Gauss on least-squares and maximum-likelihood estimation[1]

Jan R. Magnus

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

December 18, 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, but we show that his reasoning is in fact correct, given his self-imposed restrictions, and persuasive without these restrictions.

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

JEL Classification: B16, C10

An old question in probability theory is the following: 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. 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.[2] With some caution we can take 1654 as the birth year of probability theory. It took a while to understand the basic rules of probability and sixty years later, in a letter to the Swiss philosopher and mathematician Louis Bourguet dated March 22, 1714, Leibniz still maintained that is was equally likely to throw twelve with two dice than to throw eleven, because "l'un et l'autre ne ce peut faire que d'une seule manière" (one or the other can be done in only one way).

Mathematical statistics is much younger and, with similar caution, we select as its beginning the publication of Gauss' famous 1809 monograph.[3] Legendre (1805) had published his method of least squares four years earlier,

[1]I am grateful to Jan van Maanen, Chris Muris, Franco Peracchi, and Steven Tijms for useful comments.

[2]The correct answer is 25 and is obtained by showing that the equation $1-(35/36)^n = 1/2$ has the solution $n \approx 24.6$.

[3]Others would choose Laplace (1774) as the beginning of statistics, which is equally reasonable.

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; see e.g. Pearson (1978), Stigler (1986), and Gorroochurn (2016) for historical details.

Some satisfaction seems to be derived in finding mistakes in the writings of great minds, and Leibniz' error is quoted frequently. Rather than laughing at Leibniz' mistake, we should realize just how difficult the beginnings of probability theory were, and that things that we now consider easy are not easy because we are so clever but because they have sunk into common knowledge.

Similarly, most Gauss commentators have found his 1809 treatment of least squares at fault. For example, Stigler (1986, pp. 141–143) considers it a "logical aberration ... essentially both circular and non sequitur" and Gorroochurn (2016, p. 163) writes that "his reasoning contains an inherent circularity because the normal distribution emerges as a consequence of the postulate of the arithmetic mean, which is in fact a consequence of the normality assumption!" The purpose of this note is to demonstrate that it is not Gauss who is at fault but his commentators.

In modern notation, Gauss starts with the linear model

$$y = X\beta + u, \tag{1}$$

where he assumes that the errors u_i are independent and identically distributed (iid) with mean zero and common variance σ^2, which we set equal to one without loss of generality. Since the u_i are iid, 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. In other words, he wants to derive the maximum-likelihood estimator for β.

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

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

so that (under normality) maximizing the likelihood is the same as minimizing the sum of squared deviations. Gauss makes life unnecessarily difficult for himself by working in a Bayesian framework, assuming a flat bounded prior for each of the β_j, so that the posterior also has bounded support. But in essence, Gauss showed (for the first time) that in the standard linear model under normality the maximum-likelihood estimator is equal to the least-squares formula.

This is an important result in itself and Gauss could have stopped there. But he did *not* want to assume at the outset that the errors are normally distributed. 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, not because his argument is wrong (as most Gauss scholars seem to believe), but because his (correct) argument is not general, which he fully realizes.

Let us reexamine his argument. Gauss (1809, book II, section III, §177) proves the following result (in modern notation).

Proposition (Gauss, 1809): *Let $y_1, y_2, \ldots, y_n$ ($n \geq 3$) be a sequence of independent and identically distributed observations from an absolutely-continuous distribution with $\mathrm{E}(y_i) = \mu$ and $\mathrm{var}(y_i) = 1$. Assume that the n realizations of y_i take only two values with frequencies n_1 and n_2, respectively ($n_1 \geq 1$, $n_2 \geq 1$, $n_1 \neq n_2$). Then, the average $\bar{y}$ is the maximum-likelihood estimator of μ if and only if the y_i are normally distributed.*

Before we prove the proposition, some comment is in order on Gauss' assumption that the n realizations of y_i take only two values. This seems to contradict the fact that the y_i follow an absolutely-continuous distribution. Of course, there is a difference between observations (random variables) from an absolutely-continuous distribution and observations (the realized values). Some statistical concepts have two terms (estimator, estimate; predictor, prediction) to emphasize this difference, but most (like observation) don't. The random variables follow an absolutely-continuous distribution, but the realizations take on specific values, and Gauss assumes that they take one or the other of two values. This is a rather heroic assumption, but it is not inconsistent or wrong. Gauss himself simply says *supponendo itaque* (by supposing therefore) as if this were a logical continuation of his argument, and provides no further comment.

To prove the proposition, Gauss argues as follows. Let $u_i = y_i - \mu$. Since the u_i are iid, they have a common density function, say $\phi(u_i)$. First assume that ϕ is the standard-normal density. Then the loglikelihood $L(\mu)$ can be written as in (2). This is maximized if and only if the sum of squares is minimized, which occurs when $\sum_i (y_i - \mu) = 0$, that is when $\hat{\mu} = \bar{y}$. Note that the additional assumption on the realizations of y_i is not required.

Now assume that $\hat{\mu} = \bar{y}$. Gauss needs to show that this implies that ϕ is the standard-normal density. As assumed, the y_i can only take two distinct values, say z_1 (n_1 times) and z_2 (n_2 times), where $n = n_1 + n_2$. Then, letting

$$d = z_1 - z_2, \qquad r = n_1/n, \tag{3}$$

he obtains

$$y_i - \bar{y} = \begin{cases} d(1-r) & \text{if } y_i = z_1, \\ -dr & \text{if } y_i = z_2. \end{cases} \tag{4}$$

Setting $L'(\mu) = 0$ then gives

$$L'(\mu) = \sum_{i=1}^{n} \frac{\phi'(u_i)}{\phi(u_i)} = \frac{n_1 \phi'[d(1-r)]}{\phi[d(1-r)]} + \frac{n_2 \phi'(-dr)}{\phi(-dr)} = 0, \tag{5}$$

which can be rewritten as

$$f[d(1-r)] = f[-dr], \qquad f(x) = \frac{\phi'(x)}{x\phi(x)}. \tag{6}$$

For each given value of r ($0 < r < 1$, $r \neq 1/2$), this has to hold for every value of d, and it is easy to see (*unde facile colligitur* in Gauss' words) that this implies that f is a constant. (We have to exclude $r = 1/2$ because this would only imply that f is symmetric around zero.) Hence, we must solve the equation $\phi'(x) = -k\, x\phi(x)$, where k is a constant. The solution to this differential equation is

$$\phi(x) = A \exp(-kx^2/2) \tag{7}$$

for some constant A.[4] Since ϕ represents a distribution it must integrate to one which implies that the constant A takes the value $A = \sqrt{k/(2\pi)}$, as proved a few decades earlier by Laplace (1774) in a *theorema elegans*, a fact gracefully acknowledged by Gauss.[5] In our case, $\sigma = 1$ and hence $k = 1$. Hence ϕ is the standard-normal distribution, and the proof is complete.

The presented proof follows Gauss' argument closely except that he sets $n_1 = 1$ and $n_2 = n - 1$ (and tacitly assumes that $n \geq 3$). The proposition tells us how far Gauss came into proving the necessity of the normality assumption. The answer is: not very far, because his conditions are rather restrictive. Two centuries later we can get a little further. In particular, Kagan et al. (1973, Theorem 7.4.1), building on an earlier result in Kagan et al. (1965), established that, in general, linear estimators of location parameters are admissible if and only if the random variables are normally distributed;

[4]The technique for solving ordinary differential equations was well-established since Leibniz' work in the early 1690s, which is why Gauss does not provide a reference; see Katz (2009, pp. 585–586) for historical details.

[5]Gauss often uses epithets for his colleagues, typically *clarissimus*. Laplace is *illustrissimus* (Gauss abbreviates ill.). Only Newton is *summus*.

and they applied the approach through admissibility to the linear model in Kagan et al. (1973, Section 7.7).

To link the linear model $y = X\beta + u$ to the proposition, Gauss thus makes three simplifying assumptions:

1. The design matrix X has only one column, namely the vector of ones, so that we only have a constant term in the model, there is only one β to estimate, and $u_i = y_i - \beta$.

2. The realizations of y_i only take two distinct values, say z_1 (n_1 times) and z_2 (n_2 times), where $n = n_1 + n_2$ and $n_1 \neq n_2$.

3. The optimum is attained at $\hat{\beta} = \bar{y}$.

The third assumption is a perfectly reasonable assumption as we are considering iid random variables y_i with common mean β and common variance. So, unless we expect Gauss to discuss shrinkage estimators, what alternative is there to estimate β? Under these three assumptions, Gauss shows that the y_i must be normally distributed.

After establishing the proposition, Gauss argues that it is thus *reasonable* to assume normality. This is a qualitative statement which can be challenged, but it is not incorrect. Gauss was primarily interested in the justification of least squares, not in pushing the normal distribution, and he fully realized that his qualitative jump from the special to the general case was not mathematically solid. Not completely happy with his restrictive 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 resulted in what we now call the Gauss–Markov theorem and it does not rely on normality of the errors.

References

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica*, 78(1), 159–168.

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).

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

Kagan, A. M., Yu. V. Linnik, and C. R. Rao (1965). On a characterization of the normal law based on a property of the sample average. *Sankhya, Series A*, 27, 405–406.

Kagan, A. M., Yu. V. Linnik, and C. R. Rao (1973). *Characterization Problems in Mathematical Statistics* (translated from the Russian by B. Ramachandran). John Wiley, New York.

Katz, V. J. (2009). *A History of Mathematics*, 3rd ed. Addison-Wesley, Reading, MA.

Laplace, P. S. (1774). Mémoire sur la Probabilité des Causes par les Évènemens. Oeuvres Complètes, vol. 8, 27–65. (Translated as "Memoir on the Probability of Causes of Events" by S. Stigler (1986), *Statistical Science*, 1, 364–378.)

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

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). Charles Griffin, London.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. The Belknap Press of Harvard University Press, Cambridge, MA.