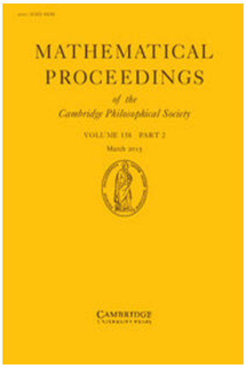

Further aspects of the theory of multiple regression

M. S. Bartlett

Mathematical Proceedings of the Cambridge Philosophical Society / Volume 34 / Issue 01 / January 1938, pp 33 - 40
DOI: 10.1017/S0305004100019897, Published online: 24 October 2008

Link to this article: http://journals.cambridge.org/abstract_S0305004100019897

How to cite this article:
M. S. Bartlett (1938). Further aspects of the theory of multiple regression. Mathematical Proceedings of the Cambridge Philosophical Society, 34, pp 33-40 doi:10.1017/S0305004100019897

FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION

By Mr M. S. BARTLETT, Queens' College

[*Received* 10 August, *read* 22 November 1937]

1. *Notation.* This paper may be regarded as a sequel to a previous paper (1) in these *Proceedings*. The vector and matrix notation of that paper used for a statistical sample is systematized somewhat further, so that while a sample S refers as before to the matrix of nm values (a sample of m observations in one variate only being a row vector), we write

$$S_2 = B_{21}S_1 + S_{2.1} \tag{1}$$

for the linear regression formula between the dependent and independent variates into which a sample is supposed partitioned (in place of equation (12) of (1)). More generally, a third submatrix S_0 is partitioned off, and its effect eliminated (corresponding to equation (13) of (1)), but without loss of generality we assume that S_2 in equation (1) above can always stand for $S_{2.0}$ if necessary. B_{21} is the *estimated* matrix of regression coefficients, given by

$$B_{21} = C_{21}C_{11}^{-1}, \tag{2}$$

where $C_{21} \equiv S_2S_1'$, $C_{11} \equiv S_1S_1'$ (dashes denoting transposed matrices). From equation (1),

$$\left.\begin{aligned} C_{22} &= B_{21}C_{11}B_{21}' + C_{22.1} \\ &= C_{21}C_{11}^{-1}C_{12} + C_{22.1}. \end{aligned}\right\} \tag{3}$$

In conformity with orthodox notation, vectors may sometimes for clarity be denoted by small letters, x being a column vector, x' a row vector. A typical member of S can thus be denoted by x.

For arguments relating to the "population" rather than to the sample, the equation for a typical member of the infinite matrix Π to which S may be allowed to tend becomes

$$x_2 = Bx_1 + x_{2.1}, \tag{4}$$

where to avoid confusion the matrix of *true* coefficients (β) is denoted simply by B. Appropriate multiplication and averaging of the elements of (4) gives the formulae

$$V_{21} = BV_{11}, \tag{5}$$

$$V_{22} = BV_{11}B' + V_{22.1}, \tag{6}$$

where $V_{21} \equiv E(x_2x_1')$. In the usual case, where S_2 denotes a sample after elimination of the general mean, V_{22} is the matrix of variances and covariances of the variates of S_2. From (5),

$$B = V_{21}V_{11}^{-1}, \tag{7}$$

a result which may be compared with (2).

2. *Multivariate selection.* The use of some of these *population* formulae may be conveniently illustrated at this stage in a problem originally proposed by Karl Pearson on the effect of selection for some statistical variates on others correlated with them. This problem has been re-examined recently by Aitken(2), but it should be noted that the conditions required are more restricted than Aitken was led to suppose.

In equation (4), $x_{2.1}$ is the residual subvector uncorrelated with x_1, but if selection operates on x_1, the assumption is now made that $x_{2.1}$ is unaffected, at least so far as means, variances and covariances are concerned. (This condition certainly holds if $x_{2.1}$ is quite independent of x_1.) If Δ is prefixed to any quantity to denote the change induced by selection, (5) and (6) immediately give

$$\left.\begin{aligned}\Delta V_{21} &= B\Delta V_{11},\\ \Delta V_{22} &= B\Delta V_{11}B'.\end{aligned}\right\} \tag{8}$$

For convenience we are supposing S_2 and S_1 to be measured from their means in this result, so that $E(x_2) = 0$, etc. If we lift this restriction, we have instead $E(x_2) = m_2$, etc., and hence we obtain the third formula

$$\Delta m_2 = B\Delta m_1. \tag{8a}$$

The linear form of (4), together with the implicit assumption of the independence of $x_{2.1}$, was probably the condition which Pearson finally had in mind(3). This condition would hold somewhat more generally if S_1 did not denote the original vector of selected variates, but a vector of appropriately chosen functions of them; but even this wider condition, while the one now usually associated in practice with a regression equation of the type (1), may not of course necessarily hold.

3. Equation (4) might also be readily related to Wright's method of path coefficients(4), and formula (6), and chain formulae of the type $m_2 = BCm_0$, where $m_1 = Cm_0$, are available in this method.

In the genetical theory of inbreeding, m can be used to denote the proportions of different possible genotypic matings or individuals (i.e. contingency frequencies rather than measures), and repeated operation with the same square selection matrix B, determined by the process of selection adopted, leads to recurrence formulae of the type

$$m_{n+1} = Bm_n, \tag{9}$$

a familiar linear difference equation which has as its solution a sum of geometric series $a_r\lambda_r^n$, where the λ_r are the roots of the characteristic equation

$$|B-\lambda| = 0. \tag{10}$$

4. Returning to the standard *sample* equation (1), with which we are most concerned in the rest of this paper, we have seen (equation (3)) that this implies a corresponding analysis of $S_2S_2' = C_{22}$. Apart from the derivation in (1) of a general

criterion Λ which was available for testing the significance of the association between S_2 and S_1 in problems where a single test of this kind might be considered useful (and which included some of the generalized tests obtained previously by Hotelling (5) and Wilks (6)), any further reduction or interpretation of C_{22} was not considered. Hotelling, (7), (8), (9) and (10), has, however, dealt generally with the linear transformation of correlated variates into uncorrelated components, linking the problem to the theory of canonical matrices, and it seems desirable to develop somewhat further the formal matrix theory of linear regression formulae given in (1), in order to show how Hotelling's theory may be incorporated. Methods of dealing with multiple measurements have also been considered recently by Fisher (see, for example, (11)), and it is of some importance to establish the interrelated nature of these investigations.

5. *Linear transformations of the dependent variates.* For any given square (non-singular) matrix A, the information contained in S may equally well be represented by

$$T = AS,$$

since S is recoverable from T by the inverse relation

$$S = A^{-1}T.$$

If we consider the analysis of one of the new set of dependent variates AS_2—$a'S_2$, say—then from (1),

$$a'S_2 = a'B_{21}S_1 + a'S_{2.1}, \tag{11}$$

whence

$$a'C_{22}a = a'C_{21}C_{11}^{-1}C_{12}a + a'C_{22.1}a. \tag{12}$$

Since the sum of squares $a'C_{22}a$ may for convenience be fixed, the problem of finding the linear function $a'S_2$ which minimizes the ratio $a'C_{22.1}a/a'C_{22}a$ is equivalently solved by finding the maximum of $a'C_{21}C_{11}^{-1}C_{12}\,a$ for given $a'C_{22}a$, of which the known solution (12) is

$$(C_{21}C_{11}^{-1}C_{12} - \lambda C_{22})a = 0, \tag{13}$$

where

$$\left.\begin{aligned} |\,C_{21}C_{11}^{-1}C_{12} - \lambda C_{22}\,| &= 0,\\ |\,C_{21}C_{11}^{-1}C_{12}C_{22}^{-1} - \lambda\,| &= 0. \end{aligned}\right\} \tag{14}$$

that is,

This problem is that of Hotelling's *most predictable criterion* (8). No restriction on the nature of S_1 is necessary, so that all analysis-of-variance problems are included. Hotelling (10) has given a solution

$$\left|\begin{array}{c|c} -\lambda C_{11} & C_{12} \\ \hline C_{21} & -\lambda C_{22} \end{array}\right| = 0,$$

but on multiplying the first column of the submatrices by C_{11}^{-1}, we get

$$\left|\begin{array}{c|c} -\lambda & C_{12} \\ \hline C_{21}C_{11}^{-1} & -\lambda C_{22} \end{array}\right| = 0,$$

that is,

$$(-\lambda)^{q-1}\,|\,\lambda^2 C_{22} - C_{21}C_{11}^{-1}C_{12}\,| = 0, \tag{15}$$

which is equivalent to (14) with λ^2 in place of λ.

Multiplying (13) on the left by a', we have

$$a'C_{21}C_{11}^{-1}C_{12}a/a'C_{22}a = \lambda, \qquad (16)$$

so that λ must be the *largest* root of (13). Hotelling proceeds to consider successively the remaining components of the general set AS_2, the system of uncorrelated components and their partners from S_1 being termed canonical variates.

6. *Principal components.* It may be of some interest to insert the corresponding analysis of the purely internal relations among p variates given by Hotelling (7). (The associated problems of sampling distributions are of course distinct.) Let a component of a sample S be

$$\gamma' = a'S. \qquad (17)$$

If, analogously to (1), we try to express S in terms of γ, we have, for the set of regression coefficients corresponding to (2), $S\gamma/\gamma'\gamma$, and for the contribution of γ to the matrix SS', $S\gamma\gamma'S'/\gamma'\gamma$, of which the trace (sum of diagonal elements) is

$$\frac{\gamma'S'S\gamma}{\gamma'\gamma} = \frac{a'(SS')^2a}{a'(SS')a}. \qquad (18)$$

If we impose the conditions that S and γ are "normalized", so that

$$\gamma'\gamma = a'(SS')a = 1,$$

and SS' is the correlation matrix R, and moreover define the principal component γ by the condition that the trace given in (18) is a maximum, then

$$(R^2-\lambda R)a = 0.$$

Hence, multiplying by R^{-1}, we have

$$\left.\begin{aligned}(R-\lambda)a &= 0,\\ \text{where}\qquad |R-\lambda| &= 0.\end{aligned}\right\} \qquad (19)$$

Hotelling (9) has proposed a useful iterative method of solving equations (19), accelerated by the device of repeated squaring of the matrix R. The analogy with repeated operation with the matrix B in (9) should be noticed; there such repetition has an actual meaning, and the straightforward evaluation of m_n from m_0 might similarly be accelerated by repeated squaring of B, and thus the limiting form

$$m_n \rightarrow a\lambda_1^n,$$

where a is a constant vector and λ_1 is the largest root of (10), could be more rapidly obtained*.

7. *Discriminant functions.* If the "most predictable criterion" were to be used as a discriminant function which maximized the value λ (for example, where S_1 corresponded to differences in the mean values of several different species or groups), equations (13) and (14) are in general the appropriate ones to consider.

* A full discussion of the evaluation of the latent roots of a matrix has been given recently by Aitken (14).

Fisher (11) has examined separately the particular case of only two species (that is, S_1 is a single row vector). The theory of this particular case deserves especial consideration, and follows most simply by inverting the relationship of S_2 and S_1, since this relationship is reciprocal. Proceeding more directly, we note that C_{11} is now a scalar quantity, and $C_{21}C_{11}^{-1}C_{12}$ a matrix of the form zz'. Hence (13) and (14), which may equivalently be written

$$\left.\begin{aligned}(C_{21}C_{11}^{-1}C_{12}-\mu C_{22.1})a &= 0,\\ |C_{21}C_{11}^{-1}C_{12}-\mu C_{22.1}| &= 0,\end{aligned}\right\} \qquad (20)$$

where $\mu = \lambda/(1-\lambda)$, become

$$\left.\begin{aligned}(zz'-\mu C_{22.1})a &= 0,\\ |zz'-\mu C_{22.1}| &= 0,\end{aligned}\right\} \qquad (21)$$

where the determinantal equation, owing to the degeneracy of zz', has only one non-zero root. From the equivalent equation

$$|zz'C_{22.1}^{-1}-\mu| = 0,$$

this root is the trace of $z(z'C_{22.1}^{-1})$; that is,

$$\mu = (z'C_{22.1}^{-1})z. \qquad (22)$$

If in the left-hand side of the equation for a, we insert this value for μ, and also put

$$a = C_{22.1}^{-1}z,$$

we obtain

$$z(z'C_{22.1}^{-1}z)-(z'C_{22.1}^{-1}z)z,$$

which is identically zero. Hence the vector a is the solution of the equation

$$C_{22.1}a = z, \qquad (23)$$

as noted by Fisher. This gives

$$\mu = z'a. \qquad (24)$$

It is also important to note that

$$\Lambda = \frac{|C_{22.1}|}{|C_{22.1}+zz'|} = \frac{1}{|1+zz'C_{22.1}^{-1}|} = \frac{1}{1+z'C_{22.1}^{-1}z} = \frac{1}{1+\mu}. \qquad (25)$$

Thus the test of significance for μ (or λ) is exactly equivalent to that of Λ, which is in this problem identical with Hotelling's generalization of Student's test. If, however, we generalize the discriminant function and so identify it with Hotelling's most predictable criterion, the test of the largest root λ_1 becomes a more special one than the Λ criterion.

This generalization may sometimes be necessary, although if the relevant information in S_1 can be condensed into a single degree of freedom from preliminary considerations, the discriminant function can not only be more simply, but more efficiently, derived. Moreover, unless the variation in the "residual sample" were found to be independent of S_1 (that is, in Hotelling's terminology, unless the remaining canonical correlations $\rho_2, \rho_3, \dots$ were zero), the function $a'S_2$ would not contain all the information on the relation of S_2 with S_1.

Thus in general, while the obvious point need not be overlooked that if a function suitable for discriminating species or groups were so devised, any subsequent use *on further data* would simply conform to orthodox analysis-of-variance lines, the more complex statistical relations of the function to the data from which it was calculated may have to be considered. The effect of selection for the largest root λ_1 makes the problem of the exact distribution of λ_1 or the other roots exceedingly complex, and a comprehensive test of Λ might sometimes prove more useful. An approximate use of the Λ test in general is considered in section 8.

8. *The χ^2 approximation for Λ.* From the relation between the limiting form of a likelihood and the χ^2 distribution (noted in (13), p. 274), it follows, on the usual assumption of the normality of $x_{2.1}$, that

$$-2\log\Lambda^{\frac{1}{2}n} \equiv -n\log\Lambda$$

tends, as $n \to \infty$, to be distributed as χ^2 with pq degrees of freedom.

Examining this approximation further, along similar lines, we obtain for the characteristic function of $-n\log\Lambda$, by an adaptation of formula (22) of (1),

$$M = \prod_{i=0}^{p-1} \frac{\Gamma\{\frac{1}{2}(n-i)\}\,\Gamma\{\frac{1}{2}(n-q-i)-nt\}}{\Gamma\{\frac{1}{2}(n-q-i)\}\,\Gamma\{\frac{1}{2}(n-i)-nt\}}, \tag{26}$$

$$K \equiv \log M = \sum_{i=0}^{p-1} K(i), \tag{27}$$

where, by expansion of the Γ-functions by Stirling's formula,

$$\left.\begin{aligned} K(i) &= qnt\left\{\frac{1}{n-i}+\frac{1}{2}\frac{q+2}{(n-i)^2}+\ldots\right\} \\ &\quad +2qn^2\frac{t^2}{2\,!}\left\{\frac{1}{(n-i)^2}+\frac{q+2}{(n-i)^3}+\ldots\right\} \\ &\quad +8qn^3\frac{t^3}{3\,!}\left\{\frac{1}{(n-i)^3}+\frac{3(q+\frac{1}{2})}{(n-i)^4}+\ldots\right\}+\ldots, \\ K &= pqt\left\{1+\frac{p+q+1}{2n}\right\} \\ &\quad +2pq\frac{t^2}{2\,!}\left\{1+\frac{p+q+1}{2n}\right\}^2 \\ &\quad +8pq\frac{t^3}{3\,!}\left\{1+\frac{p+q+1}{2n}\right\}^3+\ldots \end{aligned}\right\} \tag{28}$$

approximately. Since for the special case $p = 1$, $q = 2$ (or equivalently $p = 2$, $q = 1$), the distribution of Λ is given by

$$p(\Lambda) \propto \Lambda^{\frac{1}{2}(n-2)}\,d\log\Lambda,$$

so that

$$\chi^2 = -(n-2)\log\Lambda$$

exactly, the approximate formula

$$\chi_1^2 = -\{n-\tfrac{1}{2}(p+q+1)\}\log\Lambda \tag{29}$$

is suggested in general.

Table I gives comparisons for the case $q = 3$, for p equal to 1 or 2, the value of $\varLambda$ or $\sqrt{\varLambda}$ respectively being obtained from Fisher's z table (see (1), pp. 338–9). Formula (29) would only be used in practice of course for *both* p and q greater than 2, but while both from the nature of the correcting factor and from the table the approximation evidently begins to fail as $(p+q)/n$ increases, the use of $\varLambda$ seems unlikely for larger values of p and q except in extensive (e.g. correlational) investigations for which n would also be large and the approximation thus be valid.

TABLE I. *Values of* χ_1^2 *for* $q = 3$

p	$P = 0{\cdot}05$		$P = 0{\cdot}01$	
n	1	2	1	2
10	7·89	12·82	11·48	17·17
20	7·83	12·63	11·37	16·87
30	7·82	12·61	11·35	16·84
∞	7·815	12·592	11·341	16·812

9. In such investigations it is alternatively possible (cf. section 7) that the significance of the largest root λ_1 would be unquestioned, but a test of homogeneity of the residual variation would be required. That is, since

$$\begin{aligned}\chi_1^2 &= -\{n-\tfrac{1}{2}(p+q+1)\}\log\varLambda\\ &= \sum_{i=1}^{p} -\{n-\tfrac{1}{2}(p+q+1)\}\log(1-\lambda_i)\end{aligned}$$

($p \leqslant q$, say), we require to test

$$\varLambda' \equiv \prod_{i=2}^{p}(1-\lambda_i)$$

by the formula

$$(\chi_1^2)' = \chi_1^2 + \{n-\tfrac{1}{2}(p+q+1)\}\log(1-\lambda_1).$$

The effect of selection for the largest root might justifiably be ignored if the root λ_1 were known to be large because of a real association between S_2 and S_1. On the other hand, the components of $\varLambda$ have respectively, not the q degrees of freedom attributable to each of them if the division of S_2 were made without reference to S_1, but

$$p+q-1,\quad p+q-3,\quad \ldots,$$

and the χ^2 approximation for $\varLambda'$ has $(p-1)(q-1)$ degrees of freedom (and in general $(p-r)(q-r)$, where r is the number of components eliminated). Owing to the non-linearity of some of the restrictions, and the independence of the components being only approximate, a more detailed investigation of the range of validity of the χ^2 test for this case (i.e. after the elimination of one or more *non-zero* "canonical components") would no doubt be of value.

REFERENCES

(1) BARTLETT, M. S., *Proc. Cambridge Phil. Soc.* 30 (1934), 327.
(2) AITKEN, A. C., *Proc. Edinburgh Math. Soc.* 5 (1936), 37.
(3) PEARSON, K., *Biometrika*, 8 (1911–12), 437.
(4) WRIGHT, S., *J. Agric. Res.* 20 (1920), 557.
(5) HOTELLING, H., *Ann. Math. Stat.* 2 (1931), 360.
(6) WILKS, S. S., *Biometrika*, 24 (1932), 471.
(7) HOTELLING, H., *J. Educ. Psychol.* 24 (1933), 417 and 498.
(8) HOTELLING, H., *J. Educ. Psychol.* 26 (1935), 139.
(9) HOTELLING, H., *Psychometrika*, 1 (1936), 27.
(10) HOTELLING, H., *Biometrika*, 28 (1936), 321.
(11) FISHER, R. A., *Ann. Eugen.* 7 (1936), 179.
(12) TURNBULL, H. W. and AITKEN, A. C., *Introduction to the theory of canonical matrices* (Blackie, 1932).
(13) BARTLETT, M. S., *Proc. Roy. Soc.* 160 (1937), 268.
(14) AITKEN, A. C., *Proc. Roy. Soc. Edinburgh*, 57 (1937), 269.