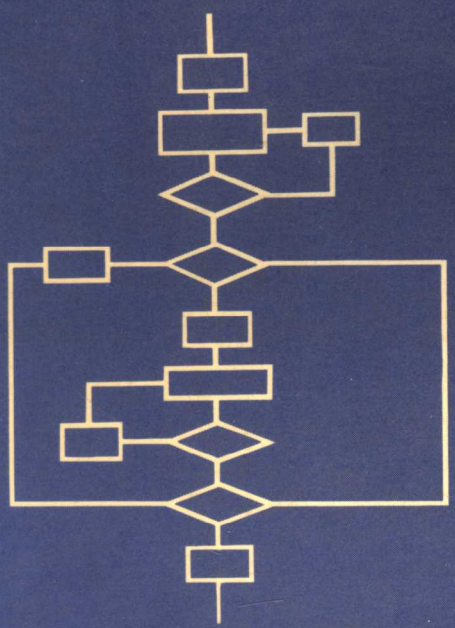
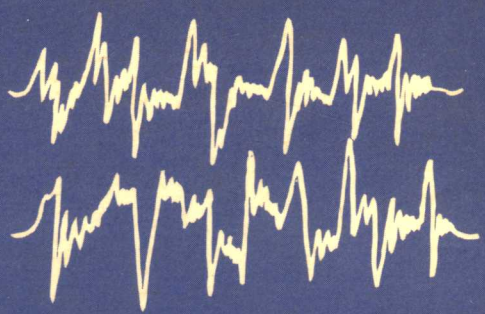
BMD

BIOMEDICAL COMPUTER PROGRAMS





HEALTH SCIENCES COMPUTING FACILITY

DEPARTMENT OF BIOMATHEMATICS

SCHOOL OF MEDICINE

UNIVERSITY OF CALIFORNIA, LOS ANGELES

UNIVERSITY OF CALIFORNIA PRESS

1977

BMD
Biomedical Computer Programs

W. J. DIXON
Editor

UNIVERSITY OF CALIFORNIA PRESS
BERKELEY · LOS ANGELES · LONDON

1977

This publication reports work sponsored under grant RR-3 of the Biotechnology Resources Branch of the National Institutes of Health.

UNIVERSITY OF CALIFORNIA PRESS,
Berkeley and Los Angeles, California

UNIVERSITY OF CALIFORNIA PRESS, LTD.
London, England

ISBN: 520-02426-5
Library of Congress Catalog Card Number: 72-98008

$9.85

Orders for this publication should be directed to one of the above addresses. Comments on programs or orders for tape copies of the programs should be addressed to:

Health Sciences Computing Facility
CHS Bldg., AV-111
University of California
Los Angeles, California 90024

Third Edition, 1973
Second Printing, 1974
Third Printing, 1976
Fourth Printing, 1977

This publication replaces
Numbers 2 and 3 of the series,
Publications in Automatic Computation

Manufactured in the United States of America

PREFACE

If you are entirely new to computers go immediately to Introduction Section V, page 60.

If you know computers but not much statistics, go to Sections I and III, pages 2 and 37.

If you know computers and statistics go to the Table of Contents or the list of programs in Section I-C, page 12.

If you have used the BMD manuals before continue with this preface.

This publication is a combination of the previous BMD and BMDX manuals with modification as follows:

1) The BMDX programs are renumbered into the BMD series.

2) Programs 3M, 6M and 1T are omitted since they are superseded by other programs. The set of programs 4S to 8S are omitted since they are now rarely used.

3) The Introduction has been extensively revised to simplify the search for appropriate programs and to ease the path for the beginner.

4) A number of updated references are included in the Introduction and in the individual program writeups.

5) The output for each program has been re-run to show the various changes made in the past few years. Some new examples are substituted (e.g., 7M).

6) An expanded explanation to the approach to 5V, 10V and 11V is included and more explanatory material has been added to 2R.

The new numbers for the X-programs are listed here.

12D	X84	8M	X72	7R	X85
13D	X70	9M	X75		
		10M	X74		
11S	X76	3T	X92	9V	X82
12S	X77	4T	X68	10V	X64
13S	X94	5T	X93	11V	X63
14S	X90			12V	X69

We appreciate the many contributions forwarded to us by users throughout the world and hope that they will continue to assist us with improvements in these programs.

It is impossible to give proper credit to them or to all of our own staff who have contributed to this manual. Some of this information is contained in the Annual Report of the facility.

Laszlo Engelman supervised the preparation of this edition.

In addition to the programs in this manual, a newer series, BMDP, is also available. The BMDP programs contain advanced features not found in the BMD programs, but do not completely replace them. References to the BMDP programs are given throughout this issue. A second Edition of the BMDP programs will be published by University of California Press in mid-1977 and will include the following programs:

P1D: Simple Data Description
P2D: Detailed Data Description, Including Frequencies
P3D: Comparison of Two Groups with t Tests
P4D: Single Column Frequencies -- Numeric and Nonnumeric
P5D: Histograms and Univariate Plots
P6D: Bivariate (Scatter) Plots
P7D: Description of Groups (Strata) with Histograms and Analysis of Variance
P8D: Missing Value Correlation
P9D: Multiway Description of Groups
P1F: Two-way Frequency Tables -- Measures of Association
P2F: Two-way Frequency Tables -- Empty Cells and Departures from Independence
P3F: Multiway Frequency Tables -- the Log-Linear Model
P1L: Life Tables and Survival Functions
P1M: Cluster Analysis of Variables
P2M: Cluster Analysis of Cases
P3M: Block Clustering
P4M: Factor Analysis
P6M: Canonical Correlation Analysis
P7M: Stepwise Discriminant Analysis
PAM: Description and Estimation of Missing Data
P1R: Multiple Linear Regression
P2R: Stepwise Regression
P3R: Nonlinear Regression
P4R: Regression on Principal Components
P5R: Polynomial Regression
P6R: Partial Correlation and Multivariate Regression
P9R: All Possible Subsets Regression
PAR: Derivative-Free Nonlinear Regression
P1S: Multipass Transformation
P3S: Nonparametric Statistics
P1V: One-way Analysis of Variance and Covariance
P2V: Analysis of Variance and Covariance, Including Repeated Measures
P3V: General Mixed Model Analysis of Variance

W. J. Dixon
March 1977

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

INTRODUCTION

The uninitiated may wish to read Section V immediately before any other earlier sections. Those with some computer experience may wish to start with Section I. Those who have used BMD programs before may need only note some of the rearrangements and modifications as indicated in the Preface and the list of programs in Section I-C.

Section I. GENERAL CLASSIFICATION AND DESCRIPTION OF PROGRAMS

Many problems in medical research require extensive analyses of large amounts of data. As far as possible, the data handling process should be made automatic and rapid. The research worker should have the appropriate tools for effecting the types of analysis his research requires.

The types of programs and their specific forms for inclusion in this manual have been guided by the demands arising in the UCLA Medical Center for statistical and mathematical procedures to assist in a wide variety of research problems. When these needs could be generalized for a "package" program, it was developed with parameters specifying the problem to be chosen by the user. Revisions and improvements have been guided by the comments received from the users of these programs from many areas of application.

This publication has been prepared with the following goals to meet a portion of these research needs:

1. To provide programs for the commonly required tasks of data processing and statistical analysis.
2. To provide these programs in "package" form so that research workers may effect their desired computations with simple coded instructions.
3. To provide these "package" programs in a general form so that a wide variety of problems may be handled by each program by specifying the appropriate parameters of the problem.

The task of preparing the basic research data for computer analysis differs greatly from problem to problem. Problems in medical research involve data having many variables for each case or many observations but fewer variables. The remaining sections of this introduction present some of the most frequently requested information about the described programs.

Each program is designed to handle a certain type of data analysis task. However, a particular result may be obtained from more than one program. An overview of various analyses and the program applicable for solution are presented in section I-B, page 5.

The programs are arranged in six classes for the convenience of organization of this manual. Because of the nature of statistics, this classification is not unique. For example, programs performing multivariate analysis of variance computations which we have placed in class V could as well have been placed in class M.

The classes are:

D	Description and Tabulation
M	Multivariate Analysis
R	Regression Analysis
S	Special Programs
T	Time Series Analysis
V	Variance Analysis

An analysis usually starts with data screening (description and tabulation) programs and then proceeds with various types of analyses, each based on the findings of the preceding analyses.

A. How to Find the Program You May Wish to Use.

A new user of statistical packages may wish to become familiar first with just a few basic programs and then learn about some of the more powerful and more specialized programs as he gains sophistication and experience.

In the list of programs in section I-C a very restricted set which are easy to set up are indicated by **. The user can obtain a considerable amount of information from these programs, yet they all require only two or three control cards with the remainder of the cards being optional. Adding the programs indicated by * to the above creates a fundamental or 'minimum' set of programs. With these programs an extensive variety of statistical tasks can be performed.

The two-star programs in the description series (1D, 2D, 4D and 7D) provide descriptive statistics, scatter-plots, frequency counts and one-way analyses of variance. Program 2R provides very general regression capabilities; 1V performs a one-way analysis of variance. Programs 1M and 5M are the simpler multivariate analysis programs, providing principal component and discriminant analysis.

There are several comprehensive and powerful programs which give the user great freedom in specifying a very wide variety of problems and options. These programs are: 7D for preliminary data screening, 2R for regression, 7M for discrimination, 8M for factor analysis, 8V for analysis of variance, 10V for a combined analysis of variance and covariance with tests of linear hypotheses, and 9S which provides for general data reformatting and transformation. In addition, 5D features additional plotting provisions, 2S analyzes contingency tables and 2T computes spectral analysis of time series.

The following section contains problem-oriented comparative statements on the various programs. This is followed by a section, arranged in alphabetical class order, giving one-paragraph statements of the purpose of each program. The program descriptions in the body of the book contain detailed instructions and examples of the use of each program. At the end of this section (page 25) there are Check Lists for the four program series, D, M, R and V, which compare the features of the programs within each series.

B. Comparative Description of Programs

Data Screening and Description. In most statistical studies a considerable amount of editing and screening of the basic data are required before the more elaborate analyses of sections R, V and M can be carried out. Most of the programs in those classes require that no observation be missing. Also, various assumptions are made in the mathematical models associated with those analyses, and examination of data characteristics appropriate to these assumptions can be assisted with the programs in class D.

After the original data have been punched on cards, various graphical and descriptive characteristics of the data can be useful in detecting gross errors in observations, coding and keypunch errors, presence of inappropriate cases, etc. An inventory, column by column, of all legal punched codes can be obtained from program 4D. The frequency counts will be useful tabulations for any data coded as single-column information.

Many research workers wish to see their data in graphical form and they plot the data by hand or by using an automatic plotter. The computer's printer can provide graphical output that is adequate for many purposes. The scatter plots automatically available in some of the R and T class programs can be obtained for the input data by using 2D or 5D. Histograms can be obtained by using 5D or 7D. The latter creates individual histograms for each subgroup or for comparisons across strata of the input data and provides for the exclusion of special values from the histograms.

Cross-tabulations are often required as a preliminary to other analyses or as a final form for reporting. Program 8D provides a statistical description of each variable and cross-tabulation for pairs of variables that can have 34 by 99 categories. Variable "stacking" and elimination of empty rows or columns are possible. Program 9D provides for 20 by 20 cross-tabulation with ability to eliminate inappropriate cases. Program 2S can be used to construct contingency tables for one or more specified sets of intervals on each variable, and a chi-square, contingency coefficient and maximum-likelihood ratio are computed for each table before and after collapsing of tables based on small expected cell values.

Programs 10D and 11D are also useful for frequency counts. These programs list together all of the cases that are alike for all of the entered variables and thereby identify the "profiles" present in the various cases and their corresponding frequencies. An important use is examination of data for completeness of determinations for the several variables considered. Dichotomies can be defined for each variable according to the presence or absence of a determination on that variable. Those cases that are complete for any given set of variables can then be readily identified, as can the patterns of completeness. Another useful application is analysis of patterns of symptoms that are jointly present.

Means, standard deviations or variances, extreme values and counts of cases having non-excluded values for each variable are reported by many programs, the simplest of which is 1D. Program 2D is somewhat more general providing this type of descriptive information as well as the correlation coefficients computed from those cases which satisfy a user specified criteria. When there are unknown (missing) values within the data matrix program 3D can be used to provide the correlation coefficients, each coefficient computed from all existing pairs of values of the corresponding variables. For problems having an extremely large number of variables, a user specifiable portion of the correlation coefficient matrix can be obtained using 12D.

Data to be analyzed are frequently categorized into subgroups or strata. Programs 6D and 7D report means and standard deviations for each variable and correlation coefficients for each pair of variables within each stratum of cases. Univariate t-tests between means of two groups are obtainable from 13D, and multivariate comparisons are computed by 4M. Program 13D also reports the F-statistics and the appropriate probability levels for the equality of variances. The univariate comparisons of means for more than two groups are obtainable from 7D and the multivariate comparisons from 5M or 7M. Various multiple range tests can be performed for group means with program 7V.

Analysis of Variance. The purpose of the analysis of variance is the comparison of means when the data are grouped or classified in one or more ways. Many problems for which an analysis of variance is desired may be done very easily without the aid of the computer if the design is balanced and the problem is small. Also, it is possible to perform an analysis of variance by proper use of one of the regression programs (2R). Each of the programs in class V is designed to provide a solution for a class of possible deviations from the standard simple problem.

For a simple one-way design the program 1V or 7D can be used. For a one-way design with covariates program 4V or 9V can be used. Both of the programs allow analysis of various re-combinations of the group of cases.

Program 2V provides an analysis for full factorial designs with equal replications (all observations present) and provides polynomial breakdowns to aid in the study of linear, quadratic, etc. components of ordered classifications. Program 3V is similar to 2V but allows the introduction of covariates. Program 8V is also similar to 2V but handles all complete balanced designs including factorial, nested, and partially nested designs, and expected mean squares are computed for arbitrary sampling from finite or infinite populations.

The general linear hypothesis method in 5V and 10V provides a very flexible approach to analysis of variance. The user specifies the design, which may be factorial, nested, etc., and may specify contrasts, orthogonal or otherwise, for which the appropriate sum of squares will be computed. Using 10V is usually much easier since the program can generate design variables automatically. However, for nested designs with unequal numbers of subgroups 5V should be used. The program 6V is similar to 5V and 10V. It handles hypotheses like 10V but requires the design variable values to be recorded for each case; cases do not have to be sorted. All three of these programs perform Model I analysis of variance tests and test user specified hypotheses.

Multivariate analysis of variance problems can be handled by 11V or 12V. Program 11V computes likelihood-ratio statistics and approximate F-statistics to test multivariate hypotheses of the form $A\beta C'=D$, where β is a matrix of regression coefficients and A, C, and D are matrices specified by the user. Program 12V performs Model I univariate and multivariate analysis of variance and covariance for any hierarchical design with equal cell sizes.

In biological research, unequal groups or missing values seem to be almost inevitable. All programs, except 2V, 3V, 8V and 12V in the class V programs can be used for unequal group sized data. If a problem originally designed for analysis by 2V, 3V or 8V has a small number of missing values the analysis may be accomplished by using 3V with the introduction of special covariates. If larger amounts of data are missing or the groups are unequal in size, larger problems can be handled by program 5V or 10V by introducing or having the program generate appropriate design variables. By specifying the design variable values and covariates for each case, program 11V can perform the multivariate analysis of variance computation for a problem which may have been designed for 12V.

Regression. A regression analysis provides a relationship between a dependent variable and one or more independent variables. A linear relation is estimated by least-squares in programs 1R, 2R, and 3R and for polynomials in 5R. In program 4R a trigonometric function is used, in 6R, the exponential function $\alpha + \beta \rho^{x}$, and in 7R a user specified functional form. Program 11V can be used to solve multivariate regression problems.

Program 1R provides an analysis for one dependent and one independent variable. Since various groups of cases may be combined by the program, the computations for the analysis of covariance are also obtained and are reported in the usual form for an analysis of covariance. This program is a generalization of program 1V. The covariance model is $y = \mu_i + \beta x + e$.

If there are several independent variables, programs 2R or 3R may be used. These programs compute least-squares estimates of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ from given values of y and $x_1, \ldots, x_p$ for the model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e$$

Program 2R proceeds in a stepwise fashion by forming successive regression analyses, including the "best" variable first, finding next the variable whose additional contribution to the first is greatest, and so on. (The definition of "best" as used here is given in the Introduction to program 2R.) The use of a stepwise procedure has the advantage not only of providing useful analytic information to the user, but also of providing protection against failures in computational accuracy. Plots are available and may be useful in detecting lack of linearity, gross errors, etc. Variables may be forced into the regression or may be automatically deleted when their F-values become too low. Regression may be performed with zero intercept. Program 3R does not order variables and does not provide information at each step as does 2R, but it does provide for various combinations of cases in successive regression problems.

Program 4R "fits" a trigonometric series up to the ninth harmonic by least squares. This program is useful in the analysis of time series with fixed periods. For other time-series analyses, see the programs in Class T. The regression model in program 4R is

$$y = a_0 + \sum_{i=1}^{n} (a_i \cos(2\pi it/k) + b_i \sin(2\pi it/k)) + e$$

The program assumes that the given values of t are equally spaced and that the wave length of the fundamental harmonic is known.

The polynomial regression model in program 5R is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$$

The fitted regression and the original data can be plotted. The description of a number of physiological processes require the use of some form of the logistic function. The function may be "fitted" with the use of a form of the exponential function in program 6R.

Program 7R is a very flexible non-linear least-squares fitting program which is appropriate for a wide variety of problems that do not lend themselves to representation by equations linear in the parameters. The program computes a least-squares fit

$$y_k = f(x_{1,k}, \ldots, x_{t,k}; \theta_1, \ldots, \theta_p) + e_k$$

of a user specified function f to data values y_k and x_k by means of stepwise Gauss-Newton iterations on the parameters $\theta_1, \ldots, \theta_p$. Values of the parameters can be constrained to be in user specified intervals. The stepwise nature of the calculation provides a means of avoiding singularity problems by not modifying, at any stage, parameters that have become redundant.

Multivariate Analysis. Many research studies encounter multivariate observations, i.e., several variables are observed for each case. Analyses of these data use the information on the interrelationships (correlations) among these variables. When cases belong to two or more groups, we may wish to find functions of the variables which separate the groups. These functions may then be useful in predicting the group membership of new cases. Programs 4M, 5M and 7M compute linear discriminant functions. Program 4M computes a discriminant function for two groups and orders the cases by functional value, thereby illustrating the degree of success of discrimination.

For more than two groups (and for two groups as well) 5M or 7M can be used. These programs compute classification functions f_i, such that a case, x, is classified in group j, if $f_j(x) > f_i(x)$ for all i different from j. Although these programs are based on the concept of a discrimination or classification problem, they serve also as effective multivariate extensions of 7D or 1V. Program 7M in particular produces results that are quite helpful in interpreting the one-way analysis of variance aspects of multivariate data from several groups.

The computation in 7M is done in a stepwise manner, and results of each step are reported. These programs assume that the variables have equal or nearly equal covariance matrices.

We frequently wish to study the interrelations of variables. Programs 1M, 2M, 8M and 9M are designed for such investigations.

Programs 1M and 2M compute principal components for the data. These new variables are linear weighted sums of the original variables. The new variables are determined in order to provide a parsimonious summary of the original variables. The principal component values for each case are ordered in terms of their sizes. Additionally, program 2M can be used to investigate the relationship of other observed variables to these principal components by providing a regression on the principal components.

Program 9M, canonical correlation analysis, addresses itself to a slightly more general problem, that of the relationship between two sets of variables. The variables in each of the two sets are transformed into two new sets of variables, which are uncorrelated with other variables in their own sets, and are determined so the new variables will be maximally correlated between the sets. This analysis may aid in obtaining a simpler relationship among the observed variables.

Another type of investigation of interrelationships of variables, factor analysis, is performed by 8M. As in principal components analysis the purpose of this analysis is to express the original data in terms of a smaller set of variables, called factors. Rotation is used to concentrate loadings of the original variables on the extracted factors.

Screening for outlying cases can be done by program 10M, by computing the Mahalanobis distance of each case from the center of the distribution of the remaining cases. If the probability of the F-statistic corresponding to the greatest distance is smaller than a user-specified value, the case is removed.

Time Series. Long sequences of correlated, frequently time dependent, observations may be analyzed by programs in the time series class. These programs extract harmonic components, estimate frequency-response functions, and analyze evoked responses.

Program 2T is a simple spectral analysis program which begins by computing auto- and cross-covariances and transforms these to obtain power spectra, cross-spectra, frequency-response functions, and coherence functions.

Program 3T also computes auto-spectra, cross-spectra, and coherence functions, but it does so by directly decomposing the input series into frequency components using the Fast Fourier Transform. This gives a much faster program with considerably greater capacity. Program 3T also contains an extensive set of filtering options and may be used for this purpose alone.

Beginning with a sequence of cross-spectral matrices, normally generated by 3T, program 4T estimates multiple-coherence functions and frequency-response functions between a set of input and a set of output series. Confidence bands are obtained for the gain and phase of the components of the frequency-response functions. Estimates of the cross-spectra of the output series partialed on the input series are also obtained.

Program 5T computes averaged evoked response and standard deviations as a function of time from a stimulus pulse. The pulses on one (or two) pulse channels act as time markers for this averaging.

Data Formatting and Transformation. Although many of the programs provide some editing features, it may be desirable to perform any extensive modifications with the use of a separate program so that the data deck can be used directly with little further modification in a number of other programs. Program 9S provides a general type of editing process permitting transgeneration and allowing input and output in the form of punched cards, BCD tape, or binary tape.