The Basis for the Criterion of Least Squares

E. RICHARD COHEN

Atomic Energy Research Department, North American Aviation, Inc., Downey, California

THERE have been many who have objected to the use of the method of least squares on the grounds that it assumes the Gaussian curve for the probability distribution of the error in any particular measurement. This is the formulation which is indicated in Sec. 3 of the preceding article.[1] The objection is indeed valid that the assumption of the Gaussian distribution is often unwarranted in many experimental configurations. However, the problem of determining a "best" set of values which can be computed from an overdetermined system of equations is essentially the problem of determining an analytic basis on which one can define the adjective "best." The condition of least squares serves as one such analytic criterion. This says nothing in itself of what could be called the *physical* interpretation of the criterion. It is recognized, in general, that the method of least squares corresponds to the "Axiom of Maximum Likelihood," if the distribution functions of all the errors are Gaussian. Gauss himself was able to justify the method on a much wider base, and in 1821 he published a theory which replaces this axiom with an "Axiom of Minimum Error" or "Axiom of Maximum Weight."[2] The definition of "best" is not to be made on the basis of that solution which is most likely to be correct, but on that solution which is most accurate and to which can, therefore, be attached the greatest statistical weight. (The statistical weight of a statistical variable is defined in the present sense as the reciprocal of the variance, i.e., the reciprocal of the mean square error, of the quantity.)

Consider then an overdetermined set of N_{Eq} equations expressing relationships between q variables. We can find an infinite number of solutions for the q variables depending on how we choose to combine the equations. We can think of this process as one in which some set of $q-1$ of the equations are used to express $q-1$ of the variables in terms of one particular variable, say x_1. These values are then to be substituted into the remaining $N_{Eq}-q+1$ equations to give a set of $N_{Eq}-q+1$ values for the variable x_1. Any of these numerical values is a possible choice for the variable x_1 and, in general, we would want to take some weighted average of these numerical values. The numerical value of such an average depends both on the weights attached to the elements which go to make up the average and on the numerical values of these elements. Ultimately, therefore, the value ascribed to x_1 depends on $N_{Eq}-q$ independent parameters which specify the *mathematical form* of the average (there is one condition to restrict the choice of the $N_{Eq}-q+1$ weights; this is the condition that their sum is unity) and N_{Eq} quantities which are either the observational numerics or quantities directly deduced from them (such as relative deviations from a set of origin values) which determine the *numerical value* of the average.

The numerical value of x_1 is thus expressed as a linear combination of N_{Eq} numerical quantities, each of which has associated with it a mean square error. If the numerical quantities are observationally independent, we can assert that the mean square error of x_1 is the sum of certain coefficients, depending on the $N_{Eq}-q$ free parameters, times the mean square errors of the N_{Eq} observational numerics. The Axiom or Condition of Minimum Error states that the "best" choice for x_1 is that one whose error is a minimum with respect to the possible variation of the free parameters. We shall show below that this condition is equivalent to the condition of least squares, although the results cannot, in general, be identified as corresponding to that set which has maximum likelihood, except in the case when the distribution function for the errors is specified to be Gaussian. This is, however, an advantage for one can easily construct distributions (for example, rectangular distributions) for which the condition of maximum likelihood has no unique solution. Furthermore, the development of the condition of minimum error is quite general in regard to the forms of the error distribution functions; all that is specified is the mean square error, so that the range of applicability of the theory of least squares is extended from Gaussian distributions to the much larger class of distributions with finite second moments.

AN EXAMPLE

As an elementary example of the method and in order to clarify the concepts involved, let us consider a problem which is perhaps the simplest possible example. We have two measurements of the quantity x; these two measurements are a_1 and a_2, in general, $a_1 \neq a_2$. We also assume that each measurement represents a single selection from a universe of values. We let the probability distribution of the first measurement be $P_1(\xi)$ such that the probability is $P_1(\xi)d\xi$, that the result of a measurement of the quantity x by the specified pro-

[1] J. W. M. DuMond and E. R. Cohen, Revs. Modern Phys. **25**, 691 (1953).

[2] E. Whittaker and G. Robinson, *Calculus of Observations* (Blackie & Sons, London, 1944), fourth ed., p. 224.

cedure shall lie between the values ξ and $\xi+d\xi$. Similarly, the second measurement of x is to be characterized by the probability distribution $P_2(\eta)$. We need not make any detailed specification of the form of the distribution functions P_1 and P_2; it is not even necessary that the two functions have similar form. We impose upon them only the restriction that the distributions have finite second moments.

If we have two measurements of x then we can take some average value to use as the "best" choice. There are, however, several different choices for this average and we can write

$$x_0(\alpha)=\alpha a_1+(1-\alpha)a_2, \tag{1}$$

where α is any real number, although intuitively we would prefer that $0\leq\alpha\leq 1$ because this is the condition that x_0 lies between the values a_1 and a_2. Defined in this way, the mean of the universe of values of x_0 is

$$\begin{aligned}\bar{x}_0&=\int\int[\alpha\xi+(1-\alpha)\eta]P_1(\xi)P_2(\eta)d\xi d\eta\\&=\alpha\int\xi P_1(\xi)d\xi\cdot\int P_2(\eta)d\eta\\&\quad+(1-\alpha)\int P_1(\xi)d\xi\cdot\int\eta P_2(\eta)d\eta.\end{aligned}\tag{2}$$

Now each probability distribution is assumed to be normalized so that

$$\int P_1(\xi)d\xi=1;\quad\int P_2(\eta)d\eta=1, \tag{3}$$

and furthermore, if there are no systematic errors in either measurement, the expectation value of each measurement is x

$$\int\xi P_1(\xi)d\xi=x;\quad\int\eta P_2(\eta)d\eta=x. \tag{4}$$

Therefore, Eq. (2) reduces to

$$\bar{x}_0=\alpha x+(1-\alpha)x=x, \tag{5}$$

so that the expectation value of the average is indeed the quantity we are trying to measure independent of the value of the parameter α which determines the particular average. But we now ask the question, "How accurate is this average; what is its standard deviation?" If we let ϵ^2 be the mean square deviation of the universe from which x_0 is extracted, we have, by the usual definition,

$$\begin{aligned}\epsilon^2&=\int\int[\alpha\xi+(1-\alpha)\eta-x]^2P_1(\xi)P_2(\eta)d\xi d\eta\\&=\int\int[\alpha(\xi-x)+(1-\alpha)(\eta-x)]^2P_1(\xi)P_2(\eta)d\xi d\eta\\&=\alpha^2\int(\xi-x)^2P_1(\xi)d\xi\int P_2(\eta)d\eta\\&\quad+(1-\alpha)^2\int P_1(\xi)d\xi\int(\eta-x)^2P_2(\eta)d\eta\\&\quad+2\alpha(1-\alpha)\int(\xi-x)P_1(\xi)d\xi\int(\eta-x)P_2(\eta)d\eta,\\\epsilon^2&=\alpha^2\int(\xi-x)^2P_1(\xi)d\xi+(1-\alpha)^2\int(\eta-x)^2P_2(\eta)d\eta.\end{aligned}\tag{6}$$

The two integrals in the last form of Eq. (6) are the variances (the mean square errors), respectively, of the first and of the second measurements. These quantities are defined to be σ_1^2 and σ_2^2. The expression for the error in the average in terms of the errors of the numbers entering into the average is, therefore,

$$\epsilon^2=\alpha^2\sigma_1^2+(1-\alpha)^2\sigma_2^2. \tag{7}$$

We find that, although the expectation value of the average is independent of α, the error in the average *is* a function of α: for $\alpha=0$ we have $\epsilon^2=\sigma_2^2$ and for $\alpha=1$ we have $\epsilon^2=\sigma_1^2$; we may reasonably ask whether a proper choice of α might not result in a value of ϵ^2 which is smaller than either of these, and indeed, what choice of α yields the minimum value of ϵ^2? By completing the square we can write Eq. (7) as

$$\begin{aligned}\epsilon^2&=\alpha^2(\sigma_1^2+\sigma_2^2)-2\alpha\sigma_2^2+\sigma_2^2\\&=\frac{\alpha^2(\sigma_1^2+\sigma_2^2)^2-2\alpha\sigma_2^2(\sigma_1^2+\sigma_2^2)+\sigma_2^4+\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\\&=\frac{\sigma_1^2\sigma_2^2+[\alpha(\sigma_1^2+\sigma_2^2)-\sigma_2^2]^2}{\sigma_1^2+\sigma_2^2}.\end{aligned}\tag{8}$$

Since α appears now only in a term which can never be negative, we see immediately that the minimum value of ϵ^2 occurs when the square bracket vanishes. This is achieved when $\alpha=\sigma_2^2/(\sigma_1^2+\sigma_2^2)$; the minimum value of ϵ^2 corresponding to this choice is

$$\epsilon_0^2=\sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2),$$

and the corresponding value of x_0 which we indicate by $\mathbf{x}_0$ is

$$\mathbf{x}_0=\epsilon_0^2\left(\frac{a_1}{\sigma_1^2}+\frac{a_2}{\sigma_2^2}\right). \tag{9}$$

The "weights" to be attached to measured quantities in order to compute the "best" average are therefore proportional to the reciprocal of the mean square error of each measurement. If we define the statistical weights

$$w_1=1/\sigma_1^2 \quad w_2=1/\sigma_2^2, \tag{10}$$

we have two important formulas

$$x_0=(w_1a_1+w_2a_2)/(w_1+w_2), \tag{9.1}$$

$$w_0=1/\epsilon_0^2=w_1+w_2. \tag{8.1}$$

The first of these essentially justifies the nomenclature of "statistical weight" for the reciprocal variance, since it is this quantity which determines the importance of the measurement in the computation of the average. The second formula shows that when this "best" average is obtained the weight of the result (computed as the reciprocal of the variance of the average) is just the sum of the weights of the components and, furthermore, that this is the maximum weight ascribable to any average. Any other linear combination of the two observations would have a weight which is less than the sum of the individual weights and is, therefore, an inefficient average which "wastes weight." The weight of the average w as a function of the weights assigned to the individual measurements and the choice of the average value is indicated in Fig. 1. The measurement x_1 has variance $\sigma_1^2=1/w_1$, and the measurement x_2 has variance $\sigma_2^2=1/w_2$. Each value of α corresponds to a particular value of x_0, and the statistical weight of this average value is a function not only of the parameter α but also of the statistical weights of x_1 and x_2. The statistical weight of the average is a maximum (equal to the sum of the weights of the individual components), when the weighting employed in computing the average is determined by the statistical weights of the components.

THE GENERALIZED THEORY*

We can now establish a generalized theory of least squares in which the individual equations are not observationally independent.[3] Let there be q variables which are connected by N equations. The numerical constants in these equations are interrelated and are determined by a set of n observationally independent quantities s_α. We shall assume that the equations can all be linearized so that the system is

$$\begin{aligned}
&a_{11}x_1+a_{12}x_2+a_{13}x_3+\cdots+a_{1q}x_q=c_1\\
&a_{21}x_1+a_{22}x_2+a_{23}x_3+\cdots+a_{2q}x_q=c_2\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&a_{N1}x_1+a_{N2}x_2+a_{N3}x_3+\cdots+a_{Nq}x_q=c_N,
\end{aligned} \tag{11}$$

where the constants c_μ are numerical quantities which are explicit functions of the quantities s_α.

In the discussion to follow it will be useful to introduce two conventions: One is the Einstein summation convention; in any product a repeated index is to be summed over all values permissible to it unless specifically indicated to the contrary. The other is the use of different characters to indicate the range of the index. Roman letters $(i, j, k, \cdots)$ shall be used for indices relating to the unknowns and, therefore, have the range

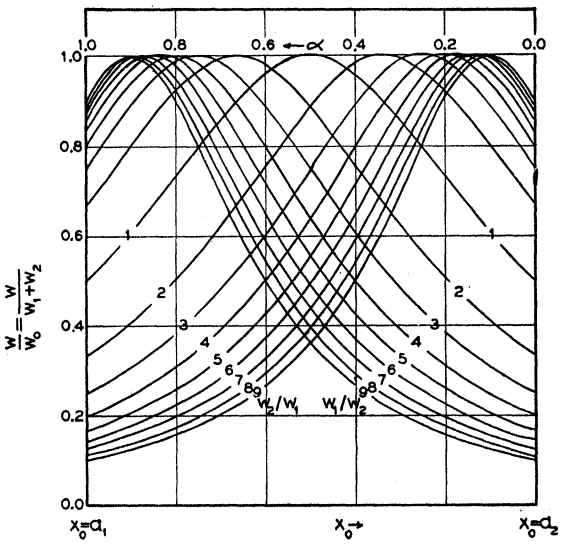

FIG. 1. The weight assignable to the average of two observed quantities as a function of the statistical weights of those quantities and the weighting parameter α of the average. The weight of the average is a maximum if α is determined by the statistical weights.

1 to q; Greek letters from the middle of the alphabet $(\mu, \nu, \sigma, \cdots)$ shall be used for indexes relating to the observational equations and have the range 1 to N; Greek letters from the start of the alphabet $(\alpha, \beta, \cdots)$ shall be used for indexes relating to the observationally independent variables and have the range 1 to n. It is thus possible to identify a quantity with two indexes as a square or a nonsquare matrix by noting whether the indices are of the same type or not.

Since the c_μ are functions of s_α the *statistically independent* variations ϵ_α in s_α will produce *statistically correlated* variations η_μ in c_μ:

$$\begin{aligned}
&\epsilon_\alpha=ds_\alpha,\\
&\eta_\mu=dc_\mu=A_{\mu\alpha}\epsilon_\alpha,
\end{aligned} \tag{12}$$

where $A_{\mu\alpha}=\partial c_\mu/\partial s_\alpha$ are the derivatives of the computed

---

* Note added in proof: Professor J. W. Tukey of Princeton University has kindly pointed out to the author that this generalized system of least squares is not new. R. L. Plackett, Biometrika **36**, 458 (1949) has traced the origins to Laplace and Markoff as well as to Gauss. A. C. Aitken, Proc. Roy. Soc. Edinburgh **55**, 42 (1938), has also presented a generalization equivalent to the present development.

[3] J. W. M. DuMond and E. R. Cohen, Report to the National Research Council on the Atomic Constants, December, 1950. See also E. R. Cohen, Phys. Rev. **81**, 162 (1951).

observational quantities with respect to the fundamental independently measured quantities s_α.

We wish to find the most accurate values of x_i which can be computed from Eq. (11). To find each x_i we need a set of multipliers $\lambda_\mu^{(i)}$ by which to multiply the N observational equations so that we obtain

$$x_i=\lambda_\mu^{(i)}c_\mu=\lambda_\mu^{(i)}a_{\mu j}x_j. \tag{13}$$

Equation (13) is to be understood as follows: Each of the N individual equations of observation (identified by a particular value of the index μ) is to be multiplied by a number λ_μ, there being a set of N numbers for each variable x_i. These numbers are to be so chosen that when the equations are then summed the coefficient of each x other than the specific x_i is zero, while the coefficient of x_i is unity. This, therefore, implies that there are q conditions on each of the q sets of N multipliers,

$$\lambda_\mu^{(i)}a_{\mu j}=\delta_{ij}, \tag{14}$$

where δ_{ij} is the usual Kronecker symbol.

One might call $\lambda_\mu^{(i)}$ the left inverse matrix to $a_{\mu i}$ by virtue of Eq. (14). However, $a_{\mu i}$ is not a square matrix and does not have an inverse in the strict sense. The matrix $\lambda_\mu^{(i)}$ contains Nq elements, while Eq. (14) imposes only q^2 conditions. The fact that $\lambda_\mu^{(i)}$ is, therefore, not uniquely defined by Eq. (14) is the basis of the entire subsequent discussion. We still have a free choice of $f=N-q$ of the $\lambda_\mu^{(i)}$ for each of the q values of $i=1, 2, 3 \cdots q$.

The error in the computed value of x_i is a result of the errors ϵ_α in the quantities s_α; let v_{ij} be the mean square value of the product $dx_i dx_j$. Then

$$\begin{aligned} v_{ij}=\langle dx_i dx_j\rangle&=\langle(\lambda_\mu^{(i)}A_{\mu\alpha}\epsilon_\alpha)(\lambda_\nu^{(j)}A_{\nu\beta}\epsilon_\beta)\rangle \\ &=\lambda_\mu^{(i)}\lambda_\nu^{(j)}A_{\mu\alpha}A_{\nu\beta}\langle\epsilon_\alpha\epsilon_\beta\rangle. \end{aligned} \tag{15}$$

The ϵ_α are, however, independently observed quantities and are completely unrelated, one to the other, so that

$$E_{\alpha\beta}\equiv\langle\epsilon_\alpha\epsilon_\beta\rangle=\overline{\epsilon_\alpha^2}\delta_{\alpha\beta}. \tag{16}$$

We must now choose the $\lambda_\mu^{(i)}$ in such a way as to minimize the value of the variance of x_i; hence we must minimize the expression

$$v_{ii}=\lambda_\mu^{(i)}\lambda_\nu^{(i)}A_{\mu\alpha}A_{\nu\beta}E_{\alpha\beta} \text{ (not summed on } i) \tag{17}$$

consistent with the restrictions imposed on $\lambda_\mu^{(i)}$ by Eq. (14). We therefore introduce q^2 Lagrangian multipliers $L_j^{(i)}$, so that we may then take independent variations of $\lambda_\mu^{(i)}$ in the expression

$$\lambda_\mu^{(i)}\lambda_\nu^{(j)}A_{\mu\alpha}A_{\nu\beta}E_{\alpha\beta}-2L_j^{(i)}\{\lambda_\mu^{(i)}a_{\mu j}-\delta_{ij}\} \text{ (not summed on } i). \tag{18}$$

Differentiating this expression and making use of the symmetry of $E_{\alpha\beta}$, we obtain

$$L_j^{(i)}a_{\mu j}=\lambda_\nu^{(i)}S_{\nu\mu}, \tag{19}$$

where

$$S_{\nu\mu}=S_{\mu\nu}=\sum_\alpha A_{\mu\alpha}A_{\nu\alpha}\overline{\epsilon_\alpha^2}.$$

$S_{\nu\mu}$ is the error matrix of the observational equations. The diagonal elements of $S_{\nu\mu}$ are simply the variances of the numerical constants c_μ. If these numbers are uncorrelated, by which we imply that any given s_α occurs in only one of the c_μ's, the product $A_{\mu\alpha}A_{\nu\alpha}$ (not summed on α) will be different from zero only for $\mu=\nu$ and $S_{\nu\mu}$ will be a diagonal matrix. This is the only case usually considered in most discussions of the theory of least squares; the recognition that $S_{\nu\mu}$ need not be diagonal constitutes what is here referred to as the "generalized theory of least squares." The weight matrix of the observational equations, $\pi_{\mu\nu}$, is defined as the inverse of the error matrix so that

$$\pi_{\mu\nu}S_{\nu\tau}=\delta_{\mu\tau}=S_{\mu\nu}\pi_{\nu\tau}. \tag{20}$$

If we therefore multiply Eq. (19) by the weight matrix $\pi_{\mu\tau}$, we obtain the result

$$\begin{aligned} L_j^{(i)}a_{\mu j}\pi_{\mu\tau}&=\lambda_\nu^{(i)}S_{\nu\mu}\pi_{\mu\tau} \\ &=\lambda_\nu^{(i)}\delta_{\nu\tau}, \\ L_j^{(i)}a_{\mu j}\pi_{\mu\nu}&=\lambda_\nu^{(i)}. \end{aligned} \tag{21}$$

This defines $\lambda_\nu^{(i)}$ in terms of the Lagrangian multipliers. Insertion of Eq. (21) into Eq. (14) gives us

$$L_j^{(i)}a_{\mu j}\pi_{\mu\nu}a_{\nu k}=\delta_{ik}, \tag{22}$$

so that $L_j^{(i)}$ must be the inverse of the symmetric matrix $P_{ij}=a_{\mu i}\pi_{\mu\nu}a_{\nu j}$ (P_{ij} is symmetric since $\pi_{\mu\nu}$ is symmetric; the symmetry of this latter matrix follows directly from the symmetry of $S_{\mu\nu}$). We have therefore established also the symmetry of $L_j^{(i)}$ ($L_j^{(i)}=L_i^{(j)}$).

We now return to Eq. (13) and write [using Eq. (21)]

$$x_i=\lambda_\mu^{(i)}c_\mu=L_j^{(i)}a_{\nu j}\pi_{\nu\mu}c_\mu. \tag{23}$$

This is the solution for the "best" choice for the x_i; Eq. (23) is, however, just the solution of the system of equations

$$P_{ji}x_i=a_{\nu j}\pi_{\nu\mu}c_\mu, \tag{24}$$

and these are exactly the same equations as one obtains from the condition of least squares. The basic quadratic form associated with Eq. (24) and from which Eq. (24) can be deduced as the condition that the quadratic form have its minimum value (i.e., the "least-squares" condition) is

$$Q=\sum_{\mu\nu}\Big(\sum_i a_{\mu i}x_i-c_\mu\Big)\pi_{\mu\nu}\Big(\sum_j a_{\nu j}x_j-c_\nu\Big). \tag{25}$$

In particular, in the case where the variables c_μ are themselves observationally independent (which is equivalent to $S_{\mu\nu}$ being diagonal), we find that $\pi_{\nu\mu}$ is a diagonal matrix and its diagonal elements, which we may write simply as p_μ, are just the usual statistical weights to be ascribed to each observational equation.

Under these conditions the quadratic form, Eq. (25), reduces to the more familiar expression

$$Q=\sum_{\mu}(\sum_{i}a_{\mu i}x_i-c_\mu)^2p_\mu, \tag{25.1}$$

and Eq. (24) reduces to the usual normal equations

$$\sum_{i}(\sum_{\mu}p_\mu a_{\mu j}a_{\mu i})x_i=\sum_{\mu}p_\mu a_{\mu j}c_\mu. \tag{24.1}$$

Since Eqs. (23) and (24) are actually the same equations as one would obtain from the condition of least squares, we have shown not only that the axiom of maximum statistical weight is equivalent to the axiom of maximum likelihood for a Gaussian distribution but, furthermore, that the usual normal equations of least squares correspond to the condition of maximum statistical weight, even in those cases where the condition of maximum likelihood is impossible of simple analytic formulation or is ambiguous. One of these ambiguous cases occurs even when the error distributions are Gaussian if the c_μ are observationally correlated in such a way that the matrix $S_{\nu\mu}$ cannot be diagonalized, which is in general the case if $n>N$. In the earlier discussion[2] of the generalized theory the present writer was not fully appreciative of this ambiguity which was left unresolved. Although the quadratic form Eq. (25) was presented as the proper generalization of the simpler expression (25.1), the proof was semi-intuitive and was based on plausibility arguments rather than a firm logical foundation. The present formulation achieves this and has the further advantage of admitting to consideration by the condition of least squares a much wider class of error distribution functions than merely the Gaussian.