Least square method

Definition

It is a mathematical method and with it gives a fitted trend line for the set of data in such a manner that the following two conditions are satisfied.

1.The sum of the deviations of the actual values of Y and the computed values of Y is zero.

2.The sum of the squares of the deviations of the actual values and the computed values is least.

This method gives the line which is the line of best fit. This method is applicable to give results either to fit a straight line trend or a parabolic trend.

The method of least squares as studied in time series analysis is used to find the trend line of best fit to a time series data.

Secular Trend Line

The secular trend line (Y) is defined by the following equation:

$Y = a + b X$

Where, Y = predicted value of the dependent variable

a = Y-axis intercept i.e. the height of the line above origin (when X = 0, Y = a)

b = slope of the line (the rate of change in Y for a given change in X)

When b is positive the slope is upwards, when b is negative, the slope is downwards

X = independent variable (in this case it is time)

To estimate the constants a and b, the following two equations have to be solved simultaneously:

$\Sigma Y = na + b\,\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

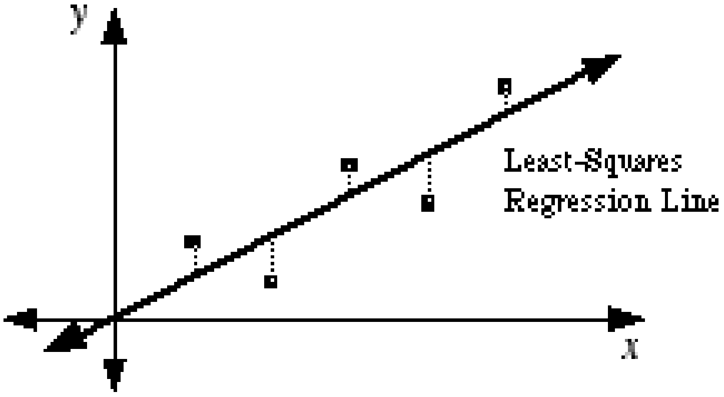
To simplify the calculations, if the midpoint of the time series is taken as origin, then the negative values in the first half of the series balance out the positive values in the second half so that $\Sigma X = 0$. In this case, the above two normal equations will be as follows:

$\Sigma Y = na$

$\Sigma XY = b\Sigma X^2$

The lengths of the vertical dotted lines are the residuals.

The least-squares regression line is the linear fit that minimizes the sum of the squares of the residuals.



Given data {(x_1, y_1), . . . ,(x_N , y_N)}, we may define the error associated to saying y = ax + b by

$$E(a,b) \;=\; \sum_{n=1}^{N} (y_n - (ax_n + b))^2 .$$

This is just N times the variance of the data set {y_1−(ax_1+b), . . . , y_N−(ax_N +b)}. It makes no difference whether or not we study the variance or N times the variance as our error, and note that the error is a function of two variables.

The goal is to find values of a and b that minimize the error. In multivariable calculus we learn that this requires us to find the values of (a, b) such that

$$\frac{\partial E}{\partial a} = 0, \quad \frac{\partial E}{\partial b} = 0.$$

Note we do not have to worry about boundary points: as |a| and |b| become large, the fit will clearly get worse and worse. Thus we do not need to check on the boundary.
Differentiating E(a, b) yields

$$\begin{aligned}\frac{\partial E}{\partial a} &= \sum_{n=1}^{N} 2\,(y_n - (ax_n + b)) \cdot (-x_n)\\ \frac{\partial E}{\partial b} &= \sum_{n=1}^{N} 2\,(y_n - (ax_n + b)) \cdot 1.\end{aligned}$$

Setting ∂E/∂a = ∂E/∂b = 0 (and dividing by 2) yields

$$\begin{aligned}\sum_{n=1}^{N} (y_n - (ax_n + b)) \cdot x_n &= 0\\ \sum_{n=1}^{N} (y_n - (ax_n + b)) &= 0.\end{aligned}$$

These equations can be used to find the error from the line of fit.

Applications:

The most common application of this method, which is sometimes referred to as "linear" or "ordinary", aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model. Some applications include processing profilograms, estimation of chemical reaction population, automatic lens design and optical correction..

Significance:

Least square method is the most efficient linear regression estimator when the assumptions hold true. The estimates of the unknown parameters obtained from linear least squares regression are the optimal estimates from a broad class of possible parameter estimates under the usual assumptions used for process modelling. Practically, linear least squares regression makes very efficient use of the data. Good results can be obtained with relatively small data sets. the theory associated with linear regression is well-understood and allows for construction of different types of easily-interpretable statistical intervals for predictions, calibrations, and optimizations.

Another benefit of satisfying these assumptions is that as the sample size increases to infinity, the coefficient estimates converge on the actual population parameters.

* Least Square Method

1) Fit a least square line for the following data.

X	1	2	3	4	5
Y	2	5	3	8	7

X	Y	XY	X^2
1	2	2	1
2	5	10	4
3	3	9	9
4	8	32	16
5	7	35	25

The equation of least square line $Y = a + bX$

$\therefore$ $\Sigma Y = na + b\Sigma X$

$25 = 5a + 15b$ —— ①

$\Sigma XY = a\Sigma X + b\Sigma X^2$

$88 = 15a + 55b$ —— ②

By Eq. ② - 3 Eq. ①,

$b = 1.3$ and $a = 1.1$

$\therefore$ Equation of least square line is $\hat{Y} = 1.1 + 1.3X$

2) Construct the simple linear regression equation of Y on X. If $n=7$,

$\sum x_i = 113$, $\sum x_i^2 = 1983$, $\sum y_i = 182$, $\sum x_i y_i = 3186$

The equation of Y on X is of the form,

$\hat{Y} = a + bX$ —— ①

$\therefore \sum y_i = na + b\sum y_i$ —— ②

$\sum x_i y_i = a\sum x_i + b\sum x_i^2$ —— ③

Substituting the values,

$7a + 113b = 182$ —— ④

$113a + 1983b = 3186$ —— ⑤

By ④×113 − ⑤×7, we get

$-1112b = -1736$

$\therefore b = 1.56$ and

$a = 0.82$

$\therefore \hat{Y} = 0.82 + 1.56X$ is the equation of line.

3) Using the method of least squares, find an equation of the form $\hat{Y} = ax + b$ that fits the following data:

x	0	1	2	3	4
y	1	5	10	22	38

Here $n = 5$,

From the given data,

x	y	xy	x^2
0	1	0	0
1	5	5	1
2	10	20	4
3	22	66	9
4	38	152	16

$$\Sigma x_i y_i = a \Sigma x_i^2 + b \Sigma x_i \quad \text{———} \quad (1)$$

$$\Sigma y_i = nb + a \Sigma x_i \quad \text{———} \quad (2)$$

$\therefore$ The normal equations are

$$30a + 10b = 243 \quad \text{———} \quad (3)$$

$$10a + 5b = 76 \quad \text{———} \quad (4)$$

On solving (3) and (4),

$a = 9.1$, $b = -3$

$$\therefore \quad \hat{Y} = 9.1x - 3$$

HW

4) Given these measurements of the two quantities x and y; find Y_7 :

x	2	4	6	8	10	12	14
y	2	4	4	5	5	7	Y_7

Soln: $\hat{Y} = \frac{29}{70} x + \frac{56}{35}$

Hint: Minimize the sum of deviation errors $(d_i = \hat{y}_i - y_i)$ by differentiating w.r.t. to both a and b separately and equating it to zero.