1986

Choosing a statistical package for a microcomputer

Philip Dixon, *Cornell University*

CHOOSING A STATISTICAL PACKAGE FOR A MICROCOMPUTER

One of the appealing benefits of buying a microcomputer is the ability to do statistical analyses without competing for access and paying for time on a mainframe computer. The difficult decision is which statistical package to buy. Just as there is no single best microcomputer because everyone's needs are different, there is no single best statistical package. This article, based on my experience reviewing statistical packages for the Statistical Computing Support Group at Cornell, discusses some of the considerations in selecting a package. Although I will concentrate on the general-purpose packages, many of the same concerns are also appropriate for specialized packages, such as those available for time-series analysis, matrix operations, or econometrics.

Statistical packages can be judged by many criteria. My list (Table 1) includes some characteristics that are required for a package to be adequate. An additional set of characteristics make the difference between an adequate package and an excellent package. Finally, the personal considerations may or may not be important depending on your needs and familiarity with computers.

The first two required criteria are obvious. The package you buy must be able to run on your computer and it must do the analyses you regularly do. Statistical packages have been written for all of the common microcomputers, but the largest and most powerful are written for IBM-PC type computers and the Macintosh. Most packages allow for some data manipulation and transformation, plot histograms and scatterplots, and calculate descriptive statistics, regressions, *t* tests, chi-square tests and simple ANOVAs. If you need nonparametric tests, high-resolution graphics, unbalanced or multi-way ANOVA, or multivariate analyses, each of these is available in at least one general purpose package; check the vendor's information or a review of package capabilities (for example Lachenbruch 1983, Carpenter et al. 1984, Fridlund 1986, or Lehman 1986).

It is harder to judge the numerical accuracy of a package. Contrary to some popular opinion, just because a number was calculated by computer does not ensure its accuracy. Both round-off and truncation errors can be serious if poorly chosen algorithms are used. The formula for the variance taught in many introductory statistics courses is very susceptible to truncation error because the last step is to subtract two relatively large numbers. The variance of the three numbers 9000, 9001, and 9002 is 1, but if computer calculations are done in single-precision arithmetic, the computed answer is 0 (Anscombe 1967). Similar sorts of sums-of-squares calculations are often used in multiple regression and analysis of variance routines. Well-designed packages do calculations in double-precision arithmetic, use algorithms that are less influenced by truncation and round-off errors, and print warnings if the calculated results are potentially inaccurate.

Two data sets are often used to check numerical accuracy of multiple regression algorithms, which may be inaccurate if the independent variables are highly correlated. The Wampler (1980) data sets are constructed with different correlations between the variables. Lesage and Simon (1984) compare the performance of a selection of packages on these data sets; some packages calculate regression coefficients accurately, but others print inaccurate numbers or refuse to do the problem. Carpenter et al. (1984) present results for regression accuracy with the Longley (1967) data set, one specific case of correlation between seven economic variables. The regression coefficients calculated by different packages range from 2 to 9 digits correct (Carpenter et al. 1984).

A more serious, but less frequent, error occurs when some packages print an "answer" when in fact there is none. For example, a regression with two correlated predictor variables has no unique answer. An adequate package will either refuse to do the problem or print a warning message and ignore one of the predictor variables. One package (NCSS version 4.1) calculates and

Table 1. Characteristics of statistical packages.

Requirements for an acceptable statistical package:
Able to do the analyses you regularly use
Able to run on your computer
Numerically accurate
Desirable features of a good statistical package:
Easy to use, with good documentation
Warnings about inappropriate statistics
Easy to enter and manipulate data
Versatile, e.g. easy to link together analyses
Can name or label columns of data (=variables)
Accepts character data
Can do analyses on subgroups specified By classification variables
Speed of analysis
Able to handle large problems
Amount of output generated
Personal considerations:
Menu or command driven
Easy to repeat similar analyses of different data sets
Quality of graphical output

prints what appears to be a unique solution by setting the coefficient for the second variable to zero without any warning messages.

Even if the calculations are accurate, they may be inappropriate. Common examples are calculating a chi-square statistic when the expected frequencies are small or a pooled *t* test when variances are unequal. An excellent package will print warnings when statistical assumptions are stretched, but the responsibility for the correctness of the analysis rests with the user and statistician.

The rest of the desirable features are those that make it easier to organize the statistical analyses. Most microcomputer packages lack the data manipulation capabilities available in mainframe packages. Better packages allow one to transform variables easily, reshape the data set, and feed the results of one calculation into another analysis. For example, one typical problem I encounter is to take data collected from leaves on plants, calculate an average for each plant, then analyze the means. This can't be done without good data management capabilities, unless you print out the plant means, then type them back into a data set.

Usually associated with good data management is the ability to separate the data set into subgroups identified by a classification variable. With this BY variable capacity, one could calculate the plant means used above by entering two columns of data, one containing the value for each leaf, and one that identifies which plant it came from, rather than entering separate columns of data for each plant. Finally, I find it a great help to be able to label columns with some descriptive name and use character variables to identify observations.

Packages differ considerably in their speed and the size of problem they can handle. In general, packages that are written in interpreted BASIC are slower and smaller than those written in other languages. The advantage to a BASIC program is that the same package is usually available for more than one type of computer. Many of the larger packages can handle problems with over 100 variables and 1000 observations, but there is a

Table 2. Reviews of microcomputer statistical software in the *American Statistician.*

Package	Version	Volume and page numbers	Date
AIDA	9/82	**39**:70–72	Feb 1985
BMDPC	2.05	**39**:213–215	May 1985
Dasy	1.5	**39**:215–219	May 1985
GAUSS	1.38	**40**:167–169	May 1986
KEY-STAT		**40**:50–51	Feb 1986
MSUSTAT	2.20	**39**:72–74	Feb 1985
NCSS	4.1	**39**:315–318	Nov 1985
PC Anova, PC Statistician		**40**:164–167	May 1986
RATS	1.13	**40**:223–225	Aug 1986
SPSS/PC+		**40**:225–228	Aug 1986
STAN	II.0	**39**:146–148	May 1985
SYSTAT	1.1	**39**:67–70	Feb 1985

difference between being capable of handling that much data and easily handling that much data. Although the computing speed of a microcomputer may be equal to that of a mainframe, microcomputer printers are usually much slower. Large data sets usually result in lots of output; a desirable feature is to be able to control the amount and detail of the output.

The remaining considerations are important to me, but may not be for you. I strongly prefer command-driven programs because I do not like responding to long series of menus; if you use a program infrequently you may prefer menus. I often do the same analysis on more than one set of data, so I like the ability to construct a command file specifying my analysis. Finally, I currently use other programs to graph results, but there are some packages that include reasonably high-quality graphics.

This article describes a series of things to consider when choosing a statistical package. It is not another review of statistical packages because packages change so frequently. Recent comparative reviews are those by Lachenbruch (1983), Carpenter et al. (1984), Goodban and Hakuta (1984), Fridlund (1985, 1986), and Lehman (1986). Other sources of information are the vendor's information, the reviews published in this *Bulletin,* and a series of package reviews published in the *American Statistician* (Table 2). The *American Statistician* reviews are done like book review, so they differ in quality and content, but they almost always summarize the features and test the numerical accuracy of the package. Be warned that packages change, so some problems mentioned in a review may be corrected in later versions of the software.

Literature Cited

Anscombe, F. J. 1967. Topics in the investigation of linear relations fitted by the method of least squares. Journal of the Royal Statistical Society Series B. **29**:1–52.

Carpenter, J., D. Deloria, and D. Morganstein. 1984. Statistical software for microcomputers. Byte (April 1984):234–264.

Fridlund, A. J. 1985. Taking the bull by the horns: four statistics packages provide a range of power and features. InfoWorld (February 11):42–50.

———. 1986. Statistics software. InfoWorld (September 1, 1986):31–37.

Goodban, N., and K. Hakuta. 1984. Statistical quintet. PC World (September):186–195.

Lachenbruch, P. A. 1983. Statistical programs for microcomputers. Byte (November 1983):560–570.

Lehman, R. S. 1986. Macintosh statistical packages. Behavior Research Methods, Instruments, and Computers 18(2):177–187.

Lesage, J. P., and S. D. Simon. 1984. Numerical accuracy of statistical algorithms for microcomputers. Pages 53–57 *in* American Statistical Association Statistical Computing Section Proceedings. New York, New York, USA.

Longley, J. W. 1967. An appraisal of least squares programs for the electronic computer from the point of view of the user. Journal of the American Statistical Association **62**:819–841.

Wampler, R. H. 1980. Test procedures and problems for least squares algorithms. Journal of Econometrics **12**:3–22.

Philip Dixon
Cornell Plantations
1 Plantations Rd
Ithaca, NY 14850

MULTIVARIATE STATISTICAL PACKAGE FOR THE IBM PC NOW AVAILABLE

MVSP, a MultiVariate Statistical Package, is a program written for the IBM PC and close compatibles. It is geared towards simple analyses of small- to medium-sized data sets. It is also available for only the price of a disk and postage.

This package contains procedures to perform various ordination and clustering analyses. These procedures include: principal components analysis, reciprocal averaging, many similarity and dissimilarity measures, average linkage cluster analysis, and diversity indices. This program is menu-driven and easy to use, with all possible options being presented to you at each step. The data files may be created and maintained using any

database, spreadsheet, or word processor program which creates plain ASCII files.

A copy of this program may be obtained from the author by sending a check to cover the cost of a disk, mailer, and postage ($5 total). Or you may send a formatted, double-sided floppy disk and the cost of return postage to:

Warren L. Kovach
Department of Biology
Indiana University
Bloomington, IN 47405