Gauss on least-squares and maximum-likelihood estimation

Jan R. Magnus

Department of Econometrics and Data Science,
Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

October 25, 2021

Abstract: Gauss' 1809 discussion of least squares, which can be viewed as the beginning of mathematical statistics, is reviewed. The general consensus seems to be that Gauss' arguments are at fault ('circular and non sequitur' in Stigler's (1986, p./ 141) words), but we interpret them as the (correct) scribbles from a genius. As such they provide a unique insight into the way Gauss' mind worked.

Key words: Gauss, least squares, maximum likelihood, history

An old question in probability theory is: suppose we throw with two fair dice, how many times do we need to throw so that the probability of at least one double-6 is at least 1/2. In the first half of the seventeenth century, the Chevalier de Méré, a well-known gambler, thought that he needed 24 throws.[1] This problem has become famous because it intrigued Pascal and Fermat, and the solution is contained in a letter of Pascal to Fermat dated July 29, 1654. With some caution we can take 1654 as the birth year of probability theory.

Mathematical statistics is much younger and, with similar caution, we select as its beginning the publication of Gauss' famous 1809 monograph.

[1] The correct answer is 25 and is obtained by solving $1 - (35/36)^n = 1/2$. This gives $n = 24.6$.

Legendre (1805) had published his method of least squares four years earlier, but he developed his method as an approximation tool and no randomness is assumed. Gauss (1809), in contrast, works in the context of random variables and distributions.[2] In modern notation, he starts with the linear model

$$y = X\beta + u,$$

where he assumes that the errors u_i are independent and identically distributed with mean zero and common variance σ^2, which we set at 1, since its value plays no role in the analysis. Since the u_i are independent and identically distributed, they have a common density function, say $\phi(u_i)$, and the logarithm of the joint density becomes $\sum_{i=1}^{n} \log \phi(u_i)$. Gauss wishes to estimate β by maximizing the joint density.

Gauss is aware of the fact that *if* he assumes normality of the errors, then the joint density will be of the form

$$\sum_{i=1}^{n} \log \phi(u_i) = a - b \sum_{i=1}^{n} u_i^2,$$

so that (under normality) maximizing the likelihood is the same as minimizing the sum of squared deviations. In other words, under normality, the maximum likelihood estimator is equal to the least-squares formula. This is a standard result in every first course in econometrics, and Gauss presented it first.

But Gauss did *not* want to assume at the outset that ϕ is the standard-

[2] I have freely drawn on the excellent historical texts by Pearson (1978), Stigler (1986), and Gorroochurn (2016).

normal density. Instead he wants to show that normality of the errors is not only sufficient but also necessary for the maximum likelihood estimator to be equal to the least-squares formula. In this attempt he fails, as many commentators have noted. His development (book II, paragraphs 175 and following) should be read as the scribbles of a genius, perhaps not entirely correct but nevertheless persuasive. In a sense, these scribbles provide an insight into his mind which a polished development might not have.

Let's try to follow Gauss' train of thought. He begins by making two simplifying assumptions. First, he considers the special case where $X = \imath$, the vector of ones (so that we only have a constant term in the model). Then β is a scalar and $u_i = y_i - \beta$. To maximize with respect to β we must have

$$\sum_{i=1}^{n} \phi'(u_i)/\phi(u_i) = 0.$$

Second, he simplifies by assuming that the y_i can only take two distinct values, namely y_1 and y_2, under the restriction that

$$y_2 = y_3 = \cdots = y_n.$$

Then, letting $\alpha = (y_1 - y_2)/n$, he obtains

$$y_1 - \bar{y} = (n-1)\alpha, \qquad y_2 - \bar{y} = -\alpha.$$

At this point he makes his third and final assumption, namely that the optimum in this special case is attained at $\hat{\beta} = \bar{y}$. Under these three assumptions we have

$$\sum_{i=1}^{n} \frac{\phi'(u_i)}{\phi(u_i)} = \frac{\phi'[(n-1)\alpha]}{\phi[(n-1)\alpha]} + \frac{(n-1)\phi'(-\alpha)}{\phi(-\alpha)} = 0,$$

which can also be written as

$$\frac{\phi'[(n-1)\alpha]}{[(n-1)\alpha]\phi[(n-1)\alpha]} = \frac{\phi'(-\alpha)}{(-\alpha)\phi(-\alpha)} = k,$$

where k is some constant. Solving the differential equation $\phi'(x) = kx\phi(x)$ gives

$$\phi(x) = A\exp(kx^2/2)$$

for some constant A. For ϕ to be a proper distribution, k must be negative (in fact, $k = -1$ since we have assumed that $\sigma = 1$) and the constant A had been determined a few years earlier by Laplace as $A = 1/\sqrt{2\pi}$, as gracefully acknowledged by Gauss.[3] Based on these simplifying assumptions, Gauss concludes that normality of the error distribution is a reasonable assumption, which is the correct conclusion. Gauss did not prove that normality of the errors is a necessary condition (and it isn't), but he did prove that normality of the errors is a reasonable condition.

Gauss' argument is not a 'proof' as it rests on a few rather dubious assumptions, as he realized himself. Not completely happy with his assumptions, Gauss (1823) considered the same model again. This time he asked a different question, namely: which linear unbiased estimator has the smallest variance? This result is what we now call the Gauss–Markov theorem.

[3] Gauss uses epithets for his colleagues, typically 'clarissimus'. Laplace is 'illustrissimus'. Only Newton is 'summus'.

References

Gorroochurn, P. (2016). *Classic Topics on the History of Modern Mathematical Statistics.* Hoboken, NJ, USA: John Wiley.

Pearson, K. (1978). *The History of Statistics in the 17th and 18th Centuries Against the Changing Background of Intellectual, Scientific and Religious Thought. Lectures by Karl Pearson given at University College London during academic sessions 1921–1933* (Edited by E. S. Pearson). London, UK: Charles Griffin.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, Massachusetts, USA and London, UK: The Belknap Press of Harvard University Press.

Gauss, C. F. (1809). Theoria Motus Corporum Coelestium. Perthes et Besser, Hamburg. (Translated as "Theory of Motion of the Heavenly Bodies Moving About the Sun in Conic Sections" by C.H. Davis (1857). Little, Brown, Boston. Reprinted in 1963, Dover, New York).

Gauss C F. (1823). Theoria Combinationis Observationum Erroribus Minimis Obnoxia. Dieterich, Göttingen. (Translated as "Theory of the Combination of Observations Least Subject to Errors" by G.W. Stewart (1995). SIAM, Philadelphia).

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