The Study of Biological Versus Statistical Variation in Multivariate Morphometrics: The Descriptive Use of Multiple Regression Analysis

Gene H. Albrecht

Systematic Zoology, Vol. 28, No. 3. (Sep., 1979), pp. 338-344.

THE STUDY OF BIOLOGICAL VERSUS STATISTICAL VARIATION IN MULTIVARIATE MORPHOMETRICS: THE DESCRIPTIVE USE OF MULTIPLE REGRESSION ANALYSIS

GENE H. ALBRECHT

Abstract

Albrecht, G. H. (Department of Anatomy, School of Medicine, University of Southern California, Los Angeles, California 90033) 1979. The study of biological versus statistical variation in multivariate morphometrics: The descriptive use of multiple regression analysis. Syst. Zool. 28:338–344.—Multivariate statistical techniques (such as canonical variate and principal component analyses) are often used to ordinate or summarize morphometric data to facilitate biological interpretation of the morphological relationships under study. While the major axes of statistical variation which are derived in such analyses may have direct biological significance, there is no *a priori* reason that the biological and statistical determinants of morphological variation necessarily be concordant. Multiple regression provides a simple means of identifying and describing the maximum degree of relationship between (1) a variable, such as size or latitude, which is thought to have some biological relevance to the problem at hand, and (2) the set of uncorrelated variables, such as canonical or principal component variates, which represent the major axes of statistical variation and which may be thought of as a convenient, analytically efficient system of reference axes describing the multivariate data space. Of particular significance is the ability to examine the full multidimensional space and detect biological information having an angular relationship to the major axes of statistical variation. [Multiple regression analysis; multivariate analysis; morphometrics; statistical and biological variation.]

Multivariate statistical techniques are often used to ordinate morphometric data so that biological parameters underlying morphological relationships among individuals or groups may be more readily discovered. Commonly used techniques include canonical variate, discriminant function, principal component, and principal coordinate analyses (see Blackith and Reyment, 1971, for definitions and examples). All have in common a primary purpose of summarizing multivariate data in a relatively few dimensions that retain the majority of information formerly dispersed among the larger array of original variables. An additional advantage shared by all is the lack of statistical correlation among the transformed variates as compared to the complex of statistical dependencies usually found among the original variables. Such a reduction in both dimensionality and correlation results in a greater probability that the investigator will be able to make biologically relevant statements about the morphometric relationships under study.

Populations (or individuals) are often found to be ordered on the first few transformed variates of a multivariate analysis according to morphological gradients suggestive of differences in size, shape, time, function, behavior, or ecology. For example, Oxnard (1967; nine measurements of the scapula of 27 genera of primates) interpreted the first canonical variate as reflecting the extent to which the shoulder is subjected to compressive or tensile forces, the second canonical variate as reflecting relative degrees of arboreality or terrestriality, and the third canonical variate as reflecting the uniqueness of the human condition. Johnston and Selander (1971; 16 measurements of the skeleton of 33 populations of North American house sparrows) interpreted the first and second principal component variates, after regression analyses involving 15 environmental variables, as reflecting classic examples of Bergmann's and Allen's ecogeographic rules, respectively. Jantz (1973; 15 measurements of the skull of five archaeological populations of Arikara Indians) interpreted the first canonical variate for males (the second for females) as reflecting directional microevolutionary changes correlated with the influx of European trade artifacts. These few examples demonstrate the biological insights that are capable of being derived from the results of multivariate analyses.

Not all multivariate results can be interpreted with the same facility as those just described. For example, Howells (1973; 70 measurements of the skull of 17 populations of recent humans) interpreted the first canonical variate as a "primary human discriminator" and the second canonical variate as a general contrast between major geographic areas. However, he was unable to detect a general size gradient lying obliquely in the plane defined by the first two canonical variates (Albrecht, in prep., as identified by the method described herein). Howells (1973), in apparent prophecy of the present work, provided a perceptive account of the interpretative problems which accompany multivariate studies; in particular, he noted the difficulty of discerning biological variation which may not coincide with the major axes of statistical variation.

In assigning biological interpretations to multivariate results there is no *a priori* reason to either (1) treat separately each axis or variate which describes the transformed multivariate data space, or (2) believe that the statistical arrangements achieved in the form of axes or variates necessarily have a precise or unique biological reality. Rather, when multivariate methods are used to summarize and describe—that is, ordinate—morphometric data, the biological determinants of morphological variation should be sought with respect to the relative and absolute positions of groups or individuals in the *full* multivariate data space. This approach relegates the major axes of statistical variation to the role of convenient reference axes which may or may not be of direct biological import; these axes take value from the analytic efficiency with which they describe and display the multivariate data. The results of Howells (1973; as reinterpreted by Albrecht, in prep.) and Albrecht (1978; see the example below) offer sufficient support of the premise that biological and statistical variation are not necessarily concordant.

The present study demonstrates how multiple regression serves as a simple exploratory tool that facilitates the interpretation of multivariate morphometric results in terms of meaningful biological relationships. Specific application is directed at describing the relationship between (1) an independently designated variable thought to have some biological relevance to the morphometric problem under consideration (the criterion variable; e.g., size, latitude, humidity, or prey size), and (2) the major axes of statistical variation as defined by multivariate statistical procedures applied to the original set of morphometric data (the predictor variables; e.g., canonical or principal component variates). Of particular significance is the ability to look in all directions of the full multivariate data space so that biological information having an angular relationship with respect to the major axes of statistical variation can be readily detected and easily described. The intent is to emphasize the descriptive, rather than the statistical, use of multiple regression analysis; the theoretical and practical aspects of hypothesis testing or other statistical elaborations may be obtained from standard textbooks such as Snedecor and Cochran (1967).

MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis allows for the determination of the degree of relationship between (1) a single criterion variable Y, and (2) a set of p predictor variables X_1, X_2, . . . , X_p (the formulations of Tatsuoka, 1971, are followed here).[1] This is accomplished by constructing a linear combination $\tilde{Y}$ from the set of predictor variables such that the difference between the criterion variable Y and this new "predicted" variable $\tilde{Y}$ are minimized. The outcome is the constant a and the set of regression coefficients $b' = [b_1, b_2, \ldots, b_p]$ such that $\tilde{Y} = a + b_1X_1 + b_2X_2 + \ldots + b_pX_p$ where the sum of squared errors $e^2 = \Sigma\,(\tilde{Y} - Y)^2$ is as small as possible. Finding the set of regression coefficients b reduces to computing the matrix product

$$b = S_{xx}^{-1}S_{xy} \qquad (1)$$

where S_{xx}^{-1} is the inverse of the $p \times p$ covariance matrix for the predictor variables and S_{xy} is the vector of covariances between the p predictor variables and the criterion variable. In practice, the sums-of-squares-and-cross-products or the correlation analogs of the above matrices yield identical solutions except for a scaling factor which affects the "predicted" variable $\tilde{Y}$. The constant a is

$$a = Y - \Sigma\,(b_iX_i) \qquad (i = 1, p) \qquad (2)$$

using the mean values for the criterion and predictor variables.

The multiple correlation coefficient R, which represents the correlation between the criterion variable Y and the "predicted" variable $\tilde{Y}$, is

$$R = \sqrt{\Sigma(B_i/r_{iy})} \qquad (i = 1, p) \qquad (3)$$

where r_{iy} is the correlation coefficient between the ith predictor variable X_i and the criterion variable Y, and B_i is the standardized partial regression coefficient of Y on X_i. The B_i's are

$$B_i = b_i(s_i/s_y) \qquad (4)$$

where s_i and s_y are the standard deviations of X_i and Y, respectively. The squared multiple correlation coefficient R^2 is the proportion of variance shared by the criterion and "predicted" variables.

APPLICATION TO MULTIVARIATE RESULTS

The above calculations are somewhat simplified for regression of the criterion variable on *uncorrelated* predictor variables. This situation is attained when analyzing the relationship between (1) some variable thought to be important in explaining the underlying structure of morphometric data (e.g., size, latitude, or another morphological character), and (2) the set of variates which represents the major axes of statistical variation as derived from the application of some multivariate statistical procedure (e.g., canonical variates, principal component variates, or any other variates whose intercorrelations are all zero).

Assume a multivariate procedure applied to some original data set yields p transformed predictor variables $X_1, X_2, \ldots, X_p$, say canonical variates, which are uncorrelated and whose variances are $s_1, s_2, \ldots, s_p$.[2] An external criterion variable Y, say latitude, has variance equal to s_y and covariances with the predictor variables equal to $s_{1y}, s_{2y}, \ldots, s_{py}$. Since the off-diagonal elements (i.e., the covariances among the predictor variables) of the matrix S_{xx} of equation (1) are zero, the elements of the vector b of regression coefficients are

$$b_i = s_{iy}/s_i^2 \qquad (i = 1, p). \qquad (5)$$

[1] Multiple regression is a restricted case of canonical correlation analysis which finds the highest degree of relationship between multiple predictor variables and *multiple* criterion variables (see Glahn, 1968; and Tatsuoka, 1971). When canonical correlation analysis is limited to a single criterion variable, the canonical correlation is equal to the square of the multiple regression coefficient. When canonical correlation analysis is further limited to uncorrelated predictor variables, the eigenvector associated with the canonical correlation is equal to the vector of multiple regression coefficients multiplied by a scalar (s_1/s_y); all other calculations leading to Table 1 remain the same.

[2] These variances are usually the eigenvalues derived from the multivariate analysis of the original set of variables. In practice it is necessary to define the exact covariance matrix which is being considered. In the case of canonical variates, the between-groups covariance matrix would most likely be of interest although the pooled within-groups dispersion also is characterized by uncorrelated variates.

The constant a is calculated according to equation (2). Substituting terms from equations (4) and (5) into equation (3) yields the squared multiple correlation coefficient R^2 as the sum of the squared correlation coefficients r_{iy}^2

$$R^2 = \Sigma\,(s_{iy}^2/\sqrt{s_i^2 s_y^2})^2 = \Sigma\,r_{iy}^2 \quad (i = 1, p) \quad (6)$$

where r_{iy} is the ratio of variances and covariances for the criterion and predictor variables

$$r_{iy}^2 = s_{iy}/\sqrt{s_i^2 s_y^2}. \quad (7)$$

The squared correlation coefficient r_{iy}^2 is the proportion of variance shared by the criterion variable and the ith predictor variable. The multiple regression coefficient R follows directly from equation (6).

When the predictor variables are considered as a descriptive system of axes for the multidimensional data space, the vector *b* whose elements are

$$\tilde{b}_i = b_i/\sqrt{\Sigma\,b_i^2} \quad (i = 1, p) \quad (8)$$

gives the direction cosines for determining the position of the new axis which corresponds to the "predicted" variable $\tilde{Y}$ (i.e., the vector *b* is normalized such that $b_i' b_i = 1$. Thus, $\tilde{b}_i$ is the cosine of the angle θ_i between the "predicted" variable $\tilde{Y}$ and the major axis of statistical variation represented by the ith predictor variable.

The variance $s_{\tilde{y}}^2$ of $\tilde{Y}$ is

$$s_{\tilde{y}}^2 = \tilde{b}' S_{xx} \tilde{b} = \Sigma\,(\tilde{b}_i s_i^2) \quad (i = 1, p). \quad (9)$$

The calculated value, as compared to the total variance of the predictor variables ($\Sigma\,s_i^2$, $i = 1, p$), measures the importance of the "predicted" variable in explaining the observed morphometric differences. This is meaningful only if the multiple correlation coefficient indicates that the "predicted" variable $\tilde{Y}$ is a reasonable estimate of the criterion variable Y.

Examination of the above equations reveals multiple regression to be based on simple calculations involving (1) the variances of the predictor variables, (2) the variance of the criterion variable, and (3) the covariances between the criterion and the predictor variables. Computations are simplified because the variances of the predictor variables (often known as eigenvalues associated with the major axes of statistical variation) are available as part of the output of the common computer programs which perform multivariate analyses.

A MULTIVARIATE EXAMPLE

Albrecht (1978, table XV), on the basis of 24 measurements of the craniofacial skeleton, calculated canonical variate means for 35 populations of adult males belonging to the Old World primate genus *Macaca*; for simplicity, only the first five of the total of 24 canonical variates are retained here. Albrecht (1978, table VIII), on the basis of a geometric approximation of skull volume, also calculated a size variable for the 35 populations of macaques. The question is whether the populations of macaques are ordered in the multidimensional canonical space according to a simple gradient of increasing size.

The five canonical variates are the predictor variables $X_1, X_2, \ldots, X_5$ whose variances s_i^2 are given in column (1) of Table 1.[3] All covariances among these five variates are zero since canonical variates, by definition and construction, are statistically independent of one another with respect to the total between-groups variation. The volumetric size variable is the criterion variable Y whose variance

[3] Care must be taken in calculating variances and covariances since some multivariate analyses may be weighted according to sample sizes. For example, canonical variate analysis is often based, as it was here, on a weighted between-groups covariance matrix of the raw data; accordingly, calculations involving the canonical variates should be similarly weighted by the sample sizes of the groups involved. In the present example, the between-groups variance of the ith canonical variate is $(\Sigma\,\bar{X}_{ij}^2 n_j)/(g - 1)$, where $\bar{X}_{ij}$ is the group mean of the jth group on the ith canonical variate expressed as a deviation score from the grand mean, n_j is the sample size of the jth group, and g is the number of groups. The variance of the size variable and the covariances between the size variable and the canonical variates are similarly weighted according to sample sizes.

TABLE 1. STATISTICS FOR MULTIPLE REGRESSION OF SIZE (CRITERION VARIABLE Y) ON CANONICAL VARIATES (PREDICTOR VARIABLES X_i) FOR MALE MACAQUES. SEE TEXT FOR EXPLANATION OF COLUMN HEADINGS.

Canonical Variate (X_i)	(1) s_i^2	(2) $s_{iy}^2 \times 10^{-2}$	(3) r_{iy}	(4) r_{iy}^2	(5) b_i	(6) $\tilde{b}_i$	(7) θ_i°
1	18.95	15.11	0.881	0.776	.7974	.7569	40.8
2	8.26	4.72	0.416	0.173	.5714	.5424	57.2
3	3.29	0.15	0.020	0.000	.0456	.0433	87.5
4	2.83	0.68	0.103	0.011	.2403	.2280	76.8
5	1.89	0.56	0.103	0.011	.2963	.2812	73.7
$s_y^2 = 1.552 \times 10^{-3}$		$a = 0.0$	$R = 0.985$	$R^2 = 0.971$	$s_{\tilde{y}}^2 = 13.59$	$\Sigma\ s_i^2 = 35.22$	

is $s_y^2 = 1.552 \times 10^{-3}$. The covariances s_{iy}^2 calculated between the size variable and each of the canonical variates are given in column (2). The correlation coefficients r_{iy} between the size variable and each of the canonical variates, calculated from the variances and covariances as per equation (7), are given in column (3); the squares r_{iy}^2 of these correlation coefficients are given in column (4). The regression coefficients b_i, calculated as the ratio of column (2) to column (1) as per equation (5), are given in column (5). The regression constant, as per equation (2), is $a = 0.0$ since all values in this analysis represent deviation scores from the grand means of the criterion and predictor variables. The normalized direction cosines $\tilde{b}_i$, as per equation (8), are given in column (6); the corresponding angles θ_i are given in column (7). The squared multiple regression coefficient, calculated by summation of column (4) as per equation (6) is $R^2 = 0.971$; the multiple correlation coefficient is $R = 0.985$. The variance of the "predicted" size variable $\tilde{Y}$, as per equation (9), is $s_{\tilde{y}}^2 = 13.59$ which represents 38.6 percent of the total between-groups variation for the five canonical variates ($\Sigma\ s_i^2 = 35.22$).

Multiple regression analysis reveals size to be an important factor in interpreting the morphometric relationships among the 35 populations of macaques. An optimal linear combination of the five canonical variates is highly correlated with size ($R = 0.985$) and explains a significant part (38.6%) of the total between-groups variation. Of the canonical variates which define the five-dimensional data space, the first two are correlated with size and contribute significantly to the determination of the size vector. The last three canonical variates, while they do account for 22.7 percent of the total between-groups variation, are minimally correlated with size and contribute little to the determination of the size vector; indeed, the multiple correlation coefficient is reduced by only 0.022 if the last three canonical variates are excluded from consideration. The study of size variation among the macaques is, therefore, reasonably limited to the plane of the first two canonical variates.

Utilization of only the first two canonical variates requires the direction cosines $\tilde{b}_i$ be renormalized as per equation (8) for which now $p = 2$. The relevant elements of *b* become $\tilde{b}_1 = 0.8128$ and $\tilde{b}_2 = 0.5825$, and the corresponding angles which relate $\tilde{Y}$ to the first and second canonical variates are 35.6° and 54.4°, respectively. Using these angles, the "predicted" size variable $\tilde{Y}$ is plotted in Fig. 1 relative to the distribution of the populations of macaques on the first two canonical variates. The plotted size variable remains highly correlated with the volumetric size variable ($R = 0.974$) and now explains 56.3% of the total between-groups variation in the plane of these two canonical variates.

As previously discussed by Albrecht (1978), the plot of the first two canonical variates demonstrates a dichotomy in the

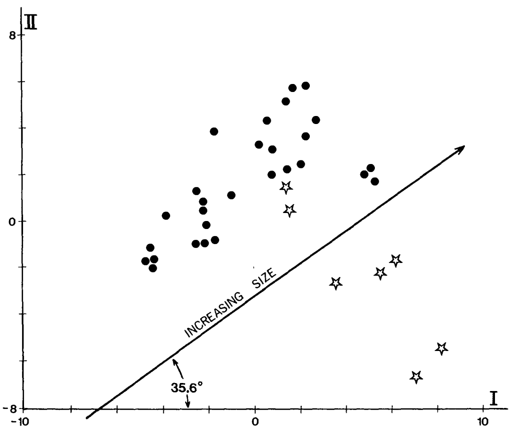

FIG. 1.—Canonical variates one and two for the analysis of 35 populations of male macaques. Only group centroids are shown with the Sulawesi macaques indicated by stars. The gradient of increasing size which corresponds to the "predicted" size variable Ŷ is shown. See Table 1 and Albrecht (1978, Table XV) for statistics.

nature of variation which characterizes (1) those macaques endemic to the Indonesian island of Sulawesi, formerly known as the Celebes, and (2) all the other macaques of the genus. The non-Sulawesi macaques are ordered parallel to the size gradient such that size differences represent the major component of variation in the craniofacial skeleton. In contrast, the Sulawesi macaques are ordered orthogonal to the size vector along an axis which may be regarded as representing size-independent shape information. Thus, given that the overall range of variation in skull morphology is comparable in magnitude, the non-Sulawesi macaques are characterized by relatively small shape differences and large size differences, and the Sulawesi macaques are characterized by relatively large shape differences and small size differences. The significance of these size-shape contrasts in the skull of the macaques relates to differing expressions of ecogeographic and speciation phenomena.

Interpretations of multivariate results are often limited to the major axes of statistical variation as determined by the particular analytic procedure employed. The present example from the skull morphology of macaques, and that mentioned earlier for human crania, demonstrate that the statistical and biological determinants of morphological differences need not necessarily be concordant. With this possibility in mind, the foregoing formulation of multiple regression analysis represents one easily applied, descriptive method by which biological parameters underlying morphometric data may be more readily elucidated and confirmed.

ACKNOWLEDGMENTS

I thank Charles E. Oxnard and Patricia L. Mensah of the University of Southern California, Dwight W. Read of the University of California at Los Angeles, and Pete E. Lestrel of the Veterans Administration Hospital, Sepulveda, California, for their critical evaluation of this manuscript. The journal reviewers (including N. A. Campbell of the Division of Mathematics and Statistics, CSIRO, Wembley, Australia) provided valuable suggestions for the revision of this work. This research was supported by a Biomedical Research Support Grant (NIH 5-S07-RR05356) to the School of Medicine of the University of Southern California.

REFERENCES

ALBRECHT, G. H. 1978. The craniofacial morphology of the Sulawesi macaques: multivariate approaches to biological problems. Contrib. Primatol. 13:1–151.

ALBRECHT, G. H. Size variation in the craniofacial skeleton of recent human populations. (Manuscript in preparation.)

BLACKITH, R. E., AND R. A. REYMENT. 1971. Multivariate morphometrics. Academic Press, New York.

GLAHN, H. R. 1968. Canonical correlation and its relationship to discriminant analysis and multiple regression. J. Atmos. Sci. 25:23–31.

HOWELLS, W. W. 1973. Cranial variation in man: a study by multivariate analysis of patterns of differences among recent human populations. Papers of the Peabody Museum of Archaeology and Ethnology, Vol. 67, Harvard University, Cambridge, Massachusetts.

JANTZ, R. L. 1973. Microevolutionary change in Arikara crania: a multivariate analysis. Amer. J. Phys. Anthrop. 38:15–26.

JOHNSTON, R. F., AND R. K. SELANDER. 1971. Evolution in the house sparrow. II. Differentiation in North American populations. Evolution 25:1–28.

OXNARD, C. E. 1967. The functional morphology of the primate shoulder as revealed by comparative anatomical, osteometric and discriminant function techniques. Amer. J. Phys. Anthrop. 26:219–240.

SNEDECOR, G. W., AND W. G. COCHRAN. 1967. Statistical methods. Iowa State University Press, Ames, Iowa.

TATSUOKA, M. M. 1971. Multivariate analysis: Techniques for educational and psychological research. John Wiley and Sons, New York.

Manuscript received February 1979
Revised May 1979