MULTIVARIATE STATISTICAL ANALYSIS
FOR MATHEMATICS MAJORS

GUDMUND R. IVERSEN
SWARTHMORE COLLEGE

TECHNICAL REPORT NO. 90-009

SLAW is supported by a grant from the Exxon Corporation
Statistics in the Liberal Arts Workshop (SLAW) is a group of educators concerned with the teaching of statistics. The workshop was initially funded by the Sloan Foundation. Continuing support has been provided by the Exxon Corporation.

Participants in the summer of 1991 are:

Donald L. Bentley
Pomona College

George W. Cobb
Mt. Holyoke College

Janice Gifford
Mt. Holyoke College

Katherine T. Halvorsen
Smith College

Homer T. Hayslett, Jr.
Colby College

Gudmund Iversen
Swarthmore College

Robin H. Lock
St. Lawrence University

David S. Moore
Purdue University

Thomas L. Moore
Grinnell College

Norean Radke-Sharpe
Bowdoin College

Rosemary Roberts
Bowdoin College

Dexter Whittinghill III
Colby College

Jeffrey Witmer
Oberlin College
Multivariate statistical analysis for mathematics majors

Gudmund R. Iversen
Swarthmore College

In the discussion above Roberts suggests the possibility of an applied statistics course for mathematics majors being a course in multivariate statistical analysis. The purpose of this paper is to expand on her proposal and to discuss ways in which such a course could be organized and taught.

The main reason for her suggestion that a course in multivariate statistics would be suitable and of great interest to mathematics majors is that multivariate statistics is so closely related to other parts of the mathematics curriculum. In particular, multivariate statistics is heavily dependent upon linear algebra; indeed some people claim that this part of statistics is nothing but linear algebra. It is also a part of statistics than can be approached at many different mathematical levels, and this flexibility means that such a course can be tailor made to the interests and backgrounds of both the students and the instructor. It is also a topic where major parts can be taught without any prior knowledge of statistics. For a full understand of multivariate statistics it is necessary to have knowledge of other parts of statistics, but that need not come until later in the life of a student.

Because of its close relationship to such a central part of the mathematics curriculum as linear algebra and no need for prior statistical knowledge, multivariate statistics lends itself particularly well to be taught to students before they take additional courses in probability and statistics. Typically, statistics students take multivariate
statistics after they have had an introduction to mathematical statistics, but that makes their multivariate course different from the one we are discussing here.

A multivariate statistics course as a first course in statistics might stimulate students to continue taking other statistics courses, and as statisticians we see such a multivariate statistics course as a way of attracting students to our field. The traditional course in mathematical statistics often becomes a mix of seemingly unrelated and different parts, and it does not get very far towards showing how statistics is actually used on real problems. The proposed course in multivariate statistics would counteract this impression of statistics.

Such a course would also be beneficial to mathematics majors who would not go on to take additional statistics courses. It would show such students that mathematical methods do not exist in a void, but that there are real applications to abstract mathematical ideas. From as early on as word-problems in calculus, students see that mathematics can be used for things other than just solving mathematical problems at the end of a chapter in a textbook. But it is not easy to go from these simplified calculus problems to showing mathematics majors how their subject of study is used to solve practical problems in the real world. The mathematics majors may have taken an occasional course in something like physics, and there they would see actual uses of mathematics. But the core of the mathematics curriculum for college majors is strangely void of real uses of mathematics and mainly seems to deal with mathematics as a very closed system only good for using theorems to prove other theorems.

Multivariate statistics typically deals with the situation where there is one dependent variable $Y$ that is thought to be influence by two or more independent variables $X_1, X_2, \ldots, X_p$. (If there is only one independent variable $X$, then we have a bivariate instead of a multivariate analysis.) For example, as a dependent variable we may have data on death rates from cervical cancer for all the counties in the country. When we glance at these 30xx numbers we see that they differ from each other, and we want to study why it
is that these numbers are different. It may be that it is thought that this variable is influenced by three other variables: degree of urbanization, percentages minorities, and percentage of the work force employed in chemical industries. We can then collect data on these variables, and if we place the data for a county in a row and the data on a variable in a column, we get a matrix with 30xx rows and 4 columns. The numbers in this matrix form the basis for a multivariate analysis of this problem.

One goal in the analysis of these data is to express each value of $Y$ as a linear combination of the $X$-values for that country plus a residual term $E$. Thus, we are looking for an equation of the type

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + e_i$$

The same $b$-values are used for all the counties. Let the column of $Y$-values be denoted by the vector $Y$, and let $X$ be a matrix with 30xx rows, the first column consisting of 1s, the second column the observations on the $X_1$ variable, the third column the values of the $X_2$ variable, and the last column the values of the $X_3$ variable. Finally, let the vector $E$ consist of the columns of $e$s and the vector $B$ consist of a column vector with the four $b$s. With these symbols the equations above can be represented by the matrix equation

$$Y = XB + E$$

It is here the multivariate analysis starts, and the first the question is how to find the best values of the four $b$-coefficients through least squares multiple regression.

If the students in this course have not had linear algebra before, this setting provides an excellent example for the needs of matrix theory, thereby leading in to determinants and the solving of simultaneous equations. In such a course the learning of the linear algebra and the developing of multivariate statistical analysis would go hand in hand.

It is more reasonable, however, to expect the students to have had linear algebra before they take the course on multivariate statistical analysis. These students will be able to go through the least squares manipulations of the matrix equation above and the
derive the estimator for the B vector. From there it is natural to move on the n-dimensional geometry of multiple regression. With each of the four variables represented by a vector in n dimension, multiple regression simply becomes a projection of the Y vector on to the hyperplane defined by the X vectors. All the ordinary sums of squares in multiple regression become the squares of lengths of the corresponding vectors in this space. The degrees of freedom become the dimensionality of the spaces spanned by the various vectors, and cosines of angles between vectors equal the corresponding correlation coefficients. With some care and using only two X variables it is even possible to illustrate the angle that gives us a partial correlation coefficient between Y and one of the X variables while controlling for the effect of the other X variable.

This kind of matrix algebra and n-dimensional geometry is a very powerful way of dealing with multivariate statistical problems. It asks the students to draw on their knowledge of linear algebra, and at the same time it gives substantive and numerical answers to actual, real world issues. Linear algebra becomes more than simply an abstract manipulation of matrices and determinants, and it comes alive in a way that is not possible to achieve in a strict mathematics course.

Such a course also exposes the students to data analysis on the computer. The students are spared the tedious numerical work of having to do computations with matrices and determinants. Instead, all this work is done by the computer. That way it becomes much more realistic to add and delete variables to the analysis and see how such additions and deletions change the results.