January 3, 1986

Criteria for Evaluating IBM-PC Statistical Packages

Paul Marsh

Institute of Statistics

North Carolina State University

Raleigh, NC 27695-8203

Abstract

With the IBM-PC family becoming the standard in microcomputer hardware, many software vendors have begun producing microcomputer statistical packages. This large influx of new packages makes it difficult for researchers to choose one and be reasonably confident that their choice is a good. Not only must hardware and cost be considered, but a wide range of topics centering on the types of potential users must also be studied. A set of criteria is presented for researchers to use when comparing and evaluating IBM-PC statistical packages.

Keywords

Microcomputers; software; algorithms;

1. INTRODUCTION

Like a modern day Diogenes, many researchers have been scanning the microcomputer advertisements and software evaluations looking for "the" statistical package that will turn their machines into true research tools. Today there exists an abundance of such packages; some were written specifically for microcomputers, others are were originally mainframe packages. How is someone supposed to choose the "right" package? Four-color ads do not relay much information concerning reliability and accuracy. Most evaluations are woefully inaccurate, mainly because the reviewers lack the background in database management and numerical analysis necessary to provide insightful criticisms. The result of this chaotic situation is that many users purchase poorly written packages which are ill-designed for their needs.

The American Statistical Association began addressing the issue of statistical package evaluation in 1975 (Francis, Heiberger, and Velleman); however, no set standards have been formalized for this process (McKenzie 1982). Creating such standards is not trivial since computer hardware is rapidly changing (e.g. the microcomputer boom spawned by the addition of floppy disk drives to early Apple II machines in 1978 and advances in supercomputer technology) and numerical software techniques are constantly evolving. These transitions cause statistical software to be a perishable item (Hamer 1981). Users should be aware of this fact when selecting software, and authors and journal editors should assist users in the selection process by ensuring that their reviews are objective, current and based on a clearly outlined set of criteria (Hamer 1981 and Nie and Norusis 1982).

The scope of this report is devoted to describing a set of criteria that can be used to evaluate the current IBM-PC/XT/AT statistical packages. It is expected that any package will have one or more points which are not applicable since the criteria represent a "pie in the sky" picture. Although the criteria were constructed specifically for the IBM-PC family, the basic ideas are transferable to other kinds of microcomputers. The criteria have been segmented into a series of "partitions" to facilitate the description of individual points. Furthermore, the term "IBM-PC" will be used throughout this paper to mean any of the members of the IBM-PC family (e.g. to include IBM-PC XTs and IBM-PC ATs) and all clones that are "sufficiently" compatible. Any differences that occur among the IBM-PC family members on a given point will be noted.

2. HARDWARE

Since hardware represents most of the initial capital investment that has either already been made or is anticipated, it constitutes from the user's viewpoint a logical beginning for the evaluation criteria. The hardware constraints of microcomputers present many problems to the software designer because there are a myriad of system configurations in use (Nash 1982). Until the IBM-PC became available, many statistical software designers were adopting a "wait-and-see" policy; that is, vendors were waiting until they saw a microcomputer on the market that alleviated their hardware problems. Regardless of the decisions made by the software designer in this area, the package should be compatible with the various versions of PC-DOS and with the multitude of adapter cards. The first hardware constraint that confronts the microcomputer statistical software designer concerns the allocation of random access memory (RAM). The amount of RAM is significantly less and the access speed to it is much slower than on larger machines. It is fairly simple to address those problems when creating a statistical package; however, those designers adopting mainframe packages to IBM-PCs are usually forced to require expanded RAM on machines using their software in order to avoid completely rewriting the package. Regardless of the origin of the package, the software designer should also consider how to utilize additional RAM if it is available. For example, as add-on memory boards which overcome the 640K RAM limit imposed by PC-DOS become increasing common, the designer should utilize more and larger buffers to speed file and printer operations. A second set of memory constraints is the handling of disk drives. Most of the modified mainframe statistical packages require a hard disk. While this means that`larger data sets can be processed faster and in more ways, it also implies that a more expensive machine must be available. Packages designed specifically for microcomputers often are oriented toward floppy drives. Many of these can be awkward to use since they were designed to accomodate one-drive systems and do not access additional drives. Also, since PC-DOS pathnames are not really relevant to floppy disk drives, floppy-based packages often do not recognize their use in file specifications and therefore are poor choices for hard-drive systems. Another hardware constraint is the short word length of the central processing unit (the Intel 8088 chip in the IBM-PC and IBM-PC XT and the Intel 80286 chip in the IBM-PC AT). The use of extended precision arithmetic within critical software routines avoids undue arithmetic rounding errors, although it does so at the expense of execution speed. Adapting the software to accomodate an optional 8087 or 80287 math coprocessor chip alleviates the latter problem. The last set of hardware constraints that must be addressed by the software designer is the handling of the user-interface peripherals. A wide range of printer drivers should be standard fare since the cornucopia of printers now available use a variety of control codes. User-interface management and graphical information control are just a pair of the problems that face the software designer when considering how to provide monitor support. With the possible exception of informing the package what monitor interfaces have been installed, the user should not be inconvenienced by this facility. In some cases, provisions should be made for incorporating mice, plotters and light pens into the user interface. Table 1 outlines the hardware considerations that the software designer must address when designing an IBM-PC statistical package. The user should check these items when selecting a statistical package to ensure that it is compatible with the prospective hardware.

Table 1

Hardware Constraints Associated with IBM-PC Statistical Packages

Constraints
Minimum RAM required
Maximum RAM utilized
8087/80827 math chip support
Hard disks, floppy drives support
Printer support
Monitor support
Miscellaneous peripheral support

3. COST

Besides hardware, the price of the statistical package and any associated software represents the remainder of the initial capital outlay. Package price itself is difficult to use as a measure of comparison. Volume or special group price schedules are sometimes available. Some packages such as SAS-PC offer only multi-workstation annual licenses. Other packages, such as SYSTAT, are purchased on a one-price-buys-everything basis. Finally, packages such as BMDP-PC offer a price based on the number of modules (subroutines) that are purchased. Besides the basic package price, the user should also be aware to look for any hidden, added costs such as the purchase of a nonstandard operating system or language. The user should check the items outlined in Table 2 when considering the purchase of a statistical package.

Table 2

Considerations When Purchasing

IBM-PC Statistical Packages

Considerations
Package price
Hidden costs

4. ENVIRONMENT

Just like hardware, the operating environments prevalent in microcomputers present many unique problems to the statistical software designer. Since users generally have little or no access to systems help, the statistical package must coordinate the hardware, software, and data activities (Nash 1982). Complicating these tasks is the fact that many microcomputer operating systems and languages are limited in scope. The usual operating system for an IBM-PC statistical package, PC-DOS, is somewhat unique and therefore the software designer will have some difficulty porting the statistical package to other operating environments. Another factor that affects the portability of a package is the choice of language. The underlying language also affects the speed, stability, and accuracy of the package and may affect the ability of the user to modify or add to the package. The remaining environmental decisions made by the software designer are more visible to the user. All packages should have the ability to query both the hardware itself and the user the first time the package is used and then store the information in a configuration file. That file then can be accessed whenever the package is used. As straightforward as this seems, there exist a few packages that require the user to describe the system hardware everytime the package is accessed. The mode in which the package communicates with the user is probably the most visible result of any of the environmental decisions made by the software designer. Menu-driven systems are useful for infrequent or inexperienced users while more sophisticated users tend to prefer a command language. Query-driven systems may be applicable in certain situations. Users that wish to analyze large datasets or do repetitive work require a batch facility. Abort ("hot-key") and interrupt keys should be provided to allow long analyses to be discontinued. The interrupt key should temporarily suspend operations while the abort key should return the user to the command level. The other environmental decisions affect the long-term use of the package. The ability to backup the package is a must. The package designer should also recognize that the microcomputer/user relationship is self-contained; that is, the user should be able to use the microcomputer system without relying on other people. The system should be proficient at detecting and diagnosing user-input errors. The package should be well-debugged. Lastly, the package should be easy to learn and easy to use. While these two considerations are far more subjective and more dependent on the individual user's expertise than the other criteria listed in this treatise (Carpenter 1984), even value judgments in these areas are useful information to potential buyers. Table 3 outlines the environmental considerations which encompass the various components of the computer and user interfaces.

Table 3

Environmental Considerations

for IBM-PC Statistical Packages

Considerations
Operating system
Language
User-interface mode
Configuration file
Archivability
Error handling
Ease to learn/Ease of use

5. DOCUMENTATION

Documentation is perhaps the single most important set of criteria. It affects how quickly a user can learn the system and in the long run, how easily he can use it. Nothing is more frustrating for the user than trying to find a point in a disorganized, incomplete manual or trying to decipher the text once the point is found. The documentation should be well-written and organized, provide plenty of clearly highlighted examples that illustrate both the important and subtle attributes of the package, and have technical references for every procedure to allow the user to know exactly what statistical test is being conducted (Francis, Heiberger, and Velleman 1975). Novice and infrequent users find tutorial and on-line help facilities accomodating. More experienced users find that a good table of contents, index, and reference card can provide needed information quickly. Some critics of statistical packages feel that the documentation should be so complete that the user could use it to learn statistical techniques; however, considering the number of books and articles written on the various aspects of statistical analysis, it is unreasonable to expect that the documentation should serve this purpose. Table 4 outlines the documentation standards.

Table 4

Documentation Standards

for IBM-PC Statistical Packages

Standards
Organization/Clarity/Completeness
Table of contents/Index
Examples
Technical references
Tutorials
Reference cards
Online help

6. DATABASE MANAGEMENT

In the mainframe market, it was the ability of SAS in the area of database management that catapulted it past older, more established packages such as BMDP and SPSS. In some ways, the handling of the data is even more critical on microcomputers since the amount of memory, both internally and on mass media, is substantially less and the speed of memory access is much slower than on larger machines. The user should note the maximum number of observations (records), variables, and data points (the product of the number of variables times the number of observations) that can be processed by the package. For most researchers, the ability to process 20 to 30 variables for 1000 or fewer observations is sufficient. Another good item for the user to note is the types of variables that can processed. Character variables can not be handled by many microcomputer statistical packages. Also, most packages do not utilize dynamic, boolean, and integer variables to conserve memory. The most important item for the user to note in a statistical package is how the package handles missing data. Many packages do not accomodate missing values at all. In some of the remaining packages, the facility clumsily requires the user to specify a missing value. A good package will have a transparent (to the user) method of handling missing points. All packages should be able to read data from either the keyboard or a diskette. If a special diskette format is used by the package, it should be described it in detail. Provisions should be made for accessing diskette files from other software packages (e.g. LOTUS 1-2-3 .WKS files, dBase .DBF files, and DIF files). Keyboard data entry should utilize a full-screen program having automated prompts and validity checks. A robust full-screen editor should also be included. The user should be aware that command structure (menu, query, or command) affects how easily the data can be manipulated once it is stored correctly. Simple operations such as dropping or renaming variables can be done with any structure; more complex operations such as array structures, repetitive loops, and conditional execution are best done with a command language approach. The user should ensure that the types of variable manipulations that will be required are supported. The user should also examine the operations that can be performed on datasets. While all the existing packages logically define a dataset as a rectangular, flat file, they are extremely different in their abilities to manipulate datasets. Concatenating datasets vertically and horizontally (merging), interleaving two sorted datasets and updating the data points of one data set with the values from a second data set are some of the features that can be available. Table 5 outlines the data management features.

Table 5.

Data Management Features

of IBM-PC Statistical Packages

Features
Minimum file size
Variable types
Missing data facility
Data entry facility
Editor
Operations on variables
Operations on datasets

6. OUTPUT

Output handling bridges both data management and statistical analysis. Whether the user is generating a sophisticated report, listing a dataset, or analyzing information, the output should be readable. A wide range of printers should be fully supported by the statistical package. The user should have the ability to specify such things as line width, character pitch, and lines per inch and per page. Bit-image graphics printers should be accomodated even if graphical interface hardware is not available for the system monitor. Users should also have the ability to specify special output formats and to redirect output to disk or other devices, either for more analysis or for inclusion in some other software product, such as a wordprocessor. Table 6 lists these criteria.

Table 6

Criteria for Output from IBM-PC Statistical Packages

Criteria
Printer support
Graphics support
Report facility

7. STATISTICAL PROCEDURES

7.1 General Characteristics

Finally, the researcher has approached what he believed he originally wished to investigate; that is, whether a particular package has the needed analytical capabilities. Here is where individual needs will surface. The statistical procedures required by a researcher are unique; however, the underlying characteristics of the procedures are similar (see Table 7). The software designer should have spent a sufficient amount of time assessing both the computational aspects of the analytical procedures and the hardware restrictions of microcomputer systems to ensure that fast, stable, and accurate algorithms were written. The argument that mainframe statistical packages are inherently faster and more accurate than their microcomputer counterparts is not entirely true (Janis 1985 and Carpenter 1984). The standard IBM-PC single- and double-precision arithmetic operations offer as much accuracy as the corresponding operations on most mainframe computers and a IBM-PC equipped with a math coprocessor chip can perform those operations more accurately. Furthermore, to argue the superiority of mainframe packages strictly from a hardware standpoint totally ignores the fact that mathematical algorithms are equally dependent on software for their speed, accuracy, and stability. The software designer should incorporate subgroup processing, weight statements, and variable lists into the package to substantially increase its utility. If these latter features are included in a package, the user can be fairly certain that the developer did more than convert some algebraic notation into a computer language such as BASIC.

Table 7

General Characteristics of Statistical Procedures

Characteristic
Speed/Accuracy/Stability
Variable lists
Weight Statements
Subgroup Processing

7.2 Statistical procedures

Choosing the proper statistical analysis procedures to include in a statistical package is a difficult task for the software designer. The designer must keep in mind that the computer resources are limited. The target audience must be defined before determining the statistical procedures to include. For this treatise, it is assumed that a basic, generalized (i. e., not slanted towards any particular application or field) statistical package is being considered. Econometric, time series, operations research, and factor analysis procedures are a few of the types of analyses that fall outside the scope of this paper.

7.2.1 Univariate procedure

A single generalized univariate procedure is seldom available in a statistical package. For example, SAS has six such subroutines and the statistics available vary among the routines. This multiplicity confuses the new user, stretches out the learning curve for the package and decreases the utility of the package. According to Bass (1985), these deficiencies can be overcome if the software is "orthogonal"; that is, there is no duplication of features among the modules within the package. In the case of a univariate descriptive statistical procedure, a software designer can develop a single module that will produce all the statistics outlined in Table 8.

Table 8

Statistics Calculable with a Univariate Procedure

Means (arithmetic, harmonic, and geometric)

Median

Mode

Variance and standard deviation

Skewness and kurtosis

Coefficient of variation

Number of valid, missing and total points

Minimum, maximum, and range

Sum

Uncorrected and corrected sum of squares

Standard errors of the moments

Frequency distributions

Percentiles

Ranks

7.2.2 Frequency and contingency tables

A second procedure that should be included in the basic statistical package is a routine to construct frequency and contingency tables. At the present time, the subroutine for this type of analysis is very poor in most microcomputer statistical packages. The procedure usually divides each distribution into categories and tabulates the data based on these subgroups. The resulting frequency table can yield a very distorted view of the true distribution. The reason most designers use this approach is to overcome problems associated with limited memory. Good packages use a tree structure for this procedure to allow multidimensional tables to be constructed. This approach not only makes the procedure distribution-independent but also allows non-numeric data to be analyzed. Table 9 outlines the statistics that should be available in a frequency procedure.

Table 9

Statistics for Frequency Procedures

Chi-squares (both for individual cells and the whole table)

Fisher's exact test

Pearson and Spearman correlation coefficients

Likelihood ratio chi-square

Kendall's tau-b

Stuart's tau-c

Somer's D

Lambda (both symmetric and asymmetric)

Uncertainty coefficients (symmetric and asymmetric)

Cochran-Mantel-Haenszel statistics

7.2.3 Planned experiments procedures

At least three procedures should be included in the basic statistical package to handle planned experiments. The first procedure in this area, the generalized linear models subroutine, should use a matrix approach. This approach easily handles various complex models such as nested, random, covariate, multivariate, split-plot and latin square, facilitates the computation of diagnostics such as a variance-covariance matrix and the Durbin-Watson statistic, and allows discrete and unbalanced data to be analyzed. Users should be able to construct models both with and without the dependent variable intercept, test hypotheses that do not involve the total model mean square error, and obtain statistics such as the standard errors and adjusted means. Additional statistics such as the coefficient estimates and the residuals of the dependent value should be available in an optional output dataset for additional analysis. Finally, the user should be able to access diagnostics such as the Durbin-Watson statistic to have some assurance that the statistics produced by the procedure are accurate. The second procedure that should be included for planned experiments is a correlation subroutine. The procedure should be able to produce the Pearson product-moment, Spearman, and Kendall correlation coefficients and the sums of squares and cross-products matrix for additional analyses. The last procedure that should be included for planned experiments should be a procedure with the ability to do both paired and unpaired t-tests.

Table 10a

Statistics Associated with Planned Experiments

Generalized linear models procedure

- Corrected sums of squares
- Mean squares
- Degrees of freedom
- F statistics
- Total R-square
- Beta estimates
- Predicted and residual values
- Nonstandard hypothesis tests
- Standard errors
- Adjusted means
- Means comparison tests
- Type I, II, III & IV sums of squares
- Sums of squares & cross-products matrix
- Correlation matrix
- Variance-covariance matrix
- Inverse of variance-covariance matrix
- Partial r's
- Durbin-Watson statistic

Table 10b

Statistics Associated with Planned Experiments

Correlation procedure

- Pearson, Spearman, and Kendall correlation coefficients
- Sums of squares and cross-products matrix

Table 10c

Statistics Associated with Planned Experiments

T-test procedure

- T statistic and its probability
- Degrees of freedom
- Group means

7.2.4 Graphical procedures

A set of graphical procedures should also be included in the basic statistical package to illustrate and illuminate data. Graphics comprise one of the areas in which microcomputers far outshine larger machines. Aside from ensuring that the proper interface hardware has been added to the microcomputer system, the user should be relieved of the cumbersome details of the drawings. Only then can the user concentrate on using scatterplots, histograms, contour plots, and other devices to elucidate data and speed analyses. All figures should be clear, accurate, and well-labeled. The user should be able to scale the figures within the allowable hardware constraints and be able to save the output to disk for future replay. Fundamental line plotter graphics should always be available. Tables 11a and 11b outline the graphical procedures and their characteristics.

Table 11a.

Graphical Procedures for IBM-PC Statistical Packages

Histograms

Bar-charts

Scatterplots

Box & whiskers

Pie charts

Table 11b.

Attributes for IBM-PC Statistical Graphics Procedures

Clear labelling

Accurate scaling

"Replay" facility

7.2.5 Interface procedures

Finally, a few "interface" procedures should be included to complete the package. There should be procedures for multi-key sorts, listing of datasets, and transporting files in and out of the statistical package. The last procedure is particularly important so that the user can easily create files that can be accessed by word-processing, data communications, and spreadsheet programs. Table 12 lists these procedures.

Table 12

Interface Procedures for IBM-PC Statistical Packages

Sort procedure

Print procedure

Transport procedure

8. CLOSING

When choosing a microcomputer statistical package, the user must realize that there are distinct areas that require attention. First, the user should scan the analytical procedures included in the package to ascertain whether the needed procedures are included. Second, the database management facilities should be reviewed carefully to ensure that there are sufficient data manipulation capabilities. Third, the user should check the hardware requirements and the price of the package. If these items are acceptable, the user has found a suitable package. The other areas outlined in this paper, namely documentation, output control, and computing environment determine the desirability of the package and should be considered if several packages are acceptable.

8. CLOSING

When choosing a microcomputer statistical package, the user must realize that there are distinct areas that require attention. First, the user should scan the analytical procedures included in the package to ascertain whether the needed procedures are included. Second, the database management facilities should be reviewed carefully to ensure that there are sufficient data manipulation capabilities. Third, the user should check the hardware requirements and the price of the package. If these items are acceptable, the user has found a suitable package. The other areas outlined in this paper, namely documentation, output control, and computing environment determine the desirability of the package and should be considered if several packages are acceptable.

REFERENCES

BASS Institute (1985) *1985 BASS Reference Manual, Version 84.2*, BASS Institute Inc., P. O. Box 349, Chapel Hill, NC 27514

Carpenter, J., Deloria, D., and Morganstein, D. (1984) "Statistical Software for Microcomputers", *Byte*, April 1984, 234-264

Francis, I., Heiberger, R. M., and Velleman, P. F. (1975) "Criteria and Considerations in the Evaluation of Statistical Program Packages", *The American Statistician*, February 1975, Vol. 29, No. 1, 52-56

Hamer, R. M. (1981) "Papers That Evaluate Computer Programs", *The American Statistician*, November 1981, Vol. 35, No. 4, 264

Janis, J. (1985) Notes from a computer short course on microcomputer statistical packages conducted at UNC-CH, Chapel Hill, NC 27514.

McKenzie, J. D. Jr. (1982) "Standards for Statistical Packages?", ASA Stat. Computing Section Proceedings, 1982

Nash, J. C. (1982) "Where to Find the Nuts & Bolts: Sources of Software" Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface

Nie, H. N. and Norusis, M. J. (1982) "More on Evaluating Computer Programs", The American Statistician, May 1982, Vol. 36, No. 2, 141