A Historical Note on the Method of Least Squares

Author(s): R. L. Plackett

Source: *Biometrika*, Dec., 1949, Vol. 36, No. 3/4 (Dec., 1949), pp. 458-460

Published by: Oxford University Press on behalf of Biometrika Trust

Stable URL: https://www.jstor.org/stable/2332682

If there is more than one unknown parameter, and the variance v of the observations z_i is finite but not a known function of the mean θ, v must be estimated as sampling proceeds, and a boundary cannot be specified in advance. Let us suppose that v can be estimated from a sample of m observations by an estimate $\hat{v}$ that is asymptotically normal with variance $O(m^{-1})$ as $m \to \infty$. When, say, half the sampling has been done that will ultimately be seen to be needed, $\hat{v}$ satisfies in probability the relation

$$\hat{v} = v(1 + O(n^{-\frac{1}{2}})), \tag{23}$$

and so the eventual sample size n and estimate $\hat{\theta}$ of θ satisfy

$$\text{var}(\hat{\theta}) \sim \hat{v}/n. \tag{24}$$

Thus the required value of n can be predicted at this half-way stage with asymptotic validity.

I am indebted to Mr D. V. Lindley for helpful criticism. The treatment given above is somewhat heuristic in places, notably in the argument used in arriving at equation (11). It seems to me, however, that the line of approach may prove to be worth pursuing.

REFERENCES

BLACKWELL, D. & GIRSHICK, M. A. (1947). A lower bound for the variance of some unbiased sequential estimates. *Ann. Math. Statist.* **18**, 277.

FINNEY, D. J. (1949). On a method of estimating frequencies. *Biometrika*, **36**, 233.

HALDANE, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222.

STEIN, C. & WALD, A. (1947). Sequential confidence intervals for the mean of a normal distribution with known variance. *Ann. Math. Statist.* **18**, 427.

TWEEDIE, M. C. K. (1945). Letter in *Nature, Lond.*, **155**, 453.

A historical note on the method of least squares

BY R. L. PLACKETT, *University of Liverpool*

1. The purposes of this note are:

(i) to summarize the justifications by Laplace, Gauss and Markoff of the method of least squares;

(ii) to suggest that Gauss was the first who justified least squares as giving those linear estimates which are unbiased of minimum variance;

(iii) to modernize and extend his proof to cover a general theorem due to Aitken.

It is not my object to provoke controversy, and I have attempted to indicate where a personal opinion is intended.

2. The method of least squares has been in use now for over 150 years. During the nineteenth century the writings of Todhunter (1865), Merriman (1877) and others gave the impression that Laplace (1812–20, collected works 1886) was largely responsible for putting the method on a theoretical basis by means of the calculus of probability, whereas the contribution of Gauss (1821, collected works 1873) was minimized or ignored. Lately, the emphasis has changed, and in recent papers and text-books Markoff (1912) is credited with justifying the method without superfluous assumptions of normality. For these reasons, it seems desirable to disentangle the various justifications proposed and to allot credit in due proportion

3. In general let $\boldsymbol{\theta}(s \times 1)$ be a vector of unknown parameters, $\mathbf{x}(n \times 1)$ a vector of observations, $\boldsymbol{\epsilon}(n \times 1)$ a vector of errors and $A(n \times s)$ a matrix of known quantities; so that

$$A\boldsymbol{\theta} - \mathbf{x} = \boldsymbol{\epsilon}.$$

Further, suppose that $W(n \times n)$ is a diagonal matrix whose elements are the reciprocals of the error variances. It is required to form an estimate $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$. The method of least squares leads to estimates which satisfy

$$A'WA\boldsymbol{\theta}^* = A'W\mathbf{x}.$$

Neither Laplace nor Gauss used matrix notation, but their results can immediately be written in that form.

4. Laplace (1812–20) discusses the method of least squares in Book 2, Chapter 4, and in the first three Supplements. He proves a series of results which are summarized—I hope fairly—in the following:

THEOREM. *Among all $s \times n$ matrices F leading to estimates of the form $FA\boldsymbol{\theta}^* = F\mathbf{x}$, the expected values of the elements of $|\boldsymbol{\theta}^* - \boldsymbol{\theta}|$ are minimized as $n \to \infty$ when $F = \mu A'W$, μ being an arbitrary multiplier.*

The proof is long but runs on these lines: if $\mathbf{u}$ is the error of $\boldsymbol{\theta}^*$ then $FA\mathbf{u} = F\boldsymbol{\epsilon}$; Laplace proceeds to determine the joint characteristic function of $F\boldsymbol{\epsilon}$ and deduces that when all errors have the same distribution, symmetrical about zero, $F\boldsymbol{\epsilon}$ has a multivariate normal distribution as $n \to \infty$; whence $\mathbf{u}$ also has a multivariate normal distribution, and the expected values of the elements of $|\mathbf{u}|$ are determined; finally, he shows that $F = \mu A'W$ implies the vanishing of the differential coefficients of these expected values with respect to the elements of F.

In more detail, Laplace first takes $s = 1$ and maximizes the probability that his estimate lies between given limits; he then notes that this is the same as minimizing $\mathscr{E}|\theta^* - \theta|$, and continues to use this criterion when $s = 2$, stating that the result can be extended to greater values of s. In the first Supplement he considers the possibility of a bias in the observations and suggests its removal by introducing an additional parameter whose coefficient is unity in all equations.

5. Gauss presented his justification in 1821. The paper is written in Latin, but a French translation was published by Bertrand in 1855 and the fundamental theorem incorporated in Bertrand's own book of 1888. In the early sections of his paper, Gauss also considers the possibility of bias in his observations and makes it clear that the preferred estimates are those with minimum variance, although of course he does not use this terminology. He begins with errors of differing variance, and by choosing suitable multipliers presents the equations in a form where the errors have the same variance. The proof of the following theorem is in Art. 20; it is implicit that he is seeking unbiased estimates:

THEOREM. *Among all the systems of coefficients $B(s \times n)$ which give $B\boldsymbol{\epsilon} = \boldsymbol{\theta} - \boldsymbol{\theta}\dagger$, the estimate $\boldsymbol{\theta}\dagger$ being independent of $\boldsymbol{\theta}$, those for which the diagonal elements of BB' are minimized are provided by the method of least squares.*

Put $$\boldsymbol{\xi} = A'\boldsymbol{\epsilon} \quad \text{so that} \quad \boldsymbol{\xi} = A'A\boldsymbol{\theta} - A'\mathbf{x}.$$

The solution of these equations is $\boldsymbol{\theta} = \boldsymbol{\theta}^* + D\boldsymbol{\xi}$, and with $DA' = E$ this becomes

$$E\boldsymbol{\epsilon} = \boldsymbol{\theta} - \boldsymbol{\theta}^* \quad \text{so} \quad (\boldsymbol{\theta}^* - \boldsymbol{\theta}\dagger) = (B - E)\boldsymbol{\epsilon}.$$

If this is true for all $\boldsymbol{\theta}$ then $(B - E)A = 0$, i.e. on post-multiplying by D', $(B - E)E' = 0$, which implies $BB' = EE' + (B - E)(B - E)'$. It is now clear that the diagonal elements of BB' are minimized when $B = E$.

6. Matrix notation has been adopted for brevity in the preceding sections, but no matrix theorems have been assumed. Taking for granted the now familiar properties of matrices regarding associative products and inverses, the preceding demonstration can be modernized and shortened.

If $\boldsymbol{\theta}^* = B\mathbf{x}$ is unbiased for all $\boldsymbol{\theta}$, then $BA = I$. With $C = A'A$ it follows that $C^{-1} = BAC^{-1}$, so

$$BB' = (C^{-1}A')(C^{-1}A')' + (B - C^{-1}A')(B - C^{-1}A')',$$

i.e. the diagonal elements of BB' are least when $B = C^{-1}A'$, which is the solution provided by least squares.

7. Markoff (1912) devotes Chapter 7 of his book to the method of least squares. He states that each observation is to be considered as a particular case of many, and as an unbiased estimate of some linear function of the unknown parameters. His determination of unbiased estimates of these parameters having minimum variance is closely followed in the paper by David & Neyman (1938).

8. Aitken (1934) has extended the theorem of Gauss by proving that with a known matrix V of variances and covariances of the observations, the minimum of $(A\boldsymbol{\theta} - \mathbf{x})'V^{-1}(A\boldsymbol{\theta} - \mathbf{x})$ provides estimates $\boldsymbol{\theta}^*$ such that $\boldsymbol{\varphi}^* = P\boldsymbol{\theta}^*$ is an unbiased estimate of $\boldsymbol{\varphi} = P\boldsymbol{\theta}$ with minimum variance. Gauss's method can be used to prove this also.

If $\boldsymbol{\varphi}^* = B\mathbf{x}$ then $BA = P$ and

$$[B]\,V[P(A'V^{-1}A)^{-1}A'V^{-1}]' = [P(A'V^{-1}A)^{-1}A'V^{-1}]\,V[P(A'V^{-1}A)^{-1}A'V^{-1}]',$$

consequently

$$BVB' = [P(A'V^{-1}A)^{-1}A'V^{-1}]\,V[P(A'V^{-1}A)^{-1}A'V^{-1}]' + [B - P(A'V^{-1}A)^{-1}A'V^{-1}]\,V[B - P(A'V^{-1}A)^{-1}A'V^{-1}]'.$$

If we consider the diagonal elements here, the second term on the right gives a positive definite quadratic form, so minimum variance is attained when

$$B = P(A'V^{-1}A)^{-1}A'V^{-1},$$

the solution given by the method indicated above.

9. It is therefore my opinion that Laplace and Gauss proved theorems which are quite different; that the justification given by Gauss is preferable; and that Markoff, who refers to Gauss's work, may perhaps have clarified assumptions implicit there but proved nothing new. It is evident that Gauss's proof is valid for all values of n, entirely free from any assumption of normality, and capable of immediate development.

REFERENCES

AITKEN, A. C. (1934). On least squares and linear combination of observations. *Proc. Roy. Soc. Edinb.* A, **55**, 42–7.

BERTRAND, J. (1888). *Calcul des probabilités*. Paris.

DAVID, F. N. & NEYMAN, J. (1938). Extension of the Markoff theorem on least squares. *Statist. Res. Mem.* **2**, 105–16.

GAUSS, C. F. (1855). *Méthode des moindres carrés* (trans. J. Bertrand). Paris.

GAUSS, C. F. (1873). Theoria combinationis observationum erroribus minimis obnoxiae. Pars prior. *Werke*, Band **4**. Göttingen.

LAPLACE, P. S., Marquis de (1886). *Théorie analytique des probabilités*, 3rd edition, Oeuvres, **7**. Paris.

MARKOFF, A. A. (1912). *Wahrscheinlichkeitsrechnung* (trans. H. Liebmann), 2nd edition. Leipzig and Berlin.

MERRIMAN, M. (1877). A list of writings relating to the method of least squares, with historical and critical notes. *Trans. Conn. Acad. Arts Sci.* **4**, 151–232.

TODHUNTER, I. (1865). *A History of the Mathematical Theory of Probability*. Macmillan.

The characteristic function of a weighted sum of non-central squares of normal variates subject to s linear restraints

BY G. I. BATEMAN

1. Suppose $x_1, x_2, \ldots, x_n$ are independent normal variables with expectations $a_1, a_2, \ldots, a_n$ respectively and with unit variance, that is to say, we suppose that

$$p(x_j) = \frac{1}{\sqrt{(2\pi)}} \exp[-\tfrac{1}{2}(x_j - a_j)^2], \quad (j = 1, 2, \ldots, n).$$

We consider a weighted non-central sum of squares of the type

$$\psi'^2 = \sum_{j=1}^{n} c_j x_j^2,$$

where the x's are as defined and the c_j $(j = 1, 2, \ldots, n)$ are constants. It is assumed that the x_j are subject to s linear restraints

$$\sum_{j=1}^{n} b_{lj}\, x_j = \rho_l \quad (l = 1, 2, \ldots, s),$$

b_{lj} and ρ_l $(l = 1, \ldots, s;\ j = 1, 2, \ldots, n)$ being given constants. The characteristic function of the joint distribution of $\psi'^2, \rho_1, \rho_2, \ldots, \rho_s$ may be written down immediately. We have

$$\phi(t, t_1, \ldots, t_s) = \prod_{j=1}^{n} \left[\int_{-\infty}^{+\infty} \frac{1}{\sqrt{(2\pi)}} \exp\left\{ -\tfrac{1}{2}(x_j - a_j)^2 + itc_j x_j^2 + \sum_{l=1}^{s} it_l b_{lj} x_j \right\} dx_j \right].$$

The evaluation of the integral is straightforward and we obtain

$$\phi(t, t_1, \ldots, t_s) = \prod_{j=1}^{n} (1 - 2itc_j)^{-\frac{1}{2}} \exp\left(it \sum_{j=1}^{n} \frac{c_j a_j^2}{1 - 2itc_j} - \frac{1}{2} \sum_{l=1}^{s} \sum_{m=1}^{s} A_{lm} t_l t_m + i \sum_{l=1}^{s} B_l t_l \right), \tag{1}$$

where

$$A_{lm} = \sum_{j=1}^{n} \frac{b_{lj} b_{mj}}{1 - 2itc_j} \quad \text{and} \quad B_l = \sum_{j=1}^{n} \frac{a_j b_{lj}}{1 - 2itc_j}. \tag{2}$$