1. Introduction

Welcome to the series of e-learning modules on Principle of Least Squares. Here we look at the principle of least square, its history, application, mathematical and stochastic models, conditional and parametric models, observation equations, and procedure of using least square method in estimating unknown quantities both linear and non linear.

By the end of this session, you will know:

- The meaning of 'Principle of Least Squares'
- Different fields where this principle is applied
- Categories of least squares
- History of least squares
- Theory of principle of least squares
- Stochastic model and mathematical model
- Conditional and parametric models
- Method of finding least square solutions both in case of linear and non linear least squares

The "Principle of Least Squares" states that the most probable values of a system of unknown quantities upon which observations have been made are obtained by making the sum of the squares of the errors a minimum.

This statement however faces criticism on two grounds being: One, it is indefinite in that the term "error" is not rigorously defined; and two, the ordinary proof of the principle is defective from the fact that the same indefinite nomenclature, and loose reasoning founded thereon, are used throughout.

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

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.
Here it should be noted that, when the problem has substantial uncertainties in the independent variable, that is 'x' variable, then the 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 of least squares.

2. Categories of 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.
A closed-form solution (or closed-form expression) is any formula that can be evaluated in a finite number of standard operations.

The non-linear problem has no closed-form solution and is usually solved by iterative refinement.
At each iteration the system is approximated by a linear one.
Thus the core calculation is similar in both the cases.

The fundamental basis for least-squares method was first described by Carl Friedrich Gauss around the year 1794 at the age of eighteen.
The method of least squares grew out of the fields of astronomy and geodesy as scientists and mathematicians sought to provide solutions to the challenges of navigating the Earth's oceans during the Age of Exploration.

This method was the culmination of several advances that took place during the course of the eighteenth century:
The combination of different observations taken under the same conditions contrary to simply trying one's best to observe and record a single observation accurately.
This approach was notably used by Tobias Mayer while studying the librations of the moon.

The combination of different observations as being the best estimate of the true value; errors decrease with aggregation rather than increase, perhaps was first expressed by Roger Cotes

The combination of different observations taken under different conditions was notably performed by Roger Joseph Boscovich in his work on the shape of the earth.

The combination of different observations taken under different conditions was also notably performed by Pierre-Simon Laplace in his work in explaining the differences in motion.
The development of a criterion that can be evaluated to determine when the solution with the minimum error has been achieved was developed by Laplace in his Method of Least Squares.

In 1822, Gauss was able to state that the least-squares approach to regression analysis is optimal in the sense that in a linear model where the errors have a mean of zero, are uncorrelated and have equal variances, the best linear unbiased estimator of the coefficients is the least-squares estimator. This result is known as the Gauss–Markov theorem.

The idea of least-squares analysis was also independently formulated by the Frenchman Adrien-Marie Legendre in 1805 and the American Robert Adrain in 1808.

3. Fundamental Principle of Least Squares

Fundamental principle of LS
Let there be a system of quantities X one X two X three etc., whose true values are unknown and suppose a set of fallible observations too have been made upon certain functions of these.
The resulting observation equations will be F one of x one x two etc is equal to O, f two of x one x 2 etc is equal to O dash etc. where O, O dash etc are observed quantities.

Now T, T dash etc., are the values of O, O dash etc., would have had if the observations had been free from accidental error, the priori probability of the occurrence of the set of observations O, O dash etc., is P is equal to c into e power h square into u square plus u square dash plus etc.
Where u is equal to T minus O, u dash is equal to T dash minus O dash etc.
Where u is equal to T minus O, u dash is equal to T dash minus O dash etc.
However, as soon as the observations have been made, the element of chance disappears, and the function P is no longer a variable, but a perfectly definite quantity, though unknown.

The choice between two different sets of observed values, after the observations have been made, is strictly a question not of a priori but a posterior probability.
However, we may assume that the set of observations which have the greater priori probability in its favour is on the whole the nearer approach to truth. Hence, of the two sets of observed values, between which we cannot otherwise choose, that one will be , so far as can be judged, the more accurate for which u square plus u square dash plus etc., is less.

Now let us suppose that we have two sets of adopted quantities M, M dash etc and M one, M one dash etc., both close approximations to T, T dash etc., between which it is desired to choose.
Since both sets of quantities are close approximations to the true values of the quantities observed, it is manifestly possible that they both might have resulted from actual sets of observations of equal precision; also so far as the question of accuracy of representation of the true values is concerned, it is immaterial whether they were actually observed or chosen in any other way.

Hence the set of the quantities is, as far as can be judged, the more accurate representation of T, T dash etc. For which delta square plus delta square dash plus etc., is less, where delta is equal to M minus T, delta dash is equal to M dash minus T dash etc.
And in adopting the quantities M, M dash etc., we have assumed no restrictions except that they shall be close approximations to T, T dash etc.
Hence, we may assume them to be defined in terms of adopted values of the unknowns x one, x two, etc., by the equations
F one dash of x one x two etc is equal to M, f two dash of x one x two etc., is equal to M dash etc.

Hence finally, the best values of the unknown quantities, x one, x two, etc are given by making delta square as small as possible.
It is to be noted that the quantities delta, delta dash etc., are the actual errors of the quantities found by substituting any assumed values of x one x two etc., in the first members of the observations equations, and as such are distinct both from errors of observation and from “residuals”

In fact, where u is an error of observation and v a residual, we have
U is equal to T minus O, v is equal to M minus O, delta is equal to M minus T and hence delta is equal to v minus u.
If we are content, as we usually must be, to obtain the best possible approximation, not to the true values of the quantities observed, but to their observed values, the above condition becomes v square is equal to a minimum, the ordinary statement of the law.

4. Stochastic and Mathematical Model

Stochastic Model

- The co-variances (including variances) and hence the weights as well, form the stochastic model
- Even an ‘un-weighted’ adjustment assumes that all observations have equal weight which is also a stochastic model
- The stochastic model is different from the mathematical model
- Stochastic models may be determined through sample statistics and error propagation, but are often a priori estimates

Mathematical model

- The mathematical model is a set of one or more equations that define an adjustment condition.
- Models also include collinearity equations in photogrammetry and the equation of a line in linear regression.
- It is important that the model properly represents reality – for example the angles of a plane triangle should total 180°, but if the triangle is large, spherical excess cause a systematic error so a more elaborate model is needed

There are two types of model namely conditional model and parametric models.

- Conditional model enforces geometric conditions on the measurements and their residuals
- Parametric model expresses equations in terms of unknowns that were not directly measured, but relate to the measurements, say for example a distance expressed by coordinate inverse.
- Parametric models are more commonly used because it can be difficult to express all of the conditions in a complicated measurement network

Observation equations:

- Observation equations are written for the parametric model
- One equation is written for each observation
- The equation is generally expressed as a function of unknown variables (such as coordinates) equals a measurement plus a residual
- We want more measurements than unknowns which gives a redundant adjustment

5. Method of Finding Least Square Solution

Now let us discuss the method of finding least square solution
The objective consists of adjusting the parameter to a model function to best fit a data set.
Let us consider a simple problem, where a data set consist of n points x i y i, where i is equal to 1, 2, etc., n, where x i is an independent variable and y i is a dependent variable whose value is found by observation.
The model function has the form f of x beta, where the m adjustable parameters are held in the vector beta. The goal is to find the parameter values for the model which best fits the data.

The least squares method finds its optimum when the sum, S of squared residuals S is equal to summation over i is equal to 1 to n, r i square.
is minimum where, r i is equal to y i minus f of x i beta, the difference between the actual value of the dependent variable and the value predicted by the model.
The example of a model is that of a straight line.
Denoting the intercepts as β_0 and the slope as β_1, the model function is given by f of x beta is equal to beta not plus beta one into x.
The data point may consist of more than one independent variable.

Here, consider a general relationship between a dependent variable and n independent variables that is linear-in-the-parameters:
Y i is equal to theta 1 into x i one plus theta 2 into x i two plus etc., plus theta n theta into x i n theta where
y i is the i^{th} observation of the dependent variable
x i j is the i^{th} observation of the j^{th} independent variable
Theta j is the coefficient associated with the j^{th} independent variable.

Say m sets of observations (measurements) of dependent and independent variables have been made, that is,
Y 1 is equal to theta 1 into x 1 one plus theta 2 into x one two plus etc., plus theta n theta into x one n
Y two is equal to theta 1 into x two one plus theta 2 into x 2 two plus etc., plus theta n theta into x two n
Etc.,
Y m is equal to theta 1 into x m one plus theta 2 into x m two plus etc., plus theta n theta into x m n.

This set of expressions can be rewritten in more compact from using matrix vector notation as follows
Column vector y one y two etc y m is equal to matrix x one 1 x one two etc x one n theta, x two one x two 2 etc x two n theta etc., x m one x m two etc x m n theta into
Column vector theta one theta two etc theta n theta.
Letting

Y is equal to Column vector y one y two etc y m
X is equal to the matrix x one 1 x one two etc x one n theta, x two one x two 2 etc x two n theta etc., x m one x m two etc x m n theta and
Theta is equal to Column vector theta one theta two etc theta n theta.

Then Y is equal to X into theta.
If the number of observations is equal to the number of unknown parameters, then X is a square matrix and then the inverse matrix x inverse exists and
Theta cap is equal to x inverse into y
However, it is usually the case that there are more observations than unknown parameters. In this case X is no longer a square matrix, and therefore X inverse does not exist. Since there are more equations than unknowns, this means that any solution will not be unique.

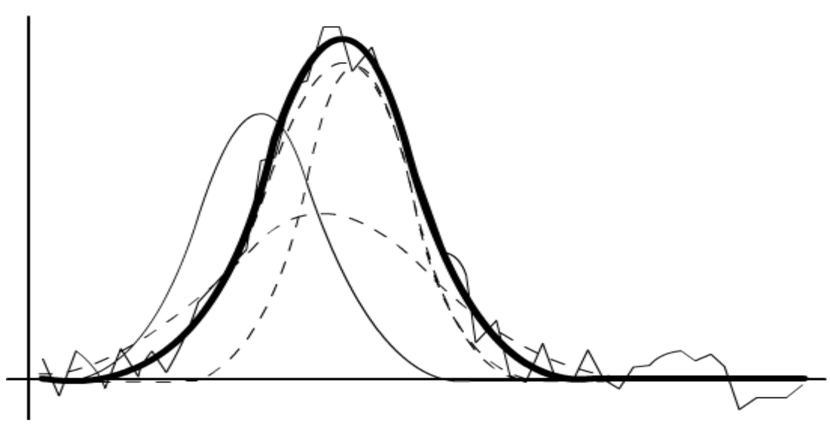
Thus, we have to determine the 'best' theta cap and one way is to find a theta cap such that the sum of the squared difference between the observed dependent variable and its estimates is a minimum, namely,

A set of the unknowns theta cap such that the sum of squared difference between the estimates, obtained using theta cap.
That is, Y i cap is equal to theta cap one into x i one plus theta cap 2 into x i two plus etc plus theta cap n theta into x i n theta.
And corresponding observed y i is a minimum. This is therefore an optimization problem where the objective is to find a theta cap that will set the sum of squared errors between observed and estimated values.

Solution of this problem yields the least square estimates of theta cap as
Theta cap is equal to x transpose into x whole inverse into x transpose into y.
This can be verified easily as follows
Y cap is equal to x into theta cap is equal to x into, x transpose into x whole inverse into x transpose y
That is y cap is equal to x into x transpose into x transpose inverse into x transpose into y is equal to y.

Let us consider the example for Non linear least squares.

Figure 1



An example of a nonlinear least squares is fit to a noisy Gaussian function,
F of x with parameters A, x not sigma is equal to A into e power minus x minus x not square divided by 2 into sigma square is shown below.
Here the thin solid curve is the initial guess, the dotted curves are intermediate iterations, and the heavy solid curve is the fit to which the solution converges.

The actual parameters are A, x not sigma is equal to 1 twenty 5. The initial guess was zero point eight, fifteen and 4, and the converged values are 1 point zero three 1 zero five, 20 point one 3 six 9, 4 point 8 six zero 2 two with R square is equal to zero point 1 four 8 five. The partial derivatives used to construct the matrix A are
D f by d A is equal to e power minus x minus x not whole square divided by 2 into sigma square,
D f by d x not is equal to A into x minus x not divided by sigma square into e power minus x minus x not whole square divided by 2 into sigma square, and
D f by d sigma is equal to A into x minus x not whole square divided by sigma cube into e power minus x minus x not whole square divided by 2 into sigma square.

Here's a summary of our learning in this session:

- The meaning of principle of least squares
- Different fields where this principle is applied
- Different categories of least squares
- History of least squares
- Theory of principle of least squares
- Stochastic model and mathematical model
- Conditional and parametric models
- Method of finding least square solutions both in case of linear and non linear