A LEAST SQUARES SOLUTION OF LINEAR EQUATIONS WITH COEFFICIENTS SUBJECT TO A SPECIAL TYPE OF ERROR

By J. K. MACKENZIE*

[*Manuscript received August 6, 1956*]

Summary

Following the classical method a least squares solution is given for the equations $\sum_{k=1}^{s} (a_{rk}+e_r b_{rk})x_k = e_r$ ($r=1, \ldots, n \geqslant s$), where the a_{rk} and b_{rk} are fixed known constants and the e_r are observed values subject to error. The solution is obtained as a series in the successive moments of the joint distribution of the e_r, and only terms up to those involving the variance are retained. In this approximation the estimated values of the x_k are biased, but, after correction for this bias and using a particular weight for each equation, the classical tests of significance for the case $b_{rk}=0$ can be applied unchanged. With suitable assumptions it is shown that the series converges more and more rapidly as $n \to \infty$ for almost all sequences of the e_r.

I. INTRODUCTION

The classical theory of the solution of a set of n linear simultaneous equations in s ($\leqslant n$) unknowns by means of least squares is due originally to Gauss (Plackett 1949). In this theory the equations are of the form

$$\sum_{k=1}^{s} c_{rk}x_k = e_r + \delta e_r \quad (r=1, \ldots, n), \qquad \ldots\ldots\ldots\ldots \quad (1.1)$$

where the coefficients c_{rk} are fixed (known) constants and the e's are observed values subject to (unknown) errors, δe_r, with zero mean. In the present paper the method of least squares is applied to the equations

$$\sum_{k=1}^{s} (a_{rk}+e_r b_{rk})x_k = e_r + \delta e_r \quad (r=1, \ldots, n), \qquad \ldots\ldots \quad (1.2)$$

where the a_{rk} and b_{rk} are fixed (known) constants and the coefficients of the x's depend linearly on the e's which are observed values subject to (unknown) errors. Such equations arise in the reduction of experimental observations leading to the determination of the direction of an invariant plane strain and a brief derivation of the equations arising in this case is given in Section II.

Section III reviews the classical procedure for solution of equations by least squares and this method is adapted, in Section IV, to a least squares solution of equations (**1.2**). The solution is given as a series in the successive moments of the joint distribution of the e's, which with suitable assumptions

* Division of Tribophysics, C.S.I.R.O., University of Melbourne.

converges for almost all sequences of the e's as $n \to \infty$. Only terms up to those containing the second moments (variance) are retained and to this degree of approximation the classical tests of significance can be applied.

II. The Physical Problem

When a change of phase of the martensitic type occurs in a solid it is accompanied by a homogeneous strain which describes the change in shape of the solid. This homogeneous strain is essentially* an invariant plane strain in which a plane (the habit plane) with unit normal **p** remains invariant and all points move in a common direction $\mathbf{d}/|\mathbf{d}|$. A point with position vector **y** moves to the point

$$\mathbf{y}' = \mathbf{y} + (\mathbf{p}.\mathbf{y})\mathbf{d}, \quad \ldots\ldots\ldots\ldots \quad (2.1)$$

while a plane with normal **n** becomes the plane with normal

$$\mathbf{n}' = \mathbf{n} - (\mathbf{n}.\mathbf{d})\mathbf{p}/(1 + \mathbf{p}.\mathbf{d}). \quad \ldots\ldots\ldots\ldots \quad (2.2)$$

The normal **p** to the habit plane can be measured readily and directly but the direction **d** is more difficult to estimate.

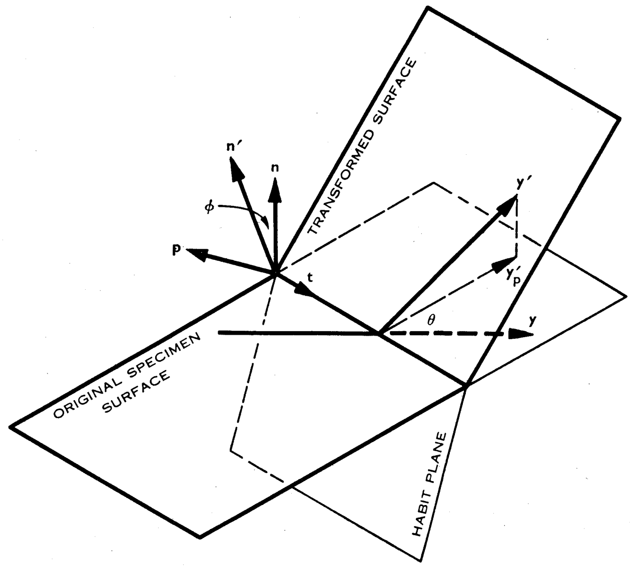

Fig. 1.—The junction between an original specimen surface and the same surface after transformation.

Figure 1 shows schematically the junction between an original specimen surface, with unit normal **n**, and the same surface after transformation, with normal **n′**. Also shown are a scratch originally in the direction of the unit vector **y** and the projection $\mathbf{y}'_p$ of the direction **y′** onto the original surface with

* According to Bowles and Mackenzie (1954) it only differs from an invariant plane strain by a small dilatation. In a recent review Bilby and Christian (1955) suggest otherwise. At present there is no direct experimental evidence that the difference is not a pure dilatation but in any case the difference is small.

normal **n**. The measurements which can be readily made are (*a*) the angle θ between **y** and the projection $\mathbf{y}'_p$ of $\mathbf{y}'$ and (*b*) the angle φ between the surface normals **n** and **n**′. In the first case it is readily shown that

$$\mathbf{p}.\mathbf{y}(\mathbf{n}\times\mathbf{y}-\tan\theta\ \mathbf{y}).\mathbf{d}=\tan\theta, \quad \ldots\ldots\ldots \quad (2.3)$$

where θ increases positively for a right-hand rotation about the normal **n**. Likewise, in the second case

$$-|\,\mathbf{n}\times\mathbf{p}\,|(\mathbf{n}-\tan\varphi\ \mathbf{t}\times\mathbf{n}).\mathbf{d}=\tan\varphi, \quad \ldots\ldots\ldots \quad (2.4)$$

where $\mathbf{n}\times\mathbf{p}=|\,\mathbf{n}\times\mathbf{p}\,|\mathbf{t}$ and φ increases positively for a right-hand rotation about the trace **t** of the habit plane in the original specimen surface.

The habit plane and other quantities appearing on the left of (**2.3**) and (**2.4**) can be measured much more accurately than the small angles θ and φ, so that the equations for the determination of **d** are effectively of the form (**1.2**).

III. The Classical Theory of Least Squares

If **C** is the $n\times s$ matrix (c_{rk}) while **x** and **e** are the $s\times 1$ and $n\times 1$ matrices (x_k) and (e_r) then the equations (**1.1**) can be written compactly in matrix notation as

$$\mathbf{Cx}-\mathbf{e}=\delta\mathbf{e}. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (3.1)$$

Then, given a symmetrical positive definite $n\times n$ weighting matrix **W**, the least squares method of solution consists in first forming the essentially positive quadratic form (weighted sum of squares, if **W** is diagonal) of the errors

$$S=(\mathbf{Cx}-\mathbf{e})'\mathbf{W}(\mathbf{Cx}-\mathbf{e}), \quad \ldots\ldots\ldots\ldots\ldots \quad (3.2)$$

and then choosing **x** so that S is minimum. This gives the normal equations

$$\mathbf{N}\hat{\mathbf{x}}=\mathbf{m}, \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (3.3)$$

where*

$$\mathbf{N}=\mathbf{C}'\mathbf{W}\mathbf{C}, \qquad \mathbf{m}=\mathbf{C}'\mathbf{W}\mathbf{e}. \quad \ldots\ldots\ldots\ldots\ldots \quad (3.4)$$

The estimate $\hat{\mathbf{x}}$ of $\dot{\mathbf{x}}$ so obtained is unbiased, since $E(\hat{\mathbf{x}})=\mathbf{N}^{-1}\mathbf{C}'\mathbf{W}\ E(\mathbf{e})=\mathbf{x}$; the symbol E denotes the expectation value of the variable in brackets and is obtained by averaging over the distribution of this variable. Further,

$$S_{\text{min.}}=\mathbf{e}'\mathbf{W}\mathbf{e}-\mathbf{e}'\mathbf{W}\mathbf{C}\mathbf{N}^{-1}\mathbf{C}'\mathbf{W}\mathbf{e}, \quad \ldots\ldots\ldots\ldots \quad (3.5)$$

and the $s\times s$ covariance matrix of the estimated values $\hat{\mathbf{x}}$ is

$$\text{Cov}\,(\hat{\mathbf{x}},\hat{\mathbf{x}})=\mathbf{N}^{-1}\mathbf{C}'\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{C}\mathbf{N}^{-1}, \quad \ldots\ldots\ldots\ldots\ldots \quad (3.6)$$

where **V** is the $n\times n$ covariance matrix of the errors $\delta\mathbf{e}$.

It was assumed above that **W** was given *a priori*, but the question arises as to the "best possible" choice for **W**. Gauss showed that of all possible

* **N** has a unique inverse provided **C** is of rank s.

H

linear combinations $\mathbf{Xe}$ of the e_r for which $\mathbf{X}$ is (functionally) independent of $\mathbf{e}$ and $\mathbf{Xe}$ is an unbiassed estimate of $\mathbf{x}$ (i.e. $\mathbf{X\,C}=\mathbf{I}$) that which gives the elements of $\hat{\mathbf{x}}=\mathbf{Xe}$ a minimum variance is given by

$$\mathbf{X}=(\mathbf{C}'\mathbf{V}^{-1}\mathbf{C})^{-1}\mathbf{C}'\mathbf{V}^{-1}. \quad \ldots\ldots\ldots\ldots\ldots \quad (3.7)$$

Usually, only the ratios of the elements of $\mathbf{V}$ are known (or estimated) so that $\mathbf{V}=\mathbf{V}_0\sigma^2$, where σ^2 is to be determined. Then, with the choice $\mathbf{W}=\mathbf{V}_0^{-1}$, an estimate of σ^2 is obtained from the equation

$$E(S_{\text{min.}})=(n-s)\sigma^2, \quad \ldots\ldots\ldots\ldots\ldots \quad (3.8)$$

and (**3.6**) simplifies to

$$\text{Cov}\,(\hat{\mathbf{x}},\hat{\mathbf{x}})=\mathbf{N}^{-1}\sigma^2. \quad \ldots\ldots\ldots\ldots\ldots \quad (3.9)$$

If it is assumed that the errors $\delta\mathbf{e}$ are normally distributed with covariance matrix $\mathbf{V}$ the maximum likelihood estimator $\mathbf{x}^*$ of $\mathbf{x}$ is the least squares estimate with $\mathbf{W}=\mathbf{V}_0^{-1}$ and the maximum likelihood estimate σ^{*2} of σ^2 is $S_{\text{min.}}/n$. Further, $S_{\text{min.}}/\sigma^2=n\sigma^{*2}/\sigma^2$ has a χ^2-distribution with $n-s$ degrees of freedom and the variables $[(n-s)/n]^{\frac{1}{2}}(\mathbf{x}^*-\mathbf{x})_r/\sigma^*[\mathbf{N}^{-1}]_{rr}$ have a t-distribution with $n-s$ degrees of freedom. For a joint test of a hypothetical solution $\mathbf{x}_h$ the variable

$$e^{2z}=\frac{n-s}{s}\,\frac{(\mathbf{x}^*-\mathbf{x}_h)'\,\mathbf{N}(\mathbf{x}^*-\mathbf{x}_h)}{S_{\text{min.}}}$$

is distributed like a variance ratio with s degrees of freedom in the numerator and $n-s$ in the denominator (Cramer 1946).

IV. Adaptation to New Problem

The method proposed for the solution of the equations (**1.2**) is as follows. First calculate the coefficients of the x_k on the left using the observed values of the e_r and then find the least squares solution $\hat{\mathbf{x}}$ as in the preceding section, treating the calculated coefficients as fixed. This is quite straightforward; the real problem is to estimate the bias in the values of $\hat{\mathbf{x}}$ so obtained and the covariance matrix of the solution after correction for bias.

The equations (**1.2**) can be written in an obvious matrix notation as

$$(\mathbf{A}_1+\mathbf{E\,B})\mathbf{x}-\mathbf{e}=\delta\mathbf{e}, \quad \ldots\ldots\ldots\ldots\ldots \quad (4.1)$$

where $\mathbf{E}$ is a diagonal matrix with diagonal elements e_r in order. It is convenient to express these equations in a form showing explicitly their dependence on the deviations of $\mathbf{e}$ from its mean value $\bar{\mathbf{e}}\equiv E(\mathbf{e})$. If $\mathbf{D}$ is the diagonal matrix $\mathbf{E}-\bar{\mathbf{E}}$ and $\mathbf{1}$ is an $n\times1$ matrix consisting of a column of ones, (**4.1**) can be written

$$(\mathbf{A}+\mathbf{D\,B})\mathbf{x}-\bar{\mathbf{e}}-\mathbf{D1}=\delta\mathbf{e}, \quad \ldots\ldots\ldots\ldots \quad (4.2)$$

where $\mathbf{A}=\mathbf{A}_1+\bar{\mathbf{E}}\mathbf{B}$. It will be assumed that the correct value of $\mathbf{x}$ is $\mathbf{x}_m$ satisfying the equations

$$\mathbf{A\,x}_m-\bar{\mathbf{e}}=0. \quad \ldots\ldots\ldots\ldots \quad (4.3)$$

Subtracting (4.3) from (4.2) gives

$$(\mathbf{A}+\mathbf{D}\,\mathbf{B})(\mathbf{x}-\mathbf{x}_m)-\mathbf{D}\mathbf{f}=\delta\mathbf{e}, \quad \ldots\ldots\ldots\ldots \quad (4.4)$$

where

$$\mathbf{f}=\mathbf{1}-\mathbf{B}\mathbf{x}_m. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.5)$$

The equations are now of the form (3.1) with $\mathbf{A}+\mathbf{D}\,\mathbf{B}$ replacing $\mathbf{C}$ and $\mathbf{D}\mathbf{f}$ replacing $\mathbf{e}$. Further, it is clear that the least squares solutions of (4.1) and (4.4) give the same value of $S_{\text{min.}}$.

The least squares solution of (4.4) is $\hat{\mathbf{x}}-\mathbf{x}_m=\mathbf{N}^{-1}\mathbf{m}$, where in the present case

$$\mathbf{N}=\mathbf{A}'\mathbf{W}\mathbf{A}+\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{B}+\mathbf{B}'\mathbf{D}\mathbf{W}\mathbf{A}+\mathbf{B}'\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{B}=\mathbf{M}+\delta\mathbf{M}, \text{ say,} \quad \ldots \quad (4.6)$$

$$\mathbf{m}=\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{f}+\mathbf{B}'\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{f}, \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.7)$$

and

$$\mathbf{M}=\mathbf{A}'\mathbf{W}\mathbf{A}+\mathbf{B}'E(\mathbf{D}\mathbf{W}\mathbf{D})\mathbf{B}. \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (4.8)$$

Now, provided all the characteristic roots of the matrix $\mathbf{M}^{-1}\delta\mathbf{M}$ are of modulus less than unity, the series expansion

$$\mathbf{N}^{-1}=\mathbf{M}^{-1}-\mathbf{M}^{-1}\delta\mathbf{M}\mathbf{M}^{-1}+\mathbf{M}^{-1}\delta\mathbf{M}\mathbf{M}^{-1}\delta\mathbf{M}\mathbf{M}^{-1}-\ldots \quad \ldots\ldots\ldots\ldots \quad (4.9)$$

is convergent (Ferrar 1951). Although the series (4.9) will not converge for some values of $\mathbf{D}$ and hence of $\delta\mathbf{M}$ it will be shown in Section V that with appropriate assumptions the series converges for almost all sequences of observations e_r as $n\to\infty$. Hence, retaining only terms of the second degree in $\mathbf{D}$, it follows that

$$\begin{aligned}\hat{\mathbf{x}}-\mathbf{x}_m=&\,\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{f}+\mathbf{M}^{-1}\mathbf{B}'\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{f}\\&-\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{B}\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{f}-\mathbf{M}^{-1}\mathbf{B}'\mathbf{D}\mathbf{W}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D}\mathbf{f}+\ldots,\end{aligned} \quad \ldots\ldots\ldots\ldots \quad (4.10)$$

so that the bias is

$$\begin{aligned}E(\hat{\mathbf{x}}-\mathbf{x}_m)=&\,\mathbf{M}^{-1}\mathbf{B}'E(\mathbf{D}\mathbf{W}\mathbf{D}-\mathbf{D}\mathbf{W}\mathbf{A}\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D})\mathbf{f}\\&-\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}E(\mathbf{D}\mathbf{B}\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}\mathbf{D})\mathbf{f}+\ldots\end{aligned} \quad \ldots\ldots\ldots\ldots \quad (4.11)$$

Terms such as $E(\mathbf{D}\mathbf{X}\mathbf{D})$ in (4.11) are simply obtained by multiplying each element of $\mathbf{X}$ by the corresponding element of $\mathbf{V}$, and to the present degree of approximation $\mathbf{M}^{-1}$ may be replaced by $\mathbf{N}^{-1}$. Clearly there will be no bias, to any degree of approximation, in the special case $\mathbf{B}\mathbf{x}_m=\mathbf{1}$.

Further,

$$\operatorname{Cov}(\hat{\mathbf{x}},\hat{\mathbf{x}})=\mathbf{M}^{-1}\mathbf{A}'\mathbf{W}E(\mathbf{D}\mathbf{f}\mathbf{f}'\mathbf{D})\mathbf{W}\mathbf{A}\mathbf{M}^{-1}+\ldots, \quad \ldots \quad (4.12)$$

$$S_{\text{min.}}=\mathbf{f}'\mathbf{D}\mathbf{W}\mathbf{D}\mathbf{f}-\mathbf{f}'\mathbf{D}\mathbf{W}\mathbf{A}'\mathbf{M}^{-1}\mathbf{A}\mathbf{W}\mathbf{D}\mathbf{f}+\ldots, \quad \ldots \quad (4.13)$$

and to the present degree of approximation these are of the same form as (3.6) and (3.5) respectively, $\mathbf{D}\mathbf{f}$ replacing $\mathbf{e}-\bar{\mathbf{e}}$. Thus, the choice

$$\mathbf{W}^{-1}=E(\mathbf{D}\mathbf{f}\mathbf{f}'\mathbf{D})/\sigma^2 \quad \ldots\ldots\ldots\ldots \quad (4.14)$$

gives the same simplifications as in the classical case. In particular, when errors in $\mathbf{e}$ are normally distributed the classical results on the various distributions follow. However, since $\mathbf{f}$ depends on $\mathbf{x}_m$, which is not estimated until the solution is complete, the appropriate $\mathbf{W}$ can only be estimated by successive approximation starting, say, with the approximation $\mathbf{f}=\mathbf{1}$. For the physical problem in Section II $\mathbf{Bx}_m$ will usually be small compared with unity and the labour of repeating the solution is probably not worth while.

The results of the present section fall short of the classical results in two respects. First, it has not been shown that the choice (**4.14**) for $\mathbf{W}$ is in any sense the best possible although it is probably not far short of this. Second, the relationship between the present solution and a maximum likelihood solution for, say, normally distributed errors remains unknown although it seems almost certain that the results after correction for bias will be the same asymptotically as $n\to\infty$.

V. Convergence of the Series (4.9)

To obtain a result concerning the convergence of (**4.9**) as $n\to\infty$ it is necessary to make some assumptions concerning the behaviour of the elements of the matrices $\mathbf{A}, \mathbf{B}, \mathbf{W}$ and about the nature of the distribution of the diagonal elements of $\mathbf{D}$, i.e. the distribution associated with the observations $\mathbf{e}$. The assumptions stated below are sufficient to ensure the convergence of (**4.9**) as $n\to\infty$ for almost all sequences of observations $\mathbf{e}$, and will be satisfied in most cases of practical importance.

It will be assumed that for all n

(i) The modulus of all the elements of the matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{W}$ are uniformly bounded by the constants a_0, b_0, and w_0 respectively,

(ii) the elements in each column of $\mathbf{A}$ have a finite mean square,

(iii) the diagonal elements Δ_j of $\mathbf{D}$ are independent and have finite variances v_j uniformly bounded by v_0,

(iv) the distributions of $\Delta_j/v_j^{\frac{1}{2}}$ are identical.

It now follows that as $n\to\infty$ the series (**4.9**) converges for almost all sequences of the $\Delta_j=e_j-\bar{e}_j$.

Since the Δ_j are independent, the choice (**4.14**) for $\mathbf{W}$ leads to a diagonal matrix. Thus, every one of the s^2 elements of the matrix $\mathbf{A'WDB}/n$ can be written in the form

$$\sum_{j=1}^{n} \alpha_j\Delta_j/n \leqslant a_0 b_0 w_0 v_0^{\frac{1}{2}} \sum_{j=1}^{n} \Delta_j/nv_j^{\frac{1}{2}}, \qquad \dots\dots\dots\dots \quad (5.1)$$

and similarly every element of the matrix $[\mathbf{B'DWDB}-\mathbf{B'}E(\mathbf{DWD})\mathbf{B}]/n$ can be written in the form

$$\sum_{j=1}^{n} \beta_j(\Delta_j^2-v_j)/n \leqslant b_0^2 w_0 v_0 \sum_{j=1}^{n} (\Delta_j^2-v_j)/nv_j. \qquad \dots\dots \quad (5.2)$$

But the variables $\Delta_j/v_j^{\frac{1}{2}}$ and $(\Delta_j^2-v_j)/v_j$, are identically distributed and have zero mean so that the strong law of large numbers (Feller 1950) applies to the sums on the right of (**5.1**) and (**5.2**). Now (ii) ensures that $\mathbf{M}/n=[\mathbf{A'WA}+\mathbf{B'}E(\mathbf{DWD})\mathbf{B}]/n$ is always a matrix with finite diagonal elements and so it follows that $\mathbf{M}^{-1}\delta\mathbf{M}$ converges to zero as $n \to \infty$ for almost all sequences of the e_j.

Thus, as n increases, (4.9) not only converges for almost all sequences e_j but does so more and more rapidly.

VI. References

BILBY, B. A., and CHRISTIAN, J. W. (1955).—Monogr. Ser. Inst. Metals, No. 18: 121.

BOWLES, J. S., and MACKENZIE, J. K. (1954).—*Acta Met.* **2**: 129.

CRAMER, H. (1946).—"Mathematical Methods of Statistics." Ch. 37. (Princeton Univ. Press.)

FELLER, W. (1950).—"Probability Theory and its Applications." Vol. 1. Ch. 10. (Wiley: New York.)

FERRAR, W. L. (1951).—"Finite Matrices." Ch. 5. (Clarendon Press: Oxford.)

PLACKETT, R. L. (1949).—*Biometrika* **36**: 458.