Method of Least Squares

Michael Dang

MATH 464WI

Date: 03/31/2023

Word Count: 4086

Carl Friedrich Gauss once said, “It is not knowledge, but the act of learning, not possession, but the act of getting there, which grants the greatest enjoyment.” The method of least squares has informed much of modern data science and the development of high performance of technology, such as weather forecasting using time series analysis and the rise of artificial intelligence, like chatGPT. There is some controversy about who discovered this method. The two mathematicians involved are Adrien-Marie Legendre (1752 - 1833) and Carl Friedrich Gauss (1777 - 1855). It was first introduced to the public by Legendre in the *Nouvelles méthodes pour la determination des orbites des comète* in 1805. However, Gauss had been using this method since about 1795, but did not release his version of the method until 1809 [8, p. 1]. Our discussion will examine the publications of both men.

Adrien-Marie Legendre (1752 – 1833) was likely born in Paris, France, though there is some evidence suggesting that he was born in Toulouse and his family later moved to Paris. He came from a wealthy family that provided him with a top quality education in mathematics and physics when he was young. He is known for Legendre polynomials which are very useful in physics and engineering. For example, they are used to determine the wave functions of electrons in the orbits of atoms. He was also known as a contributor to number theory and an elegant proof that π is irrational. In 1782, Legendre’ essay, *Recherches sur la trajectoire des projectiles dans les milieux résistants,* won the Berlin Academy prize and this launched him into his research career. One of Legendre’s major works is the *Eléments de géométrie* which was published in 1794 and was the leading elementary text on the topic for around a hundred years. In 1824, Legendre refused to vote for the government’s candidate for the Institute National and as a result his pension was stopped, and he died in poverty after a suffering from a long and painful illness in 1833 [1].

Johann Carl Friedrich Gauss (1777 – 1855) was born in Brunswick, Germany into a poor family. He exhibited his intelligence in his early teens by performing astonishing proofs. In 1792, Gauss entered Brunswick Collegium Carolinum after receiving financial aid from the Duke of Brunswick, Wolfenbüttel. Later in 1795, Gauss left Brunswick to study at Göttingen University. However in 1798, he left Göttingen to return to Brunswick without a diploma. He is considered one of the greatest mathematicians of all time along with Archimedes and Newton. He made significant contributions to number theory, geometry, probability theory, etc. In statistics, he is famous for the *Gaussian distribution*, which is consider one of the most important distributions of all time [5]. In physics, particularly electromagnetism, Gauss's law for electricity states that the electric flux across any closed surface is proportional to the net electric charge enclosed by the surfaces. In algebra, at 22 he constructed a polygon of 17 sides by ruler and compass, the greatest advance in this field since the time of Greek mathematicians and this leads to the foundation ideas of Galois theory. He also gave a proof of the fundamental theorem of algebra as his doctoral dissertation at the University of Helmstedt. In 1801, he published *Disquisitiones Arithmeticae* which was a significant contribution to number theory. He wrote: "Mathematics is the queen of the sciences and number theory is the queen of mathematics." Nevertheless, he developed the fields of potential theory and real analysis. He died of heart disease in 1855 [2].

During the Age of Discovery, from the 15th century to the 18th century, attempts to navigate the Earth's oceans lead to the development of the least squares method. Mathematicians tried to calculate the behavior of celestial bodies, objects in space such as the sun, moon, planets, and stars, to assist ships with sailing the sea. The technique is all about fitting the best line to a random data set, in other words to optimize the distance from each data point to the line. The technique of adding independent observations and finding the mean had appeared by the seventeenth century; however, it wasn't until the eighteenth century that mathematician Roger Cotes (1682-1716) gave the generalized version. Two methods were introduced independently, one of them is by Leonhard Euler (1707 – 1783) and Tobias Mayer (1723 – 1762) by dividing the observations into some groups with known parameters, then observing the total of each group; another method was proposed by Euler and Johann Heinrich Lambert (1728 – 1777) in which the parameter should minimize the absolute value of the largest deviations α and β. In other words, we have a line of best fit. This is denotated by: $y_i = \alpha - \beta x$.

Boscovich (1711 – 1787) proposed there are two conditions that α and β must satisfy: the sum of deviations is zero $[\alpha + \beta = 0]$ and the sum of the absolute values of the deviations $[|\alpha| + |\beta|]$ is a minimum. Overall, both methods work for geodetic data, with each case comparing the largest residual errors to which the observations are susceptible ([8, p.1]). The method of least squares is considered one of the most important statistical methods of the nineteenth century. Legendre was the first to publish this method in 1805 [6, p.819] but Gauss began using it in 1795 without publicly publishing his method. Gauss stated this issue would become "one of the most famous priority disputes in the history of science ..." [9]. Gauss would eventually be credited as the founder of the method, but not without a fight.

We will first examine Legendre's paper that first introduced the method of least squares in 1805. This is a document translated from the French by Professor Henry A. Ruger and Professor Helen M.Walker, Teachers College, Columbia University, New York City in 1929 [10]. Comments in square brackets are mine.

On the Method of Least Squares

In the majority of investigations in which the problem is to get from measures given by observation the most exact result which they can furnish, there almost always arises a system of equations of the form

$$[E =]\ a + bx + cy + fz + \&c.\ [\text{etc.}]$$

in which *a, b, c, f, &c*. are the known coefficients which vary from one equation to another, and *x, y, z, &c*. are the [finitely many] unknowns which must be determined in accordance with the condition that the value of *E* [the error] shall for each equation reduce to a quantity which is either zero or very small.

If there are the same number of equations as unknowns *x, y, z* &c. there is no difficulty in determining the unknowns, and the error *E* can be made absolutely zero. But more often the number of equations is greater than that of the unknowns, and it is impossible to do away with all the errors.

In a situation of this sort, which is the usual thing in physical and astronomical problems, where there is an attempt to determine certain important components, a degree of arbitrariness necessarily enters in the distribution of the errors, and it is not to be expected that all the hypotheses shall lead to exactly the same results; but it is particularly important to proceed in such a way that extreme errors, whether positive or negative, shall be confined within as narrow limits as possible.

Of all the principles which can be proposed for that purpose, I think there is none more general, more exact, and more easy of application, [than] that of which we made use in the preceding researches, and which consists of rendering the sum of squares of the errors a minimum. By this means, there is established among the errors a sort of equilibrium which, preventing the extremes from exerting an undue influence, is very well fitted to reveal that state of the system which most nearly approaches the truth.

[Legendre sets the finite sums

$$E = a + bx + cy + fz + \cdots$$

$$E' = a' + b'x + c'y + f'z + \cdots$$

$$E'' = a'' + b''x + c''y + f''z + \cdots .]$$

The sum of the squares of the errors $E^2 + E'^2 + E''^2 + \&c.$ being

$$(a + bx + cy + fz + \&c.)^2$$

$$+ (a' + b'x + c'y + f'z + \&c.)^2$$

$$+ (a'' + b''x + c''y + f''z + \&c.)^2$$

if the *minimum* [of the sum $E^2 + E'^2 + E''^2 + \cdots$] is desired, when *x* alone varies, the resulting equation will be

$$o[0] = \int ac + x \int b^2 + y \int bc + z \int bf + \&c.,$$

in which by $\int ab$ we understand the sum of similar products, i.e., $ab + a'b' + a''b'' + \&c;$ by $\int b^2$ the sum of the squares of the coefficient of *x*, namely $b^2 + b'^2 + b''^2 + \&c.,$ and similarly for the other terms.

[For illustration, let

$$E = a + bx + cy + fz$$

$$E' = a' + b'x + c'y + f'z$$

$$E'' = a'' + b''x + c''y + f''z$$

We wish to find the minimum of the sum $E^2 + E'^2 + E''^2$ with respect to *x*. By elementary calculus, this means we wish to solve $0 = \frac{d}{dx}(E^2 + E'^2 + E''^2)$.

Observe that,

$$E^2 = (a + bx + cy + fz)^2$$

$$= a^2 + abx + acy + afz$$

$$+ bxa + b^2x^2 + bxcy + bxfz$$

$$+ cya + cybx + c^2y^2 + cyfz$$

$$+ fza + fzbx + fzcy + f^2z^2$$

and

$$\frac{d}{dx}(E^2) = ab + ba + 2b^2x + bcy + bfz + cyb + fzb$$

$$= 2\,(ab + xb^2 + ybc + zbf).$$

Hence, we require

$$0 = ab + xb^2 + ybc + zbf.$$

Thus, by similar calculations for E'^2 and E''^2 we see that

$$0 = \frac{d}{dx}(E^2) + \frac{d}{dx}(E'^2) + \frac{d}{dx}(E''^2)$$

requires that

$$0 = ab + \ xb^2 + ybc + zbf$$

$$+ a'b' + \ xb'^2 + yb'c' + zb'f'$$

$$+ a''b'' + \ xb''^2 + yb''c'' + zb''f''$$

$$= (ab + \ a'b' + \ a''b'') + x(b^2 + b'^2 + b''^2)$$

$$+ y(bc + b'c' + \ b''c'') + z(bf + b'f' + b''f'')$$

which in Legendre's notation becomes

$$0 = \ \int ab + x \int b^2 + y \int bc + z \int bf \ \ .]$$

Similarly, the minimum with respect to *y* will be

$$o[0] = \ \int ac + x \int bc + y \int c^2 + z \int fc + \&c.,$$

And the minimum with respect to *z*,

$$o[0] = \ \int af + x \int bf + y \int cf + z \int f^2 + \&c.,$$

in which it is apparent that the same coefficients $\int bc$, $\int bf$ &c. are common to two equations, a fact which facilitates the calculation.

In general, to form the equation of the minimum with respect to one of the unknowns, it is necessary to multiply all the terms of each given equation by the coefficient of the unknown in that equation, taken with regard to its sign, and to find the sum of these products.

The number of equations of minimum derived in this manner will be equal to the number of the unknowns, and these equations are then to be solved by the established methods. But it will be well to reduce the amount of computation both in multiplication and in solution, by retaining in each operation only so much signification of figures, integers or decimals, as are determined by the degree of approximation for which the inquiry calls.

Even if by a rare chance it were possible to satisfy all the equations at once by making all errors zero, we could obtain the same result from the equations of minimum; for if after having found the value of *x, y, z* &c. which make $E, E', \&c.$ equal to zero, we let *x, y, z* vary by $\delta x, \delta y, \delta z$ $\&c.$, it is evident that E^2, which was zero, will become by that variation $[(b\delta x + c\delta y + f\delta z + \&c.)^2$, a correction] .The same will be true for $E'^2, E''^2, \&c.$ Thus we see that the sum of squares of the errors will by variation become a quantity of the second order with respect to $\delta x, \delta y, \delta z$ $\&c.$, which is in accord with the nature of a minimum.

[Meaning if we substitute $x + \ \delta x$ for *x*, $\ y + \ \delta y$ for *y*, $z + \ \delta z$, for *z* into E^2, when *x, y, z* are the values that make $0 = E = a + bx + cy + fz$,

then we see that, $\left(a + b(x + \delta x) + c(y + \delta y) + f(z + \delta z)\right)^2$

$$= ([a + bx + cy + fz] + [b\delta x + c\delta y + f\delta z])^2$$

$$= (a + bx + cy + fz)^2 + 2(a + bx + cy + fz)(b\delta x + c\delta y + f\delta z)$$

$$+ (b\delta x + c\delta y + f\delta z)^2$$

$$= \ (b\delta x + c\delta y + f\delta z)^2.]$$

If after having determined all the unknowns *x, y, z* &c., we substitute their values in the given equations, we will find the value of the different errors $E, E', E'', \&c.$, to which the system gives rise, and which cannot be reduced without increasing the sum of their squares. If among these errors are some which appear too large to be admissible [outliers], then those equations which produced these errors will be rejected, as coming from too faulty experiments, and the unknowns will be determined by means of the other equations, which will then give much smaller errors. It is further to be noted that one will not then be obliged to begin the calculations anew, for since the equations of minimum are formed by the addition of the products made in each of the given equations, it will suffice to remove from the addition those products furnished by the equations which would have led to errors that were too large.

The rule by which one finds the mean among the results of different observations is only a very simple consequence of our general method, which we will call the method of least squares.

Indeed, if experiments have given values $a, a', a'', \&c.$ for a certain quantity *x*, the sum of squares of the errors will be $(a' - x)^2 + (a'' - [x])^2 + (a''' - x)^2 + \&c.$, and on making that sum a minimum, we have

$$0 = (a' - x) + (a'' - [x]) + (a''' - x) + \&c.$$

from which it follows that

$$x = \frac{a' + a'' + a''' + \&c.}{n},$$

n being the number of the observations.

[For example, if

$$0 = (a' - x) + (a'' - x) + (a''' - x)$$

then

$$x = \frac{a' + a'' + a'''}{3}.]$$

In the same way, if to determine the position of a point in space, a first experiment has given the coordinates a', b', c'; a second the coordinates a'', b'', c''; and so on, and if the true coordinates of the point are denoted by *x*, *y*, *z*; then the error in the first experiment will be the distance from the point (a', b', c') to the point (*x*, *y*, *z*). The square of this distance is

$$(a' - x)^2 + ([b'] - y)^2 + ([c' - z])^2,$$

If we make the sum of the squares of all such distances a minimum, we get three equations which give

$$x = \frac{\int a}{n}, \qquad y = \frac{\int b}{n}, \qquad z = \frac{\int c}{n},$$

n being the number of points given by the experiments.

[Here Legendre means $\int a = a' + a'' + a''' + \cdots$, and so on.]

These formulas are precisely the ones by which one might find the common centre of gravity of several equal masses situated at the given points, whence it is evident that the centre of gravity of any body possesses this general property.

If we divide the mass of a body into particles which are equal and sufficiently small to be treated as points, the sum of the squares of the distances from the particles to the centre of gravity will be a minimum.

We see then that the method of least squares reveals to us, in a fashion, the centre about which all the results furnished by experiments tend to distribute themselves, in such a manner as to make their deviations from it as small as possible. The application of which we are now about to make of this method to the measurement of the meridian will display most clearly its simplicity and fertility.

We will now examine Gauss' discussion of the method of least squares published in his book, *Theory of the Motion of the Heavenly Bodies Moving About the Sun in Conic Sections* in 1809, translated by Charles Henry David in 1857 [4]. Again, comments in square brackets are mine.

Determination of an Orbit from three Complete Observations

. . .

Paragraph 175, lines 3 – 6

Let V, V', V'', etc. be functions of the unknown quantities $p, q, r, s,$ etc., μ [mu] the number of these functions, v [nu] the number of the unknown quantities; and let us suppose that the values of the functions found by direct observation are $V = M, V' = M', V'' = M''$, etc.

. . .

Paragraph 179, lines 6 – 13

Therefore, that will be the most probable system of values of the unknown quantities, p, q, r, s, etc., in which the sum of the squares of the differences between the observed and computed values of the functions V, V', V''*, etc. is a minimum* [each of $V, V', V'', \ldots$ is a computed function of the unknown quantities $p, q, r, s, \ldots$, so $V = V(p, q, r, s, \ldots)$, etc.], if the same degree of accuracy [precision] in to be presumed in all observations. This principle, which promises to be of most frequent use in all applications of the mathematics to natural philosophy [science], must, everywhere, be considered an axiom with the same propriety as the arithmetical mean of several observed values of the same quantity is adopted as the most probable value.

. . .

Paragraph 180, lines 1 – 19

The principle explained in the preceding article derives value also from this, that the numerical determination of the unknown quantities is reduced to a very expeditious algorithm, when the functions V, V', V'', etc. are linear.

Let us suppose, ["computed values of the functions" – "observed values of the functions" are]

$$V - M = v = -m + ap + bq + cr + ds + etc.$$

$$V' - M' = v' = -m' + a'p + b'q + c'r + d's + etc.$$

$$V'' - M'' = v'' = -m'' + a''p + b''q + c''r + d''s + etc.$$

etc., and let us put [suppose we have linear combinations of those differences as follows]

$$av + a'v' + a''v'' + etc. = P$$

$$bv + b'v' + b''v'' + etc. = Q$$

$$cv + c'v' + c''v'' + etc. = R$$

$$dv + d'v' + d''v'' + etc. = S$$

etc. Then the v [nu] equations of article 177, from which the values of the unknown quantities must be determined, will, evidently, be the following:

$$P = 0, \qquad Q = 0, \qquad R = 0, \qquad S = 0, etc.$$

provided we suppose the observations equally good; to which case we have shown in the preceding article how to reduce the others. We have, therefore, as many linear equations as there are unknown quantities to be determined, from which the values of the latter will be obtained by common elimination.

. . .

Paragraph 186, all

In conclusion, the principle that the sum of the squares of the differences between the observed and computed quantities must be a minimum may, in the following manner, be considered independently of the calculus of probabilities.

When the number of unknown quantities [v, nu] is equal to the number of the observed quantities depending on them [μ, mu] , the former may be so determined as exactly to satisfy the latter. But when the number of the former is less than that of the latter [$v < \mu$], an absolutely exact agreement cannot be obtained, unless the observations possess absolute accuracy. In this case care must be taken to establish the best possible agreement, or to diminish as far as practicable the differences. This idea, however, from its nature, involves something vague. For, although a system of values for the unknown quantities which makes *all* the differences

$[V - M, V' - M', V'' - M'', \ldots]$ respectively less than another system, is without doubt to be preferred to the latter, still the choice between two systems, one of which presents a better agreement in some observations, the other in others, is left in a measure to our judgment, and innumerable different principles can be proposed by which the former condition is satisfied. Denoting the differences between observation and calculation by $\Delta,\ \Delta', \Delta''$, etc., the first condition will be satisfied not only if $\Delta\Delta + \Delta'\Delta' + \Delta''\Delta'' +$ etc. [the sum of the squares of the differences], is a minimum (which is our principle), but also if $\Delta^4 + \Delta'^4 + \Delta''^4 +$ etc., or $\Delta^6 + \Delta'^6 + \Delta''^6 +$ etc., or in general, if the sum of any of the powers with an even exponent becomes a minimum. But of all these principles ours is the most simple; by the others we should be led into the most complicated calculations.

Our principle, which we have made use of since 1795, has lately been published by Legendre in the work *Nouvelles méthodes pour la determination des orbites des comète, Paris,* 1806 [actually, 1805], where several other properties of this principle have been explained, which, for the sake of brevity, we here omit.

If we were to adopt a power with an infinite even exponent, we should be led to that system in which the greatest differences become less than in any other system.

Laplace (1749 – 1827), made us of another principle for the solution of linear equations the number of which is greater than the number of the unknown quantities, which had been previously proposed by Boscovich, namely, that the sum of the errors themselves taken positively, be made a minimum. It can be easily shown, that a system of values of unknown quantities, derived from this principle alone, must necessarily exactly satisfy as many equations out of the number proposed, as there are unknown quantities, so that the remaining equations come under consideration only so far as they help to *determine the choice:* if, therefore, the equation $V = M,$ for example, is of the number of those which are not satisfied, the system of values found according to this principle would in no respect be changed, even if any other value N had been observed instead of $M,$ provided that, denoting the computed value by $n,$ the differences $M - n,\ N - n,$ were affected by the same signs. Besides, Laplace qualifies in some measure this principle by adding a new condition: he requires, namely, that the sum of the differences, the signs remaining unchanged, be equal to zero. Hence, it follows, that the number of equations exactly represented may be less by unity [1] than the number of unknown quantities; but what we have before said will still hold good if there are only two unknown quantities.

Two hundred years later, the method of least squares has been implemented in various industries. In data science, this concept is used to minimize error in model fitting and in deep learning involving computer vision such as image recognition, classification, etc. In finance, this method helps quantify the relationship between two or more variables, such as a stock's share price and its earnings per share, or it can forecast stock market trends. In medicine, it can help with decision making and forecasting the likelihood of cancers or diseases. In education, it can analyze a students' failures in university.

More specifically, in the book, *Data Driven Science and Engineering* by Brunton and Kutz, the least square method is used in curve fitting, described in mathematical language as *regression*, meaning attempts to estimate the relationship between variables. Its core value is optimization; thus its ultimate goal is to minimize the error and execute the best model. In image processing, it reduces degradation and noise in images based on the mean and variance of the degradation and noise. The book, *Applied Linear Regression Models* by Kutner, Nachtsheim, and

Neter, provides the application of the method of least squares to determine the optimum lot size for a production company that manufactures refrigeration equipment, where the method can help to minimize the cost. With this application people can save a lot of money, material, time, etc.

However, there are still some disadvantages about the method of least squares. For censored data, which is data that we do not know the exact time, the method of least squares is not applicable. The line is sensitive when it comes to outliers, which are data points that separated far from others. Hence, when encountering that situation, it is best to ignore or take out the outliers then perform the least squares method. Also, the starting point is as sensitive as the outlier. For example, if the starting point is beyond the first point, the performance line will not capture most of the data. The method of least squares has less desirable properties compared to the maximum likelihood, which is another great method introduced in 1912 by R.A. Fisher (1890 – 1962) [3]. The main disadvantages of linear least squares are limitations in the shapes of the linear models over the long term. Another issue when it comes to nonlinear terms it is very difficult to find a linear model that fits the data well as the range of the data increases [7].

Legendre did not receive much credit because one of the most brilliant scientific minds was working on the same problem. However, he did introduce the method first publicly. But Gauss just did it better and provided more justification. Thus, from that point of view, the method of least squares should be attributed to Gauss. Legendre and Gauss conclude their argument is complete from both mathematical and logical perspectives, and so will I. After all, as Adrien-Mari Legendre said “All the truths of mathematics are linked to each other, and all means of discovering them are equally admissible” [8].

References

1. *Adrien-Marie Legendre - biography* (no date) *Maths History*. Available at: https://mathshistory.st-andrews.ac.uk/Biographies/Legendre/ (Accessed: March 30, 2023).

2. *Carl Friedrich gauss summary* (no date) *Encyclopædia Britannica*. Encyclopædia Britannica, inc. Available at: https://www.britannica.com/summary/Carl-Friedrich-Gauss (Accessed: March 30, 2023).

3. *The epic story of maximum likelihood - arxiv* (no date). Available at: https://arxiv.org/pdf/0804.2996.pdf (Accessed: March 31, 2023).

4. Gauss, K.F., Davis, C.H. and Gauss, K.F. (1963) *Theory of the motion of the heavenly bodies moving about the sun in conic sections*. New York: Dover Publications.

5. Insider, B. (2016) *The 17 equations that changed the course of history*, *ScienceAlert*. Available at: https://www.sciencealert.com/the-17-equations-that-changed-the-course-of-history (Accessed: March 30, 2023).

6. Katz, V.J. (2018) *A history of mathematics: An introduction*. New York, NY: Pearson.

7. *Linear Least Squares Regression* (no date) *4.1.4.1. Linear least squares regression*. Available at: https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm#:~:text=The%20main%20disadvantages%20of%20linear,properties%2C%20and%20sensitivity%20to%20outliers. (Accessed: March 31, 2023).

8. Plackett, R.L. (1972) “Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares,” *Biometrika*, 59(2), p. 239. Available at: https://doi.org/10.2307/2334569.

9. *Quotes by Adrien-Marie Legendre: A-Z quotes* (no date) *A*. Available at: https://www.azquotes.com/author/29331-Adrien_Marie_Legendre (Accessed: March 31, 2023).

10. Ruger, H. and Walker, H. (no date) *Legendre - University of York*. Available at: https://www.york.ac.uk/depts/maths/histstat/legendre.pdf (Accessed: March 17, 2023).

11. Stigler, S.M. (1986) *The history of Statistics: The measurement of uncertainty before 1900*. Cambrigde, MA: Belknap Press of Harvard University Press.