Algorithm Used: Least Square Method Algorithm

The method of **least squares** is a standard approach in regression analysis to the approximate solution of over determined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in the results of every single equation.

The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. When the problem has substantial uncertainties in the independent variable (the *x* variable), then simple regression and least squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least squares problems fall into two categories: linear or ordinary least squares and non-linear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The non-linear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Polynomial least squares describes the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

When the observations come from an exponential family and mild conditions are satisfied, least-squares estimates and maximum-likelihood estimates are identical.[1] The method of least squares can also be derived as a method of moment's estimator.

Step 1: Calculate the mean of the x-values and the mean of the y-values.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\overline{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Step 2: The following formula gives the slope of the line of best fit:

$$m = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}$$

Step 3: Compute the y-intercept of the line by using the formula:

$$b = \overline{Y} - m\overline{X}$$

Step 4: Use the slope m and the y-intercept b to form the equation of the line.

Simple explanation for plotting the line of best fit is given below :

Each point has coordinates (X, Y).
We read through all the points in the set and calculate the following:

Count = the number of points
SumX = sum of all the X values
SumY = sum of all the Y values
SumX2 = sum of the squares of the X values
*SumXY = sum of the products X*Y for all the points*

Now we can find the slope M and Y-intercept YFit of the line we want:

XMean = SumX / Count
YMean = SumY / Count
*Slope = (SumXY - SumX * YMean) / (SumX2 - SumX * XMean) YInt = YMean - Slope * Xmean*

The equation for the line is:
Y = Slope * X + YFit

References:

1. *Charnes, A.; Frome, E. L.; Yu, P. L. (1976). "The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family".Journal of the American Statistical Association **71** (353): 169.doi:10.1080/01621459.1976.10481508.*
2. ***Jump up^*** *Bretscher, Otto (1995). Linear Algebra With Applications (3rd ed.). Upper Saddle River, NJ: Prentice Hall.*
3. ***Jump up^*** *Stigler, Stephen M. (1981). "Gauss and the Invention of Least Squares".Ann. Statist. **9** (3): 465–474. doi:10.1214/aos/1176345451.*
4. *http://www.efunda.com/math/leastsquares/lstsqrmdcurve.cfm*
5. *http://www.enlistq.com/analyzing-nyc-traffic-data-using-pandas/*
6. *http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit*
7. *https://www.youtube.com/watch?v=aq8VU5KLmkY*
8. *https://www.youtube.com/watch?v=k_OB1tWX9PM*
9. *https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95*
10. *https://realpython.com/blog/python/analyzing-obesity-in-england-with-python/*
11. *https://www.researchgate.net/post/Whats_the_difference_between_training_set_and_test_set*
12. *http://faculty.cs.niu.edu/~hutchins/csci230/best-fit.htm*
13. *http://hotmath.com/hotmath_help/topics/line-of-best-fit.html*