8 Least squares

0 Texts

Harter and Stigler are the best general texts for the history of least squares. There are tens of thousands of papers on the topic of least squares.

General Texts:

Todhunter,I.(1865*) A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace* Macmillan, London

Plackett,R.L.(1958) Studies in the History of Probability and Statistics. VII. The principle of the arithmetic mean. *Biometrika* 45:130-135.

Harter,H.L. (1974) The Method of Least Squares and Some Alternatives *International Statistical Review* 42:147-174

Stigler, Stephen(1986) *The History of Statistics : The Measurement of Uncertainty Before 1900*

Significant papers:

Laplace, Pierre Simon(1774) Mémoire sur la probabilité des causes par les événemens. *Mémoires de l'Académie royale des sciences presentés par divers savans* 6:621-56. Translated in Stigler(1986).

Stigler, Stephen (1986) Laplace's 1774 memoir on inverse probability. *Statistical Science* 1

Euler,L.(1749) *Pièce qui a Remporté le Prix de l'Academie Royale des Sciences en 1748, sur les Inegalités de Mouvement de Saturn et de Jupiter.* Paris

Mayer,J.T.(1750) *Abhandlung über die Umwälzung des Mondes um seine Axe Kosmographische Nachrichten und Sammlungen for 1748* 1 52-183

Legendre,A.M.(1805) *Nouvelles Méthodes pour la Détermination des Orbites des Comètes* . Courcier, Paris

Gauss,C.F.(1809) *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium Frid*. Perthes et I H Besser, Hamburgi.

Laplace, Pierre Simon (1812) *Théorie analytique des probabilités*. Paris: Courcier

Edgeworth,F.Y.(1885) Methods of Statistics *Journal of the Royal Statistical Society Jubilee* 181-127

Jeffreys,H.(1932) An alternative to the rejection of observations *Proceedings of the Royal Society of London*, A 137: 78-87

Stein,C.(1956) Inadmissibility of the usual estimator for the mean of a multivariate normal population. *Proceedings of the Third Berkeley Symposium*, **1**,197-206

Tukey,J.W.(1962) The future of data analysis *Annals of Mathematical Statistics* 33: 1-67

Huber,P.J.(1964) Robust estimation of a location parameter *Annals of Mathematical Statistics* 35: 73-101

www.Britannica.com

1 Eighteenth century beginnings and the method of averages

Linear models are used to predict a variable as a linear function of other variables. We need to know how to select the linear function and how reliable the prediction is.

In the 18th century, the availability of accurate telescopes led to a great growth in astronomy, and much of the early work in linear models arose from the need to combine discrepant astronomical observations of a celestial object at different times and by different observers. Plackett(1958) found no evidence of the use of the arithmetic mean in the work of the ancient Babylonian and Greek astronomers. Perhaps the first formal consideration of the combination of observations is due to *Galileo* (1632), who considers the question of combining the observations of 13 observers of the elevations of a star in order to determine the distance of the star from the earth.

Mayer(1750) developed the Method of Averages for fitting a linear equation to observed data:

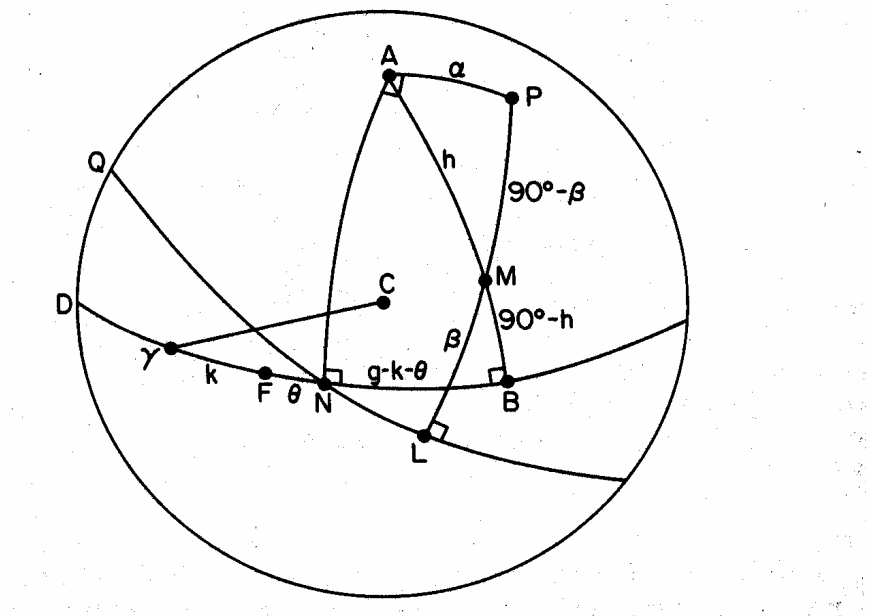


Figure 1.2. The moon. M, The crater Manilius; NL, the moon's equator; P the moon's equatorial pole; NB, the circumference parallel to the plane ecliptic; A, the pole of NB; Cγ, the direction of the equinox from the moon center C; F, the node of the moon's orbit and the plane of the ecliptic. See [illegible] for more details.

From Stigler's description:

Mayer's method for the resolution of inconsistent observational equa:ions can be discerned in his discussion of the position of the crater Manilius. Figure 1.2 represents the moon, which Mayer considered as a sphere. The great circle QNL represents the moon's true equator, and P is the moon's pole with respect to this equator, one end of its axis of revolution. The great circle DNB is that circumference (or apparent equator) of the moon that is seen from earth as parallel to the plane of the ecliptic, the plane of the earth's orbit about the sun, and A is the pole of the moon with respect to DNB, its apparent pole as viewed by an earthbound astronomer oriented by the ecliptic. The point γ (the point on the circle DNB in the direction from the moon's center C toward the equinox) was taken as a reference point. The circle DNB and the pole A will vary with time, as a result of the libration of the moon, but they form the natural system of coordinates at a given time. The equator QNL and the pole P are fixed but not observable from earth. Mayer's aim was to determine the relationship between these coordinate systems and thus accurately determine QNL and P. He accomplished this by making repeated observations of the crater Manilius. Now, in Figure 1.2, M is the position of Manilius, and PL and AB are meridian quadrants through M with respect to the two polar coordinate systems. Mayer was able to observe the position of M on several occasions with respect to the constantly changing coordinate system determined by DNB and A; that is, subject to observational error he could at a given time measure the arcs AM = h and γB = g.

To determine the relationship between the coordinate systems, Mayer sought to find the fixed, but unknown, arc length AP = α, the true latitude of Manilius β = ML, and the distance θ between the unknown node or point of intersection of the two circles (N) and the known point of intersection of the plane of the orbit of the moon and the circle DNB. He let k = γ F be the observed longitude of F. Then g, h, and k were observable and varied from observation to observation as a result of the motion of the moon (and observational error); and α, θ and β were fixed and unknown, to be determined from the observations. Because N AP forms a right angle, a basic identity of spherical trigonometry implies that these quantities are related nonlinearly by the equation

$$\sin\beta = \cos\alpha\cosh + \sin\alpha\sinh\sin(g - k - \theta)$$

These equations are linearized when β and h are small. Mayer observed Manilius on 27 days as the moon rotated. There is an equation for each observation. The equations are divided into three groups according as the coefficient of α is close to 1,-1, or 0.

Table 1.1. Mayer's twenty-seven equations of condition, derived from observations of the crater Manilius from 11 April 1748 through 4 March 1749.

Eq. no.	Equation	Group
1	$\beta - 13°10' = +0.8836\alpha - 0.4682\alpha \sin\theta$	I
2	$\beta - 13°8' = +0.9996\alpha - 0.0282\alpha \sin\theta$	I
3	$\beta - 13°12' = +0.9899\alpha + 0.1421\alpha \sin\theta$	I
4	$\beta - 14°15' = +0.2221\alpha + 0.9750\alpha \sin\theta$	III
5	$\beta - 14°42' = +0.0006\alpha + 1.0000\alpha \sin\theta$	III
6	$\beta - 13°1' = +0.9308\alpha - 0.3654\alpha \sin\theta$	I
7	$\beta - 14°31' = +0.0602\alpha + 0.9982\alpha \sin\theta$	III
8	$\beta - 14°57' = -0.1570\alpha + 0.9876\alpha \sin\theta$	II
9	$\beta - 13°5' = +0.9097\alpha - 0.4152\alpha \sin\theta$	I
10	$\beta - 13°2' = +1.0000\alpha + 0.0055\alpha \sin\theta$	I
11	$\beta - 13°12' = +0.9689\alpha + 0.2476\alpha \sin\theta$	I
12	$\beta - 13°11' = +0.8878\alpha + 0.4602\alpha \sin\theta$	I
13	$\beta - 13°34' = +0.7549\alpha + 0.6558\alpha \sin\theta$	III
14	$\beta - 13°53' = +0.5755\alpha + 0.8178\alpha \sin\theta$	III
15	$\beta - 13°58' = +0.3608\alpha + 0.9326\alpha \sin\theta$	III
16	$\beta - 14°14' = +0.1302\alpha + 0.9915\alpha \sin\theta$	III
17	$\beta - 14°56' = -0.1068\alpha + 0.9943\alpha \sin\theta$	III
18	$\beta - 14°47' = -0.3363\alpha + 0.9418\alpha \sin\theta$	II
19	$\beta - 15°56' = -0.8560\alpha + 0.5170\alpha \sin\theta$	II
20	$\beta - 13°29' = +0.8002\alpha + 0.5997\alpha \sin\theta$	III
21	$\beta - 15°55' = -0.9952\alpha - 0.0982\alpha \sin\theta$	II
22	$\beta - 15°39' = -0.8409\alpha + 0.5412\alpha \sin\theta$	II
23	$\beta - 16°9' = -0.9429\alpha + 0.3330\alpha \sin\theta$	II
24	$\beta - 16°22' = -0.9768\alpha + 0.2141\alpha \sin\theta$	II
25	$\beta - 15°38' = -0.6262\alpha - 0.7797\alpha \sin\theta$	II
26	$\beta - 14°54' = -0.4091\alpha - 0.9125\alpha \sin\theta$	II
27	$\beta - 13°7' = +0.9284\alpha - 0.3716\alpha \sin\theta$	I

Source: Mayer (1750, p. 153).
Note: One misprinted sign in equation 7 has been corrected.

Now, the coefficients for the 9 observations in each group are averaged to obtain just one equation for each group; three unknowns are now determined by solving the three equations:

Table 1.2. Mayer's three equations, as derived from Table 1.1 by adding equations 1, 2, 3, 6, 9, 10, 11, 12, and 27 in group I, equations 8, 18, 19, 21, 22, 23, 24, 25, and 26 in group II, and the rest in group III.

Group	Equation
I	$9\beta - 118°8' = +8.4987\alpha - 0.7932\alpha \sin\theta$
II	$9\beta - 140°17' = -6.1404\alpha + 1.7443\alpha \sin\theta$
III	$9\beta - 127°32' = +2.7977\alpha + 7.9649\alpha \sin\theta$

Source: Mayer (1750, p. 154).

Mayer wrote

"These equations [Table 1.2] can take the place of the foregoing totality of equations [Table 1.1] because each of these three equations has been formed in the most advantageous manner (die vortheilhaftigsteArt). The advantage consists in the fact that through the above division into three classes, the differences between the three sums are made as large as is possible. The greater these differences are, the more accurately (richtiger)one may determine the unknown values of α, β, and θ."(Mayer, 1750, p. 154)

Boscovich(1757) is the first to propose fitting a straight line by requiring that the average residual be zero, and that the sum of absolute residuals be minimized subject to that constraint.

Lagrange(1774) determines the distribution of a mean when the individual errors have density

(1) uniform, $f(x) = \frac{1}{2}\{|x| \le 1\}$

(2) parabolic , $f(x) = \frac{2}{3}(1-x^2)^+$,

(3) cosine , $f(x) = \frac{\pi}{4}\cos(\pi x/2)\{|x| \le 1\}$

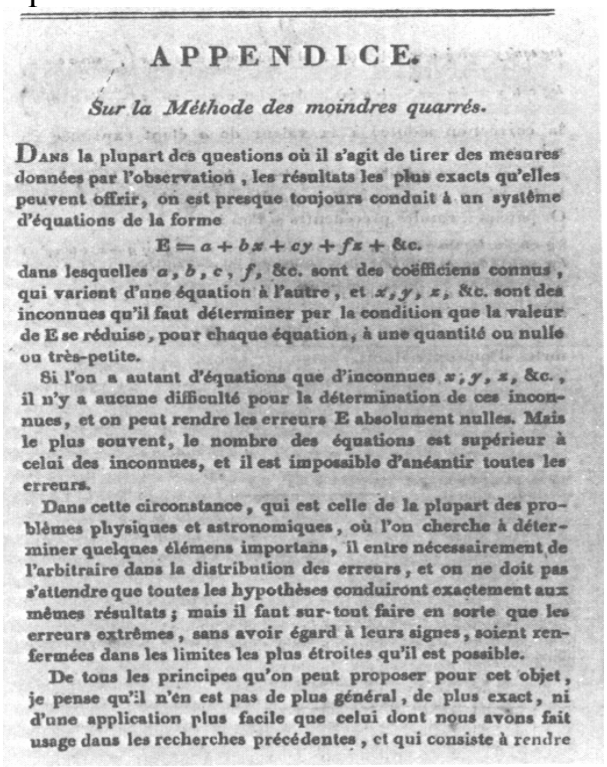
Laplace(1774) suggests combining observations so that the estimated value has minimum expected deviation from the true value, and has equal probability of falling above and below the true value. He solves this problem when the individual errors have density $f(x) = \frac{1}{2}\exp(-|x|)$ which, oddly enough, is now called Laplace's density.

Bernoulli(1778) suggests using the method of maximum likelihood to determine the 'average' of a set of observations when the error density is semi-circular , $f(x) = \frac{2}{\pi}(1-x^2)^{\frac{1}{2}}\{|x| \le 1\}$.

Laplace(1786) fits a straight line by minimizing the maximum residual.

2 Nineteenth Century Least Squares: Legendre

Legendre(1805) is the first to publish the method of Least squares, which determines the coefficients in a set of equations to make the sum of squared errors a minimum.



APPENDICE.

Sur la Méthode des moindres quarrés.

DANS la plupart des questions où il s'agit de tirer des mesures données par l'observation, les résultats les plus exacts qu'elles peuvent offrir, on est presque toujours conduit à un système d'équations de la forme

$$E = a + bx + cy + fz + \&c.$$

dans lesquelles $a, b, c, f,$ &c. sont des coëfficiens connus, qui varient d'une équation à l'autre, et $x, y, z,$ &c. sont des inconnues qu'il faut déterminer par la condition que la valeur de E se réduise, pour chaque équation, à une quantité ou nulle ou très-petite.

Si l'on a autant d'équations que d'inconnues $x, y, z,$ &c., il n'y a aucune difficulté pour la détermination de ces inconnues, et on peut rendre les erreurs E absolument nulles. Mais le plus souvent, le nombre des équations est supérieur à celui des inconnues, et il est impossible d'anéantir toutes les erreurs.

Dans cette circonstance, qui est celle de la plupart des problèmes physiques et astronomiques, où l'on cherche à déterminer quelques élémens importans, il entre nécessairement de l'arbitraire dans la distribution des erreurs, et on ne doit pas s'attendre que toutes les hypothèses conduiront exactement aux mêmes résultats; mais il faut sur-tout faire en sorte que les erreurs extrêmes, sans avoir égard à leurs signes, soient renfermées dans les limites les plus étroites qu'il est possible.

De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre

From Stigler,

Legendre used least squares to estimate the length of the metre, the new French basic unit of length, 1/10000000 times the length of the meridian arc through Paris. The French measured, with sophisticated geodetic survey methods, the lengths of four arcs between Dunkirk, Paris, Evaux, Carcassone, and Mountjoy (near Barcelona). These arcs were expressed in terms of the length of the meridian arc (a function of C) and the ellipticity α of the earth. Assuming that the latitudes of the 5 cities were determined with errors E^I, Legendre derived four equations:

$$E^I - E^{II} = 0.002923 + C(2.192) - \alpha(0.563)$$
$$E^{II}\ I - E^{III} = 0.003100 + C(2.672) - \alpha(0.351)$$
$$E^{III} - E^{IV} = -0.001096 + C(2.962) + \alpha\ (0.047)$$
$$E^{IV} - E^{V} = -.001808 + C(1.851) + \alpha\ (0.263)$$

Legendre has five errors and four equations; he handled that problem by taking E^{III} to be an unknown to be estimated. Thus he has four equations in three unknowns for his least squares problem; that was lucky, he still has 1 degree of freedom for error!
The meter was determined by Legendre to be 3.280 feet. The actual meter used was taken from Laplace, who used also an arc measured in Peru; it was 3.281 feet.

3 Gauss connects the gaussian distribution with least squares

Gauss(1809) shows that the normal (or Gaussian) law of error is necessary if the arithmetic mean is to be the most probable value of an unknown quantity based on several equally good observations, when the prior distribution of the unknown is uniform. The arithmetic mean is 'generally acknowledged' as an excellent way of combining observations, so the use of the normal error curve is justified. This is a first use of a normative Bayesian argument, saying that people must implicitly believe a certain distribution because they act in a certain way.

And then the method of least squares is a consequence of assuming Gaussian errors in fitting problems. Laplace (1810) argued that the normal error curve was justified by the Central Limit Theorem, in that each error was made up of the sum of a larger number of small errors.

From Stigler,

Or was it Legendre's prInciple(Gauss deeply affronted Legendre by referring to the method of least squares as "our principle" (Principium nostrum) and by claiming that he, Gauss, had been using the method since 1795. The ensuing priority dispute, and another one involving the law of quadratic reciprocity of number theory, exacerbated the relationship between the two men. The heat of the dispute never reached that of the Newton-Leibniz controversy, but it reached dramatic levels nevertheless. Legendre appended a semianonymous attack on Gauss to the 1820 version of his Nouvelles methodes pour la determination des orbites , and Gauss solicited reluctant testimony from friends that he told them of the method before 1805. Plackett (1972) reviews most of evidence. A recent study of this and further evidence (Stigler, 1981) suggests that, although Gauss may well have been telling the truth about his prior use of the method, he was unsuccessful in whatever attempts he made to communicate it before 1805. In addition, there is no indication that he saw its great general potential before he learned of Legendre's work. Legendre's 1805 appendix, on the other hand, although it fell far short of Gauss's work in development, was a dramatic and clear proclamation of a general method by a man who had no doubt about its importance.

Laplace(1810) shows that arithmetic means are asymptotically normal under general conditions, and that for normal error laws least squares, maximum likelihood, and his own 1774 criterion (requiring minimum expected absolute deviation from the true value, while having median at the true value) all produce the arithmetic mean. His great work on probability, Laplace(1812) contains this material and much more about least squares.

4 Early suggestions for modern methods

Gauss(1816) shows that estimating the residual variance by root mean squared deviations has superior asymptotic efficiency to using any other nth root of sums of nth powers. (The same accuracy is achieved with 114 observations at n=1, 100 at n=2, 109 at n=3, 133 at n=4,178 at n=5, 251 at n=6, 249 if the median of absolute values of the errors is used.)
Quetelet(1846) suggests using the interquartile range, the difference between the two points that contain the middle 50% of the observations.

Fourier(1824) fits a linear equation to (1) minimize the maximum absolute deviation and (2) minimize the average absolute deviation. He sets the solution up to solve systems of inequalities, that is, in the form of a linear programming problem. The method is now called Fourier's method of descents.

Cauchy(1837) proposes a variable selection procedure in which variables are not included in the fitting process if they do not materially reduce the residual sum of squares.

Pierce(1852) proposes rejecting observations when the probability of all observations is less than the probability of observations retained multiplied by the probability of making the rejected set of observations.

Chauvenet(1863) proposes rejecting the observation X in n observations if we expect 0.5 out of n normal observations to be greater than X.

5 Early Twentieth Century

Edgeworth (1885), the first modern author, compares the mean and the median, and the standard deviation and the interquartile range, deriving asymptotic distributions of these quantities. Edgeworth(1887) proposes a method for minimizing average absolute deviations, in which medians are substituted for weighted averages in the calculations. Rhodes (1930) gives a more detailed description of Edgeworth's method. Harris (1950) relates the method to linear programming, and gives a clear explanation of it. See also Wagner(1959)

Jackson (1924) shows that there is a unique solution to fitting linear equations minimizing the average pth power of error for every $p > 1$.

Jeffreys(1932) handles outliers by allowing errors to be distributed as a mixture of normals, and provides a method of solution for determining the 5 parameters of the mixture.

Paulson(1940) finds the distribution of the median of a sample of size $2n+1$ from an arbitrary distribution. Wilson(1940) identifies cases where the usual asymptotic formula for the standard error of the median is incorrect.

Nair(1948) studies the distribution of the maximum deviation of observations from the arithmetic mean, and of the same quantity divided by the standard deviation, for the purpose of evaluating a criterion for rejection of outliers.

Tukey(1949) studies the behaviour of various estimators when samples are drawn from normal mixtures with the same mean. In particular he considers trimmed means, in which a fixed fraction of observations is dropped from either end of the sample.

Dixon(1960) studies Winsorized means, in which suspected outliers are replaced by the next largest order statistic,(or next smallest if the outlier is small), rather than being rejected entirely.

Ferguson(1961) considers locally optimal tests for identifying outliers, and compares these procedures to previously described methods.

Tukey(1962) asserts that neither mean nor variance is likely to be a wise choice for making estimates from a large sample. He suggests trimmed or Winsorized means, truncated variances or mean deviations.

Huber(1964) shows that the minimax estimator when a normal error is contaminated by an arbitrary symmetrically distributed error is the maximum likelihood estimator for the Huber density

$$f(x) = C[\exp(-x^2/2)\{|x| \le k\} + \exp(k|x| - k^2/2)\{|x| > k\}]$$

The end of least squares! Stein(1956) showed that least squares was inadmissible for normal errors in more than 2 dimensions.. that is there are other procedures that have uniformly smaller risk,whatever the unknown parameter values.. And yet, people still use it.