit the data in the *least square sense* and is called a *least square curve.* Thus a line having this property is called a *least square line.*

One may also formulate a line which minimizes the distance between two points, i.e., where

$$| D_1 | + | D_2 | + ... + | D_n | \text{ is a minimum}$$

A line having this property is said to fit the data in a *least absolute deviation sense*, and is called a *least deviations* line. This is the type of line used by Dr. Bill Gray in his seasonal hurricane forecasts. One may also formulate other combinations which minimize the distance between two points, i.e., a *least square parabola*, etc.

THE LEAST SQUARE LINE

The least square line approximating the set of points (X_1, Y_1), (X_2, Y_2),..., (X_n, Y_n) has the equation

$$Y = mX + b$$

where the constants *b* and *m* are determined by solving simultaneously the equations

$$\sum Y = bn + m\sum X$$

$$\sum XY = b\sum X + m\sum X^2$$

These equations are derived by applying calculus and finding the equations which minimize the least squares distance. The derivation may be found in most statistics and calculus books.

One may solve for *b* and *m* by using algebra on the two above equations:

$$b = \frac{\left(\sum Y\right)\left(\sum X^2\right) - \left(\sum X\right)\left(\sum XY\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$m = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

Example

Fit a least square line to the following data using *x* as the independent variable

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Answer

The work involves the summation of terms as shown below:

X	Y	X^2	XY
1	1	1	1
3	2	9	6
4	4	16	16
6	4	36	24
8	5	64	40
9	7	81	63
11	8	121	88
14	9	196	126
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$

One method (using algebra)

Since there are 8 values of *X* and *Y*, *n=8*, the equations become

$$40 = 8b + 56m$$
$$364 = 56b + 524m$$

which then may be solved algebraically for *b* and *m*, which gives

$$b = \frac{6}{11} \text{ or } .545$$

$$m = \frac{7}{11} \text{ or } .636$$

The least squares equation is:

$$Y = \frac{7}{11}X + \frac{6}{11}$$

or

$$Y = .636X + .545$$

Second method (used in spreadsheets and computer programs)

Solve the following expressions

$$b = \frac{\left(\sum Y\right)\left(\sum X^2\right) - \left(\sum X\right)\left(\sum XY\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{(40)(524) - (56)(364)}{(8)(524) - (56)^2} = \frac{6}{11} \text{ or } .545$$

$$m = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2} = \frac{(8)(364) - (56)(40)}{(8)(524) - (56)^2} = \frac{7}{11} \text{ or } .636$$

which gives

$$Y = .636X + .545$$

as before

CONCLUDING REMARKS

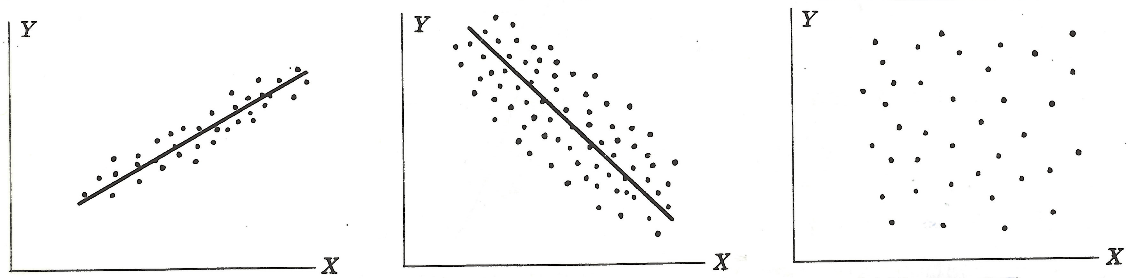
The practice of fitting data to a line or curve is also known as *regression*. However, one must also determine how well the data "fits" the line equations, which is the subject of the next section.

Correlation theory

We now consider how well an equation describes or explains the relationship between variables, or in other words, what the *correlation* is. If all values of the variables satisfy an equation exactly, the variables are *perfectly correlated*. As shown at the beginning, we can consider a variety of relationships such as lines, curves, exponentials, etc.

If X and Y denote the two variables under consideration, a *scatter* diagram shows the location of points (X, Y) on a rectangular coordinate system. If all points in this scatter diagram seem to lie near a line, as in Fig. 4a and 4b, the correlation is *linear*. If Y tends to increase as X increases, as in 4a, the correlation is called *positive* or *direct correlation*. If Y tends to decrease as X increases, as in 4b, the correlation is called *negative* or *inverse correlation*. If all points seem to lie near some curve, the correlation is called *nonlinear*, and a nonlinear equation (such as a curve, exponential, logarithm, etc.) is required for regression.

If there is no linear relationship between the variables X and Y, as shown in Fig. 4c, there is *no linear correlation* between them, i.e., they are *linearly uncorrelated.*



(a) Positive Linear Correlation **(b) Negative Linear Correlation** **(c) ~~No Correlation~~** No Linear Correlation

Fig. 4

Sometimes a transformation of coordinate systems may show a *nonlinear correlation*, but this is a more complicated subject reserved for other classes.

EXPLAINED AND UNEXPLAINED VARIATION

The *total variation* of Y is defined as $\sum (Y - Y_{est})^2$, i.e., the sum of squares of the deviations of the values of Y form the mean $\bar{Y}$. This can be written

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{est})^2 + \sum (Y_{est} - \bar{Y})^2$$

The first term on the right is called the *unexplained variation* while the second term is called the *explained variation*, so called because the deviations $Y_{est} - \bar{Y}$ have a definite pattern while the deviations $Y - Y_{est}$ behave in a random or unpredictable manner.

COEFFICIENT of CORRELATION

The ratio of the explained variation to the total variation is called the *coefficient of determination*. If there is zero explained variation, i.e., the total variation is all unexplained, this ratio is zero. If there is zero unexplained variation, i.e., the total variation is all explained, the ratio is one. In other cases, the ratio is between zero and one. The quantity, *r*, is called the *coefficient of correlation*, and is given by

$$r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum(Y_{est} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}}$$

and varies between -1 and $+1$. The signs $\pm$ are used for positive linear correlation and negative linear correlation, respectively. Note that r is a dimensionless quantity, i.e., it does not depend on the units employed.

The above equation may be rewritten in the following manner which is easier to solve on a computer:

$$r = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\left(\sum X^2\right) - \left(\sum X\right)^2\right]\left[n\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}}$$

While r quantifies the linear correlation, it does *not* quantify the variability explained. This is done by squaring the linear correlation, and multiplying by 100%

Definition: the *explained variability* (or *variance*) between X and Y is computed as $r^2 \times 100\%$

The explained variability will range from 0 to 100%, with 0 denoting no explained variance, and 100% explaining all the variance. Note that seemingly large linear correlations actually translate to small variance numbers. For example, $r = 0.5$ only explains 25% of the variance, and $r = 0.8$ only explains 64% of the variance.

EXAMPLE

Compute the correlation, and the explained variance, for the exercise above.

ANSWER

Make a table as before, but now compute Y^2 as well.

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
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum XY = 364$	$\sum Y^2 = 256$

Compute r

$$r = \frac{n\left(\sum XY\right)-\left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\left(\sum X^2\right)-\left(\sum X\right)^2\right]\left[n\left(\sum Y^2\right)-\left(\sum Y\right)^2\right]}} = \frac{8(364)-(56)(40)}{\sqrt{\left[8(524)-(56)^2\right]\left[8(256)-(40)^2\right]}} = .977$$

which shows that there is a strong, positive linear correlation between X and Y (recall that the maximum possible is 1.0).

The explained variance is:

$$r^2 \times 100\% = (.977)^2 \times 100\% = 95.45\%$$

Which shows that X explains all but 4.54% of the variability of Y.

PREDICTING *Y* AS A FUNCTION OF *X*

Now that we know that X explains a large variability of Y, we may use the regression equation to compute Y for any value of X (in the range 1 to 14; applying regression outside this range of X may not be valid since the regression equations were derived only in this range).

Exercise

If $X=10$, compute Y

Answer

$Y=.636X+.545=(.636)(10)+.545=6.905$

APPLICATIONS TO WEATHER FORECASTING

Regression is applied in a wide variety of meteorological applications. For example, they are used to correct for biases in the numerical models. Consider the hypothetical case where the NGM model consistently predicts too warm a temperature in Minnesota when the wind is from the north. This might be because Minnesota is in a north-south oriented valley where cold air tends to collect at night, and the model's resolution is too coarse to capture the valley. One could develop a regression equation for temperature based on wind direction to correct for this warm temperature bias, where X is wind direction and Y is temperature.

This is what *Model Output Statistics* (MOS) does. However, usually it's computed as a *multiple linear regression equation* of the form:

$$Y = m_1 X_1 + m_2 X_2 + \ldots + m_n X_n + b$$

where there are n values of X, each with its own *slope parameter m*. The different values of X may be a variety of meteorological variables such as temperature, wind, pressure, moisture, moisture convergence, Q-vectors, etc. Independent values (Y) include wind, temperature, and QPF. Other forecast applications using regression include:

1) NHC uses multiple regression to forecast hurricane tracks. These schemes include CLIPER where X includes persistence, latitude, and longitude; and NHC90 where X includes model winds and geopotential height.
2) NHC uses multiple regression to forecast hurricane intensity. Examples include SHIFOR where X includes persistence, latitude, and longitude; and SHIPS, where X includes SSTs, wind shear, 200-mb temperature, etc.
3) Dr. William Gray uses it to predict Atlantic seasonal hurricane activity, where X includes El Niño, African rainfall, SST, QBO, Caribbean sea pressure, wind shear, etc.

The computation of multiple linear regression equations is much more complicated, and usually requires a statistical package such as SAS or IMSL. There are also several issues regarding how to pick the optimum variables without "overfitting" the equation with too many variables (overfitting tends to make forecasts worse, not better). Please refer to a statistical book for details.

Also software such as R.
and MATLAB