Dept. of Agricultural Economics
Staff Paper No. 84-40

1984

NEW STATISTICAL COMPUTING TOOLS ON MICROCOMPUTERS: SEARCH CRITERIA, EVALUATION METHODS, AND AVAILABLE SOFTWARE

James W. Pease and Robert D. Stevens

Michigan State University

Programming (Computers)

AAEA paper presented at its annual meetings, 1984

ABSTRACT

"New Statistical Computing Tools on Microcomputers: Search Criteria, Evaluation Methods, and Available Software." James W. Pease and Robert D. Stevens (Michigan State University).

Search and evaluation costs have limited dissemination of powerful statistical packages now available for four popular microcomputer operating systems. A three-phase investigation identified 184 statistical analysis packages, selected 30 of the more comprehensive for further investigation, and carried out detailed evaluation of six packages.

NEW STATISTICAL COMPUTING TOOLS ON MICROCOMPUTERS: SEARCH CRITERIA, EVALUATION METHODS, AND AVAILABLE SOFTWARE

by

James W. Pease and Robert D. Stevens*

I. INTRODUCTION

Both the development of powerful microcomputers and (rather more recently) the growing number of sophisticated statistical software packages have provided analysts with a significant increase in statistical computing power for many data sets. These tools will increase researchers' productivity and accelerate student learning in the United States and abroad, and will be of particular value in developing nations where access to mainframe computers is generally difficult and costly in time and resources. Search and selection of optimal microcomputer software and hardware for particular tasks is especially costly in this early adoption phase due to the highly variable performance characteristics of the available statistical software and the incompatibility of operating systems.

Central to high performance in applied science is sound theory and efficient data management and analysis. The immense potential of the microcomputer for more cost-effective research, teaching, and extension efforts has begun to be recognized (Diesslin; Fuller; Litzenberg; Weber, et al.). Some inventories of available software for agricultural economists have also appeared (see Strain). To date, however, little detailed evaluative and comparative information is available for most software. This paper reports

---

*The authors appreciate comments made on earlier drafts by Michael T. Weber, Lester Manderscheid, and Valerie Kelly.

selected results of research on new statistical software. The research was carried out in three phases: (1) identification of the software available; (2) selection and general review of the more comprehensive packages; and (3) detailed evaluation of six selected packages. We have sought to reduce, through a series of publications, the costs of search and selection for those seeking cost-effective statistics packages. Although software and hardware decisions are to some extent complementary, space limitations prohibit discussion of hardware selection issues examined in many commercial publications and by Wolf.

II. ECONOMICS OF INFORMATION GENERATION AND SEARCH

Much economic literature views the economics of information as part of the literature on uncertainty. The allocation of resources for information generation, search, and dissemination is primarily caused by uncertainty, as shown in the recent survey by Hirshleifer and Riley. They explore the consequences of taking informational actions to reduce uncertainty and thus provide a formal economic framework for the analysis of search activities.

Reliable, comprehensive, and comparative information about microcomputer software is scarce for at least five reasons. First, this information has public good characteristics. The seller of this information cannot effectively limit dissemination to those who have paid for it. Second, the value of information about microcomputer software is extremely time-dependent, as new products are introduced monthly and established products are continually revised. Third, in an uncertain new market with a product of high complexity, price provides limited guidance. This research has found little correlation between price and product performance. Fourth, the value of software information is dependent upon the value of decisions it will affect and on the credibility of the information source from the perspective of the receiver.

Finally, the difficulties of identifying scattered prospective purchasers in a fairly small market whose applications promise benefits which outweigh information costs discourage information generation.

Generation of comparative information by an individual user is impractical. An obvious, but costly, technique for selecting the best software package for a particular application is to purchase all available packages, evaluate them in light of the application, and choose the most appropriate. Individuals or institutions can seldom afford the costs of this process. Potential purchasers of microcomputer software instead tend to obtain most information about product characteristics by word of mouth, magazine articles, advertising, and books. Commercial advertising by software producers gives notice of the existence of software with certain general characteristics. But it seldom indicates limitations of the package and almost never provides reliable comparative information about important factors such as program accuracy, speed, or ease of use.

Magazine articles which (1) review single programs or (2) compare and evaluate several similar programs are currently the principal source of relatively impartial and more comprehensive information about microcomputer software. But most attempts at comparative evaluations of statistics packages appearing in microcomputer magazines or books have been cursory, highly subjective, and generally not very useful sources of information for serious research applications. The only widely available comprehensive evaluations, both financed by public funds, have been carried out by Michigan State University and the Bureau of the Census (Carpenter et al.). To justify the expenditure of public resources, the costs must be outweighed by the expected social benefits to potential purchasers. The generation of comprehensive and objective information about microcomputer statistical software is an example of the production of public goods which reduce uncertainty and enable more cost-effective research.

III. THE RANGE OF MICROCOMPUTER STATISTICAL SOFTWARE

The initial phase of the investigation focused on identifying the characteristics of available microcomputer statistics software. At the beginning of the research in 1982, a decision was made to focus on programs for the most popular microcomputer operating systems and to restrict examination to programs able to operate on machines with at least 48 kilobytes of random access memory (RAM). The selected operating systems were Apple-DOS, the TRS-DOS for Models I, II, and III, CP/M-80, and MS-DOS.

Over approximately a year, statistical packages were identified through examination of magazine articles, reviews, and advertisements in microcomputing periodicals, including many of those listed by Stilwell (1983a); directories of software, such as included in Stilwell (1983b); and lists of statistics programs, in particular the one by Neffendorf which included 48 packages. A population of 184 statistics software packages was identified, 163 of which had a variety of statistical analysis routines for cross-sectional analysis and 42 of which contained time series routines. Twenty-one packages contained both types of statistical analysis.[1] Additional related software programs (134) of interest to researchers were also identified which had routines for related purposes such as survey data processing, graphing, and linear programming analysis. For each package, a brief description was prepared which included hardware requirements, operating system(s) needed, program language, source, and price. Upon further examination, many of the less expensive, less comprehensive statistics programs were found to be severely limited because, for example, they required

---

[1] Since this investigation was completed, the mainframe statistical packages SPSS and BMDP have become available for certain powerful microcomputers with hard disks.

re-entry of data for each calculation or session (Kelly, Stevens, Stilwell, and Weber).

IV. THE MORE COMPREHENSIVE STATISTICAL PACKAGES

In the second phase of the research, criteria were established to narrow examination to statistical programs with high potential cost-effectiveness for cross-sectional socioeconomic research. Time series analysis packages were not considered further because they were more complex and specialized, and thus merit a separate detailed evaluation. The specific criteria for selection were that the software include: (1) routines for entering and editing data; (2) a method of storing and reloading data; and (3) the following statistical routines: descriptive statistics, cross-tabulations, correlation analysis, and regression. Many of the packages identified with these features also included analysis of variance and minimal graphing routines. Table 1 provides information on 21 of the 30 comprehensive packages identified.

For each of the comprehensive statistics packages, vendor information, user manuals, and review articles were studied to provide (1) a page of specific information on the program, including any interfacing capability between programs for data transfer, limits on file and calculation capacities, and comments on the program based on personal experience and other reviewers' statements; and (2) one or more pages of additional information and sample printouts (Kelly, et al.).

V. EVALUATION OF SELECTED COMPREHENSIVE STATISTICS PROGRAMS

Objectives, Methods, and Criteria

The third phase of the research (reported in Pease and Lepage) sought to make detailed comparisons of six of the most comprehensive software for cross-sectional analysis. A concurrent objective was to develop a methodology for evaluating statistical software. Principal elements of the methodology

Table 1. Selected Comprehensive Statistical Software (E = Expected)

Software Name	Cost	Computer[a/]				Addit. Features		
		A	T	C	I	Time Series	Survey	Graph.
A-STAT	$175	X		X				
ABSTAT	$395			X	X	X		X
AIDA (Seattle, Wash.)	$235	X				X		
COMPSTAT	$1000			X		X		X
DB Master STAT PAK	$99	X						
ELF	$200	X		X		X		X
HSD Stats, Anova, Regress	$247	X			X			X
MaxiStat	$199		X		X		X	
MicroStat (3.0)	$395		X	X	X	X		
Number Cruncher	$200		X	X	X	X		X
SL-Micro	$250		X	X	X		X	
STAN	$300	(any micro with UCSD Pascal)						X
Stat-Systems	$295	X	X		X			
Stat. Package for Micros (SPM)	$119	X	X	E	E			X
Stat. Processing System (SPS)	$300	X	X	X	X		X	X
Statistician, The	$125		X	X	E			X
Statistics with Daisy	$79	X				X		X
StatPac (Wolonick Assoc.)	$400			E	X		X	
STATPAK (NW Analytical, Inc.)	$495			X	X			X
Statpro	$1995	X			E	X	X	X
Stats Plus	$200	X						X

[a/] A = Apple, T = TRS-80, C = CP/M systems, I = IBM-PC.

are: (1) establishment of detailed subjective and objective criteria for comparisons of routines; (2) determination of functional categories of procedures within data processing routines; (3) evaluation by several reviewers with varying microcomputer experience; (4) utilization of typical field research data in standardized data processing exercises; and (5) tests of computational accuracy.

We sought to select "comprehensive" packages, such that a researcher would be able to complete all basic data entry, data management, and statistical tasks with one program. We also wanted packages for microcomputer hardware commonly used in both industrialized and developing nations. In addition to the minimum criteria specified for phases 1 and 2, packages selected for evaluation were required to: (1) permit variables with at least seven digits; (2) allow data file sizes of at least forty variables and one hundred cases; (3) read and write standard ASCII-coded data files; (4) allow normal algebraic and logarithmic transformations, grouping of continuous variables, subsetting of cases, and joining of files; (5) handle missing values in all statistical routines; (6) include routines for descriptive statistics, frequency distributions, cross-tabulations, one- and two-way analysis of variance, and multiple regression analysis with at least ten independent variables and the ability to save residual values; and (7) not be copy-protected. The statistical packages selected were ABSTAT, AIDA, A-STAT, MICROSTAT, NUMBER CRUNCHER, and SPS. After preliminary examination, it was discovered that most programs did not completely satisfy the minimum criteria, principally by their inability to write ASCII output files.

Data files from two MSU field research projects were utilized and typical data entry, data transformation, and statistical processing exercises were formulated. Evaluation procedures and forms covering both subjective and objective elements were developed for the data processing routines and for the

documentation. Each reviewer read and evaluated the documentation, became acquainted with program operation, and completed the data processing exercises. Certain exercises were timed for comparison of the packages. Each was reviewed by one or two persons in addition to the team leader, who reviewed all packages.

Comparison of Program Routines

Space limitations restrict comprehensive discussion of the routines evaluated for each program (data entry and editing, data management, descriptive statistics, frequency distributions, cross-tabulations, ANOVA, and multiple regression). We summarize here the results of the evaluation of multiple regression routines and of computational accuracy.

In the multiple regression routines, all six packages allowed at least ten independent variables. Data management routines allowed transformations such as logs, reciprocals, exponential values, and algebraic manipulations with varying degrees of user ease. Multivariate regression is provided with all packages, while stepwise and weighted regression are each supported by only two packages. Since all packages allow residual values to be saved, multiple stage regression is also possible. Although several packages can calculate a "time series-like" regression, only MICROSTAT has explicit time series routines with plots of standardized residuals from regression plus calculation of moving averages and data smoothing.

An acceptable set of regression output statistics includes estimated coefficients, standard errors or t-values of the coefficients, coefficient of determination, standard error and F-ratio of the regression, and the number of valid and missing cases. Several programs fail to report the number of missing cases and thus do not satisfy the statistical requirements (see Table 2). In some cases, printed regression output is not well organized and labeled. MICROSTAT

Table 2. Summary of Regression Routine Features

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICRO-STAT	NUM. CRUN.	SPS
A. CAPABILITIES (Y=Yes; N=No)						
Select Subset of: Variables	Y	Y	Y	Y	Y	Y
Select Subset of: Cases	Y	Y	Y	Y	N	N
Max. Independent Variables	19	12	24	10	20	10
Residual Plot	N	N	N	Y	N	N
Output to Disk	Y	Y	N	N	N	N
Output: File Identification	Y	N	Y	Y	Y	Y
Output: Job Identification	Y	N	N	Y	N	N
Output: Variable Labels	Y	Y	Y	Y	N	Y
B. STATISTICS (Y=Yes; N=No)						
Estimated Coefficients	Y	Y	Y	Y	Y	Y
Standard Errors of Coeff.	N	N	Y	Y	Y	Y
R-square	Y	Y	Y	Y	Y	Y
Standard Error of Estimate	Y	Y	Y	Y	N	Y
F-ratio of Regression	Y	Y	Y	Y	Y	Y
Number of Valid Cases	Y	Y	Y	Y	N	N
Number of Missing Cases	N	N	N	Y	N	N
C. EVALUATION RATING (U=Unsatisfactory; A=Adequate; S=Superior)						
Documentation	A	A	A-U	A	A-S	A
Ease of Learning	S	S	U	S	A	A
Ease of Use	S	S	U	A	A	A-U
Printed Output--Presentation	S	U	A	S	A	U
Printed Output--Statistics	U	U	U	S	U	U

has by far the best printed output for multiple regression analysis, but deficiencies in data transformations, sorting, and subsetting data files detract from its ease of use. ABSTAT and NUMBER CRUNCHER also have easy to use multiple regression routines with well-presented output.

Computational Accuracy

The approach in examining computational accuracy was that of Velleman, et al., who describe five "ill-conditioning factors" which tend to cause problems for regression routines. Regression data sets were generated which were classified as moderately ill-conditioned, severely ill-conditioned, and perfectly collinear; and which had known coefficients.

Performance and errors for the moderately ill-conditioned data were generally similar for all programs. Many routines calculated the coefficients with virtually zero error, and the worst relative errors deviated from the exact coefficient by about 2 percent. Consequently, it appears that all the programs reviewed can handle the modest degree of ill-conditioning likely to be encountered in research applications. The severely ill-conditioned and perfectly collinear data sets were designed to stress the programs to the breaking point. To illustrate how the programs fare against each other and against mainframe statistical programs, Table 3 illustrates results from the ill-conditioned test problem posed by Wampler.

The mainframe statistical programs, Statistical Analysis System (SAS) and the Statistical Package for the Social Sciences (SPSS), would not estimate all three coefficients. It should also be noted that three microcomputer statistical programs calculated the sum of squared residuals more accurately than SPSS. ABSTAT and NUMBER CRUNCHER performed particularly well in all accuracy tests. The issue of microcomputer computational precision is far from settled, but these results indicate that accuracy may be more a result of the

Table 3. Estimated Regression Coefficients Using Wampler Data[a/]

Statistical Program	Coefficients B0	B1	B2
Mainframe			
SAS	100000.000120	[b/]	[b/]
TROLL	100000	-63818100	64018100
SPSS	100000	200000	[b/]
Microcomputer			
ABSTAT	100000	100000	100000
A-STAT	90916.995	295548.539	-89493.203
MICROSTAT	100000	100000	100000
N. CRUNCHER	100000	100000	100000
SPS	100000	253377.225	-53377.2256

[a/] Correct value for each coefficient is 100,000.

[b/] Program was unable to compute these coefficients.

mathematical maturity and professionalism of programmers than of computer characteristics.

VI. CONCLUSIONS

A rapidly growing body of statistical software, of increasing computational power and sophistication, is now available for the most popular microcomputer operating systems. Unfortunately, the costs of search and identification of adequate software are very high, inhibiting the diffusion of this new technology in our profession.

This investigation found considerable variations in software performance. However, many of the programs identified or reviewed are relatively inexpensive, are easy to use, provide a very satisfactory range of statistical procedures, and generate very accurate statistical results. Thus, microcomputers and associated software form an intermediate low-cost technology, suitable for certain data sets and statistical tasks. Larger, more complex statistical analysis will, however, continue to require a mainframe computer.

REFERENCES

Carpenter, James, Dennis Deloria, and David Morganstein. "Statistical Software for Microcomputers." Byte 9(1984):234-64. April, 1984.

Diesslin, H.G. "The Computer-Extension Delivery System of the Future." Amer. J. Agr. Econ. 63(1981):863-67.

Fuller, Earl J. "Microcomputers: Useful in All of Agricultural Economics and Extension." Amer. J. Agr. Econ. 64(1982):978-85.

Hirshleifer, Jack and John G. Riley. "The Analysis of Uncertainty and Information - An Expository Survey." J. of Econ. Lit. 17(1979):1375-421.

Kelly, Valerie, Robert D. Stevens, Thomas Stilwell, and Michael T. Weber. "An Annotated Directory of Statistical and Related Microcomputer Software for Socioeconomic Data Analysis," 165 pp. East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 12, 1983.

Litzenberg, Kerry K. "Computer Use in the Agricultural Economics Classroom." Amer. J. Agr. Econ. 64(1982):970-77.

Neffendorf, Hugh. "Statistical Packages for Microcomputers - An Update." Originally published in the Am. Statistician 37(1983):83-86.

Pease, James W. and Raoul Lepage. "An Evaluation of Selected Microcomputer Statistical Programs," 187 pp. East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 15, 1984.

Stilwell, Thomas C. "Periodicals for Microcomputers: An Annotated Bibliography." East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 6, 1983a.

Stilwell, Thomas C. "Software Directories for Microcomputers: An Annotated Bibliography." East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 9, 1983b.

Strain, J. Robert and Sherry Fieser. "Updated Inventory of Agricultural Computer Programs." Gainesville, FL: Food and Resource Economics Department, University of Florida, March, 1982.

Velleman, P., J. Seaman, and E. Allen. "Evaluating Package Regression Routines." Proceedings of the Statistical Computing Section, American Statistical Association (1977):82-92.

Wampler, R. "Test Problems and Test Procedures for Least Squares Algorithms." National Bureau of Standards (1978).

Weber, Michael T., James Pease, Warren Vincent, Eric W. Crawford, and Thomas Stilwell. "Microcomputers and Programmable Calculators for Agricultural Research in Developing Countries." East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 5, 1983.

Wiggins, Steven N. and W.J. Lane. "Quality Uncertainty, Search, and Advertising." Am. Econ. Rev. 73(1983):881-94.

Wolf, Chris. "Guidelines for Selection of Microcomputer Hardware." East Lansing, MI: Department of Agricultural Economics, Michigan State University, MSU International Development Working Paper No. 13, 1983.