AN EVALUATION OF SELECTED MICROCOMPUTER STATISTICAL PROGRAMS*

By

James W. Pease and Raoul Lepage

with

Valerie Kelly, Rita Laker-Ojok, Brian Thelen, and Paul Wolberg

1984

*This paper is published by the Department of Agricultural Economics, Michigan State University, under the "Alternative Rural Development Strategies" Cooperative Agreement DAN-1190-A-00-2069-00, U.S. Agency for International Development, Office of Multi-Sectoral Development, Bureau of Science and Technology, Washington, D.C. The Economics Program of the International Maize and Wheat Improvement Center (CIMMYT) also contributed resources for the evaluation.

AN EVALUATION OF SELECTED MICROCOMPUTER STATISTICAL PROGRAMS

by

James W. Pease and Raoul Lepage
with
Valerie Kelly, Rita Laker-Ojok,
Brian Thelen, and Paul Wolberg

Working Paper No. 15 1984

MSU INTERNATIONAL DEVELOPMENT PAPERS

Eric W. Crawford, Carl K. Eicher, and Carl Liedholm,
Co-Editors

The MSU International Development Paper series is designed to further the comparative analysis of international development activities in Africa, Latin America, Asia, and the Near East. The papers report research findings on historical, as well as contemporary, international development problems. The series includes papers on a wide range of topics, such as alternative rural development strategies; nonfarm employment and small-scale industry; housing and construction; farming and marketing systems; food and nutrition policy analysis; economics of rice production in West Africa; technological change, employment, and income distribution; computer techniques for farm and marketing surveys; and farming systems research. While the papers mainly convey the research findings of MSU faculty and visiting scholars, a few papers will be published by researchers and policy makers working together with MSU scholars on research and action programs in the field.

The papers are aimed at teachers, researchers, policy makers, donor agencies, and international development practitioners. Selected papers will be translated into French, Spanish, or Arabic. A list of available papers and their prices may be obtained from: MSU International Development Papers, Department of Agricultural Economics, Agriculture Hall, Michigan State University, East Lansing, Michigan 48824, U.S.A. Individuals and institutions in Third World countries and U.S. Agency for International Development officials may receive copies free of charge.

PREFACE

There is a worldwide revolution in small computer technology underway and scientists are struggling to find ways to utilize this new technology to help solve development problems in the Third World. We are pleased to announce that a number of papers on microcomputers in international agriculture will be published in our International Development Papers series. The aim of these papers is to provide timely information about the rapidly changing state of the new microprocessor technology and its use in research. The papers are also intended as guides to agricultural and social scientists on choosing, installing, and maintaining microcomputer hardware and software systems in developing countries.

Some of the papers will also document field experiences of selected established projects using new data processing hardware and software. Other papers will concentrate on developing guidelines for establishing and maintaining successful microcomputer and/or programmable calculator installations for agricultural research in developing countries.

The present paper is the ninth of these new papers. It is based on work by faculty, staff, and graduate students of the Department of Agricultural Economics, Michigan State University, on cost-effective data collection, management, and analysis techniques for developing country applications. This activity is carried out under the terms of reference of the Alternative Rural Development Strategies Cooperative Agreement--DAN-1190-A-00-2069-00--between the Office of Multi-Sectoral Development, Bureau of Science and Technology of the United States Agency for International Development and the Department of Agricultural Economics at Michigan State University. The Economics Program of the International Maize and Wheat Improvement Center (CIMMYT) also contributed funds for the evaluation.

The first author of this paper, James Pease, is a Graduate Research Assistant in the Department of Agricultural Economics at Michigan State University. He has worked on many computer-based simulation and data analysis projects, several involving data collection in developing countries. He has also assisted faculty and students in selecting microcomputer hardware and software and in exploring the capabilities of this equipment in research activities. Raoul Lepage is Professor of Statistics and Probability at Michigan State University. His major research interests include computer-intensive statistical methods and sequential analysis. Valerie Kelly and Rita Laker-Ojok are graduate research assistants and Ph.D. candidates in the Department of Agricultural Economics. Brian Thelen is a graduate research assistant and Ph.D. candidate in the Department of Statistics; and Paul Wolberg is a senior programmer in the Department of Agricultural Economics.

Readers are encouraged to submit comments about these new papers on microcomputers and to inform us of their activities in this area. Write directly to: Dr. Michael T. Weber, Acting Director, Alternative Rural Development Strategies Cooperative Agreement, Department of Agricultural Economics, Michigan State University, East Lansing, Michigan 48824-1039.

Eric W. Crawford, Carl K. Eicher, and Carl Liedholm
Co-Editors
MSU International Development Papers

LIST OF TABLES

TABLE OF CONTENTS

I. INTRODUCTION

A. Evaluation Objectives

The development of relatively inexpensive microcomputer technology has placed immense computing power within the grasp of the small research institution or project. In developing countries, the new technology presents a viable alternative between the pocket calculator, which is inexpensive but has limited capacity, and the mainframe installation, which is powerful but often inaccessible.

Unfortunately, many small computers are purchased without adequate consideration of hardware and software factors which affect the overall cost-effectiveness of microcomputer-based data processing. An element often overlooked is the availability of suitable statistical software. Information about commercial statistical software for specialized research applications is usually not obtainable in developing countries, and researchers sometimes decide to write their own statistical analysis programs. The objective of this evaluation is to help researchers assess the capacity of commercial statistical programs. It is hoped that comparative information about such software will help researchers to avoid "reinventing the wheel" and make cost-effective decisions about data processing.

Identification of a suitable statistical package requires a "search" phase, followed by an "evaluation" phase to compare available programs with each other and with analysis requirements. Even the first step can be very difficult for researchers in developing countries, and "evaluation" often consists of little more than comparison of hearsay and rumors about the software. A companion report in the MSU International Development Working Paper Series addresses the search problem, presenting information on virtually all commercial statistical programs.[1/]

This evaluation goes into greater depth to aid researchers in determining adequacy of "comprehensive" statistical programs for analysis of small- or medium-size data sets. Since most socioeconomic research projects in developing countries do not analyze time series or experimental design data,

---

[1/] Valerie Kelly, Robert D. Stevens, Thomas Stilwell, and Michael Weber. "An Annotated Directory of Statistical and Related Microcomputer Software for Socioeconomic Data Analysis." Michigan State University International Development Papers, Working Paper No. 12, 1983.

our comparative analysis was directed specifically at evaluation of the selected programs for cross-sectional research.[1/] We have also concentrated on programs suitable for relatively small surveys because larger projects often require (and can more easily justify) the preparation of customized software or use of a mainframe computer.

Software evaluations found in popular microcomputer magazines are typically oriented towards home or business applications. Many tend to be very superficial, and to more closely reflect the biases of the author than the characteristics of the program. We contend, however, that the most important aspects of statistical software can be carefully defined and often quantified in some fashion. Therefore, a second objective of the present endeavor is to develop more systematic procedures and criteria which could provide a framework for future comparisons of statistical software. But this is not to say that subjective evaluation of abstract characteristics such as "ease of use" is not necessary. Objective and subjective assessment is combined to allow the reader to gain a broader perspective of program operation and of adequacy vis-a-vis explicit minimum requirements elaborated for important program routines.

Evaluation of the computational precision of selected microcomputer statistical routines is a third specific objective. Can microcomputer statistical software produce precise results with problematic data sets? This portion of the evaluation (the procedures of which are documented in detail in Annex 2) was sorely needed, since most available information about this issue has been anecdotal or the result of unsystematic analysis.

B. Selection of Statistical Programs for Comparison

Kelly et al. (Ibid.) has identified 269 commercial microcomputer programs which perform statistical functions. Obviously, not all those programs could be evaluated. Programs selected for evaluation were required to fulfill a set of minimum criteria based on research experiences of faculty and students at Michigan State University. Most researchers would impose much more stringent requirements for one criterion or another. Nevertheless, the following served for selection purposes and were felt to represent a "lowest common denominator" of specifications for cross-sectional research.

[1/] Nevertheless, some observations may assist the reader interested in time series or experimental design analysis with these packages.

1. The program must be available for microcomputers which use either the CP/M-80, CP/M-86, MS-DOS, or Apple DOS operating systems.
2. Both keyboard data entry and permanent disk file storage must be supported within the program.
3. Variables of at least seven digits must be permitted. For example, a value of at least 99,999.99 should be allowable.
4. The maximum allowable data file size should be no less than 40 variables and 100 cases.
5. The program must be able to read files created by other software using American Standards Code for Information Interchange (ASCII) codes and produce ASCII-coded output data files.
6. Data management capability must include at least the following:
   a) Algebraic, logarithmic, and exponential variable transformations.
   b) Grouping of continuous variables.
   c) Capability to select a subset of variables or cases which can then be analyzed separately or written to disk as a new file.
   d) Capability to append (concatenate) two or more files to form a new file.
7. All statistical routines must have the capability to handle missing values.
8. At least the following statistical routines must be included:
   a) Univariate descriptive statistics.
   b) Frequency distributions.
   c) Cross-tabulations (contingency tables).
   d) One-way and two-way analysis of variance.
   e) Multiple linear regression with at least seven independent variables, plus the capability to save residual values.
9. It must be possible to make legal backup or archival copies of the program. Some software authors incorporate copy protection schemes into their programs which make normal copy programs inoperative. There is a clear danger that overseas users of copy-protected programs could find themselves without a workable program if the original disks were damaged.

Table 1 lists the programs selected, the software distributors, and hardware configurations used for the evaluation. Table 2 lists general system characteristics of the selected programs. Please note that all evaluation results and ratings refer only to the version available for these operating systems and that other versions may have somewhat different capabilities or limitations.

The selection criteria were not "etched in stone." If a program failed to satisfy a particular requirement, it could still be chosen for evaluation. In fact, since program selection was sometimes based on less than complete secondary information, subsequent examination revealed that programs thought to be completely acceptable did not satisfy all requirements. Program limitations are listed below:

Program	Limitation
1. AIDA	Seven-digit values not allowed. No two-way analysis of variance. Will not output an ASCII-coded file. Not completely copyable.
2. A-STAT	No two-way analysis of variance.
3. MICROSTAT	No capability to group values. Will not output an ASCII-coded file.
4. NUMBER CRUNCHER	Will not output an ASCII-coded file.
5. SPS	Limited data file size. No two-way analysis of variance.

Many of the selected programs offer a wider range of statistical routines than specified under our minimum evaluation criteria. Table 3 documents all available statistical routines of the selected packages.[1/] Readers should note, however, that the present evaluation was limited to data entry, data management, descriptive statistics, frequency distributions, cross-tabulations, analysis of variance, and multiple regression routines.

Several popular statistical programs were not evaluated. In some instances, these programs clearly did not meet our requirements, while time and resource constraints prevented the evaluation of other packages. Three

---

[1/] Several software distributors of the selected programs have subsequently introduced new versions which appear to remedy certain limitations identified in this evaluation. For a description of changes known to the authors of this report, see Annex 4.

Table 1. Programs Selected for Evaluation

Program	Author/Distributor	Hardware Configuration for Evaluation
ABSTAT	Anderson-Bell P.O. Box 191 Canon City, Colorado 81212 Tel. 303-275-1661	IBM Personal Computer 320 Kb Random Access Memory MS-DOS Operating System Two 5 1/4 inch disk drives
AIDA	Action-Research Northwest 11442 Marine View Drive, S.W. Seattle, Washington 98146 Tel. 206-241-1645	Apple II+ 64 Kb Random Access Memory DOS 3.3 Operating System Two 5 1/4 inch disk drives
A-STAT	Rosen Grandon Associates 7807 Whittier Street Tampa, Florida 33617 Tel. 813-985-4911	Apple II+ 64 Kb Random Access Memory DOS 3.3 Operating System Two 5 1/4 inch disk drives
MICROSTAT	Ecosoft, Inc. P.O. Box 68602 Indianapolis, Indiana 46268 Tel. 317-255-6476	TRS-80 Model II 64 Kb Random Access Memory CP/M Operating System Four 8 inch disk drives
NUMBER CRUNCHER	Dr. Jerry Hintze 865 East 400, North Kaysville, Utah 84037 Tel. 801-546-0445	IBM Personal Computer 320 Kb Random Access Memory MS-DOS Operating System Two 5 1/4 inch disk drives
STATISTICAL PROCESSING SYSTEM (SPS)	Southeast Technical Associates 1302 Greendale Drive Blacksburg, Virginia 24060 (Apple Version) DATABASIC Software 102 South Main Mt. Pleasant, Michigan 48858 (CP/M and MS-DOS Versions)	Apple II+ 64 Kb Random Access Memory DOS 3.3 Operating System Two 5 1/4 inch disk drives

Table 2. General System Requirements of Evaluated Programs

1.	Package Name	ABSTAT	AIDA	A-STAT
2.	Version	3.0 1982	9/82	79.6 1981
3.	Price	$395	$235	$175
4.	Type of CPU/Computer	8080/Z80 CPU 1/ IBM PC or compatible	Apple II+ Apple III-emulation mode	Apple II+
5.	Operating System	CP/M-80 MS-DOS CP/M-86	DOS 3.3 (modified)	DOS 3.3
6.	Language	Compiled PASCAL MT+	Applesoft BASIC	Compiled Applesoft BASIC
7.	Minimum Memory Requirement (RAM)	CP/M-80 56K MS-DOS 128K CP/M-86 128K	48K	48K
8.	Disk Drive:			
	Minimum	1	1	1
	Recommended	2	2	2
9.	Printer	Optional	Optional	Optional
10.	Monitor (min. requirement)	80 columns/24 rows	40 columns/24 rows	40 columns/24 rows

1.	Package Name	MICROSTAT	NUMBER CRUNCHER	SPS
2.	Version	3.0	3.0 1983	4.2 1983
3.	Price	$395	$200	$300
4.	Type of CPU/Computer	Z80 CPU	8080/Z80 CPU 1/ IBM PC or compatible	Apple II+
5.	Operating System	CP/M	CP/M-80 MS-DOS	DOS 3.3
6.	Language	BAZIC 2/	BASIC-80 (CP/M)	Applesoft BASIC
7.	Minimum Memory Requirement (RAM)	48K	3/	48K
8.	Disk Drive:			
	Minimum	1	1	1
	Recommended	2	2	2
9.	Printer	Optional	Optional	Optional
10.	Monitor (min. requirement)	64 columns/16 rows	80 columns/24 rows	40 columns/24 rows

1/ Most machines that run CP/M.

2/ BAZIC is a Z80-coded version of North Star BASIC.

3/ After loading BASIC, requires a minimum of 31K free bytes to load and execute.

Table 3. Statistical Routines of Evaluated Programs

	PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
1.	Descriptive Statistics	Y	Y	Y	Y	Y	Y
2.	Frequency Distribution	Y	Y	Y	Y	Y	Y
3.	Probability Statistics	Y	N 1/	N 1/	Y	Y	Y
4.	Hypothesis Testing						
	(T-test) Means	Y	Y	Y	Y	Y	Y
	(T-test) Proportions	N	N	N	Y	N	N
5.	Crosstabs/Contingency Tables	Y	Y	Y	Y	Y	Y
6.	Correlation						
	Simple Correlation	Y	Y	Y	Y	Y	Y
	Non-parametric Correlation	Y	Y	N	Y	N	Y
7.	Regression						
	Multiple Regression	Y	Y	Y	Y	Y	Y
	Stepwise	N	N	N	Y	Y	N
	Weighted	N	Y	N	N	Y	N
	Path Analysis	N	N	Y	N	N	N
	Robust	N	N	N	N	Y	N
	Principal Components	N	N	N	N	Y	Y
	Factor Analysis	N	N	N	N	N	Y
	Image Analysis (Guttman)	N	N	N	N	N	Y
	Time Series Capability	Y	Y	N	Y 2/	Y	N
8.	Analysis of Variance						
	One-Way	Y	Y	Y	Y	Y	Y
	Two-Way	Y	N	N	Y	Y	N
	Three-Way/Four-Way	N	N	N	N	Y	N
	Randomized Block	N	N 3/	N	Y	N	Y
	Split Plot	N	N	Y	N	N	N
	Repeated Measures	N	N	N	N	Y	N
	General Linear Models	N	N	N	N	N	Y
9.	Non-parametric Tests	Y	Y	N	Y 4/	Y	Y
10.	Combinations, Permutations, Factorials	N	N	N	Y	N	N
11.	Monte Carlo Data Sets	N	N	N	N	N	Y

1/ Not a separate routine, but reported in most statistical routine output.
2/ Includes moving averages and data smoothing. MICROSTAT is the only package with explicit time series routine.
3/ AIDA offers a non-parametric randomized complete block analysis.
4/ Includes extensive non-parametric tests.

Y = Yes N = No

comprehensive statistical programs which seem to satisfy most or all of the requirements include STATPAK (Northwest Analytical), STATPAC (Wolonik Associates), and STATPRO (Wadsworth Publishing). The latter company chose not to contribute to an evaluation copy of their program, while time constraints prevented evaluation of STATPAK and STATPAC.

C. Evaluation Methods

In order to reflect actual data processing tasks, data files from two Michigan State University research projects were selected, and typical data entry, data transformation, and analytical exercises were formulated. Two files (from a Cameroon livestock marketing analysis) consist of 602 cases with 12 variables and 260 cases with 13 variables (respectively), while a third file (from a year-long survey of Honduran rural small-scale industry) contains 200 cases with 17 variables. Variables in all files range from nominal to ratio measurements, include positive and negative real numbers, and have up to nine digits with three decimal places.[1/]

Reviewers completed standardized tasks for each routine, timed the operation of selected exercises, and documented any errors encountered. In addition, detailed evaluation forms were developed in which each reviewer systematically recorded both personal impressions and objective details about program documentation, operation, and performance. The printed output was then assembled and evaluated for content and presentation. The capabilities of the program module, time required to complete the exercises, and other specific information are summarized in the evaluation tables of each routine.

Subjective assessment is nonetheless necessary to evaluate other important program aspects not amenable to quantification, such as clarity of documentation, ease of learning the routine, ease of using the routine once learned, and adequacy of error handling. Ratings of selected subjective features are also presented in the summary evaluation tables. Programs are evaluated against the stated requirements for a particular routine and listed as superior, adequate, or unsatisfactory.

The background of the evaluation team members includes considerable field research experience in developing countries and microcomputer

---

[1/] This number of decimal places required values to be truncated for two packages.

experience ranging from minimal up to the level of professional programmer. Each package was evaluated by one or two reviewers. In addition, the first author evaluated all programs.

The methods used to evaluate computational accuracy are described in Annex 2. In general, data sets were generated with the intention of pushing the programs to their limits. Tests were constructed for the descriptive statistics, analysis of variance, and regression analysis routines. Special effort was directed towards testing the computational precision of regression analysis output. Most previous evaluations of microcomputer software have simply compared regression coefficients using existing research data thought to pose computational difficulties with "high precision" results generated by a major mainframe package. In contrast, the present evaluation follows the approach of Velleman et al. (1977)[1/] in which data sets were generated with known attributes reflecting five statistical problems which affect the precision of regression output. These factors are:

1. High degree of collinearity among the independent variables;
2. Low variation in an independent variable;
3. Low proportion of the dependent variable variation attributable to regression;
4. Large absolute values of the independent variables; and
5. Large differences of magnitude between the dependent variable and independent variables.

If factors one or two are present, the independent variable matrix is nearly singular and computation of accurate regression coefficients is difficult. If extreme values of these factors exist (perfect collinearity), a "good" routine should produce a clear message informing the user that computation of regression parameters is not possible. If the third condition exists, accurate prediction of dependent variable values may not be possible. Large absolute values of the independent variables may cause computer memory overflow errors through accumulation of large sums of squares or cross-products. Finally, the number of digits in coefficients listed by regression routines is affected by differences in magnitude between the dependent variable and independent variables, thereby causing differences

---

[1/] Paul Velleman, J. Seaman, and E. Allen. "Evaluating Package Regression Routines." Proceedings of the Statistical Computing Section of the American Statistical Association. 1977, pp. 82-91.

between the reported coefficient and an "exact" coefficient with more digits.

A major innovation of the Velleman research was the development of Fortran routines which (1) generate data consistent with pre-stipulated measures of these so-called "ill-conditioning" factors, and (2) compute theoretically exact regression coefficients directly from these parameters. As stated by Velleman, "Each factor can be varied along a continuum from a situation in which no problem is likely to one in which the problem is either statistically meaningless, computationally unsolvable, or both" (Velleman et al., 1977:82). The regression coefficients are known with certainty, and regression analysis output can be evaluated against the theoretical benchmark. The Velleman techniques were implemented for the present evaluation on an IBM Personal Computer, using the Intel 8087 co-processor and the powerful mathematical programming language APL. For comparative purposes, an independent routine prepared in APL was also used to calculate regression results from the data sets.

Adapting the methods used for regression analysis evaluation, data sets were also generated to test one-way analysis of variance accuracy. Precision of descriptive statistics output was evaluated by forcing calculation of sums of squares as large as 9.0E+12, which could cause problems for poorly written programs. The results of the respective tests are briefly discussed and summarized in the sections of this report pertaining to comparison of program routines. For the interested reader, a more detailed report of the computational accuracy evaluation is presented in Annex 2.

II. COMPARISON OF DOCUMENTATION AND OVERVIEW OF PROGRAM OPERATION

A. Documentation

No program is so well written, so "user-friendly" that good documentation is not required to first instruct the new user in program operation and later to serve as a reference manual for specific details on statistical routines, program modification, error handling, and technical issues. New users often prefer a chatty, "friendly" manual that reads something like a "how to" article. Others claim this style is not efficient for the experienced computer user and prefer a straightforward reference manual which is organized in a logical manner and simply states options and procedures. Exceptional documentation is likely to walk a thin line between the two approaches.

The writing style of documentation must above all be clear and avoid excessive use of computer jargon. A particularly dangerous tendency of some manuals is to use computer "buzz words" familiar only to the experienced user of the particular hardware or operating system for which the program is written. Other terminology errors involve excessive use of "shorthand" terms unique to the program and use of misleading or poorly defined statistical terms.

At a minimum, the manual should provide sufficient explanation so that the user can quickly learn to operate any routine and understand (at least in general terms) what is implied by the available options. But complete documentation is not sufficient. If the information is not easily located in the manual, many users feel it may as well not be documented. A detailed table of contents can help, but we wish to add our voices to the user clamor for a detailed index in every program manual. Given the availability of text indexing software, there is little excuse for the absence of an index.

Table 4 summarizes information about program manuals and presents our evaluation of documentation style, content, and ease of use. In all summary tables, "Y" and "N" mean "Yes" and "No". Our evaluation of program features is abbreviated by "S" = Superior, "A" = Adequate, and "U" = Unsatisfactory. Hyphenated ratings indicate that program features fall somewhere between these categories. If a limitation (e.g., "maximum allowable number of cases") was not encountered in the documentation or during the exercises, "NAL" indicates that the feature has no apparent limit

Table 4. Program Documentation Features of Evaluated Programs

	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
A. DOCUMENTATION FEATURES						
1. Items in Tables of Contents	66	83	42	60	20	20
2. Pages of Introduction/Overview	10	10	39 1/	1	1	4
3. Pages of Main Body	64	64	48	44	71	94
4. Pages of Appendix	16	8	7	54	0	7
5. Index	Y	Y	N	N	N	N
6. "Getting Started" Sample Session	Y	Y	Y	N	N	N
7. Sample User Input Per Routine	All	All	Some	N	Some	Some
8. Sample Output Per Routine	All	All	All	All	All	Some
9. Sample Data Sets in/on:	Disk	Disk	Manual	Disk	Manual	None
10. Interactive Command Summary	Y 2/	Y 2/	N	3/	3/	3/
11. Error Message Reference Table	Y	N	N	N	N	N
12. Missing Value Treatment Described for each Routine	Y	N	N	Y	N	N 4/
13. Computational Formulas	Y	N	N	Y	N	Y
14. Statistical References	N	Y	N	N	N	Y
B. EVALUATION MEASURES						
Writing Style and Terminology	S	A	U	S	A-S	U
Content and Organization	S	S	U	A	A	U
Ease of Use	S	S	A	S	U	U

1/ A-STAT has two manuals--a "user's manual" and a "language reference manual." The entire 39 pages of the former manual is essentially a "getting started" session.

2/ The command summary is also accessible in interactive sessions.

3/ Not applicable, since the program is menu-driven.

4/ Treatment is unique in that user must select out missing value codes before submitting data to routine.

Y = Yes N = No A = Adequate S = Superior

U = Unsatisfactory

(subject to overall package or hardware limitations). The following brief comments describe the documentation of each package.

Many programs provide sample data printed in the manual or stored on diskettes and detailed instructions which assist the new user to learn program features. Sample input responses and printed output are generally very helpful to the new user. An example of a complete sample session with details of user input and printed output complements the written documentation and adds to its usefulness for learning. The reader should also be aware of the trend towards "on-line" documentation, whereby information can be displayed on the monitor while operating the program. The new user can therefore be guided through the program having read little more than a "getting started" chapter in the printed documentation.

ABSTAT

The new user is guided through a "getting started" session which illustrates basic aspects of program operation. On-line "help" documentation is available at virtually any point in the program, providing invaluable assistance to new users. This documentation is organized by program commands within general headings such as "Data Set Commands" or "Statistical Commands." Sample input and output are listed with each command description. Tables and appendices provide ample reference information on missing value treatment, computational formulas, and error codes.

The writing style, use of terminology, content, and overall ease of use of this program documentation are superior. If any fault can be found, it is that the documentation is a bit terse. More complete explanations of routine options and significance of test statistics would also be helpful.

AIDA

The AIDA documentation is organized in five chapters, of which a single chapter presents the description of all commands. A command summary is also available during program operation. A complete sample session is given in the appendix, and each command description presents a sample of routine use and output.

Some elements lacking in the manual are a complete description of missing value treatment by routine, an error message summary, and computational formulas. The manual generally presents very good information in a format that is quite easy to use, but the writing style suffers somewhat from excessive use of terminology recognizable only to experi enced Apple users and shorthand terms unique to AIDA. The choice of command terms is also sometimes misleading. For example, ONEWAY calculates frequency distributions and MEANS calculates one-way ANOVA.

A-STAT

Two manuals instruct the user in A-STAT operation. The first (called the "user's manual") is actually a long "getting started" session, and is quite useful if readers have the necessary patience to follow it through 39 pages. The second document is called the "language reference manual," which is organized according to the A-STAT procedural commands. The command syntax descriptions are quite similar in style to that found in mainframe statistical program manuals, and just as difficult for the new user to understand.[1/] No interactive command summary is available, and new users will find the program virtually impossible to operate without frequent reference to the manual. No sample output is listed, treatment of missing values is not described for many routines, no error messages are referenced, no computational formulas are presented, and the manual does not include an index. Some program limitations listed in the manual do not agree with those encountered during program operation.

Given the above deficiencies, the manual content must be judged inadequate. The opaque writing style also detracts from the usefulness of the documentation.

MICROSTAT

This manual's table of contents is well organized, with sufficient detail to compensate somewhat for the lack of an index. No sample sessions are presented to acquaint the new user with program operation. However, sample output is listed for all statistical routines and all input menu items are well explained and consistently treated in the text.

The writing style is quite clear and straightforward, and in general the manual is easy to use. However, the content of the documentation has two

---

[1/] A-STAT has been developed as a functional subset of the P-STAT statistical package available on many mainframe computers. Users familiar with P-STAT will have few difficulties with the syntax.

important drawbacks: (1) the documentation of installation procedures for the program is completely inadequate,[1/] and (2) no listing of error messages is presented.

NUMBER CRUNCHER

Combined with the lack of appendices or an index, the terse and somewhat disorganized table of contents makes this manual difficult to use. In addition, the manual contains no references to error messages, missing value treatment for each routine, or computational formulas. Some documented default values do not agree with those encountered in program operation. The new user will find this documentation somewhat unhelpful, and the experienced user will be frustrated by the difficulty in locating specific information.

On the other hand, the writing style and some explanations of routines are very well done. Particularly commendable are the explanations of statistical procedures and test statistics. The written documentation is supplemented with fairly extensive explanations of procedures displayed on the screen. However, experienced users may be irritated by the number of keyboard responses necessary to proceed past these continually repeated explanations.

SPS

Given 94 pages of text in the main body, an index or a detailed table of contents is essential for effective access to this documentation. Instead, only a short and unhelpful table of contents is presented in which, for example, all explanations and instructions for the important descriptive statistics, frequency distributions, and histograms routines are referenced as "Descriptive Statistics and Plots." In addition, certain routines have confusing and unorthodox labels, such as "R x C Contingency Table Statistics." Because no sample sessions or data sets are presented, learning how to operate the program is difficult.

The user is referred to the Applesoft BASIC manual for explanations of error messages, and confusing (or simply incorrect) use of terms further detracts from manual content and readability. For example, a procedure

---

[1/] MICROSTAT was the only program evaluated which had to be "installed," i.e., configured for the particular hardware system.

which selects a data subset based on the value of a specific variable is incorrectly labeled "SORT," even though the data set is not rearranged in any manner.

The writing style is broad and rambling, making access to reference information quite difficult. Given a more coherent style, the documentation might have been written in a fraction of the pages.

B. Overview of Program Operation

General operating characteristics of the selected programs are briefly described in this section. But first, a series of alternative approaches to data processing are explained, and some "buzz words" are defined.

Most programs operate as either "interpreted" or "compiled" versions of a particular programming language. When an interpreted program is run, a language interpreter must convert each line of the program to executable microprocessor instructions. A compiled version of a program is already composed of executable machine language codes and operates much faster than the same program in an interpreted language. On the other hand, an interpreted program is much easier to modify and takes up less disk space. Among the evaluated programs, ABSTAT and A-STAT are compiled programs and the rest are interpreted. The reader should note that this does not indicate that the evaluated programs which are compiled complete data processing tasks faster than their interpreted competitors. Because of time-consuming input procedures, actual computation time is often a small proportion of the total time necessary to carry out an analysis.

Whether the packages are "command-driven" or "menu-driven" is another important difference between programming approaches. A menu-driven program presents the user with a list of allowable alternative operations, and the user responds by pressing one or more keys corresponding to the desired action. A command-driven program simply presents the user with a command level prompt (e.g., CMD=>; OK-; READY-), and the user must enter acceptable command "words" and "parameters." New users can sometimes operate a well-written menu-driven program with little or no previous study of the documentation. On the other hand, more complex procedures can be accomplished with a few short commands in a good command-driven program. In general, one might say that menu-driven programs may be easier to learn, while command-driven packages may be more efficient and possibly easier to use once the commands are learned. Among the evaluated statistical programs, ABSTAT, AIDA, AND A-STAT are command-driven and the rest are menu-driven.

Some command-driven programs permit what are often called EXEC or COMMAND files. These consist of a series of data handling or statistical commands (often prepared with a separate word processing program) which are submitted to be executed as a "batch job" by the statistical program. No further user input is then necessary while the commands are executed. All the evaluated command-driven programs allow such EXEC files.

Program operation is significantly affected by the method used to read data from disk file and to hold data in memory. Certain programs (ABSTAT and AIDA) read data from the disk file into memory only once and do not require a separate read operation from disk for each subsequent statistical routine. This approach will usually allow the user to more rapidly complete a statistical analysis because accessing data from disk is much slower than from memory. Since both the statistical program and the data must fit in memory, however, data file size may be restricted. Both ABSTAT and AIDA seek to alleviate this problem by allowing the user to read only particular variables from a disk file into memory. A-STAT and MICROSTAT, on the other hand, must read the data from disk file for each statistical routine. Thus, the latter programs allow much larger file size because processing is carried out as each data case is read. The approach of NUMBER CRUNCHER and SPS falls somewhere between these two methods.

Usually attributable to incorrect user input or defects in programming, some errors always occur when operating a computer program. "Error trapping" refers to programming instructions which detect the error and direct program execution back to a previous point. If errors are not trapped, the program "crashes" and the user may find herself/himself at the language interpreter or operating system level (or the computer might "freeze up"). Even though errors may be trapped by the program, the error messages displayed may not provide the user with enough clear information to determine the cause or solution of the problem. Poor error handling is reflected in vague error messages (e.g., "overflow error") often generated by the program interpreter or the operating system. In a package with good error handling, very few errors require re-starting the program. Program operation should resume at the beginning of the current routine with no loss of data. Unambiguous error information ought to be displayed so that a user with moderate microcomputer experience can determine the cause (if not the solution) of the problem. More extensive explanations of error messages should be listed in the program manual.

A final important operating characteristic is the procedure for printing output. Some programs print exactly what is displayed on the screen (a "screen dump"), while others may produce a more presentable report format with headings that identify the data file, the current date, or a user-supplied description of the analysis at hand. One approach requires the user to "toggle on" the printer before data processing begins, with subsequent output usually routed to both the screen display and the printer. If the user forgets to toggle on the printer, the routine must be run again to obtain printed output, with consequent loss of time. A second type of output procedure (used by NUMBER CRUNCHER and MICROSTAT) initially lists all results on the screen display and provides the subsequent option to list output in a report format on the printer without re-running the routine. This approach invariably saves time, effort, and paper.

This section has so far described the major elements which differentiate operational features of programs. A brief description of program operation of the evaluated packages is presented below.

ABSTAT

This is a compiled program written in the programming language PASCAL MT+. Yet, as demonstrated by the timed data processing exercise (listed in the summary tables of the following chapter), program operation is not particularly fast. Commands are typed by the user following the prompt "Which Command?". To access data handling or statistical routines, the user enters a short command (such as "FETCH" to read a disk file or "FREQ" to run frequency distributions) followed by options (generally prompted by the program) to indicate which variables, files, or other parameters are to be used in the procedure. Selected variables are held in memory between commands, significantly reducing read operations from disk. Because an ABSTAT data file can contain only 4,000 data points (number of cases times number of variables) it may be necessary to split data among several files and pull selected variables into memory with repeated read commands. ABSTAT allows the user to execute "batch jobs" with EXEC files.

ABSTAT error handling is completely independent of the type of microcomputer or specific operating system. In most cases, if an error is detected, a tiny arrow is displayed indicating the specific incorrect character in the command, followed by an error number and brief explanation of the problem. (Additional error reference information is contained in the manual.) The command prompt is repeated, and operation can continue without loss of data.

To route output to the printer, the user must execute the print command before typing a statistical command. Each page of printed output lists the file name, plus optional date and identification of the current analysis. Output can also be routed to a disk file for later incorporation into research reports.

AIDA

The program is first started up ("booted") with the non-copyable system disk which loads a specially modified version of the Apple operating system. Program operation is then initiated by inserting the copyable program disk which contains the data handling and statistical routines. Although all routines are written in interpreted Applesoft BASIC, program operation is surprisingly fast. Short commands (e.g., "READ" for read a file) are entered in response to the prompt "COMMAND?", and two or three subsequent prompts identify variables or allow the user to select options. AIDA also allows use of EXEC files. As with ABSTAT, selected variables are held in memory between commands, which reduces disk file read operations and consequently saves the user's time. A sophisticated "virtual memory" procedure automatically deletes unneeded variables or reads new variables to accommodate the limits of computer memory. This approach allows AIDA to handle relatively large files.

Errors are generally trapped by the program, a short error message is displayed indicating the type of error and the program line number where operation halted, and the user is usually returned to the command prompt without loss of data. Although certain errors may cause the program to crash, the documentation indicates a procedure whereby operation can often be resumed without loss of data. Error messages are not listed in the manual.

Output can optionally be routed to the screen display, a printer, or a disk file with the "PRINT" command. Before initiating a statistical routine for which printed results are desired, the user must remember to toggle on the printer. The data file name and date of analysis are printed only once, when the "PRINT" command is executed. Subsequent printed output appears exactly as on the screen display (i.e., a "screen dump").

A-STAT

A-STAT routines are contained on two disks, and the user is not prompted to change disks when a requested routine resides on the other disk. The user enters desired procedures and parameters at the command prompt "??". A-STAT utilizes a much more complex procedural language than other command-driven programs.[1/] For example, a command line requesting a frequency distribution of a variable called VAR1 might appear as follows: ?? FREQ IN=DATAFILENAME (IF VAR1 .GT. 20 DELETE) (C VAR1)$. EXEC files can be used for "batch" processing.

Although the program is compiled in Applesoft BASIC, program operation is quite slow. This is principally attributable to the program's data handling method, which requires the complete data file to be read from disk for each command. Only A-STAT among the evaluated packages stores data files in fixed format with ASCII codes. This makes A-STAT data files easily accessible by other programs such as mainframe statistical programs. On the other hand, this approach requires more storage space on disk and slows the process of reading the data.

Errors are generally trapped by A-STAT and the program is re-started automatically. Since files are read from disk for each procedure anyway, only time is lost. Applesoft BASIC error numbers and cryptic messages are usually displayed, which does little to assist in identification of the problem. Many errors cause the program to crash with no indication concerning the source of error. The documentation does not provide reference information for error messages.

If printed results are desired, the user must not forget to toggle on the printer before a statistical command is entered. Only the file name

---

[1/] As mentioned earlier, A-STAT is a language-compatible subset of the mainframe statistical package P-STAT. Note that the recently released A-STAT 83 allows selection of menu- or command-driven operation (see Annex 4 for publicized changes in new versions of the evaluated programs).

identifies the analysis at hand in printed output. Rather than a report format (such as NUMBER CRUNCHER), printed A-STAT output appears exactly as on the screen (i.e., a "screen dump").

MICROSTAT

Most interpreted programs for CP/M operating systems are written in BASIC-80, the popular BASIC dialect developed by Microsoft. Although a version of MICROSTAT written in BASIC-80 is available, the evaluated version is written in BAZIC, a dialect which reputedly offers speed and precision advantages.

Program operation begins with the main menu, from which the user can branch to data or statistical procedures. Prompts follow a consistent sequence in all routines, and menu items or subsequent options are selected by pressing only one key (rather than the more common "KEY + RETURN" sequence). Default values are usually adequate and can be selected by simply pressing RETURN. The data file must be read from disk for each routine and the file read operation is consequently rather slow.

Error trapping of MICROSTAT is inconsistent. If an unacceptable value is entered in response to an option, the screen flashes "ERROR" and the user must enter an allowable value. With memory overflow, print format overflow, or other such errors, the program crashes, leaving the user stranded in unknown BAZIC territory. Error messages are not referenced in the manual.

All output is displayed initially on the screen. The user can then select the option to print an output report without repeating the statistical routine. The file name, additional information identifying the file, and optional information identifying the current analysis are also printed.

NUMBER CRUNCHER

This program is written in the interpreted BASIC language supplied with the IBM Personal Computer. From the main menu, program operation branches to file handling, data transformation, or statistical procedures. A file must be "opened" with a selection from the file handling sub-menu before it can be accessed by any data processing or statistical routine. Although the data must be read from disk for each routine, program operation is relatively fast.

Few errors are trapped by the program. Generally, errors cause the program to crash to the programming language level, and a terse BASIC error message is displayed. The program must then be re-started. No explanation of error messages is found in the documentation.

Output is displayed initially on the display screen. A series of options follow which permit the user to print an output report without repeating the statistical procedure. The printed output includes the date and hour plus the data file name and additional information to identify the file.

SPS

As with A-STAT, SPS routines are contained on two disks. However, the user is prompted to change disks when a selected routine resides on the other disk. The program is written in interpreted Applesoft BASIC.

The main menu presents all statistical options plus an option to branch to a data file handling menu, which in turn can provide access to a data transformation menu. Data from the disk file must be re-read into memory as each routine executes, which significantly slows program operation. Because the entire data set must fit into computer memory, disk file size is restricted to approximately 3,400 data points (number of cases times number of variables). Therefore, it may be necessary to split survey data among several files. Unfortunately, SPS does not allow selected variables from separate files to be joined in memory. Instead, variables must be extracted or "stripped" from an existing file, written to a new disk file, and finally merged with a file of the "stripped" variables from other files. Not surprisingly, the process is slow.

Errors are usually trapped by SPS, and the current routine is re-started. However, data generally must be re-read into memory. Applesoft BASIC error numbers and a short list of general areas to investigate are displayed, but these are of little help in identifying the specific problem. The documentation provides no reference information about error messages.

Before execution of a statistical routine, SPS prompts the user to select either printed or displayed output. The printed results appear exactly as on the display screen, with no information to identify the file or analysis.

III. COMPARISON OF SELECTED ROUTINES

A. Introduction

This chapter describes and evaluates data handling, file management, and statistical routines of the selected programs. In the first section, data entry, editing, and listing are discussed as a functionally related group, while file management and data transformations form a second interrelated category. In the remaining sections, the descriptive statistics, frequency distributions, cross-tabulations, analysis of variance, and regression analysis routines of the evaluated packages are compared.[1/] A summary table which indicates relevant features of the programs and which presents certain objective and subjective comparisons is presented in each section.

Certain notation and terminology is common to all the summary tables and should be explained at the outset. For all routines which produce printed output, the type of "Output Information" is indicated. "File" information identifies the current data file in printed output, while "job" information identifies the results of a particular statistical routine (e.g., "Frequency Distribution of Variable Sex"). Both the above serve to identify output as the researcher's memory fades. "Variable names" (e.g., Variable One = SEX) identify particular variables in printed output whereas "value labels" indicate the meaning of particular value codes for frequency distributions and cross-tabulations (e.g., Variable Sex: 1 = female, 2 = male).

Most tables indicate whether a subset of the data file can be analyzed. Whereas selecting a subset of variables implies that only certain variables are processed, selecting a subset of cases means that the user specifies cases for analysis on the basis of a case range (e.g., Case 1 thru Case 50) or on the base of a conditional value of a variable (e.g., analyze only cases in which variable Sex = 1).

All "default maximum" values refer to the unmodified parameters of the evaluated program. In some cases, parameters such as the maximum allowable number of value codes for a frequency distribution table can be increased by the user, but there was no attempt to assess actual limits.

[1/] Although the classification presented here does not always correspond to distinct routines of a particular program, the categories are functional for conceptual and evaluation purposes.

The "Evaluation Measures" portion of most tables indicates the minimum number of user responses, the time (in minutes) required to complete the standard data processing exercises, and the rating of several aspects of program documentation, operation, and performance. "Minimum user responses" refers to the number of keystrokes necessary to complete the simplest possible analysis with that routine. Standardized exercises were developed (using the Cameroon and Honduras research data mentioned earlier) to compare the time necessary to complete a typical analysis. For example, the timed frequency distribution exercise consisted of frequencies for four variables with discrete values plus a frequency distribution of a continuous variable with values grouped in the exercise.

The ratings listed in the tables do not explicitly compare or rank the programs. Instead, the programs are evaluated with respect to desired features and statistical output elaborated for each routine. Documentation is evaluated against the features described in Chapter II--namely, writing style and terminology, content and organization, and ease of use. "Ease of learning" presents our evaluation of the level of difficulty likely to be encountered by a person who has only moderate experience with microcomputers. "Ease of use" refers to the facility with which data analysis can be accomplished once the user has learned the program. There may well be a conflict between ease of learning and ease of use. A program may assist the new user with extensive screen explanations and prompts which become redundant and time-consuming as the user gains experience. On the other hand, another program may give little guidance to the new user but be very efficient once procedures are learned.

An extremely important factor which affects both ease of learning and use is the ability to access variables by reference to a user-assigned acronym or "variable name," rather than by a program-assigned (and much less meaningful) sequential variable number. The "user friendliness" of AIDA and NUMBER CRUNCHER suffers from their use of the latter approach.

Evaluation of printed output is based on output presentation and scope of printed statistics. Our criteria of "good" presentation is based on adequacy of information identifying both the analysis and variables at hand and the readability of output. Statistics scattered across the page, unaligned decimal points, and numbers which run together are all aspects of formatting which detract from output readability. The range of calculated statistics, however, may be the "bottom line" of package performance for many users. Although each researcher will have different requirements, we have developed for each routine a set of statistics considered to be the minimal acceptable for cross-sectional research. A routine which generates the minimal statistics is assured an "adequate" rating. Calculation of the required statistics plus others considered useful for economic research warrants a "superior" rating for this feature.

In order that the reader can compare formats and statistics, printed output of a sample analysis is reproduced for each routine. It should be noted that formatting problems discussed in the text may not be obvious because of the very simple procedure chosen for sample output. Finally, note that only conspicuous strengths or weaknesses of the routines will be discussed in this chapter. If a program feature is not treated, it is considered to be adequately handled by the package.

B. Data Entry, Editing, and Listing

Data entry is not a statistical function, and a statistics program cannot be expected to have the extensive capabilities available in a specialized data entry or data base management package. Nevertheless, since data entry, editing, and variable manipulation are usually accomplished within the statistics package, such routines must be relatively fast, easy to use, and have sufficient capacity to handle typical data processing tasks.

Of critical importance to the researcher are program limitations on the allowable range of data values and the maximum variables, cases, or data points in a file. In some cases, restrictions are attributable to hardware or software limitations, while in others program design is the limiting factor. For example, if the program is written so that the data must all be resident in memory before execution of a statistical routine, the data file size will obviously be limited to the available random access memory (RAM) capacity. If cases are read one at a time during execution of a routine, the maximum file size may well be much larger. In any case, cross-sectional research will often involve analysis of data values of at least seven digits, with a data file size of at least four thousand data points (e.g., forty variables and one hundred cases). The data value restriction is virtually unmodifiable. However, practical limitations on data file size often depend more on the ease with which large data sets can be split into separate disk files (certain variables of which are later joined in memory for statistical analysis) than on stated file size limits.

Unambiguous screen prompts displaying user-specified variable names are recommended to avoid unnecessary errors in data entry, and the user should be permitted to correct at once the inevitable mistakes that occur. Highly desirable features are range checking (which compares an entered data value against a table of user-specified allowable data ranges) and verification (which checks the consistency of re-entered data against the original file).

Data editing is usually necessary to correct errors in coding or data entry. Since the user frequently will not know the exact case number in which an error has occurred, it is desirable that the edit routine include a procedure to quickly scan portions of the data file. Once the case is located, it should be possible to edit only the incorrect data point rather than respond to edit prompts for every variable in the case. Prompts in the routine should use variable names for clarity. Some routines make changes in the disk file immediately after each data point is edited (which can be time-consuming). Others hold the changes in memory and prompt the user to save the file when the editing session is completed. The danger in the latter procedure is that the user will forget to save the data to disk and thereby lose the corrections.

Often the researcher wishes to add new cases or variables to a data set or to delete cases or variables no longer required for the analysis. Besides the overall package limit on number of variables, addition of new variables is often subject to a limit specified by the user at the time of file creation (to take advantage of variable-case limit tradeoffs). Since it is not uncommon to underestimate the additional number of variables to be created with transformations, the ease of modifying such limits is important. Addition of new cases is sometimes subject to the same type of restriction. The inability to delete unwanted variables is usually little more than a nuisance, since the user can simply ignore such variables in the statistical routines. However, the inclusion of unwanted cases in the analysis would certainly distort results.

Finally, the capacity to easily list selected variables or cases to the screen or printer is an important feature. The researcher will often need to visually review the values of certain variables in a selected case range or according to the conditional value of another variable. If the edit routine does not permit the user to quickly scan the data, a good list procedure may serve the same purpose. Listing to the screen (rather than to the printer) therefore must be fast and easy to accomplish. In both cases, the variables should be clearly identified by their user-specified names.

Table 5 presents a summary of data entry, editing, and listing features of the evaluated programs.

ABSTAT

Data files in ABSTAT are limited to a maximum 64 variables within the general restriction that files contain less than 4,000 data values. For example, a data file with 64 variables could contain only about 60 cases. A file with only 10 variables, however, could have about 400 cases. This constraint is mitigated by the fast procedure for joining in memory variables or cases from separate files. The program allows easy conversion (both ways) between its own files and comma-delimited ASCII files (plus files of dBASEII, a popular data management package).

New users may be a bit confused by the file creation procedure. This requires the user to define a temporary file in memory which (after entry of data) can be saved to permanent disk file. A single routine is employed for all data entry and editing of variable names and values. The terse command syntax (e.g., "EDIT V1C2") makes this procedure somewhat more difficult to learn, but its flexibility, power, and speed are excellent. Addition or deletion of variables/cases is accomplished in seconds. Although errors in data entry can be quickly corrected, the data entry routine provides no facility for range checking or verification (which would support more efficient data entry). While the edit routine provides no simple method to scan portions of the file, the "PRINT" command accomplishes the same objective in admirable fashion.

AIDA

Each AIDA variable is stored in a separate file on disk, which allows relatively fast data access. Procedures are provided which read (but not write) fixed format ASCII files created by other programs.

Table 5. Features of Data Entry, Editing, and Listing Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
A. CAPABILITIES						
Data Entry						
1. Input File Type	Binary	Binary	ASCII	Binary	Binary	ASCII
2. Default Max. Data Points Per File	4,000	ND 1/	ND	ND	12,000	3,400
Default Max. Cases	ND	ND	ND	ND	400	ND
Default Max. Variables	64	255	45	30	32	ND
3. Allowable Input Characters	Numeric	Numeric	Num./Alpha.	Numeric	Numeric	Numeric
4. Scientific Notation Input	Y	Y	N	Y	Y	Y
5. Range of Allowable Values	10±E17	±32,767	>10±E16	>10±E16	>10±E16	10±E7
6. Range Checking	N	N	N	N	N	N
7. Verification	N	Y	N	N	N	N
8. Case Number Variable Auto Generated	N 2/	N 2/	Y	N	N	N 2/
9. Correct Prior Entries Within Routine	Y	Y	N	N	N	N
10. Max. Characters/Variable Name	8	10	NAL	5	NAL	4
11. Prompts Use Variable Names	Y	Y	N	Y	N	Y
Data Editing						
12. Scan File Within Routine	N	N	N	N	N	N
13. Edit Single Data Point	Y	Y	Y	N	Y	Y
14. Prompts Use Variable Names	Y	N	Y	Y	N	N
15. Changes Made In: Memory or Disk File	Memory	Memory	Disk	Disk	Disk	Memory
16. Add Variables	Y	Y	Y	N	Y	Y
17. Add Cases	Y	Y 3/	Y	Y	Y	Y
18. Delete Variables	Y	Y	Y	N	N	Y
19. Delete Cases	Y	Y 4/	Y	Y	N	Y

(continued)

Table 5. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
Data Listing						
20. Select Subset of:						
Variables	Y	Y	Y	Y	Y	Y
Cases (Conditional Basis)	Y	Y	Y	N	N	N
Cases (Range Interval)	Y	Y	Y	Y	N	Y
21. User-defined Format	N	N	N	Y	Y	N
22. Output to Disk	Y	Y	N	N	N	N
23. Output Information:						
File Identification	Y	N 5/	Y	Y	Y	Y
Job Identification	Y	N	N	Y	N	N
Variable Names	Y	N	Y	Y	N	Y
B. EVALUATION MEASURES						
Rating of:						
Documentation	A-S	S	U	A-S	U	A
Ease of Learning	A	S	U	A	U	A
Ease of Use	S	S	A	U	U	A
Presentation of Printed Output	S	A	U	S	A	U

1/ Manual states that more than 11,000 data points can be held in memory.

2/ Case numbers can be generated with transformation routines.

3/ Very difficult to add cases beyond the current file size.

4/ Cases can be "temporarily" deleted from statistical routines.

5/ Information presented only at time "PRINT" is invoked.

Y = Yes N = No

U = Unsatisfactory A = Adequate S = Superior

NAL = No Apparent Limitation ND = Not Documented

Data value restrictions are a serious problem in AIDA. No data value may have more than five digits (no matter where the decimal point is placed) and the maximum allowable range of integers is ± 32,767. For example, a value of $1,000.01 could not be represented in AIDA.

Data files are created as temporary files in memory and must be explicitly saved to disk by the user. The data entry procedure requires the user to indicate the maximum number of cases which will be entered in a new file and automatically calculates the maximum allowable number of variables for the file. Once specified, this limit is very difficult to modify. Given the relatively large allowable file sizes (at least 11,000 data values), few users will find the variable/case tradeoff limiting. Data entry is very well organized and provides the ability to easily correct previously entered values. Although range checking is not permitted, AIDA provides the only data verification procedure among the evaluated packages.

While the user must know the specific case number and variable to be changed, data editing is otherwise straightforward. Cases can be easily added up to the limit specified upon file creation, but permanent deletion of unwanted cases is a complicated procedure. Although data files cannot be scanned within the edit routine, screen display of portions of the data set is relatively easy with the list routine. Annoying features of the list routine are the lack of variable names and the poor formatting of printed output.

A-STAT

A-STAT data is stored in fixed ASCII format and has extensive facilities for reading and writing files in the popular Data Interchange Format (DIF) and in the format of the Apple File Cabinet program. The ASCII storage format, however, has the drawback of slow data access within A-STAT. Data value and file size limits are probably more than adequate for most research efforts. Alphanumeric characters are valid data values, which may provide facility for expanded table labeling and listing.

Since there are three unrelated routines which can be used, data entry is very confusing in this package. The routines DATA (fixed format entry) and READ (free format entry) are as difficult to learn (and just as unfriendly) as are many mainframe routines. The authors of A-STAT seem to hint that the user should employ her/his own text editor for data entry.

Addition of cases is possible with the "EDIT" routine (the third data entry method), but new variables must be entered into a new file and joined to the old file with yet another procedure. Variables or cases are deleted with permanent file modifications (the "NO.OP" routine). The data file cannot be scanned within "EDIT," and the listing procedure is slow and cumbersome.

MICROSTAT

MICROSTAT stores data in files which can only be accessed by programs written in BAZIC, the North Star BASIC-compatible language of this version. A utility is provided to convert comma-delimited ASCII files created by other programs, but conversion of MICROSTAT files to a format accessible by other software requires complicated programming. Files are limited in size only by disk capacity. The default limitation of 30 variables can be increased if the user re-installs the program.[1/]

Data entry, editing, and listing are all accomplished through a single Data Management System sub-menu, thereby facilitating these operations. Although entry is relatively straightforward, entered values cannot be corrected within the routine. Instead, the user must note the case number of the error and make the correction later with the edit routine.[2/]

While the package provides no method to scan the data while editing, the list procedure allows easy viewing of a subset of variables or cases. Since the user must edit all variables in a case rather than a single variable, editing is a time-consuming activity. Once a file is created, it is not possible to subsequently add or delete variables with the edit procedure. The MICROSTAT list procedure provides the most extensive options for formatted data listings among the evaluated packages.

NUMBER CRUNCHER

Comma-delimited ASCII files may be converted to the binary files which Number Cruncher uses internally, but no reverse conversion is possible. Limits on the number of variables or cases will usually prove non-restrictive, and can be easily modified if a larger file is needed.

[1/] The ultimate limit is not documented and was not determined in the evaluation.

[2/] The case number is displayed as part of the data entry prompt. However, a case number variable is not automatically created.

The data entry procedure of this package is a "bare bones" routine with limited flexibility. The user must contend with confusing terminology which defines cases as "rows" and variables as "columns." No variable names are displayed in prompts. In addition, entered data values cannot be corrected within the data entry routine. The edit routine requires the user to know the specific case and variable number to be edited and does not provide any easy method to scan portions of the data set.[1/] The "LIST" (screen output) and "PRINT" (printer output) routines provide little help since the data file can only be viewed from beginning to end. The printer output routine requires user knowledge of BASIC, as variables must be formatted with the "##.#" notation of that language.

<u>SPS</u>

This program includes extensive capacity for file conversion. Users can convert SPS files to ASCII fixed format, to DIF files, or to a format usable by the popular VISIPLOT program. Fixed format, comma-delimited, and DIF files can also be converted to SPS files.

Although the range of allowable data values is somewhat smaller than that of most other programs (7 digit values), it is not likely to cause serious problems for most researchers. Because the data file is restricted to approximately 3,400 data points and because there is no easy method to join variables from separate files for analysis, SPS could be cumbersome for large surveys.

Data entry is relatively clear, although it is difficult to comprehend why this routine is invoked by a control key sequence rather than by the otherwise consistent menu selection procedure. No verification or range checking is possible, and entered data values cannot be changed within the routine.

Editing is also generally straightforward, although the routine requires the user to specify the program-assigned sequence number of the variable to be corrected rather than the user-assigned variable name. The list routine for scanning the data file is wonderfully flexible, allowing the user to easily identify errors by scrolling backwards and forwards through the data. The printer output suffers from the absence of file, job, and variable identification, and from poor formatting.

[1/] The case number is displayed as part of the data entry prompt. However, a case number variable is not automatically generated.

C. Data Management

File management and data transformations will probably occupy much more of the researcher's time than statistical procedures and are just as important for the final outcome of the analysis. The data (once entered and edited) in most instances must be transformed, recoded, and manipulated in several ways before being submitted to a statistical routine. The ease of using data management procedures will have an important influence on the overall user-friendliness of the package.

Minimum transformations should include the ability to add, subtract, multiply, or divide variables either by a user-specified constant or by other variables in the data set, plus the capacity to perform logarithmic and exponential transformations. Probably the most critical requirement for cross-sectional socioeconomic research is that re-coding of continuous variables to discrete interval codes (grouping) be possible and relatively easy to accomplish. This capability is especially important for frequency distribution and cross-tabulation analysis.

Two general methods of user input are employed in transformation routines. One presents a menu of available transformations and prompts for user input of the variables to be transformed. The other method requires the user to enter the variables and transformation operators in some variant of an equation form. The new user will usually find the first procedure more "user-friendly," while the latter is often more flexible and powerful for the experienced user. Since the details of how variables were transformed could easily be forgotten, the package should provide in either instance some procedure by which the user can obtain a hardcopy listing of transformation operations. To obtain such a listing for the evaluated packages, it is necessary to toggle the printer so that all prompts and user responses are sent to both screen and printer. This method is inefficient and slow.

Other data management procedures necessary in a statistical package include the capability to select and write to disk subsets of the data, to sort the file by one or more variables, to join two or more data files, and to execute operating system functions. The first procedure is necessary if the user desires to produce a new file which contains a selected group of variables or cases to be analyzed separately. This procedure may be less important if the package provides a fast method of selecting such subsets within the various statistical routines.

The user will often wish to view or print the data in ascending or descending order according to one or more sort variables. The sort procedure must be used if the package only allows selection of subsets by case ranges, rather than by values of an "index" variable (i.e., "conditional" selection). MICROSTAT is the only evaluated program which exhibits this limitation. Some sort routines alter the order of cases in memory while others alter the original disk file or produce a new sorted file on disk.

Joining two or more files is often necessary if data have been entered in separate sessions or on separate microcomputers. Files may have to be appended vertically (adding cases from one file to another) or horizontally (adding variables from one file to another).

Finally, a good data management routine will execute essential operating system commands within the program or will allow a temporary exit to the operating system. Critical system operations to be performed include viewing file directories plus deleting, renaming, and copying files.

A summary of data management features of the evaluated programs is contained in Table 6.

ABSTAT

All desired data transformations are input by the user in equation form with left-hand variables calculated as the result of algebraic manipulations of right-hand constants or variables (e.g., [X2] = [X1] + 10). Certain pre-established functions such as "MEAN," "LN" (natural log), and "CASNUM" (sequential case number) make these manipulations easier. Grouping continuous values into discrete codes is quite simple. An exceptional feature of the routine is the ability to stipulate up to five conditions (connected by a logical "AND") which restrict the application of a particular transformation to a subset of cases. Although the experienced user will appreciate the flexibility and speed of this routine, the command syntax is somewhat demanding.

Selecting subsets, sorting data, and appending files are all fast and easy to accomplish, with no limitations of any practical consequence. A weak point in data management of this package is the absence of any procedure to issue system commands while operating the program. If a file name is forgotten, the ABSTAT user must exit to the operating system level in order to obtain a file directory.[1/]

[1/] The CP/M-80 version of ABSTAT allows the user to view the disk directory.

AIDA

AIDA allows an exceptional combination of pre-established functions and user-specified equations for transformation of variables. For example, the "INTERVAL" function provides a simple and effective method to group continuous data. Another function lags data values for time series-type analysis. Equation transformations are more difficult to learn but are limited in power only by the imagination of the user and the restrictions of Applesoft BASIC. It is necessary that the user have some familiarity with BASIC in order to exploit the potential of this routine.

Albeit with considerable difficulty, most file management functions can be accomplished with AIDA. For example, appending (joining) two files vertically requires the user to re-start the system and run a separate program. There seems to be no procedure to join AIDA files horizontally. Operating system commands can be executed while operating AIDA.

A-STAT

A-STAT employs a (poorly documented) "variable transformation language" to carry out data manipulations. The necessary conversions are usually entered as options in the statistical procedure command line (e.g., (IF SEX .EQ. 1 SET SEX .EQ. 0)) and can be made permanent by creating a new file with the file modification procedure. Transformation formulas are limited to 255 total characters. In addition, the procedure seemed to be easily overwhelmed by manipulations of only moderate complexity. Since both the sub-program and the data are read from disk for every routine, transformations are very slow. If a syntax or typing error occurs, all routines must complete their read procedures before the error is detected and the program re-started. This can cause considerable time to be lost.

MICROSTAT

In this package, a menu of transformation options is presented to the user. Unfortunately, the available transformations may be inadequate for some research purposes. In particular, only algebraic transformations are permitted, and the lack of a recode function makes the grouping of continuous variables impossible.[1/] Another drawback is the difficulty of

---

[1/] The recently released version 4.0 advertises grouping (recoding) capability and increased transformation capability (see Annex 4 for publicized changes in new versions of the evaluated programs).

Table 6. Features of Data Management Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
A. CAPABILITIES						
Select						
1. Write to File Subset of:						
Variables	Y	Y	Y	Y	N	Y
Cases (Conditional)	Y	Y	Y	N	N	Y
Cases (Range)	Y	Y	Y	Y	N	Y
Sort						
2. Maximum Key Variables for Sort:	NAL	1	NAL	NAL	1	No Sort
3. Ascending/Descending Sort Order	Ascending	Ascending	Ascending	Both	Both	No Sort
4. Sorts File in:	Memory	Memory	New Disk File	Same Disk File	New Disk File	No Sort
Append Files						
5. Vertically	Y	Y	Y	Y	N	Y
6. Horizontally	Y	Y	Y	Y	Y	Y
Interface With Operating System						
7. Temporary Exit to Operating System	N	Y	Y	N	N	N
8. System Commands Within Program:						
View Disk Directory	N	N	N	Y	Y	Y
Delete Files	N	N	Y	Y	Y	Y
Rename Files	N	N	N	Y	N	N
Copy Files	N	N	N	Y	N	N

(continued)

Table 6. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
Data Transformations						
9. Form of User Input	Equation	Equation	Equation [1]	Menu	Menu	Menu
10. Hardcopy List of Trans. Formulas: [2]	N	N	N	N	N	N
11. Add/Sub/Mult/Div:						
Constant (e.g. X1 + 10)	Y	Y	Y	Y	Y	Y
Variables (e.g. X1 + X2)	Y	Y	Y	Y	Y	Y
12. Standardize	N	N	N	N	N	Y
13. Anti-log	Y	Y	N	Y	Y	Y
14. Exponential	Y	Y	Y	Y	Y	Y
15. Logarithmic						
Common	N	N	N	Y	N	Y
Natural	Y	Y	Y	Y	Y	Y
16. Geometric	Y	Y	Y	N	Y	N
17. Lag/Lead	Y	Y	N	Y	Y	N
18. Moving Average	Y	Y	N	Y	Y	N
19. Grouping (Recoding)	Y	Y	Y	N	Y	Y
20. Rank-ordering	N	Y	N	Y	N	Y
B. EVALUATION MEASURES						
1. Minimum User Responses (Keystrokes)	17	18	15	18	9	13
2. Minutes Required to Complete Exercise	20	18	55	45	30	111
3. Rating of:						
Documentation	S	S	U	A-S	U	A
Ease of Learning	A	A	U	A	A	A
Ease of Use	S	S	U	U	A	A

[1] All transformations done in statistical or file routines.

[2] In all packages, hardcopy list of transformations possible only by toggling on the printer.

Y = Yes N = No

U = Unsatisfactory A = Adequate S = Superior

NAL = No Apparent Limitation

combining two or more variables to form a third variable. The routine is also not particularly easy to learn or to use once learned. Since MICROSTAT is menu-driven, many user responses are required. Thus, transformation operations are quite slow.

Writing a subset of data to a new disk file on the basis of a conditional value is possible but not feasible in practical applications. To accomplish such a task, the user must first sort the data by the variables which reflect the selection procedure and then write the desired range of cases to a new disk file. The sort procedure was found to be slow and fatally flawed. In several instances during the sort operation, some cases were dropped entirely from the data file, while other cases were duplicated.[1/]

The data management sub-menu contains extensive operating system commands which allow the user to easily accomplish most necessary disk operations from within the program. Files can be renamed, copied, or deleted while operating MICROSTAT.

NUMBER CRUNCHER

Since the documentation of this routine is woefully inadequate (many transformations are not even mentioned in the manual), the user is left to her/his own devices to learn the procedures. Nonetheless, an extensive set of transformation options are available, encompassing the operations necessary for most research efforts. Although not difficult to learn, the menu-driven transformation procedures require numerous and repetitious user responses.

File management operations are insufficient. It is not possible to create a permanent subset of the data variables or cases, and files cannot be vertically joined.[2/] The file management procedures of copying and renaming files are also not accessible while operating the program.

[1/] We carefully examined this problem with six different data files and documented the errors. Using the same data as our exercises, the program author has informed us that the sort procedure functioned correctly and that no other user has reported such a problem. In any case, the new Release 4.0 is written with another programming language and allows selection of data subsets without sorting. (See Annex 4 for a summary of changes in the new versions of the evaluated programs.)

[2/] The program author has advertised that subset selection has now been implemented (see Annex 4).

SPS

The transformation functions of SPS are also presented in a menu format, with a scope adequate for most research tasks. The user will find the data transformation procedures relatively easy to learn. However, most functions require repeated (and slow) file read and write operations for each type of data transformation. The extremely slow operation of this routine significantly limits its usefulness for research applications.

Although a function labeled "SORT" is presented in the transformation menu, the procedure actually selects a subset of cases by a conditional value. There is no sorting routine in the program. Moreover, the user is sure to be confused by the poor use of terminology. The program also lacks the ability to rename or to copy files.

D. Descriptive Statistics

The descriptive statistics routine is the first and most critical part of the three standard data description routines, and as such it must be fast, efficient, and easy to use. Although data transformations are generally not required, the user should be allowed to easily select a subset of variables or cases to analyze. In particular, it is often necessary to select a subset of cases according to the conditional value of an "index" variable. Such a feature may allow the user to simulate the Statistical Package for the Social Sciences (SPSS) procedure BREAKDOWN (often used in socioeconomic analysis) which calculates means, standard deviations, sums, and number of cases for subsets of the data set.

Minimum required statistics for this routine are:

1. Sample mean
2. Sample standard deviation
3. Number of missing cases
4. Number of valid cases

Accuracy of the descriptive statistics routines was tested by forcing the routines to calculate very large sums of squares (up to 9E+12). All programs produced the correct results. (For a discussion of the evaluation methods for testing precision, see Annex 2.)

Table 7 presents a summary of features in the descriptive statistics routines of the evaluated programs.

Table 7. Features of Descriptive Statistics Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	DESC	DESCRIBE	DES 1/	DESC. STATS.	UNIVAR. STATS.	DESCRIBE
A. CAPABILITIES						
1. Select Subset of:						
Variables	Y	Y	N 2/	Y	Y	Y
Cases (Conditional)	Y	Y	Y	N	Y	N
Cases (Range)	Y	Y	Y	Y	N	N
2. Output to Disk	Y	Y	N	N	N	N
3. Output Information						
File Identification	Y	Y 3/	Y	Y	Y	N
Job Identification	Y	N	N	Y	N	N
Variable Names	Y	Y	Y	Y	Y	Y
4. Simulate SPSS-like Breakdown:	I	I	Y 2/	I	Y	I
B. STATISTICS						
1. Mean (sample)	Y	Y	Y	Y 4/	Y	Y
2. Std. Dev. (sample)	Y	Y	Y	Y	Y	Y
3. No. of Missing Cases	Y	N	N	N	N	N
4. No. of Valid Cases	Y	Y	Y	Y	Y	Y
5. Variance (sample)	Y	N	N	Y	Y	Y
6. Std. Error of Mean	Y	N	N	Y	Y	N
7. Mode	Y	N	N	N	N	Y
8. Median	Y	N	N	N	N	Y
9. Sum	N	N	Y	Y	Y	N
10. Min./Max.	Y	N	Y	Y	N	Y
11. Range	Y	N	N	N	N	N
12. Coefficient of Variation	Y	N	N	Y	N	N
13. Skewness	Y	N	N	Y	N	Y
14. Kurtosis	Y	N	N	Y	N	Y
15. Confidence Intervals	N	N	N	N	Y	N
16. Sum of Squared Dev.	N	N	N	Y	N	N
C. EVALUATION MEASURES						
1. Minimum User Responses (Keystrokes)	7	5	11	6	9	10
2. Minutes Required to Complete Exercise	32	16	32	61	16	38
3. Rating of:						
Documentation	A	A	U	A	A	U
Ease of Learning	S	S	U	S	A	A-U
Ease of Use	S	S	U	A	A	U
Printed Output						
-Presentation	S	A	U	S	S	A-S
-Statistics	A	U	U	U	U	U

1/ DES not a routine. Descriptive statistics file generated when data file created or transformed.
2/ Can be done with separate routine "SUBSTATS".
3/ File name listed only when "PRINT" invoked.
4/ Only MICROSTAT also lists estimated standard deviation and variance for population.

Y = Yes N = No I = Incomplete

U = Unsatisfactory A = Adequate S = Superior

1. Description of Evaluated Routines

ABSTAT

Although not particularly fast, this routine excels in ease of learning and use. An interactive HELP facility, clear documentation, and a "getting started" sample session guide the new user through the learning phase. Once the data is read into memory, seven keystrokes will generate some basic statistics for all variables in the data set. Selection of a subset of variables or cases is simple with the "SELECT" or "TRANS" commands.

The printed output is easy to read, lists file and job identification, and is presented in a well-constructed format. However, the routine does not output the sum of the analyzed variable, which makes a BREAKDOWN-like output unattainable.

AIDA

The interactive command summary, the clear documentation, and the sample session output will make learning this routine simple for most users. A short command sequence will produce statistics on a selected variable within a subset of cases. Especially easy to use are the fast and powerful transformation and case selection procedures, accessible either within the routine or as a separate module. Although the output is quite presentable, the statistics printed are fewer than for any other evaluated package. In addition, printed output does not include the number of missing values for the analyzed variable. Combined with the overall program data value restrictions, these shortcomings may limit the usefulness of the routine for many applications. A complete BREAKDOWN-like procedure is not possible.

A-STAT

Upon completion of data entry, A-STAT automatically generates a separate disk file containing descriptive statistics for all variables in the main data file. The absence of a distinct descriptive statistics routine may confuse the new user. The further absence of documentation on this unusual feature will not alleviate such confusion.

Unlike other A-STAT transformations, the selection of a subset of variables or cases for descriptive statistics is a separate routine which requires the user to generate an entirely new descriptive statistics file. To its great credit, this separate routine (called "SUBSTATS") produces a very good replicate of a BREAKDOWN procedure.

The output is very poorly formatted and includes only limited statistics for all variables in the descriptive file. The annoying "Hit RETURN for more" prompt can be eliminated from printed output, but results are then impossible to review on the display screen.[1/]

MICROSTAT

Most new users could produce descriptive statistics using this routine without reading the manual. The menu-driven routine is very clearly presented and provides adequate default values at each option. A severe limitation is that selection of a subset of cases by an index variable is possible only through the slow and unreliable SORT routine (see the section on data management).

The output is clearly exceptional in both quality of presentation and quantity of statistics reported. However, the number of missing cases for the analyzed variable is not listed.

NUMBER CRUNCHER

This routine is quite functional and is not difficult to learn. The menu options are clearly stated and well presented with screen highlighting. The new user may be confused by the requirement of referring to variables by a number representing their sequence in the data file. The experienced user will likely become further frustrated with the number of required responses to program prompts.

An excellent BREAKDOWN-like output can be produced using multiple index variables. The output is extremely well organized and formatted, but the statistics listed lack the number of missing cases.

SPS

Documentation for this routine is sparse and unhelpful. No sample runs are presented in the manual, and terms like "leptokurtic" will surely

---

[1/] The "Hit RETURN for more" prompt pertains to a program operation which temporarily halts output to the screen or printer after listing a specified number of output lines.

intimidate the beginner. Screen prompts are difficult to understand and are often misleading. Although possible, selection of a subset of cases requires time-consuming creation of a new disk file.

The printed output is a "screen dump" which provides no file or job identification. The poor organization and formatting also contribute to make the output unsuitable for research reports. The statistics presented do not include the number of missing cases.

2. Sample Output of Evaluated Routines

ABSTAT Descriptive Statistics Output

```
  ABSTAT 3.00
FILE: B:AB-EVAL3  REV# 0                                              PAGE  1

COMMAND: DESC


THERE ARE 17 VARIABLES AND  200 CASES
```

VARIABLE	MEAN	STD.DEV.	VARIANCE	STD ERROR OF MEAN	COEFF OF VARIATION
6 ASS	3157.93	13109.9	1.71870E+008	927.013	415.144
7 WRK	1.91000	1.94701	3.79085	.137675	101.938
8 PRD	5853.15	12503.4	1.56334E+008	884.121	213.618

VARIABLE	MEDIAN	MODE	MINIMUM	MAXIMUM	RANGE
6 ASS	916.650	NONE	9.50000	1.53517E+005	1.53508E+005
7 WRK	1.00000	1.00000	1.00000	16.0000	15.0000
8 PRD	1458.26	NONE	17.8800	98965.3	98947.5

VARIABLE	SKEWNESS	KURTOSIS
6 ASS	9.51349	100.693
7 WRK	3.63647	20.4739
8 PRD	4.60567	29.0193

AIDA Descriptive Statistics Output

```
- - - - - - - - - - - - - - - - - - -

COMMAND ? DESCRIBE
  V=6^8


V6 ASSETS

MIN =           0     MAX =       1535
N   =         200     MEAN=       31.1
VAR.=   17188.221     S.D.=    131.104

V7 WORKERS

MIN =           1     MAX =         16
N   =         200     MEAN=       1.91
VAR.=       3.791     S.D.=      1.947

V8 PROD

MIN =           1     MAX =       9896
N   =         200     MEAN=    584.845
VAR.= 1563364.55      S.D.=   1250.346

- - - - - - - - - - - - - - - - - - -
```

Note: Some variables were truncated to conform with AIDA data value restrictions.

A-STAT Descriptive Statistics Output

```
??LIST DES=EVALDAT3$



     A-STAT LIST PROCEDURE.

DES FILE=EVALDAT3
ADES.EVALDAT3,S6,D2,V0 READ, NV=17
 HIT RETURN FOR MORE...?

   DESCRIPTION FILE EVALDAT3,S6,D2,V0
  VARNAME         MIN        MAX        SUM
  -------         ---        ---        ---

1 ID                 1       1135    111981
2 TYP                1         60      1461
3 LOC          7030001   14160001 2.13191564E+09
4 REG                1          3       394
5 STR                1          3       459
6 ASST             9.5   153517.5  631585.3
7 WRKS               1         16       382
8 PROD           17.88   98965.34 1170629.09
9 MAT                0   53521.13 379559.43
10 SAL               0   21532.81 188453.62
11 CP20           7.79   42204.64 170484.33
12 VLP           12.05    4112.24 169794.79
13 VLF               0    6085.88  24139.23
14 VAL         -923.49   74447.11 787024.87
15 CASH       -7768.63   54992.85 553392.591
16 EC20     -14137.69   31728.13 234151.36
17 PD            56.07   22994.66 470371.65

HIT 'RETURN' FOR MORE...?

   DESCRIPTION FILE EVALDAT3,S6,D2,V0
  VARNAME         MEAN       STD         N
  -------         ----       ---        ---

1 ID           559.905   395.603       200
2 TYP            7.305    10.951       200
3 LOC       10659578.2 3530590.77          200
4 REG             1.97     1.002       200
5 STR            2.295      .728       200
6 ASST        3157.927 13109.946         200
7 WRKS            1.91     1.947       200
8 PROD        5853.145 12503.363         200
9 MAT         1897.797  5469.255         200
10 SAL        942.268  3012.977        200
11 CP20       852.422  3642.351        200
12 VLP        848.974  935.992        200
13 VLF        120.696  558.948        200
14 VAL       3935.124   8530.689         200
15 CASH      2766.963   6376.593         200
16 EC20      1170.757    4112.6        200
17 PD        2351.858   3374.205         200
```

MICROSTAT Descriptive Statistics Output

```
---------------------- DESCRIPTIVE STATISTICS --------------------
HEADER DATA FOR: B:EVAL3C     LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200   NUMBER OF VARIABLES: 17


          DESC STATS ON ASSETS, WORKERS AND PRODUCTION


           VARIABLE NAME: ASSET      N = 200

           ARITHMETIC MEAN = 3157.926497

          SAMPLE STD. DEV. = 13109.94586
           SAMPLE VARIANCE = 171870680.5
  COEFFICIENT OF VARIATION = 415.1441103%

      POPULATION STD. DEV. = 13077.12993
       POPULATION VARIANCE = 171011327.1
  COEFFICIENT OF VARIATION = 414.1049496%

STANDARD ERROR OF THE MEAN = 927.0131614

                   MINIMUM = 9.5
                   MAXIMUM = 153517.5

                       SUM = 631585.2994
            SUM OF SQUARES = 3.619676537E+10
              DEVIATION SS = 3.420226542E+10

   1ST MOMENT = 0
   2ND MOMENT = 171011327.1
   3RD MOMENT = 2.127516857E+13MOMENT COEFFICIENT OF SKEWNESS = 9.513398597
   4TH MOMENT = 2.944713412E+18MOMENT COEFFICIENT OF KURTOSIS = 100.6916088

NORMAL DISTRIBUTION GOODNESS OF FIT TEST:
```

Cont. next page.

```
--------------------- DESCRIPTIVE STATISTICS ---------------------
HEADER DATA FOR: B:EVAL3C     LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200   NUMBER OF VARIABLES: 17

          DESC STATS ON ASSETS,WORKERS AND PRODUCTION

           VARIABLE NAME: WORKR        N = 200

           ARITHMETIC MEAN = 1.909999999

          SAMPLE STD. DEV. = 1.947011626
           SAMPLE VARIANCE = 3.790854271
  COEFFICIENT OF VARIATION = 101.9377815%

      POPULATION STD. DEV. = 1.942137997
       POPULATION VARIANCE = 3.7719
  COEFFICIENT OF VARIATION = 101.6826177%

STANDARD ERROR OF THE MEAN = .1376745123

                   MINIMUM = 1
                   MAXIMUM = 16

                       SUM = 381.9999998
            SUM OF SQUARES = 1483.999999
              DEVIATION SS = 754.3799999

   1ST MOMENT = 0
   2ND MOMENT = 3.7719
   3RD MOMENT = 26.63914204   MOMENT COEFFICIENT OF SKEWNESS = 3.63646974
   4TH MOMENT = 291.2867114   MOMENT COEFFICIENT OF KURTOSIS = 20.47388841

NORMAL DISTRIBUTION GOODNESS OF FIT TEST:
```

Cont. next page.

```
---------------------- DESCRIPTIVE STATISTICS ----------------------
HEADER DATA FOR: B:EVAL3C     LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200   NUMBER OF VARIABLES: 17


          DESC STATS ON ASSETS, WORKERS AND PRODUCTION


           VARIABLE NAME:  PROD       N = 200

           ARITHMETIC MEAN = 5853.14545

          SAMPLE STD. DEV. = 12503.36265
           SAMPLE VARIANCE = 156334077.5
  COEFFICIENT OF VARIATION = 213.617836%

      POPULATION STD. DEV. = 12472.06508
       POPULATION VARIANCE = 155552407.2
  COEFFICIENT OF VARIATION = 213.0831223%

STANDARD ERROR OF THE MEAN = 884.1212514

                   MINIMUM = 17.88
                   MAXIMUM = 98965.34

                       SUM = 1170629.09
            SUM OF SQUARES = 3.796234376E+10
              DEVIATION SS = 3.111048143E+10

   1ST MOMENT = 0
   2ND MOMENT = 155552407.2
   3RD MOMENT = 8.934756598E+12MOMENT COEFFICIENT OF SKEWNESS = 4.605402785
   4TH MOMENT = 7.021129401E+17MOMENT COEFFICIENT OF KURTOSIS = 29.01706648

NORMAL DISTRIBUTION GOODNESS OF FIT TEST:
```

Number Cruncher Descriptive Statistics Output

```
---------------------------- Number Cruncher (c)----------------------------
Date:               08-23-1983
Time:               00:05:28
Data File Name:     eval3
File Description:   HONDURAS SSE FILE
```

Results for Column (6)
ASSETS

1.	Number of non-missing values	200
2.	Sum	631585.3
3.	Mean (AVERAGE)	3157.926
4.	Variance	1.718707E+08
5.	Standard deviation	13109.95
6.	Standard error	927.0132
7.	Upper 95% con. limit of mean	4974.872
8.	Lower 95% con. limit of mean	1340.981
9.	Upper 99% con. limit of mean	5545.912
10.	Lower 99% con. limit of mean	769.9402

Results for Column (7)
WORKERS

1.	Number of non-missing values	200
2.	Sum	382
3.	Mean (AVERAGE)	1.91
4.	Variance	3.790854
5.	Standard deviation	1.947012
6.	Standard error	.1376745
7.	Upper 95% con. limit of mean	2.179842
8.	Lower 95% con. limit of mean	1.640158
9.	Upper 99% con. limit of mean	2.26465
10.	Lower 99% con. limit of mean	1.55535

Results for Column (8)
PRODUCTION

1.	Number of non-missing values	200
2.	Sum	1170629
3.	Mean (AVERAGE)	5853.146
4.	Variance	1.563341E+08
5.	Standard deviation	12503.36
6.	Standard error	884.1212
7.	Upper 95% con. limit of mean	7586.023
8.	Lower 95% con. limit of mean	4120.268
9.	Upper 99% con. limit of mean	8130.642
10.	Lower 99% con. limit of mean	3575.649

SPS Descriptive Statistics Output

```
~D ESCRIPTIVE STATISTICS FOR ASST

MEAN = 3157.92           MEDIAN = 916.65

ASST HAS NO MODE

VAR. = 171870681           S.D. = 13109.945

MIN = 9.5           MAX = 153517.5

N = 200

2ND MOMENT = 171011327

3RD MOMENT = 2.12751686E+13

4TH MOMENT = 2.94471343E+18

COEF. OF SKEWNESS = 9.51339856
COEF. OF KURTOSIS = 100.691609

WANT ANOTHER DESCRIPTION (Y/N) Y
DESCRIPTIVE STATISTICS FOR WRKS

MEAN = 1.91           MEDIAN = 1

MODE = 1

VAR. = 3.79           S.D. = 1.947

MIN = 1           MAX = 16

N = 200

2ND MOMENT = 3.7719

3RD MOMENT = 26.639142

4TH MOMENT = 291.286711

COEF. OF SKEWNESS = 3.63646974
COEF. OF KURTOSIS = 20.4738884
```

Cont. next page.

```
WANT ANOTHER DESCRIPTION (Y/N) Y
DESCRIPTIVE STATISTICS FOR PROD

MEAN = 5853.14            MEDIAN = 1458.255

PROD HAS NO MODE

VAR. = 156334078            S.D. = 12503.362

MIN = 17.88            MAX = 98965.34

N = 200

2ND MOMENT = 155552407

3RD MOMENT = 8.93475663E+12

4TH MOMENT = 7.02112947E+17

COEF. OF SKEWNESS = 4.6054028
COEF. OF KURTOSIS = 29.0170666
```

E. Frequency Distributions

After simple descriptive statistics, frequency distributions are likely to be the most commonly used statistical routine in cross-sectional research. The researcher will require frequency distributions for some variables with discrete codes and for others with continuous values; the routine must be flexible and efficient in both instances. There should be no restrictions on allowable data values; that is, all values allowed within program limits should be acceptable for frequency distributions with no requirement of recoding. Further, the program should automatically tabulate all discrete values. The user should not be required to specify all discrete codes to be tabulated. If a continuous variable must be grouped into intervals, the package should provide for user-specified intervals either within the frequencies routine or with an external transformation routine. User-specification of intervals is preferable to the procedure used by some programs in which equal intervals are automatically calculated and tabulated. Although some concessions must be made to computer memory limitations, 25 tabulated intervals or discrete values is probably the minimum necessary for most research. In order to identify discrete codes or re-coded intervals in printed output, the routine should allow user input of value labels. Additional desirable features include histograms and weighted frequencies.

The minimum statistics required are:

1. Absolute frequency (count)
2. Relative frequency (percent)
3. Cumulative frequency (count)
4. Cumulative frequency (percent)
5. Number of valid cases
6. Number of missing cases

A summary of frequency distributions features is given in Table 8.

1. Description of Evaluated Routines

ABSTAT

The clear documentation, interactive HELP procedure, and brief commands make this a superior routine to learn and to use. Unique among the evaluated routines, there appear to be no apparent limit on the number of values that can be tabulated in the frequency table. Selection of subsets

Table 8. Features of Frequency Distributions Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	FREQ	ONEWAY HIST.	FREQ	FREQ. DIST.	HIST.-FREQ.	HIST.
A. CAPABILITIES						
1. Select Subset of:						
Variables	Y	Y	N 1/	Y	Y	Y
Cases (Conditional)	Y	Y	Y	N	N	N
Cases (Range)	Y	Y	Y	Y	N	N
2. Max. Variables Per Freq. Request	NAL	NAL	NAL	1	1	1
3. Treatment of:						
Non-integer Values	Tabulated	Truncated	Truncated	Tabulated	Tabulated	Tabulated
Negative Values	Tabulated	Excluded	Set to zero	Tabulated	Tabulated	Tabulated
4. Group Values:						
Within Routine	N	N	Y	Y	Y	Y
With Transform	Y	Y	Y	Y	Y	Y
5. Max. No. of Value Codes	NAL	101	101	50	40	40
6. Range of Allowable Value Codes:	NAL	0-100	0-100	NAL	NAL	NAL
7. Histogram	Sep. Routine	Sep. Routine	N	Y	Y	Y
8. Weighted Frequencies	N	Y	Y	N	N	N
9. Output to Disk	Y	Y	N	N	N	N
10. Output Information:						
File Identification	Y	Y 2/	Y	Y	Y	N
Job Identification	Y	N	N	Y	N	N
Variable Names	Y	Y	Y	Y	Y	Y
Value Labels	N	Y	N	N	N	N

(continued)

Table 8. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	FREQ	ONEWAY HIST.	FREQ	FREQ. DIST.	HIST.-FREQ.	HIST.
B. STATISTICS						
1. Absolute Count	Y	Y	Y	Y	Y	Y
2. Relative Frequency	Y	Y	Y	Y	Y	Y
3. Cumulative Count	Y	N	N	Y	Y	N
4. Cumulative Frequency	Y	N	Y	Y	Y	N
5. Valid No. of Cases	Y	Y	Y	Y	Y	N
6. No. of Missing Cases	N	N	N	Y	Y	Y
7. Mean	N	N	N	N	Y	N
8. Variance	N	N	N	N	Y	N
9. Standard Deviation	N	N	N	N	Y	N
10. Range	N	N	N	N	Y	N
11. Min./Max.	N	N	N	N	Y	Y
12. Range	N	N	N	N	Y	N
13. Median	N	N	N	N	Y	N
14. Mode	N	N	N	N	N	N
15. Percentile Limits	N	N	N	N	Y	N
16. Number of Cases Outside Intervals	Not Applicable	N	N 3/	Y	N 4/	Not Applicable
17. Z-Score	Y	N	N	N	N	N
C. EVALUATION MEASURES						
1. Minimum User Responses (Keystrokes)	7	10	11	10	11	10
2. Minutes Required to Complete Exercise	18	5/	5/	48	25	5/
3. Rating of:						
Documentation	A-S	A-S	U	A	U	U
Ease of Learning	S	S	U	A	A	A
Ease of Use	S	A	U	U	U	U
Printed Output						
-Presentation	A	A	A	S	A	U
-Statistics	U	U	U	A	S	U

1/ Although the function is documented, it does not work correctly.
2/ File information given only when "PRINT" invoked.
3/ Values outside allowable range are included in the last interval.
4/ Program will crash if values are not contained within maximum 40 groups.
5/ Not possible due to value code limitations.

Y = Yes N = No
U = Unsatisfactory A = Adequate S = Superior
NAL = No Apparent Limitation

and grouping are accomplished quickly and easily with the "SELECT" and "TRANS" procedures.

The printed output is well organized, but the automatic seven-digit formatting of frequency values is annoying. Value labels are not permitted (which adversely affects the quality of presentation) and missing values are not reported.

AIDA

Because this routine has good documentation, an interactive command summary, and brief commands, the beginner will likely find it very easy to learn. Selection of subsets and grouping are accomplished efficiently with the "GROUP," "CASES," and "TRANS" commands. However, the user will probably feel restricted by the limited allowable range of value codes (0-100), the truncation of positive non-integers (e.g., 3.13 is tabulated as 3), and the exclusion of negative values from the tabulation.

Value labels to identify discrete codes or intervals are permitted with the separate "LABELS" routine, thus greatly improving the readability of printed output. The straightforward presentation of output would be superior if file or job identification were also printed. Listed statistics are clearly inadequate, however, since no cumulative count or percent is calculated nor is the number of missing cases printed.

A-STAT

The difficult command syntax and poor documentation make any more than the most elementary frequency distribution analysis difficult to learn. Data values to be tabulated must range from zero to one hundred. Positive real values are truncated to integers for tabulation (as in AIDA) and negative values are simply set to zero (both are undocumented features). The variable selection procedure does not function. On the positive side, selection of subsets of cases and grouping of values are easily accomplished within the frequencies routine by variable transformations once the syntax is mastered. Weighted frequency distribution is an additional nice feature. Typing or syntax errors are very time-consuming, however, making this a difficult routine to use.

Unless the user relinquishes the capability to view output on the display screen, the readability of printed output is adversely affected by the "Hit

RETURN for more" prompt. Value labels are not permitted. Statistics presented are substandard, lacking the cumulative count and the number of missing cases.

MICROSTAT

The clear documentation and well-structured menus make this routine relatively easy to learn. Yet, analysis of survey data may well prove tedious because of the number of required user responses. If frequencies are desired for a variable with discrete values, for example, the user must laboriously enter all values to be tabulated. For grouped frequencies, the routine automatically calculates and tabulates intervals. However, the forced calculation of equal intervals for groups may not be adequate for the purpose at hand. This deficiency, plus the lack of an adequate re-coding capability (see section on data management) may thus leave the researcher with no usable frequency distribution capability. Specification of data subsets for analysis is possible only by case ranges, which requires a time-consuming and unreliable "SORT."

Although value labels are not permitted, the printed output is very well organized and formatted. In addition, the automatically generated histogram is a very nice feature.

NUMBER CRUNCHER

Since there are only three lines of documentation for this routine, the new user must learn the procedures by trial and error. The menus are, however, relatively straightforward and guide the learner through the routine.

As discussed above, a good frequency distribution routine should allow tabulation of both codes of discrete variables and intervals (groups) of continuous variables. NUMBER CRUNCHER unfortunately permits only tabulation of intervals. By itself, this restriction would not be so serious since the user could stipulate an interval of one for discrete codes. Regrettably, the routine merely prompts for the desired number of groups and automatically calculates up to 40 groups (the documentation incorrectly states "up to 10 groups") with equal intervals. For example, if the data values range from 100 to 200 and the user specifies 40 groups, intervals of 2.5 units are tabulated. Thus, tabulation of discrete codes with range exceeding 40 or tabulation of unequal intervals force the user to re-code the data with the data transformation module. Both time and information may be lost. Selection of data subsets for analysis is not feasible within the frequency routine or through the transformation routine.

The printed frequency labels are sequential group numbers assigned by the routine. Although the listed midpoint of the interval can be used to identify the group, the frequency labels are confusing.[1/] Value labels are the obvious solution. Simple job information would also aid in identifying the analysis. The statistics reported are by far the most comprehensive of the evaluated routines, and the automatic generation of a histogram is a very nice feature.

SPS

Documentation for this routine is difficult to locate in the manual (it is called "Histograms") and generally unhelpful. No sample prompts or sample output are given in the manual for this routine. The poor documentation and ambiguous explanation of input procedures make the "HIST" procedure somewhat complex to learn.

Both selection of data subsets and transformations are very time-consuming and tedious operations. Selection of subsets requires separate files to be written to disk, while transformations necessitate extensive read operations and user input in the data management module. Continuous values can be grouped within the frequencies routine. However, SPS imposes the same restrictions (equal interval widths and 40 maximum groups) as does NUMBER CRUNCHER. Researchers may find this too inflexible and be forced to carry out extensive re-coding. Frequency distributions of discrete variable values are attainable only for variables with a range of less than 40 units.

The printed output lacks any file or job identification and includes the annoying "<R> for continuation of table" every 12 lines. Value labels are not permitted. In addition, total valid cases, cumulative count, and cumulative frequency are not reported. Overall, the output lacks both readability and sufficient statistical measures.

[1/] NUMBER CRUNCHER is the only routine to list the midpoint of frequency distribution intervals.

2. Sample Output of Evaluation Routines

ABSTAT Frequency Distributions Output

ABSTAT 3.00
FILE: B:AB-EVAL3 REV# 0 PAGE 2

COMMAND: FREQ

VARIABLE: 4 REG

VALUE	FREQ	CUM FREQ	%	CUM %	Z SCORE
1.00000	103	103	51.5	51.5	-.968008
3.00000	97	200	48.5	100.0	1.02788
TOTAL	200	200	100.0	100.0	

VARIABLE: 5 STA

VALUE	FREQ	CUM FREQ	%	CUM %	Z SCORE
1.00000	32	32	16.0	16.0	-1.77777
2.00000	77	109	38.5	54.5	-.404974
3.00000	91	200	45.5	100.0	.967819
TOTAL	200	200	100.0	100.0	

AIDA Frequency Distributions Output

- - - - - - - - - - - - - - - - - - - -

```
COMMAND ? ONEWAY
  V=2,4,5
  OPT=NONE


V4 REGION

VALUE        FREQ.    % OF 200

  1 REG1      103        51.5
  3 REG3       97        48.5

V5 STRATA

VALUE        FREQ.    % OF 200

  1 STR1       32          16
  2 STR2       77        38.5
  3 STR3       91        45.5
```

- - - - - - - - - - - - - - - - - - - -

Note: Some variables were truncated to conform with AIDA data value restrictions.

A-STAT Frequency Distributions Output

```
??FREQ IN=EVALDAT3 (C    4 5)$


     A-STAT FREQ PROCEDURE.

INPUT FILE=EVALDAT3

DICTIONARY READ.  NV=17

END OF ADATA.EVALDAT3,S6,D2,VO

RECORDS READ= 200

VARIABLE NAME= REG

 CODE    FREQ        %       CUM %

 1        103      51.5      51.5
 3         97      48.5       100

TOTAL GOOD CASES= 200

MORE TABLES--HIT RETURN..?

VARIABLE NAME= STR

 CODE    FREQ        %       CUM %

 1         32        16        16
 2         77      38.5      54.5
 3         91      45.5       100

TOTAL GOOD CASES= 200
```

MICROSTAT Frequency Distributions Output

```
-------------------- FREQUENCY DISTRIBUTIONS --------------------
HEADER DATA FOR: B:EVAL3C     LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200    NUMBER OF VARIABLES: 17


     VARIABLE: 4. REGIO

              TEST ON EVAL3C VARIABLE #4 (REGION)

                                                 ....CUMULATIVE...
  ===== VALUE =====      FREQUENCY    PERCENT   FREQUENCY PERCENT
        1.00                  103      51.50       103      51.50
        3.00                   97      48.50       200     100.00

                         TOTAL  200    100.00


  ===== VALUE =====      FREQUENCY .............................
        1.00                  103 |==============================
        3.00                   97 |============================


-------------------- FREQUENCY DISTRIBUTIONS --------------------
HEADER DATA FOR: B:EVAL3C     LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200    NUMBER OF VARIABLES: 17


     VARIABLE: 5. STRAT

              TEST ON EVAL3C VARIABLE #5 (STRATA)

                                                 ....CUMULATIVE...
  ===== VALUE =====      FREQUENCY    PERCENT   FREQUENCY PERCENT
        1.00                   32      16.00        32      16.00
        2.00                   77      38.50       109      54.50
        3.00                   91      45.50       200     100.00
                         TOTAL  200    100.00

  ===== VALUE =====      FREQUENCY .............................
        1.00                   32 |==========
        2.00                   77 |========================
        3.00                   91 |==============================
```

Number Cruncher Frequency Distributions Output

---------------------------- Number Cruncher ----------------------------

Date: 08-23-1983
Time: 00:22:12
Data File Name: EVAL3
File Description: HONDURAS SSE FILE
Column number: 4 REGION

Frequency Table

Group	Freq	%Freq	Cum.Freq.	%Cum.Freq.	Midpoint	Histogram
1	103	51.5	103	51.5	15.00E-01	:*********************
2	0	0.0	103	51.5	25.00E-01	:
3	97	48.5	200	100.0	35.00E-01	:*******************

Histogram Interval Width			1.00E+00		
25th-%tile	1.00E+00	Median	1.00E+00	75th-%tile	3.00E+00
Minimum	1.00E+00	Maximum	3.00E+00	Range	2.00E+00
Mean	1.97E+00	Variance	1.00E+00	Std. Deviation	1.00E+00
Number of missing-values			0		

---------------------------- Number Cruncher ----------------------------

Date: 08-23-1983
Time: 00:24:15
Data File Name: EVAL3
File Description: HONDURAS SSE FILE
Column number: 5 STRATA

Frequency Table

Group	Freq	%Freq	Cum.Freq.	%Cum.Freq.	Midpoint	Histogram
1	32	16.0	32	16.0	15.00E-01	:********
2	77	38.5	109	54.5	25.00E-01	:******************
3	91	45.5	200	100.0	35.00E-01	:**********************

Histogram Interval Width			1.00E+00		
25th-%tile	2.00E+00	Median	2.00E+00	75th-%tile	3.00E+00
Minimum	1.00E+00	Maximum	3.00E+00	Range	2.00E+00
Mean	2.30E+00	Variance	5.31E-01	Std. Deviation	7.28E-01
Number of missing-values			0		

SPS Frequency Distributions Output

```
FREQUENCY DISTRIBUTION OF REG
----------------------------------------
MIN = 1    MAX = 3

INTERVAL              FREQUENCY RELATIVE %
----------------------------------------
1 -> 2.49                 103     .514
2.5 -> 3.99               97      .484

TOTAL MISSING = 0
% OF TOTAL CASES MISSING = 0


FREQUENCY DISTRIBUTION OF STR
----------------------------------------
MIN = 1    MAX = 3

INTERVAL              FREQUENCY RELATIVE %
----------------------------------------
1 -> 1.99                 32      .159
2 -> 2.99                 77      .385
3 -> 3.99                 91      .454

TOTAL MISSING = 0
% OF TOTAL CASES MISSING = 0

<R> FOR HISTOGRAM
```

F. Cross-tabulation

Cross-tabulation (or contingency tables) completes the triad of descriptive measures generally required of a statistics package. Rather than impose limits on acceptable values, a good cross-tabulation routine will accept all positive and negative real values. Memory limits may require restrictions on the maximum allowable rows or columns of the cross-tab table, but a reasonable minimum for research is 25 rows by 10 columns.

Since economic researchers often analyze continuous variables, a flexible and simple procedure should be included to allow grouping of continuous variables. As long as the procedure is relatively fast and easy to use, this could be handled either within the routine or in a separate transformation routine. Groups should be identified in printed output by value labels or at least described by the lower and upper limit of the interval.

Highly desirable optional features include: (1) an easy to use procedure to subset the data by case ranges or by the value of an index variable (i.e., a "conditional" value); (2) the capability to request cross-tabulation tables for several variables in a single command (e.g., variable one by variable two, variable one by variable three, etc.); and (3) the capability to generate one table for each value of a third variable (often called three-way cross-tabs).

Statistics required for cross-tabulation output are:

1. Cell/row/column counts
2. Cell/row/column percentages
3. Number of missing cases
4. Number of valid cases
5. Overall chi-square test statistic with degrees of freedom

A summary of cross-tabulation features is listed in Table 9.

1. Description of Evaluated Routines

ABSTAT

Documentation (both written and interactive), ease of learning, and ease of use are again very strong points of ABSTAT. The straightforward documentation and parsimonious commands make the routine as easy to learn as a menu-driven procedure. ABSTAT imposes no limitations on acceptable values for the routine. It does, however, limit the table size to

Table 9. Features of Cross-Tabulations / Contingency Tables Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	XTAB	TWOWAY	TABLES	CROSS-TAB	CROSS-TAB	CROSS-TAB
A. CAPABILITIES						
1. Select Subset of:						
Cases (Conditional)	Y	Y	Y	N	N	N
Cases (Range)	Y	Y	Y	Y 1/	N	N
2. Group Values:						
Within Routine	Y	N	Y	N	N	Y
With Transform	Y	Y	Y	N	Y	Y
3. Treatment of:						
Negative Values	Tabulated	Exclude	Not Allowed	Not Allowed	Not Allowed	Tabulated
Non-integer Values	Tabulated	Truncate	Truncate	Tabulated	Truncate	Tabulated
4. Default Max. Table Size (Row x Column)	NAL x 10 2/	12 x 12	15 x 15	20 x 5	32 x 32	10 x 10
5. Range of Allowable Values	NAL	1-12	1-15	All positive	1-60 3/	NAL
6. Weighted Cross-tabs	N	Y	Y	N	N	N
7. Output to Disk	Y	Y	N	N	N	N
8. Output Information						
File Identification	Y	Y 4/	Y	Y	Y	N
Job Identification	Y	N	N	Y	N	N
Variable Names	Y	Y	Y	Y	N	Y
Value Labels	N	Y	N	N	N	Y
B. STATISTICS						
1. Cell Count	Y	Y	Y	Y	Y	Y
2. Row/Column Count	Y	Y	Y	Y	Y	Y
3. Cell Percent (of total)	Y	N	Y	Y	Y	N
4. Cell Percent (of row/column)	Y	Y	Y	N	Y	Y
5. Row/Column Percent (of total)	Y	Y	Y	Y	Y	Y
6. Valid Number of Cases	Y	Y	Y	Y	Y	Y
7. Number of Missing Cases	Y	N	N	Y	N	Y
8. Expected Count	N	N	Y	Y	Y	N
9. Expected Percent	N	N	N	Y	N	N
10. No. Values Out of Bounds	Not Applicable	Y	N 5/	Y	N	N 6/
11. Chi-square (total or cell)	Y	Y	Y	Y	Y	Y
12. Degrees of Freedom	Y	Y	Y	Y	Y	Y
13. Probability of Chi-square	N	Y	Y	Y	Y	N

(continued)

Table 9. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	XTAB	TWOWAY	TABLES	CROSS-TAB	CROSS-TAB	CROSS-TAB
14. Cramer's V	N	N	Y	N	Y	Y
15. Contingency Coefficient	N	N	Y	N	Y	Y
16. Tschuprow's T	N	N	Y	N	Y	N
17. Lambda	N	N	Y	N	Y	N
18. Tau (B and/or C)	N	N	Y	N	Y	N
19. Gamma	N	N	Y	N	Y	N
20. Somer's D	N	N	Y	N	N	N
21. Kendall's S	N	N	Y	N	N	N
22. Phi Coefficient	N	N	N	N	Y	Y
23. Spearman's Rho	N	N	N	N	Y	N
24. Pearson's Prod. Mom. Corr.	N	N	N	N	Y	N
25. Kolmogorov-Smirnov Goodness of Fit	N	N	N	Y	N	N
C. PERFORMANCE MEASURES						
1. Minimum User Responses (Keystrokes)	9	10	28	16	14	24
2. Minutes Required to Complete Exercise	31	66	46	7/	42	73
3. Rating of:						
Documentation	A-S	A-S	A	A	A	U
Ease of Learning	S	A	U	U	U	U
Ease of Use	S	A	A	U	U	U
Printed Output						
-Presentation	A	A	A	A	U	U
-Statistics	A	U	U	U	U	U

1/ Although documented, this function does not work properly.
2/ 19 columns can be printed in a table with a 132-column printer.
3/ Upper limit is approximate.
4/ Information given only when "PRINT" invoked.
5/ Out of bounds values tabulated as value = 15.
6/ Included in Number of Missing Values.
7/ Not possible due to value code limitations and error in cross-tab routine.

Y = Yes N = No

U = Unsatisfactory A = Adequate S = Superior

NAL = No Apparent Limitation

the number of table columns which can fit across the printed output. Since this restriction is apparently not related to memory limitations, the correct treatment would be continuation of the table on a separate page or line. The user may specify that the routine tabulate all discrete values or automatically calculate groups of equal intervals. Transformations for groups and creation of subsets are also easily accomplished by the separate "TRANS" command. Three-way cross-tabulation tables are easily generated with the "SELECT" command.

Since value labels are not permitted, the meaning of groups created by transformations can easily be lost in printed output. Overall, the printed output is poorly formatted, very cluttered, and barely adequate.

AIDA

This routine imposes serious restrictions on the default maximum allowable table size (12 rows by 12 columns) and acceptable data values (only positive integers ranging from 1 to 12).[1/] Negative values are tabulated as missing values, and positive non-integers are truncated. These restrictions seriously limit the usefulness of the routine for research purposes.

Except for the value restrictions, the routine would be extremely easy to learn and use. Although the reader is apt to become confused by the use of "shorthand" terms such as "NC" (Number of Columns), "NR" (Number of Rows), and "DF" (Degrees of Freedom), the documentation is generally very good. Since grouping is easy with the separate "TRANS" command, some of the problems caused by the table and value restrictions are alleviated. Three-way cross-tabulations are easily obtained with the "GROUPS" option. In addition, multiple cross-tabulation tables can be requested in a single command.

A welcome feature is the separate "LABELS" routine which allows the user to input value labels to be printed in the cross-tabulation output. Unfortunately, the output is difficult to read due to poor formatting and an important statistic ("cell percent of total") is missing.

[1/] The document states that both defaults can be "moderately" increased by an experienced user with the separate "AIDA CHANGES" routine.

A-STAT

This routine presents another example of serious value restrictions which will often imply considerable re-coding in socioeconomic research. The only acceptable values are positive integers from 1 to 15, and the table size is restricted to 15 rows by 15 columns.

Many users will find the routine difficult to learn due to the writing style of documentation and the complexity of command syntax. Experienced users will be frustrated by the value restrictions mentioned above, although they will often be able to work around the restrictions through variable transformations. Subsets are easy to create, and re-coding intervals to discrete codes can be accomplished within the cross-tabulation command. Again, mistakes attributable to syntax or typing errors are very time-consuming. Multiple pair-wise tables can be requested with a single command, and "N-way" tables are feasible with multiple commands using the variable transformation procedure.

The output is unique (and welcome) among those evaluated for its border surrounding cells of the table. The statistics presented in printed output are more extensive than for most routines but do not include the number of missing cases.

MICROSTAT

A good transformation capability would allow this routine to supersede the restrictions on table size (20 rows by 5 columns), on data value range (positive values only), and on discrete values (no intervals allowed). Unfortunately, the transformations available are virtually useless for such a purpose. In addition, the user must laboriously enter all data values to be tabulated.

Although the printed output is quite presentable, the generation of one table for cell counts and another for all percentages may displease some researchers. Tables of expected counts and percentages are also listed. Cell percentages of row and column totals are not reported.

NUMBER CRUNCHER

Data restrictions make this routine difficult to learn and use. The only allowable data values are positive integers ranging from 1 to somewhat greater than 60 (the documentation and program give conflicting information on the maximum allowable value). The user may find that the program crashes if unacceptable data values are encountered. The transformation routine necessary to accommodate the data to the restrictions is somewhat awkward and slow. Analysis of a subset of cases is not feasible.

The output gives no identification of the tabulated variables and no labels of individual table rows and columns. It is particularly difficult to associate row and column numbers on the output with values in the data set. The user may be further confused by inconsistent use of the terms "rows" and "columns" (throughout the rest of the program a "row" is a case and a "column" is a variable). Although the statistics presented are more extensive than those of other packages, the number of cases with missing values is not included.

SPS

Although the maximum table size is restricted to 10 rows by 10 columns, there is no apparent limitation on the allowable data values for the routine. Re-coding to avoid the table size restriction is quite easy with a grouping procedure in the routine. Re-coding with the separate transformation routine is much less efficient. Unfortunately, if variables have discrete codes, the user must input all values to be tabulated. Analysis of subsets is only possible through creation of new disk files.

The documentation is wordy and very confusing, and the new user will find the routine difficult. Both beginning and experienced users will probably dislike the output, which combines poor formatting and a total absence of file or job identification. Printed statistics do not include the "cell percent of total." A very nice feature, however, is the option to print value labels for table rows and columns.

2. Sample Output of Evaluated Routines

ABSTAT Crosstabulation Output

```
ABSTAT 3.00
FILE: B:AB-EVAL3  REV# 0                                           PAGE  3

COMMAND: XTAB


            *** CROSS TABULATION ***

 5 STA                      4 REG      : DISCRETE

            1.00000
  DISCRETE        3.00000TOTAL
     1.00000    18     14     32
      % TOT   9.0%  7.0% 16.0%
      % ROW  56.2% 43.7%
      % COL  17.5% 14.4%

     2.00000    54     23     77
      % TOT  27.0% 11.5% 38.5%
      % ROW  70.1% 29.9%
      % COL  52.4% 23.7%

     3.00000    31     60     91
      % TOT  15.5% 30.0% 45.5%
      % ROW  34.1% 65.9%
      % COL  30.1% 61.9%

      TOTAL    103     97    200
             51.5% 48.5%100.0%

CHI SQUARE =       22.0621    DEGREES OF FREEDOM = 2
```

AIDA Crosstabulation Output

```
- - - - - - - - - - - - - - - - - - -

COMMAND ? TWOWAY
  V=4,5
  OPT=R%,C%,TESTS


TWOWAY TABLE    N= 200
ROW = 4 REGION
          COL = 5 STRATA
       1     2     3
     STR1  STR2  STR3
REG11   18    54    31 :   103
   C%  56.   70.   34. :   51.
   R%  17.   52.   30.
REG33   14    23    60 :    97
   C%  44.   30.   66. :   49.
   R%  14.   24.   62.
        32    77    91
       16.   39.   46.

CHI SQ.= 22.062 DF= 2    P= .0000

- - - - - - - - - - - - - - - - - - -
```

Note: Some variables were truncated to conform with AIDA data value restrictions.

A-STAT Crosstabulation Output

```
END OF ADATA.EVALDAT3,S6,D2,V0

RECORDS READ= 200


  TABLE: REG BY STR


CELL CONTENTS:
FREQ
ROW PERCENT
COLUMN PERCENT
EXPECTED VALUE
CELL CHI-SQUARE


HIT RETURN FOR MORE...?

         1          2          3      TOTALS
    I---------I---------I---------I
  1 I      18I      54I      31I     103
    I  17.48I  52.43I  30.1I     51.5
    I  56.25I  70.13I  34.07I
    I  16.48I  39.66I  46.86I
    I    .14I    5.19I    5.37I
    I---------I---------I---------I
  3 I      14I      23I      60I      97
    I  14.43I  23.71I  61.86I     48.5
    I  43.75I  29.87I  65.93I
    I  15.52I  37.34I  44.14I
    I    .15I    5.51I    5.71I
    I---------I---------I---------I
           32        77        91      200
           16      38.5      45.5
HIT RETURN FOR MORE...


CHI-SQUARE = 22.0621336
WITH 2 DEGREES OF FREEDUM
PROB OF CHI-SQUARE = 7.12102609E-07
CRAMER'S V = .332130499
CONTINGENCY COEFFICIENT = .315200212
T = 1.05028886
LAMBDA SYMMETRIC = .252427184
LAMBDA PREDICTING ROWS = .298969072
LAMBDA PREDICTING COLUMNS = .211009174
TAU-B = .254520316
TAU-C = .2831
GAMMA = .426548139
SOMER'S D PREDICTING ROWS = .228619882
SOMER'S D PREDICTING COLUMNS = .283355017
SOMER'S D SYMMETRIC = .253061589
KENDALL'S S = 2831
```

MICROSTAT Crosstabulation Output

------------------ CROSSTAB / CHI-SQUARE TESTS ------------------

HEADER DATA FOR: B:EVAL3 LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200 NUMBER OF VARIABLES: 19

TEST CROSSTABULATION OF REGION BY STRATA --HONDURAS DATA

ROW VARIABLE= REGIO COL. VARIABLE= STRAT

OBSERVED FREQUENCIES

	1	2	3	TOTAL
1	18	54	31	103
3	14	23	60	97
TOTAL	32	77	91	200

CHI-SQUARE = 22.062, D.F.= 2, PROB. = .0001

------------------ CROSSTAB / CHI-SQUARE TESTS ------------------

HEADER DATA FOR: B:EVAL3 LABEL: HONDURAS SSE DATA
NUMBER OF CASES: 200 NUMBER OF VARIABLES: 19

TEST CROSSTABULATION OF REGION BY STRATA --HONDURAS DATA

ROW VARIABLE= REGIO COL. VARIABLE= STRAT

OBSERVED PERCENTAGES

	1	2	3	TOTAL
1	9.00	27.00	15.50	51.50
3	7.00	11.50	30.00	48.50
TOTAL	16.00	38.50	45.50	100.00

CHI-SQUARE = 22.062, D.F.= 2, PROB. = .0001

Number Cruncher Crosstabulation Output

----------------------------- Number Cruncher (c) -----------------------------

Date: 10-29-1983
Time: 20:47:02
Data File Name: eval3
File Description: HONDURAS SSE FILE
Table Type : Raw Data

	Col (1)	Col (2)	Col (3)	Total
Row (1)	18.0	54.0	31.0	103.0
Row (2)	0.0	0.0	0.0	0.0
Row (3)	14.0	23.0	60.0	97.0
Total	32.0	77.0	91.0	200.0

Table Type : Row Percentages

	Col (1)	Col (2)	Col (3)	Total
Row (1)	17.5	52.4	30.1	100.0
Row (2)	0.0	0.0	0.0	0.0
Row (3)	14.4	23.7	61.9	100.0
Total	16.0	38.5	45.5	100.0

Table Type : Column Percentages

	Col (1)	Col (2)	Col (3)	Total
Row (1)	56.3	70.1	34.1	51.5
Row (2)	0.0	0.0	0.0	0.0
Row (3)	43.8	29.9	65.9	48.5
Total	100.0	100.0	100.0	100.0

Table Type : Table Percentages

	Col (1)	Col (2)	Col (3)	Total
Row (1)	9.0	27.0	15.5	51.5
Row (2)	0.0	0.0	0.0	0.0
Row (3)	7.0	11.5	30.0	48.5
Total	16.0	38.5	45.5	100.0

SPS Crosstabulation Output

```
CO  L.#1=> REG1
-----------------------------------
STR1 18:17.4%
-----------------------------------
STR2 54:52.4%
-----------------------------------
STR3 31:30%
-----------------------------------
COL. TOTALS=103:51.4%
<R> TO CONT.
COL.#2=> REG2
-----------------------------------
STR1 14:14.4%ROW TOTAL = 32:15.9%
-----------------------------------
STR2 23:23.7%ROW TOTAL = 77:38.5%
-----------------------------------
STR3 60:61.8%ROW TOTAL = 91:45.5%
-----------------------------------
COL. TOTALS=97:48.5%
<R> TO CONT.
   R X C CONTIGENCY STATISTICS


TOTAL CASES = 200
MISSING CASES = 0
CHI SQUARE = 22.0621336 :2 D.F.
PHI COEFFICIENT = .332130499
CRAMER'S V = .332130499
CONTINGENCY COEFFICIENT = .315200212
```

G. Analysis of Variance

Computational accuracy is a very important criterion for a researcher contemplating the use of a particular package and associated analysis of variance routine. Some of the many design and operational features which differentiate equally accurate Analysis of Variance (ANOVA) routines are described below. The precision evaluation (using testing procedures adapted from the evaluation of regression analysis routines) assessed the accuracy of one-way analysis of variance calculations. Since all packages accurately calculated the expected statistics, results are not summarized here.[1/] Two-way analysis of variance routines were not tested due to time constraints imposed by the widely varying ANOVA input requirements of the evaluated packages. Table 10 presents a summary of ANOVA features.

The analyses of variance procedures necessary for cross-sectional economic research are far less complex than those required by physical scientists for analysis of experiments. Social scientists cannot first fix certain control variables at various levels and subsequently measure the responses of an experimental unit. Instead, characteristics of the phenomena are described and associated with measurements on a variable of interest. In our simple one-way analysis of variance exercise, we tested whether profits of small firms vary significantly by the population size of the locality in which they operate. Population size is generally termed a "factor" (or "control" variable) while the specific values or codes of the factor are called "factor levels." Profit in this case is the "dependent" or "response" variable. Other social science disciplines define these terms differently, often calling the factor variable a "group" or "sample" variable.

A two-way analysis of variance includes the effect of a second factor. Following the above example, this might be an examination of firm profit variance with respect to both population size and geographical region. Since there may be an "interaction effect" between the two factors, the ANOVA routine should give the option to include this interaction as a source of variation in the dependent variable.

Cross-sectional survey data files are typically organized as cases (consisting of all measurements on a particular unit of analysis, such as one

---

[1/] The procedures followed in the evaluation of precision are documented in Annex 2.

Table 10. Features of Analysis of Variance Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	ANOV1 ANOV2	MEANS RBLOCK RMEANS	ANOVA	ANALYSIS OF VARIANCE	ONE-WAY TWO-WAY 3-WAY 4-WAY	ANOVA
A. CAPABILITIES						
1. Types of ANOVA:						
One-way	Y	Y	Y	Y	Y	Y
Two-way	Y	N	N	Y	Y	N
3-way, 4-way	N	N	N	N	Y	N
Randomized Block	N	Y 1/	N	Y	N	Y
Split-plot	N	N	Y	N	N	N
Repeated Measures	N	N	N	N	Y	N
General Linear Models	N	N	N	N	N	Y
2. Select Subset of:						
Cases (Conditional)	Y	Y	Y	N	N	N
Cases (Range)	Y	Y	Y	Y	N	N
3. Unbalanced Data Allowed:						
One-way	Y	Y	Y	Y	Y	Y
Two-way	N	NA	NA	N	Y	NA
4. Maximum Levels Per Factor	NAL	12	8	NAL	NAL	11
5. Allowable Value Range of Factors	NA	1-255	1-8	NA	NAL	NAL
6. Weighted ANOVA	N	Y	Y	N	N	N
7. Output to Disk	Y	Y	N	N	N	N
8. Output Information						
File Identification	Y	Y 2/	Y	Y	Y	N
Job Identification	Y	N	N	Y	N	N
Factor Level Labels	Y	N	N	N	Y	Y

(continued)

Table 10. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
ROUTINE NAME	ANOV1 ANOV2	MEANS RBLOCK RMEANS	ANOVA	ANALYSIS OF VARIANCE	ONE-WAY TWO-WAY 3-WAY 4-WAY	ANOVA
B. STATISTICS (1-way & 2-way)						
1. Factor Level Means	Y	Y	Y	Y	Y	Y
2. Factor Level Std. Dev.	Y	N 3/	N	N	N 4/	Y
3. No. Cases/Factor Level	Y	Y	Y	Y	Y	Y
4. Significance of F-ratio	N	Y	Y	Y	Y	N
5. Confidence Intervals	N	N	N	N	Y	N
6. Bartlett's Chi-square	N	N	N	N	N	Y
7. Eta-square	N	N	N	N	N	Y
8. Interaction Effect	N	NA	NA	Y	Y	NA
9. Covariate Effects	N	N	N	N	Y	N
10. Apriori/Aposteriori Contrasts	N	N	N	N	Y	Y
C. PERFORMANCE MEASURES						
1. Minimum User Responses (Keystrokes)	8	11	27	13	14	18
2. Minutes Required to Complete Exercise	14	7	12	22	29	22
3. Rating of:						
Documentation	U	U	U	U	S	A
Ease of Learning	U	S	U	U	A	A
Ease of Use	A	S	A	U	A	A
Printed Output						
-Presentation	A	A	U	A	S	A
-Statistics 5/	A	A	U	U	S	A-S

1/ RBLOCK computes a non-parametric one-way ANOVA based on a randomized complete block description. RMEANS computes a one-way non-parametric ANOVA.
2/ File information presented only when "PRINT" invoked.
3/ Factor level variance is listed.
4/ Factor level standard error is listed.
5/ Rating based on one-way ANOVA.

Y = Yes N = No
U = Unsatisfactory A = Adequate S = Superior
NAL = No Apparent Limit
NA = Not Applicable

firm) and variables (each one of which is the measurement of a particular characteristic, such as firm profits). Factors and dependent variables are usually represented in one of two ways for ANOVA routines. For one-way ANOVA, the first method (most appropriate for cross-sectional research) uses the data value of one variable in a case to represent a factor level, while the data value of a corresponding second variable represents the dependent variable value. In the second method (often found in statistics texts), each variable implicitly represents exactly one factor level, while the data values of any particular variable correspond to the dependent variable value for that factor level. Transformation of cross-sectional data will be necessary for ANOVA routines using this input procedure.

For either method, at least six factor levels would seem to be a reasonable minimum number of allowable factor levels. It should be possible to label factor levels in printed output. Other desirable features include tests for homogeneous variance and calculation of confidence intervals for factor level means.

A common restriction of two-way ANOVA routines is that the data be "balanced," meaning that each combination of factor levels have an equal number of observations. Although this restriction might be reasonable for some physical science research applications, it will usually make the routine unsuitable for cross-sectional research.

Because of differing terminology and the numerous variants of ANOVA experimental designs, documentation must clearly define the program's use of terms. Although a complete explanation is beyond the scope of necessary program documentation, the manual should also briefly discuss the meaning of calculated test statistics. Minimum statistics required for one-way ANOVA include factor level means, standard deviations, number of cases, and a standard analysis of variance table with F-statistic. Two-way ANOVA should include the above plus the interaction sum of squares (if requested) and respective F-statistic. It should be noted that the statistics produced by one-way ANOVA are very similar to those required for a BREAKDOWN procedure and could be used for that purpose.

1. Description of Evaluated Routines

ABSTAT

For the new user unfamiliar with analysis of variance terminology and procedures, the terse documentation for this routine will likely pose serious problems. The routine is in fact not difficult to operate, but certain rigidities limit its usefulness for cross-sectional research. First, there is no explicit factor variable in the routine. Instead, each factor level (called "sample" in ABSTAT) is represented by a distinct variable with the data values representing dependent variable measurements for that level. Although this "column-wise" treatment of ANOVA data is widely utilized in statistics textbooks, analysis of cross-sectional survey data requires tedious transformations to accomplish one-way ANOVA, while two-way ANOVA requires re-entry of data. Unbalanced data are allowed for one-way ANOVA. However, two-way ANOVA is made overly restrictive by the requirement of balanced data (each combination of factor levels must have the same number of dependent variable observations). There is no apparent limit on the number of factor levels allowable for either type of ANOVA.

Although the two-way ANOVA output does not include an interaction term, the organization of printed output and the statistics calculated are adequate for the most part. The "column-wise" treatment of data permits the explicit labeling of factor levels by their respective variable names.

AIDA

Four short paragraphs of documentation for the one-way ANOVA routine (AIDA has no two-way ANOVA) are not likely to be sufficient for the user unfamiliar with analysis of variance. The experienced researcher will likely find the routine very easy to learn and use once the "GROUPS" option (to define factor levels) is mastered. The user is prompted by the "GROUPS" option for the factor variable and the discrete values which define factor levels. Missing values and values not explicitly selected with "GROUPS" are ignored, and unbalanced data are allowed. Weighted ANOVA is also supported. Analysis of several dependent variables is easily accomplished with a single command. Analysis of cross-sectional survey data is much easier than with some other packages, since the representation of ANOVA data is the factor-dependent variable format described in the introductory portion of this section.

The printed output suffers from the ubiquitous formatting problems of AIDA. Factor level labels are not permitted. Although factor level standard deviations would be preferable to the listed variances, the statistics presented are adequate. Nice features include the significance level of the F-statistic and the Student's T-statistic (calculated if only two factor levels are selected).

A-STAT

The terminology used in this documentation is the most confusing of all the evaluated routines. The user must somehow learn that "one-between" indicates a simple one-way analysis of variance, while "one-between one-within" refers to a split-plot analysis. The commands may also be confusing to the new user, as the ANOVA is requested in the following form: "Dependent variable = Factor." Data for a one-way ANOVA can be arranged either "column-wise" (as in ABSTAT) or in a factor-dependent variable format (as in AIDA and SPS). Simple two-way ANOVA does not appear possible. Up to eight factor levels are allowed for one-way ANOVA. Values of the factor variable must be numbered consecutively from one to eight, which will often imply extensive data transformations for cross-sectional data. The routine has two very nice features: multiple analyses can be requested with one command and weighted ANOVA is possible.

The output is poorly organized, difficult to read, and poorly formatted. Factor level labels are not permitted. Although the significance level of the F-ratio is a welcome statistic, the omission of standard deviations of each factor level makes the overall statistical output inadequate.

MICROSTAT

In an otherwise well-written manual, the ANOVA documentation of this program is completely unhelpful. The new user may find this the most difficult MICROSTAT routine. Principally because factor levels are defined by consecutive ranges of cases, the one-way ANOVA is complex and time-consuming. For example, the user might define factor level one as cases 1-35, while factor level two might be cases 36-50. With a typical cross-sectional data set, this would require the user to sort by the (implicit) factor variable, to determine relevant case ranges by listing the data, and finally to run the ANOVA routine. This process is very slow. We also found the sorting routine to be unreliable (see the section on data management).

The two-way ANOVA routine is unlikely to be useful for cross-sectional research because of its restrictive data input requirements. For this routine, the levels of factor one must be represented by separate variables and the levels of factor two must be individual cases. Besides one-way and two-way ANOVA, randomized complete block ANOVA is permitted.

The printed output is very readable and well organized (as is virtually all MICROSTAT output). It is unfortunate that factor level labels are not permitted since the user might easily forget the definition of factor levels. Output includes the significance level of the F-ratio and (unique among the evaluated packages) the grand mean. The critical standard deviations of factor levels, however, are not listed.

NUMBER CRUNCHER

This routine includes the most extensive and well-written ANOVA documentation of any evaluated program. The new user is guided through the basic steps with an adequate explanation of terms and the meaning of test statistics. Although many users will not understand all the complex types of ANOVA possible with NUMBER CRUNCHER, the manual refers the reader to a suitable statistics text for guidance.

It is very unfortunate and confusing, however, that the various ANOVA routines are scattered without rhyme or reason throughout the manual. If the new user utilizes the one-way ANOVA routine, she/he finds the "column-wise" treatment of factor levels described in previous sections as inappropriate for cross-sectional data. One-way ANOVA is also possible in the two-way ANOVA routine with the data in the factor-dependent variable format. The two-way ANOVA procedure, however, requires that factor levels be numbered consecutively starting from one, which will often imply re-coding of cross-sectional data. An important feature not often encountered in the evaluated packages is that unbalanced data are allowed for either one-way or two-way ANOVA. In addition, an extremely flexible set of options is available to calculate interaction effects, random/fixed effects, factor contrasts, and covariate effects. The procedures are complicated, yet provide considerable analytical power for the advanced user. Three-way, four-way, and repeated measures ANOVA are also supported.

The output of both the one-way and two-way ANOVA routines is generally well set up and readable. Unfortunately, only in the one-way sub-program are variable names printed to identify factor levels. A particularly nice feature of one-way ANOVA output is the listing of factor level means by order of ascending magnitude (which allows quick comparisons). This routine also generates confidence intervals with simple graphs at a user-specified probability level. All required statistics are presented. The manual does warn that factor level standard errors are calculated from the model mean square error rather than directly from the factor level data.

SPS

Documentation for the one-way ANOVA routine is interspersed with that of the T-test procedure but generally is not difficult to follow. Printed examples will also help the new user understand the procedures. One sore point is consistent use of the confusing term "sort code" to signify selection of discrete factor level values.

Many users will be happy to discover not only a factor-dependent variable format but also optional definition within the ANOVA routine of value ranges to represent factor levels. For example, it is possible to analyze the variation of profits per unit of land by farm size without a variable transformation, even though farm size was entered as a continuous variable. Either fixed or random effects models can be analyzed, aposteriori comparisons of factor level pairs can be calculated, and hypotheses can be tested (for two factor levels) which assume either homogeneous or non-homogeneous variance.

For two-way ANOVA, the user must employ the General Linear Models routine. Although this sub-program imposes no apriori constraints on experimental design, neither does it provide a great deal of guidance for the user. Many researchers would reject two-way ANOVA with this routine as simply too much bother. Another experimental design available in SPS is randomized complete blocks ANOVA.

For one-way ANOVA, factor level labels are explicitly prompted and printed in output tables. No information is given which identifies either the file or the particular analysis at hand. Poor formatting is also a recurrent problem in SPS. All required statistics are presented, along with the useful Bartlett's Chi-square test for homogeneous variance among groups.

2. Sample Output of Evaluated Routines

ABSTAT ANOVA Output

```
ABSTAT 3.00
FILE: B:NEWEVAL3   REV# 4                                            PAGE  1

COMMAND: ANOV1

MISSING VALUE TREATMENT: VARWISE

                *** 1 WAY ANALYSIS OF VARIANCE ***

      SAMPLE                SIZE          MEAN        STD DEV
       3 STR1                 32       636.235        1250.44
       4 STR2                 77       1488.74        2772.55
       5 STR3                 91       4597.81        8744.98

                      DEGREES OF       SUM OF        MEAN OF
SOURCE OF VARIANCE     FREEDOM        SQUARES        SQUARES           F TEST
AMONG SAMPLES             2        5.76117E+008   2.88058E+008        7.55083
WITHIN REPLICATIONS     197        7.51540E+009   3.81492E+007
TOTAL                   199        8.09152E+009
```

AIDA ANOVA Output

```
- - - - - - - - - - - - - - - - - - - -

COMMAND ? MEANS
  GROUP VAR;VALUES= 5;1^3
  V=15

ANALYSIS OF V 15 CASHMARG

GR.     N         MEAN     VARIANCE

  1    32        63.25    15632.065
  2    77      148.403    76844.007
  3    91      459.275   764693.313
TOT   200      276.225   406561.955

SOURCE       SS          DF      MS          F
GRPS     5758692.21       22879346.11    7.548
ERR      75147136.8    197381457.547
TOT        80905829    199
                                        P= .001

- - - - - - - - - - - - - - - - - - - -
```

Note: Some variables were truncated to conform with AIDA data value restrictions.

A-STAT ANOVA Output

```
??ANOVA IN=EVALDAT3$


     A-STAT ANOVA PROCEDURE.

***WARNING: NO OPEN PARENS>NO MODS MADE.

INPUT FILE=EVALDAT3


DICTIONARY READ.  NV=17
IN FILE=EVALDAT3,S6,D2,VO

ENTER ANOVA REQUEST IN THE FORM:
DEP VAR LIST = <GROUP VAR>
      BY <WITHIN NAME(LEVELS) >
*?CASH=$TR$
ANOVA REQUEST:
CASH=STR$
END OF ADATA.EVALDAT3,S6,D2,VO

RECORDS READ= 200

GROUP FACTOR=>STR
ANALYSIS NUMBER=>1

      ANALYSIS OF VARIANCE TABLE
---------------------------------------
SOURCE  DF       SS      MS  F-TEST   PROB
---------------------------------------
BTWN/SS 199 8.0915269E+09 40660939.2
GRPS      2 576117735 288058867   7.551 0
WTHN GP 197 7.51540916E+09 38149285.1
---------------------------------------
TOTAL   199 8.0915269E+09 40660939.2

HIT 'RETURN' FOR MORE...?
GOOD N FOR GROUP 1 =32
GOOD N FOR GROUP 2 =77
GOOD N FOR GROUP 3 =91

HIT 'RETURN' FOR MORE...?
MEAN FOR GROUP 1 =636.235
MEAN FOR GROUP 2 =1488.73766
MEAN FOR GROUP 3 =4597.80516

HIT 'RETURN' FOR MORE...?
```

MICROSTAT ANOVA Output

```
---------------------- ANALYSIS OF VARIANCE ----------------------
HEADER DATA FOR: B:REGR3C     LABEL: SUBSET HONDURAS DATA
NUMBER OF CASES: 200    NUMBER OF VARIABLES: 5

                        ONE-WAY ANOVA

                  ANOVA CASHMARGIN BY STRATA

                  GROUP        MEAN       N
                    1       636.235      32
                    2      1488.738      77
                    3      4597.805      91

             GRAND MEAN    2766.963     200

 VARIABLE  5:  CASH

  SOURCE     SUM OF SQUARES  D.F.  MEAN SQUARE    F RATIO    PROB.
 BETWEEN      576117733.000    2  288058866.500     7.551    .0007
 WITHIN      7515409161.000  197   38149285.080
 TOTAL       8091526894.000  199
```

Number Cruncher ANOVA Output

---------------------------- Number Cruncher (c)----------------------------

Date: 09-13-1983
Time: 00:39:41
Data File Name: NEWVAL3
File Description: HONDURAS SSE FILE

Analysis of Variance Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	Probability
Among Groups	2	5.761178E+08	2.880589E+08	7.55	0.001
Within Groups	197	7.51541E+09	3.814929E+07		
Adj Total	199	8.091527E+09			

The columns used in this analysis are

Column 1 with the label : STR1
Column 3 with the label : STR2
Column 4 with the label : STR3

Cont. next page.

```
---------------------------- Number Cruncher (c)----------------------------
Date:               09-13-1983
Time:               00:40:06
Data File Name:     NEWVAL3
File Description:   HONDURAS SSE FILE
```

Table of Means & Errors

Column #	Mean	Std. Error	No. Observations
1	636.235	1091.863	32
3	1488.738	703.8788	77
4	4597.805	647.4743	91

```
---------------------------- Number Cruncher (c)----------------------------
Date:               09-13-1983
Time:               00:44:05
Data File Name:     NEWVAL3
File Description:   HONDURAS SSE FILE
```

99 % Confidence Interval Values

Column	Lower limit	Mean	Upper limit	Label
1	-2203.982	636.235	3476.452	
3	-342.2321	1488.738	3319.708	
4	2913.558	4597.805	6282.052	

Plot of Treatment Confidence Intervals

```
Col (  1)    --------------O--------------
Col (  3)               ----------O----------
Col (  4)                                   ---------O---------
```

SPS ANOVA Output

A N O V A   T A B L E

SOURCE	SS	D.F	MS
TREAT	576117738	2	288058869
ERROR	7.51540916E+09	197	38149285.1
TOTAL	8.0915269E+09	199	*******

F VALUE = 7.55 ;2,197 D.F.
ETA SQUARED = .0712001264

<R> TO CONT.

TREAT	N	MEAN	V	S.D.
1	32	636.23	1563603.7	1250.44
2	77	1488.74	7687037.92	2772.55
3	91	4597.81	76474695.2	8744.98

BARTLETT'S CHI SQUARE = 156.845 ;2 D.F.

<R> TO CONT.

H. Regression Analysis

The most critical requirement of a regression analysis routine for cross-sectional research is that the accuracy of parameter estimates approach that attainable with a reliable mainframe statistical package. The precision evaluation tested results of regression routines using specially generated "problematic" data sets. While the evaluation procedures were briefly discussed in the first chapter, a more complete discussion is presented in Annex 2. Using "modestly problematic" types of data likely to be encountered in most research efforts, the results indicate that the accuracy of all regression routines is adequate. A composite "Mean Accuracy Error" is presented in the following regression analysis summary table, but the reader should note that differences between packages are not considered significant.[1/] Nonetheless, some packages have value restrictions, programming errors, and/or a lack of diagnostic procedures which can affect the routine's usefulness for many applications. Further testing with data sets designed to break down regression routines demonstrated that ABSTAT and NUMBER CRUNCHER generated quite accurate results in the most stressful of circumstances. ABSTAT performed particularly well with all the test data. A-STAT and MICROSTAT, on the other hand, began to break down with the most problematic data sets and generated inaccurate results.[2/]

A second requirement is that the regression routine be reasonably fast and easy to use. A good package will also facilitate transformations (such as logs or reciprocals) and selection of subsets of cases. Missing values encountered in the data should cause the case to be omitted from analysis only if the missing data occur in the dependent variable or one of the

---

[1/] Using all the coefficients calculated from moderately ill-conditioned data sets, the Mean Accuracy Error is computed as follows:

$$\sum_{j=1}^{k} \sum_{i=0}^{3} \left| \frac{B_{ij} - \hat{B}_{ij}}{B_{ij}} \right| / 4k$$

Where B_{ij} (i = 0 to 3 for each data set j) is the theoretically exact coefficient and $\hat{B}_{ij}$ is the corresponding estimated coefficient. For any particular package, results were obtained for "k" data sets, so that "4k" is the total number of estimated coefficients.

[2/] AIDA and SPS were not tested with the most stressful data because of their input value restrictions.

independent variables in the current regression. It should be possible to save at least the regression residuals, if not the predicted values of the dependent variable. Variable names should be listed in printed output in order to identify dependent and independent variables.

Ordinary Least Squares multivariate regression is a minimum requirement, although many researchers may also require step-wise regression, weighted least squares, or time series capability. Few of the evaluated packages permit step-wise or weighted regression. Some have a "time series-like capability," including at least the ability to log variables, save regression residuals, generate plots, and calculate the Durbin-Watson statistic. Since all evaluated packages allow regression residuals to be saved to disk, multi-stage least squares regression is also feasible. Ten independent variables would seem to be a reasonable minimum requirement for most research. There should be no constraint on the number of cases other than that mandated by the overall package limitation.

Regression output should include at least the following:

1. Estimated coefficients
2. Standard errors of coefficients (or t-values)
3. Coefficient of determination (R^2)
4. Standard error of the regression
5. F-ratio of the regression
6. Number of valid cases
7. Number of missing cases

Additional features which would be useful include the previously mentioned weighted and step-wise regression, time series and forecasting capability, and a standardized residual plot.

Regression routine features are presented in Table 11.

1. Description of Evaluated Routines

ABSTAT

Although the documentation may seem rather terse, this regression routine is extremely easy to learn and to use. The interactive "HELP" procedure is always available in case of confusion. Selection of subsets and transformations are simple and fast. Although predicted values are not automatically generated, calculation through the transformation routine is

Table 11. Features of Regression Routines

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS 1/
ROUTINE NAME	REGR	MULT	REGRESSION	REGRESSION	REGRESSION	MULTI
A. CAPABILITIES						
1. Regression Routines:						
Multivariate	Y	Y	Y	Y	Y	Y
Weighted	N	Y	N	N	Y	N
Stepwise	N	N	N	Y	Y	N
Path Analysis	N	N	Y	N	N	N
Principal Components	N	N	N	N	Y	N
Robust	N	N	N	N	Y	N
Time Series-like Capability 2/	Y	Y	N	Y	Y	N
2. Select Subset of:						
Variables	Y	Y	Y	Y	Y	Y
Cases (Conditional)	Y	Y	Y	N	N	N
Cases (Range)	Y	Y	Y	Y	N	N
3. Maximum Allowable:						
Independent Variables	19	12	24	>10 3/	20	10
Cases	NAL	NAL	NAL	NAL	NAL	NAL
4. Save Residuals	Y	Y	Y	Y	Y	Y
5. Save Predicted Values	N	N	Y	N	Y	Y
6. Residual Plot	N 4/	N 4/	N	Y	N 4/	N 4/
7. Output Regression Report to Disk	Y	Y	N	N	N	N
8. Output Information						
File Identification	Y	N 5/	Y	Y	Y	Y
Job Identification	Y	N	N	Y	N	N
Variable Labels	Y	Y	Y	Y	N 6/	Y
B. STATISTICS						
1. Estimated Coefficients	Y	Y	Y	Y	Y	Y
2. Standard Errors of Coefficients	N	N	Y	Y	Y	Y
3. R-square	Y	Y	Y	Y	Y	Y
4. Standard Error of Estimate	Y	Y	Y	Y	N 7/	Y
5. F-ratio of Regression	Y	Y	Y	Y	Y	Y
6. Number of Valid Cases	Y	Y	Y	Y	N	N
7. Number of Missing Cases	N	N	N	Y	N	N
8. Standardized Coefficient	Y	Y	Y	N	Y	N
9. Significance of Coefficients	N	N	Y	Y	Y	N
10. Student's T-test	N	N 8/	N	Y	Y	Y
11. Adjusted R-square	N	N	Y	N	Y	Y
12. Mean of Dependent Variables	N	Y	N	Y	N	N

(continued)

Table 11. (continued)

PACKAGE NAME	ABSTAT	AIDA	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS 1/
ROUTINE NAME	REGR	MULT	REGRESSION	REGRESSION	REGRESSION	MULTI
13. Means, Std. Dev. of All Variables	N	Y	N	Y	N	N
14. Significance of F-ratio	N	Y	Y	Y	Y	N
15. Multiple R	Y	Y	Y	Y	N	N
16. ANOVA Table	Y	N	Y	Y	Y	Y
17. Var.-covar. Matrix	N	N	N	N	N	N
18. Correlation Matrix	Y 9/	Y 10/	Y	N 4/	N 4/	N 4/
19. Confidence Intervals of:						
Coefficients	N	N	N	N	N	N
Predicted Variable	N	N	N	N	N	N
20. Durbin-Watson Statistic	Y	N	Y	Y	Y	Y
21. Von Neumann Ratio	N	N	N	N	N	N
C. EVALUATION MEASURES						
1. Minimum User Responses (Keystrokes)	13	14	34	13	18	13
2. Minutes Required to Complete Exercise	16	9	40	30	30	40
3. Mean Accuracy Error	2.8784E-3	2.8633E-2	2.9583E-3	3.0158E-3	2.5400E-3	6.3332E-3
4. Rating of:						
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
-Statistics	U	U	U	S	U	U

1/ Refers only to the Multiple Linear Regression routine except for the Mean Accuracy Error, which refers to the General Linear Models routine.
2/ Lag transformation, plots, capability to save residuals, and Durbin-Watson statistic considered minimal necessary. Only MICROSTAT has an explicit time series routine.
3/ Maximum independent variables are not documented.
4/ Function can be accomplished with a separate routine.
5/ Information provided only when "PRINT" invoked.
6/ Variable name listed only for dependent variable.
7/ Presents the Mean Square Error of the regression.
8/ Presents the F-test of coefficients.
9/ The simple correlation of the independent variables with the dependent variable is reported.
10/ The correlation matrix of independent variables is reported.

Y = Yes N = No

U = Unsatisfactory A = Adequate S = Superior

NAL = No Apparent Limitation

very fast. Simple time series-like analysis and forecasting can also be carried out with this routine.

The printed regression report is very well organized and formatted, with file and job identification plus names for all variables. However, the statistics do not include the standard errors of the coefficients. This omission (from an economist's viewpoint) will make the routine unusable for many research applications. The number of cases with missing values is also not reported.[1/]

AIDA

Although generally informative, the documentation tends to confuse the reader with "shorthand" terms such as "IV," "DV," and "PART." The routine, however, is very easy to learn and use. A few simple commands can produce (in a remarkably short time) regression reports for several different dependent variables. The powerful and easy to use data management procedures permit rapid calculation of transformations or selection of subsets. Predicted values (not calculated in the regression routine) can be quickly generated from the actual values and residuals. It should be noted that the routine allows weighted regression and could also be used for simple time series-like analysis.

The printed output is very poorly organized and presented. File identification is listed only when the printer is toggled on, no job identification is printed, and numbers are often run together. On the other hand, the output presents some very useful statistics not offered by most other packages, such as the correlation matrix of independent variables, the means and variances of all variables, the inverse matrix of independent variables, and the significance level of the equation F-ratio. The number of cases with missing values, however, is not reported. It should also be noted at this point that the data value restrictions of the package (for practical purposes, values are limited to four digits) may severely limit the usefulness of this routine for many research applications.

[1/] The recently released version 3.03 reports standard errors and t-tests of regression coefficients (see Annex 4 for publicized changes in new versions of the evaluated programs).

A-STAT

This unique regression procedure requires the user to generate a descriptive statistics file and a correlation matrix file with separate routines before entering the regression module. The documentation offers little assistance and may further confuse the user with its poor writing style. The two-step procedure, the demanding typing and syntax requirements, and the slow execution make the routine very difficult to learn and to use.

Although the output is generally well presented, many users will find confusing the terminology referring to BETA and B-WEIGHT regression coefficients. Poor formatting is a problem (particularly in the ANOVA table) as numbers often run together. An important statistic listed is the value of the determinant, which gives the user a quick check for the effect of multicollinearity. The significance level of regression coefficients is reported, but the number of cases with missing values is not printed.

MICROSTAT

New users will find this menu-driven routine quite easy to follow. As experience is gained, however, the user may become frustrated with the constant attention required for the numerous obligatory menu responses. The lack of capability for recoding variables, the cumbersome process of subset selection, and the poor error handling also detract from the routine's usefulness. The missing value treatment is particularly unfortunate, in that cases with missing values for any variable in the data file (not just any variable in the regression) are omitted. MICROSTAT includes a step-wise regression procedure and (alone among the packages evaluated) has explicit routines for time series regression, moving averages, and data smoothing.

The output of the routine is superb, equal in quality to that of virtually any mainframe package. Except for the adjusted R-square, every statistic commonly required for regression analysis is presented. It is difficult to imagine how such an output could be improved, which makes the other recoding and subset selection limitations of the routine all the more unfortunate.

NUMBER CRUNCHER

The documentation for regression analysis is voluminous and does a better job than most packages of explaining input procedures and output statistics to the new user. Unfortunately, documentation for the different variants of regression analysis is scattered throughout the manual.

The new user will find the menu options fairly easy to follow, but (as in all other NUMBER CRUNCHER routines) may find annoying the requirement that variables be selected by a sequential variable number rather than by a label. The experienced user may regard the large number of obligatory user responses as tedious. Transformations can be accomplished without great difficulty. Subset selection, however, is not feasible. NUMBER CRUNCHER permits the most extensive set of regression procedures available in the evaluated packages. Weighted regression, step-wise regression, robust regression, principal components analysis, and simple time series-like analysis are all possible with the package.

The output is generally well organized and formatted. Quizically, however, the statistics presented in screen output differ rather significantly from those listed in printed output. For example, standard errors of coefficients are not displayed but are printed. In neither instance are independent variables identified by variable names. The number of cases with valid values and with missing values is also not reported.

SPS

The manual recommends that the ordinary regression routine (MULTI) be used for "initial data snooping" and that more detailed analysis be carried out with the General Linear Models routine. Although the latter routine was utilized for the precision analysis evaluation, the present general evaluation concerns only the MULTI routine.

The documentation is adequate, and the new user should have little difficulty learning the routine. Transformations are time-consuming and cumbersome, requiring at least one additional read operation from the disk file, while selection of subsets requires a new data file to be written to disk. Of concern to some users will be the maximum of ten independent variables allowable in the regression. The output is poorly formatted and organized, with statistics running together and the " R to cont." prompt spoiling the presentation. Although the number of missing cases is not listed, the statistics are otherwise adequate.

2. Sample Output of Evaluated Routines

ABSTAT Regression Output

```
ABSTAT 3.00
FILE: B:NEWEVAL3  REV# 1                                                     PAGE  2

COMMAND: REGR

                    *** MULTIPLE LINEAR REGRESSION ***

DEPENDENT VARIABLE:  4 CASH                 200 VALID CASES

COEFF OF DETERMINATION:    .635513    ESTIMATED CONSTANT TERM:      -1054.30
MULTIPLE CORR COEFF:       .797190    STANDARD ERROR OF ESTIMATE:    3879.08

ANALYSIS OF VARIANCE FOR THE REGRESSION:
                      DEGREES OF     SUM OF      MEAN OF
SOURCE OF VARIANCE     FREEDOM      SQUARES      SQUARES       F TEST
      REGRESSION          3       5.142E+009   1.714E+009     113.914
      RESIDUALS         196       2.949E+009   1.504E+007
      TOTAL             199       8.091E+009
                                            CORRELATION
               REGRESSION   STANDARDIZED        WITH
VARIABLE       COEFFICIENT   COEFFICIENT     DEPENDENT
 1 STA            927.161        .105916       .254005
 2 ASS            .343871        .706982       .783892
 3 WRK            318.067      9.711E-002      .560269
```

AIDA Regression Output*

```
COMMAND ? MULTI
  V=15,5^7
  OPT=MATRIX


CORR. MATRIX

  5  .253952169
  6  .783804937  .167331265
  7  .560331949  .305804841  .609324756
VARIANCES
    406561.955  .530628142  17188.2211  3.79085428
MEANS
    276.225  2.295  31.1  1.91
VARIABLES
     15CASHMARG    5STRATA      6ASSETS      7WORKERS

INVERSE I.V. MATRIX

5  1.1038636
6  .0333643328  1.59153313
7  -.357896546  -.97996351  1.70656252

REGRESSION ON V15 CASHMARG

    I.V.       BETA     B   PART     F

  5 STRATA       .10692.731    .01  5.466
  6 ASSETS       .707 3.438  .314168.752
  7 WORKERS      .09731.849 6E-03  2.979
    CONSTANT          -104.333

MULT. R =    .797  R SQ=   .635
S.E. EST=387.942  F   =113.861
      N = 200      DF=   3/196  P= <.0001
```

*Note: Some variables were truncated to conform with AIDA data value restrictions.

A-STAT Regression Output

```
??REGRESSION IN=EVALDAT3 COR=EVALDAT3 DES=EVALDAT3$



A-STAT REGRESSION PROCEDURE.


INPUT FILE=EVALDAT3


INPUT MATRIX= EVALDAT3


DES FILE= EVALDAT3

ADES.EVALDAT3,S6,D2,V0 READ, NV=17


VAR LABELS


TYP
STR
ASST
WRKS
CASH


DICTIONARY READ.  NV=5


INPUT REGRESSION EQUATIONS

IN THE FORM:

VNAME = VNAME + VNAME + VNAME..
OR OUT=FNAME FOR RESIDUALS

*?CASH=STR+ASST+WRKS

DETERMINANT=.569603336
```

Cont. next page.

```
***NOTE: STATISTICS BASED ON N FOR EACH VAR

DEPENDENT VARIABLE=>CASH
MINIMUM N=200

               BETA    SIMPLE R
               ----    ------ -
 STR           .106      .254
 ASST          .707      .784
 WRKS          .097       .56

HIT 'RETURN' FOR MORE...

DEPENDENT VARIABLE=>CASH

               B-WT     STD-ERR    PROB
               ----     -------    ----

 STR        927.161  396.609        .02
 ASST          .344     .026 0
 WRKS       318.067   190.41       .096
INTERCEPT -1054.301

HIT 'RETURN' FOR MORE...

MULT. R=    .797190423
R-SQUARED= .635512571
S.E. OF EST=4003.57168
ADJUSTED R-SQR=.605798922

HIT RETURN TO CONTINUE..?

      ANALYSIS OF VARIANCE TABLE
---------------------------------------
SOURCE DF   SS         MS     F-TEST  PROB
---------------------------------------
REG  15  5.14226708E+09 342817805  21.388 0
RES 184  2.94925985E+09 16028586.1
---------------------------------------
TOT 199  8.09152693E+09

HIT RETURN TO CONTINUE..?
```

MICROSTAT Regression Output

----------------------- REGRESSION ANALYSIS -----------------------

HEADER DATA FOR: B:REGR3C   LABEL: SUBSET HONDURAS DATA

NUMBER OF CASES: 200   NUMBER OF VARIABLES: 5

REGRESSION CASHMARGIN ON STRATA, ASSETS AND WORKERS

INDEX	NAME	MEAN	STD.DEV.
1	TYPE	7.305	10.951
2	STRAT	2.295	.728
3	ASSET	3,157.926	13,109.946
4	WORKR	1.910	1.947
DEP. VAR.:	CASH	2,766.963	6,376.593

DEPENDENT VARIABLE: CASH

VAR.	REGRESSION COEFFICIENT	STD. ERROR	T(DF=196)	PROB.	PARTIAL r↑2
STRAT	927.1611	396.6090	2.338	.0204	.0271
ASSET	.3439	.0265	12.996	.0000	.4629
WORKR	318.0674	184.4891	1.724	.0863	.0149
CONSTANT	-1054.3013				

STD. ERROR OF EST. = 3879.078

R SQUARED = .636

MULTIPLE R = .797

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F RATIO	PROB.
REGRESSION	5142267060.000	3	1714089020.000	113.914	.0001
RESIDUAL	2949259834.000	196	15047244.050		
TOTAL	8091526894.000	199			

Number Cruncher Regression Output

------------------------------ Number Cruncher ------------------------------

Date: 10-29-1983
Time: 20:38:38
Data File Name: b:eval3
File Description: HONDURAS SSE FILE

Multiple Regression Report

Dependent Variable : Col (15) CASHMARG

Analysis of Variance Table

		----------Sequential----------			----------Last----------		Simple
Source	df	Sum-Sqr	R2	F-Ratio	Sum-Sqr	F-Ratio	R2
Mean	1	1.531217E+09					
Col (5)	1	5.220542E+08	6.45	34.69	8.223228E+07	5.46	6.45
Col (6)	1	4.575488E+09	63.00	304.07	2.541322E+09	168.89	61.45
Col (7)	1	4.472535E+07	63.55	2.97	4.472535E+07	2.97	31.39
Model	3	5.142267E+09	63.55	113.91			
Error	196	2.94926E+09					
Total	199	8.091527E+09					

Mean Square Regression 1.714089E+09

Mean Square Error 1.504724E+07

Parameter Estimation

Variable	Coefficient	Std. Coefficient	Std. Error
Col (0)	-1054.301	0	0
Col (5)	927.1611	.105916	396.609
Col (6)	.3438714	.7069818	.0264603
Col (7)	318.0675	9.711786E-02	184.4892

SPS Regression Output

```
*F ILE=STRIPDAT3     Y=CASH

COEFFICIENTS (196 D.F. FOR T TESTS)
------------------------------------------
B0=-1054.302S.E.=914.582      T=-1.16
B1=927.161   S.E.=396.608     T=2.33
B2=.343      S.E.=.026        T=12.99
B3=318.067   S.E.=184.489     T=1.72

STANDARD ERROR = 3879.07773

B1=STR           B2=ASST
B3=WRKS
<R> TO CONT.
                A N O V A

SOURCE    SS          D.F.     MS
-------------------------------------------
REG.    5.14226706E+093    1.71408902E+09
ERROR   2.94925983E+09196 15047244.1
-------------------------------------------
TOTAL   8.09152689E+09199
-------------------------------------------
F = 113.913818
R SQUARE = .635512571
ADJ. R SQUARE = .631812191
DURBIN WATSON = 2.12725228
<R> TO CONT.
```

IV. CONCLUSIONS

Using data typical of that analyzed in socioeconomic research in developing countries, selected routines of six microcomputer statistical packages have been evaluated. To establish some benchmark for comparison of routines, desirable functions and a set of minimum statistical output have been specified. Programs were then assigned a subjective classification based upon these criteria. It is hoped that program features and assessment criteria discussed here can assist other researchers in the evaluation of microcomputer statistical software. We emphasize to the reader that the criteria and our assessment of program features are based on research and programming experience which may not reflect a specific user's statistical processing requirements. Some researchers will make much more demanding requirements of statistical packages, but it will be difficult for the non-programmer to accomplish high-caliber research with routines that do not at least produce the minimum statistical output specified in this report.

Weber et al. (1983) have described how microcomputer use in developing country research situations offers both advantages and disadvantages compared with utilization of alternative equipment such as programmable calculators or mainframe computer installations.[1/] For statistical software, microcomputer programs also seem to offer some advantages to the researcher in a developing country. First, a microcomputer and an adequate statistical package can provide the capability to carry out more timely data analysis than possible at a distant inaccessible mainframe installation. The purchase price of the evaluated packages is low compared to typical statistical processing costs incurred at a mainframe installation. However, actual total costs for microcomputer statistical analysis are a complex function of software, hardware customs duties, maintenance, repair, and training costs; all of which should be taken into account to evaluate relative cost-effectiveness. Less tangible (but sometimes as important) may be the benefits accrued from a closer interaction of the researcher with the data.

[1/] Michael T. Weber, James Pease, Warren Vincent, Eric Crawford, and Tom Stilwell, "Microcomputers and Programmable Calculators for Agricultural Research in Developing Countries," International Development Working Paper No. 5, Michigan State University, 1983.

Some observers consider the researcher-microcomputer interaction in microcomputer statistical analysis to be conducive to better research.

Most of the statistical programs evaluated here suffer from common shortcomings which users accustomed to sophisticated mainframe packages will likely find somewhat frustrating:

1. Data file size restrictions. This may be a limit on maximum allowable data points, maximum number of variables, or maximum number of cases in a file. Data files must always be small enough to fit on one disk, which (all else equal) makes a large disk capacity advantageous. The restrictions are made less rigid in some packages by file management functions which allow specified variables from separate files to be either input to computer memory for processing or to be transferred to a new disk file. The ABSTAT "PULL" function is a good example of how data file size restrictions can be mitigated. The largest data set used in the evaluation (12 variables and 602 cases) was split into two disk files for both ABSTAT and SPS. All other packages could handle the data in one disk file.[1/]

2. Missing values. Researchers need the ability to move particular variable codes into and out of the "missing" category, automatic exclusion of missing values for specified variables in all transformation and statistical routines, and reporting of the number of excluded missing cases. AIDA is somewhat better than other programs for changing the codes which routines regard as missing. Only SPS does not allow a specific "missing value" code which is automatically excluded from the operation of all routines, while ABSTAT and MICROSTAT do a good job of consistently reporting the number of missing cases. ABSTAT also allows a flexible set of user options for treatment of missing values in its routines.

3. Error handling. Errors caused by faulty program operation or by unacceptable data should be "trapped" within the program. The ultimate in poor performance is an error message generated by the program language interpreter/compiler or by the computer

---

[1/] Default restrictions were modified in some packages according to directions in the program manuals.

operating system. Messages should convey enough unambiguous information so that the user can determine the cause of errors. A reference table in the manual should present a more detailed explanation of such error messages. When an error is encountered, program operation should return to the beginning of the current routine without loss of data.

ABSTAT has no equal in error handling. As mentioned earlier, the program seems virtually impossible to "crash." Other packages were adequate to poor, although to be fair it must be mentioned that error handling in some cases may be more limited by programming language of the package than by any oversights by the programmers.

4. Data entry. Since mainframe statistical programs usually do not allow interactive data entry, microcomputer packages generally offer more flexible and powerful data entry functions.[1/] Microcomputer statistical programs in turn have limited capacity when compared with dedicated data entry or data base management software. Such programs are particularly useful for the design of interactive "screen formats," the comparison of entered values against a table of user-specified acceptable values for each variable (range checking), and the verification of accurate keypunching by re-entry of data values. No evaluated statistical program allows screen design for data entry or range checking of entered values. AIDA is the only package which supports data verification. If all data are entered by a person familiar with the data, the simple data entry functions available in the evaluated statistical software may be sufficient. If a non-researcher enters the data, it may be worthwhile to consider the purchase of a data entry or data base management program (with due care for the ease of transferring data to the statistics program).

5. Output. Some of the evaluated programs generate poorly organized and poorly formatted output. Printed output with numbers run together or with badly aligned decimal points is difficult to

[1/] Most mainframe installations provide some type of separate data entry software for input of data to the statistical package.

read and requires re-typing for research reports. The user should also have the option of routing output to a disk file where it (ideally) could be incorporated directly into word processing text.

Most microcomputer programs lack adequate identification of output. The file name, a user-specified file description, and the date of analysis are critical requirements for later output identification. Also necessary is optional "job identification" information (like an SPSS RUN NAME identification) by which the researcher is reminded of the particular characteristics of previous analyses. All variables in printed output should be identified by variable names or labels--variable numbers are simply inadequate. For some routines such as frequency distributions or cross-tabulations, value labels are necessary which permit extended identification of codes used in the variables. Finally, no package did an adequate job of listing variable transformations carried out prior to statistical operations. Since it is much too difficult to remember specific variable transformations, printing of transformation formulas is the obvious solution.

No package satisfies all the above criteria for output. MICROSTAT output is very well organized and readable, but print formatting errors may cause the program to crash. ABSTAT and NUMBER CRUNCHER also generate well-formatted and identified output. Most programs do a fair job of identifying variables with at least a four-character variable name. Only AIDA and SPS allow value labels for cross-tabulation output, and only AIDA permits such labels for frequency distributions.

The adequacy of printed statistics may also be a problem for some research applications. Although the minimum statistical output requirements proposed in the routine comparisons of Chapter III are not particularly onerous, no package satisfied the criteria for all routines. Two packages (AIDA and SPS) generated minimum statistics for only one routine, while A-STAT did not satisfy the requirements for any routine. The most common deficiency is that statistical routines of most of the evaluated programs did not report the number of cases with missing values. If this shortcoming is disregarded, NUMBER CRUNCHER satisfies all remaining requirements, while A-STAT satisfies the criteria for three of the five evaluated statistical routines. Table 12 indicates program fulfillment of statistical output criteria with and without the missing cases statistic.

Adequate printed statistics would be considered by some researchers as the "bottom line" for evaluation of program performance. Other important dimensions which affect program adequacy include the organization and presentation of output, program documentation, error handling, and overall ease of use. Table 13 presents our subjective ranking of these attributes.

Some readers may be frustrated that they can find no indication in this evaluation of a recommendation for one statistical package over another. We have deliberately avoided such a prescription for two reasons: (1) choice of the "right" statistical package depends on the constraints and data processing requirements of each researcher, and (2) software development increases at such a pace that our recommendation would be outdated before you read this evaluation. For example, software distributors of at least three of the programs evaluated here (ABSTAT, A-STAT, and MICROSTAT) have introduced new versions of their packages which appear to be significantly different from the versions evaluated for this report.

What we hope to promote is that researchers utilize the systematic evaluation methods described in this report to aid their selection of statistical software. Any such selection process should begin by identifying the human and physical constraints associated with the environment in which the software is to be used. Next, minimum performance criteria should be established for the selected software according to the data analysis tasks to be performed. The following questions may help to organize the process:

1. What are the constraints imposed by existing hardware?
   a. Computer brand and memory capacity
   b. Disk drive capacity
   c. Availability of printer

2. What are the constraints imposed by existing software?
   a. Type of operating system
   b. Availability of language interpreters/compilers
   c. Need to interface with other programs (e.g., mainframe programs, word processing programs, spreadsheets, data base management programs, or graphics programs)

Table 12. Minimum Statistics Printed By Evaluated Routines*

Program	ABSTAT		AIDA		A-STAT		MICROSTAT		NUMBER CRUNCHER		SPS	
	Output Includes:		Output Includes:		Output Includes:		Output Includes:		Output Includes:		Output Includes:	
Routine	Number of Missing Values	All Other Minimum Statistics	Number of Missing Values	All Other Minimum Statistics	Number of Missing Values	All Other Minimum Statistics	Number of Missing Values	All Other Minimum Statistics	Number of Missing Values	All Other Minimum Statistics	Number of Missing Values	All Other Minimum Statistics
Descriptive Statistics	Y	Y	N	Y	N	Y	N	Y	N	Y	N	N
Frequency Distributions	N	Y	N	N	N	N	Y	Y	Y	Y	Y	N
Cross-tabulations	Y	Y	N	N	N	Y	Y	N	N	Y	Y	N
Analysis of Variance (One-way)	NA	Y	NA	Y	NA	N	NA	N	NA	Y	NA	Y
Regression Analysis	N	N	N	Y	N	Y	Y	Y	N	N	N	Y

* The "minimum statistics" refer to a set of desired statistics established in this evaluation for each routine. In some cases, nearly equivalent statistics are accepted (e.g., variances for standard deviations).

Y = Yes
N = No
NA = Not Applicable

Table 13. Relative Ranking of Statistical Programs*

	ABSTAT	AIDA	A-STAT	MICRO-STAT	NUMBER CRUNCHER	SPS
Documentation	1	1	3	2	2	3
Error Handling	1	2	3	3	3	2
Ease of Use	1	1	4	2	3	4
Presentation of Output	1	2	3	1	1	3

* Each program is ranked relative to all other evaluated packages. Program dimensions with lower numbers are considered superior to those with higher numbers. If two programs receive the same ranking, our subjective evaluation does not distinguish one as better than the other. Taking the first program dimension as an example, the documentation of ABSTAT and AIDA is roughly equal and is better than the documentation of MICROSTAT and NUMBER CRUNCHER.

3. What are the human constraints likely to be faced?
   a. Language skills (if native language of user is not English)
   b. Previous computer experience or knowledge of computer programming
   c. User knowledge of statistics
   d. Specialization of data processing tasks (particularly implications for data entry if researcher does not input the data)

4. What are the characteristics of the data to be analyzed (now and in the future)?
   a. Type of variables (character or numeric)
   b. Number of variables and cases

5. What is the minimum statistical analysis required?
   a. Essential data management capabilities (transformations, recoding, sorting, etc.)
   b. Essential statistical routines
   c. Essential statistics required in each routine
   d. Minimum accuracy required for routines such as regression

Having answered these questions, the reader could identify the package from this evaluation which satisfies the most critical needs. Obviously, trade-offs will probably be necessary. A simple illustration of this process would be an Apple II+ user who wants a descriptive statistics routine providing the mean, mode, and standard deviation as well as the capability of sending output to disk file for subsequent incorporation into word processing files. A review of Table 7 reveals that SPS is the only Apple-based program reviewed which provides the mode, while AIDA is the only Apple-based program evaluated which can route output to disk file. A trade-off is necessary if the user must buy an Apple-based program. ABSTAT performs both tasks, but it requires (since ABSTAT is a CP/M-based program) purchase of a CP/M memory card for the Apple. If the descriptive statistics and output storage requirements are essential, purchase of the memory card and ABSTAT may be the best choice.

The six evaluated programs are only a small subset of software packages with statistical capability. Having answered the questions above, the reader should be able to select several promising programs from a general directory of statistical software.[1/] Inquiries could then be made asking software distributors very specific questions about the capacity of their program to satisfy carefully defined requirements.

[1/] Such as Kelly et al., "An Annotated Directory of Statistical and Related Microcomputer Software for Socioeconomic Development," MSU International Development Working Paper No. 12, 1983.

It is somewhat more difficult to obtain information on the ease of use and computational accuracy of a statistical package. We recommend that the program manual of promising programs be purchased to evaluate whether the manual is written at a level commensurate with the skills of computer operators. At a minimum, the prospective purchaser should request samples of documentation and output from the software distributor to evaluate program operation and to determine whether output can be easily read and interpreted. If regression analysis or analysis of variance is required, the prospective purchaser should ask some pointed questions about computational accuracy. In particular, the software distributor should be asked whether tests have been carried out to evaluate program performance with "unstable" or "ill-conditioned" data sets. Programs which have been successfully tested with unstable data sets will likely prove more satisfactory overall than those which have not.

There is no easy solution to the problem of selecting computer software--each package has both strengths and weaknesses. Selection of microcomputer software is complicated by the number of new programs entering the market and the highly variable, relatively untested quality. On the other hand, the investment in microcomputer statistical packages is usually a small fraction of the total cost for the microcomputer installation or of the costs associated with analysis on a mainframe computer. If an initial selection fails to provide the full range of desired capabilities, it is often possible to purchase small add-on programs to enhance performance. In addition, software authors are constantly updating and improving their products. Owners of existing versions are encouraged to make recommendations for improvements and are usually provided with updated versions for a nominal fee.

In summary, making a successful purchasing decision in the microcomputer software market requires a mixture of careful analysis and intuition. If programs are purchased without a systematic search process, frustration with both the program and the microcomputer may be the result. On the other hand, searching or waiting for the "ideal" program may deprive the researcher of a useful tool for data analysis. It is hoped that the information and procedures contained in this evaluation can aid in making balanced, reasonable software purchases.

ANNEX 1.

GLOSSARY OF TERMS

ANNEX 1. GLOSSARY OF TERMS

Applesoft	Apple Computer version of BASIC.
ASCII file	American Standards Code for Information Interchange codes for letters and numbers.
BASIC	Most popular programming language for microcomputers.
BASICA	IBM Personal Computer extended version of BASIC.
Batch	Mode of program operation in which user pre-specifies all tasks and options. No user input is allowed during actual program operation.
Binary file	A microcomputer-specific machine representation of data.
Bit	Smallest unit of data accessible by computer.
Boot	Re-start the microcomputer system.
BREAKDOWN	Procedure in SPSS to compute sums, means, s.d., and number of cases within subsets.
Byte	Eight bits.
Cases	For our purposes, equivalent to records or observations.
Comma-delimited ASCII file	File using ASCII codes with commas separating variables in each case.
Command-driven	Requires the user to type in symbols, words, or phrases to request functions.
Compiled programs	Programs are translated to machine code before program execution.
CP/M-80	Popular operating system for 8-bit microcomputers.
CP/M-86	Version of CP/M-80 available for 16-bit microcomputers.
CPU	Central Processing Unit, the "brains" of the microcomputer.
Data points	Individual values of a variable for a particular case.
DIF	Data Interchange Format, developed by VISICORP, for transferring data between programs.
DOS	Disk Operating System--the computer programs which handle "housekeeping" functions.
File cabinet	A public domain APPLE file management program.
Fixed-format ASCII file	File using ASCII codes with variables occupying the same columns in each case.

Free format ASCII file	File using ASCII codes with variables in a case separated by "delimiters," such as commas or spaces.
Hardware	The physical components of the computer system.
Interactive	"Interaction" between the program and the user. Each user entry can cause a response by the program.
Interpreted programs	Program commands executed one line at a time.
K or Kb	Kilobyte, actually 1,024 bytes.
Menu-driven	Requires the user to choose from a menu of alternatives for each program function.
MS-DOS	Popular operating system for 16-bit microcomputers.
PC-DOS	Equivalent to MS-DOS for the IBM Personal Computer.
Random Access Memory	Computer Memory in which user-input programs can be run.
Screen dump	Operation which lists to the printer what is currently displayed on the screen display.
Software	Programs.
SPSS	Statistical Package for the Social Sciences.
Toggle	To change a "software switch" to one of two positions (usually off/on).
TRS-80	Tandy/Radio Shack Corporation Computer.
Value label	Text identifying a particular code (value) of a variable.
Variable	Equivalent to a data field.
Variable name	Text identifying a particular variable.
VISIPLOT	Plotting program developed by VISICORP, the producers of VISICALC.
Z80A/8080	Popular 8-bit microprocessors, generally using CP/M operating system.

ANNEX 2.

PRECISION EVALUATION OF SELECTED MICROCOMPUTER STATISTICAL PROGRAMS

ANNEX 2: PRECISION EVALUATION OF SELECTED MICROCOMPUTER STATISTICAL PROGRAMS

INTRODUCTION

This is a report of an evaluation of six microcomputer statistical software packages (SSP), performed for the International Maize and Wheat Improvement Center (CIMMYT) by a team consisting of Raoul LePage (Professor of Statistics and Probability), Brian Thelen (Ph.D. Candidate and Graduate Assistant in Statistics), and Jim Pease (Ph.D. Candidate and Graduate Assistant in Agricultural Economics), all of Michigan State University.

Our selection of six microcomputer statistical software packages (SSP) for this study was based on a screening of eleven SSP available for APPLE, IBM-PC, and CP/M-based microcomputers. The six SSP evaluated were:

SSP	COMPUTER	OPERATING SYSTEM
AIDA	APPLE II+	DOS
ABSTAT	IBM-PC	MS-DOS
A-STAT	APPLE II+	DOS
MICROSTAT	TRS-80 MODEL II	CP/M
NUMBER CRUNCHER	IBM-PC	MS-DOS
SPS	APPLE II+	DOS

To understand the need for careful evaluation of SSP on computers, it helps to know that there are no standards for such implementations. Statistics exist principally in the form of mathematical formulas. The SSP go a step beyond Statistics in that they forge particular links to computation and impose often ad-hoc conventions for input/output, formatting, and the important command structure through which routines are called for execution. Thus, each SSP places its own specialized demands on the user, requiring that a host of things be learned which are not strictly a part of Statistics, and which may be done quite differently by another SSP. There are no real standards for any of this, but by using the SSP and comparing it with others, the user can see how sensibly the job has been done. "Felicity" is the term we will use to describe this aspect of the SSP.

Invisible to the user (but no less important) are legions of highly specialized algorithms which have been substituted for the mathematical expressions of theory. For example, the SSP does not directly compute the standard deviation, but rather computes a result with a particular programming algorithm which will, under given conditions, agree with the theoretical expression to a given accuracy. There are no standards for selecting these algorithms, and the typical SSP provides no documentation on this point. This is a serious matter, since improper selection of algorithms can be the source of inaccuracies and inefficiencies. "Fidelity" is the term we will use when referring to this aspect of the SSP.

Felicity and fidelity of an SSP are a direct reflection of the creativity and mathematical maturity of the programmers. Unfortunately, there are few norms by which the SSP may be objectively judged. One approach is to look for flaws which might seriously detract from the usability of an SSP or call into question the maturity and/or professionalism of the programmers. Accordingly, we severely tested the computational accuracy and stability of the algorithms used by each SSP for multiple linear regression (MLR), since this is where improper algorithms would most likely be found. As a further check, we tested the Analysis of Variance (ANOVA) algorithm. In addition, we severely tested the calculation of standard deviation in a way designed to detect a common flaw in algorithms used for this purpose. Our tests were aimed at the fidelity issue. Yet inevitably our work, involving much data-entry and computation, exposed strengths and weaknesses in the felicity of these SSP as well.

We requested from all SSP authors information about the algorithms used for the multiple linear regression routine. Although some authors did stipulate the type of algorithm used, it is usually impossible to evaluate the adequacy of the particular implementation without tracing the program listing itself. This involves an often unacceptable (from the software distributor's viewpoint) request for proprietary information. Thus, to evaluate SSP fidelity, it was necessary to adopt a "black box" approach which evaluates the response of SSP routines to ill-conditioned data sets.

There are some major differences between these SSP which make a truly meaningful comparison difficult. Two main differences are varying restrictions on input magnitudes and varying precision of output. AIDA had the most severe input restrictions, requiring numbers between -32,767 and 32,767 (i.e., 16 bits with 1 bit used for the sign). Although decimal values are allowed, acceptable value ranges are thus further restricted (e.g., with one decimal place, values are restricted to $\pm$3,276.7). SPS strongly recommends input of numbers with no more than seven digits. The remaining packages allow at least 14-digit input. Our solution was to prepare each data set in two precisions: one set of 14-digit data to be run on the 4 packages which could handle it, and a second set rounded to 5 digits to be run on all 6 packages.

Some packages print seven digits for each coefficient in a regression, while others output only four digits. We decided to separate the issue of accuracy of computation from that of reporting-precision of the output. Thus, if the theoretically exact coefficient, for example, rounded off to four digits agrees with the four-digit coefficient reported by an SSP, then the error is considered to be zero. Table 2.1 reports restrictions on input and a description of output for the three evaluated routines.

The following sections will report test results. Major attention is given to the methods employed to generate data sets with specified magnitudes of ill-conditioning factors known to cause problems for Multiple Linear Regression routines and the response of the evaluated SSP to such data. Additional sections report test results for ANOVA and descriptive statistics. A final brief section sets out our general conclusions about the SSP.

DATA SETS FOR MULTIPLE LINEAR REGRESSION

An Extreme Data Set

R. Wampler (1978) posed the following ill-conditioned test problem for regression, where Y indicates the dependent variable, X_0 is the constant column, and X_1 and X_2 are independent variables:

Y	X_0	X_1	X_2
200,000	1	1.00001	0.99999
600,000	1	2.00001	1.99999
400,000	1	1.99999	2.00001
400,000	1	0.99999	1.00001

The theoretically exact regression coefficients are $B_0 = B_1 = B_2 =$ 100,000, Sum of Squared Residuals (SSRESID) = 4×10^{10}. When run on the IBM Personal Computer using A Programming Language (APL) and the INTEL 8087 math co-processor, we obtained B_0 = 100,000, B_1 = 100,000.0071, B_2 = 99,999.99286, SSRESID = 4×10^{10}, and fitted values

Table 2.1. Description of Input Restrictions and Output of Regression, ANOVA, and Descriptive Statistics Routines

Statistical Package	Regression Output	ANOVA Output	Descriptive Stats Output	Restrictions on Input	Comments
1. ABSTAT	7 digits	6 digits	6 digits	17 digits	1
2. AIDA	3 dec. places	3 dec. places	3 dec. places	16 bits	0,3,4
3. A-STAT	3 dec. places	9 digits	3 dec. places	>16 digits	2
4. MICROSTAT	4 dec. places	3 dec. places	4 dec. places	>16 digits	
5. N. CRUNCHER	8 digits	7 digits	7 digits	>16 digits	1
6. SPS	9 digits	9 digits	3 dec. places	7 digits	5

0. Does not give standard error for regression.
1. Does not give T-statistic or related p-value for regression coefficients.
2. Requires a two-step procedure for regression.
3. Does not give sums of squares of residuals as part of regression routine.
4. Reports F-value for regression coefficients rather than T-statistic.
5. The SPS manual recommends the General Linear Models (GLM) routine be used for precise results, rather than the Multiple Linear Regression routine of the package. Accordingly, al results of SPS were obtained with the GLM routine.

3×10^5, 5×10^5, 5×10^5, and 3×10^5. As can be seen in Table 2.2, the IBM-based NUMBER CRUNCHER and ABSTAT did about as well as APL, as did the TRS-80-based MICROSTAT. The APPLE-based A-STAT and SPS gave coefficients which at first seem far from correct, but which are nonetheless close to being least squares solutions as evidenced by the fact that their SSRESID is relatively close to that of least squares. Results could not be calculated with AIDA due to its input restrictions.

Such inaccuracies are not confined to SSP on microcomputers. As reported by Flynn (1983), the mainframe statistics package, Statistical Analysis System (SAS), running on an IBM 360, gave $B_0 = 10000.00012000$, but was unable to compute B_1 or B_2. Also on the IBM 360, the package TROLL gave $B_0 = 100000$, $B_1 = -63818100$, $B_2 = 64018100$, and a somewhat greater SSRESID than SPS or MICROSTAT. The Statistical Package for the Social Sciences (SPSS), version 8.3, running on the Control Data Corporation Cyber 750 at Michigan State University, gave $B_0 = 100000.00$, $B_1 = 200000.00$, $B_2 = 0$ ("insufficient tolerance"), and SSRESID = 40000000016. 00024. A similar result was obtained with the MINITAB mainframe statistics package.

While it is true that the Wampler data set has exactly one regression solution, it is also true that very slight changes in the independent variables create a data set with infinitely many solutions. This is a situation which makes the scientist feel uncomfortable, but it remains a possibility and exposes difficulties whose seriousness we must assess.

Multiple Linear Regression Overview

We have studied the literature of statistical computing, with emphasis on multiple linear regression and other computations that present special difficulties (particularly on 8-bit computers) to the authors of SSP. As previously described, our plan was to generate data sets which would subject SSP to stress and expose weaknesses in fidelity, such as inaccurate/misleading answers or incomplete/improper error messages. As regards these objectives, the approach of Velleman, Seaman, and Allen (1977) to generating data sets for multiple linear regression seems useful. Their attempt to isolate distinct contributing sources of ill-conditioning is commendable and reasonably well described. By paralleling their work, we checked their approach and were able to see whether troublesome data sets from their study are a problem for the SSP of our own evaluation.

Table 2.2. Results of Tests on Wampler Data

	ABSTAT	A-STAT	MICROSTAT	NUMBER CRUNCHER	SPS
B_0	100000.0	90916.995	100000.000	100000	100000
B_1	1.000E5	295548.539	100000.0000	100000	253377.225
B_2	100000.0	-89493.203	100000.0000	100000	-53377.2256
SSRESID	4E10	Error[a]	3.87889333E10	4E10	4.0000000E10
Fitted Y					
- Case 1	300000	296976.1814	300000	300000	300003.8669
- Case 2	500000	503031.5174	500000	500000	500003.8663
- Case 3	500000	50302.8166	500000	500000	499996.9313
- Case 4	300000	296968.4806	300000	300000	299996.9319

[a] "Format error in line 5350."

Some small ambiguities and inaccuracies in Velleman et al. forced us into a "ground-up" analysis in which all aspects of generating these key data sets were studied and in which all formulae were brought into agreement with assumptions. The net result of our work is a prescriptive method for generating data sets in APL on the IBM-PC that can be submitted to the multiple linear regression routines of a SSP. Each data set is specified by five ill-conditioning parameters as described in Velleman et al. These are:

Parameter	Velleman et al. description
K_i^2	Measure of increasing multicollinearity (Squared spectral condition number) of the design matrix X (without constant column).
K_i^2	Condition of X (with constant column appended) due to small coefficient of variation K_i^2 of a column of X.
R	The correlation coefficient of the dependent vector Y with its regression on the columns of X. 1/
L	The absolute magnitude of the numbers.
M	The ratio of the magnitude of Y to that of X.

For the Wampler data set, in terms of $\delta = 10^{-5}$,

$$X'X = \begin{bmatrix} 4 & 6 & 6 \\ 6 & 10+4\delta^2 & 10-4\delta^2 \\ 6 & 10-4\delta^2 & 10+4\delta^2 \end{bmatrix}$$

The eigenvalues of

$$\begin{bmatrix} 10-4\delta^2 & 10+4\delta^2 \\ 10-4\delta^2 & 10+4\delta^2 \end{bmatrix}$$

are $\gamma_1 = 20$, $\gamma_2 = 8\delta^2 = 8 \times 10^{-10}$. The ratio of the largest eigenvalue to the smallest eigenvalue is thus 2.5×10^{10}. This is called the squared spectral condition number, and is written K_i^2. The smallest coefficient of variation for any (non-constant) column of X is denoted by K_i^2, which is equal to .16 for the Wampler data set. The fitted values 300000, 500000, 500000, 300000 have correlation $R = 1/\sqrt{2} = .707$ with the actual Y-values 200000, 600000,

---

1/ We use "Y" to represent the dependent vector, where Velleman et al. use "y".

400000, 400000. The quantities L and M are used by Velleman et al. as scale factors in their particular method of generating data sets. For the Wampler data, M (the relative magnitude of Y to X) is of order of magnitude 100000, where L is under 10.

A choice was made to generate a number of data sets on the IBM-PC. The particular system has 320K RAM, an 8087 co-processor, and runs APL*PLUS/PC under PC-DOS 1.0. The elegance, speed, and 16-decimal accuracy of APL (combined with unusually fine formatting and file-creation capabilities of this Scientific Time Sharing Corporation implementation) saw us part-way through many specialized input demands of the various packages. Additional input format requirements were satisfied by modifying the format of test data with a text editor.

Our software, written by LePage in APL, accepts a 5-vector of ill-conditioning parameters, generates a data set for multiple linear regression from this vector, and independently computes the theoretically correct regression coefficients and estimated variance. It then checks these coefficients by solving the regression and finally outputs various formatted versions of the data sets to disk. In none of our moderately stressful runs (reported in Tables 2.3, 2.7A, 2.7B, and 2.7C) did the APL solution differ from the theoretical to 10 digits. In our severely stressful runs (reported in Tables 2.4 and 2.8), there were some instances of disagreement in the third digit.

Generation of Data Sets for Multiple Linear Regression

Following Velleman et al., we have created computer routines (in APL) which accept a five-place vector of ill-conditioning parameters, and compute a data set for regression consisting of a 101 x 4 design matrix X and a 101 x 1 dependent vector Y.[1/] Unfortunately, Velleman et al. discuss fragments of their method in isolation and do not write everything out together. Nonetheless, we are confident their method has been reproduced. Our routines calculate the regression coefficients B from formulas (given below) that express B in terms of the ill-conditioning parameters directly. As explained earlier, we added a routine to regress Y on X, thus checking our work. According to our reading:

[1/] For the Wampler data, $\log_e (K_i^2) = 32$; $K_j^2 = .16$; $R^2 = .5$; L = 2; M = 200,000.

(1) $D = 3 \times 3$ diagonal matrix with $D_{11}D_{33} = D_{22} = 1, D_{11}{}^4 = K_i{}^2$

(2) $U = Z^{(1)}, Z^{(2)}, Z^{(3)}$

(3) V = 3 x 3 random orthonormal matrix uniformly distributed on mutually perpendicular points of the sphere in 3-space, and V is independent of Z.

(4) $X_0 = UDV' = X_0{}^{(1)}, X_0{}^{(2)}, X_0{}^{(3)}$ (the columns of X_0)

(5) $\underset{\sim}{1} = 101 \times 1$ vector of one's

(6) $a_j = 201\ K_j / 10((K_j{}^2 + 1)/\sqrt{101})$

(7) $X = \underset{\sim}{1}, L[a_j\ X_0{}^{(1)} + \sqrt{(1-a_j)/101}\ \underset{\sim}{1}, X_0{}^{(2)}, X_0{}^{(3)}]$

(8) $Y = LMRU/\sqrt{3} + \sqrt{(1-R^2)}\ Z^{(4)}$

We derived formulas for the theoretical solution B of the regression of Y on X which appear to agree with Velleman et al., but which are in any case consistent with the above and were confirmed by APL data runs. These are:

$$B_0 = -LB_{01}\ \sqrt{(1-a_j)}/a_j\sqrt{3}$$

$$B_1 = B_{01}/a_j$$

$$B_2 = B_{02}$$

$$B_3 = B_{03}$$

where the 3-space vector $B_0 = (B_{01}, B_{02}, B_{03})$ is given by:

$$B_0 = VD^{-1}/\sqrt{3}$$

For completeness, our APL routines are reproduced below with a few comments. We made no attempt to optimize these routines, but rather chose to parallel some mathematical notes worked up in the development phase.[1/]

[1/] Certain ambiguous or incorrect statements are made in the Velleman et al. article. The most important are:

(a) If a matrix U has entries which are jointly normally distributed, but not constant, then it is not true that normalizing the column of U (in the sense of unit length) results in normally distributed variables, as incorrectly stated on page 84.

(b) The remark: "Finally, both X and Y are scaled to desired magnitudes by multiplying them by selected constants" (Ibid., p. 84) was difficult for us to interpret mathematically. Subsequent remarks with respect to "relative" and "absolute" magnitudes seemed to clarify the point.

```
      )VARS
AJ      APLBETA BETA    BETAO   CHAT    COUNT   CSPEC   D       DDD     DIGXX
H       HOLDXX  KJ2     L       LARXX   N       OCOL    OCOLUMN OCOUNT  ODIM
OMAT    OMAT2   OMATRIX OZ      P       RK2     SPEC    SSRESID TTT     U
ULOOP   UU      V       VLOOP   VV      W       X       XO      XX      Y
ZU      ZCOUNT          )FNS
BOTHDATASETS    CREATE  CREATED CREATEU CREATEV DATASET DATASETS
ORTHONORM       SOLUTIONCHECK
      ∇ORTHONORM[⎕]
    ∇ ORTHONORM MATRIX
[1]   ODIM←(⍴MATRIX)[2]
[2]   OMATRIX←MATRIX
[3]   OMATRIX[;1]←MATRIX[;1]÷((+/(MATRIX[;1]*2))*0.5)
[4]   OCOUNT←1
[5]   OMAT2←OMATRIX[;⍳OCOUNT]
[6]   OMAT←(⍉OMAT2)+.×OMAT2
[7]   OCOL←OMAT2+.×(⌹OMAT)+.×(⍉OMAT2)+.×OMATRIX[;1+OCOUNT]
[8]   OCOLUMN←OMATRIX[;1+OCOUNT]-OCOL
[9]   OCOLUMN←OCOLUMN÷((+/OCOLUMN*2)*0.5)
[10]  OMATRIX[;1+OCOUNT]←OCOLUMN
[11]  OCOUNT←1+OCOUNT
[12]  →(OCOUNT<ODIM)/5
    ∇
[13] ∇
```

ORTHONORMALIZE A GIVEN MATRIX

```
      ∇CREATED[⎕]
    ∇ CREATED
[1]   D←⍳P
[2]   D←D∘.=D
    ∇
[3]  ∇
```

CREATE MATRIX D

```
      ∇CREATEU[⎕]
    ∇ CREATEU
[1]   U←(N,P+6)⍴0
[2]   ULOOP←0
[3]   UU←(N,P+6)⍴10000
[4]   U←U+(?UU)÷10000
[5]   ULOOP←ULOOP+1
[6]   →(ULOOP<12)/3
[7]   U←U-6
[8]   U[;1]←N⍴1
[9]   ORTHONORM U
[10]  OZ←OMATRIX
[11]  U←OZ[;1+⍳P]
    ∇
[12] ∇
```

CREATE MATRIX U

```
      ∇CREATEV[⎕]
    ∇ CREATEV
[1]   V←(P,P)⍴0
[2]   VLOOP←0
[3]   VV←(P,P)⍴10000
[4]   V←V+(?VV)÷10000
[5]   VLOOP←VLOOP+1
[6]   →(VLOOP<12)/3
[7]   V←V-6
[8]   ORTHONORM V
[9]   V←OMATRIX
    ∇
[10] ∇
```

CREATE MATRIX V

```
    ∇CREATE[⎕]
    ∇ CREATE NUMBER
     CSPEC←(NUMBER,4)⍴1000
     CMAT←(?CSPEC)+(10000×?CSPEC)+(100000000×?CSPEC)+(1000000000000×?CSPEC)
     CMAT[;1]←NUMBER⍴1
     ORTHONORM CMAT
     'CALLED BY OMATRIX'
    ∇
     ∇

    ∇DATASETS[⎕]
    ∇ DATASETS SPEC
     P←3
     N←101
     CREATEU
     CREATEV
     CREATED
     COUNT←1
     ('B:LEP2',H[COUNT],'.AID') ◊ ('B:LEP2',H[COUNT],'.AID')⎕NCREATE(-COUNT)
     SPEC[COUNT;]
     ZCOUNT←P+1+COUNT
     DATASET SPEC[COUNT;] ◊ HOLDXX←XX
     XX[1 101 ;] ◊ XX←('4P(O)M(-)K7G(Z99999999.9999999,),P(O)M(-)K7G(Z99999999.9999999)' ⎕FMT XX)
     XX←XX,⎕TCNL ◊ XX←XX,⎕TCLF
     XX ⎕NAPPEND(-COUNT) ◊ ⎕NUNTIE(-COUNT)
     SSRESID←+/((Y-X+.×BETA)*2)
     'SSRESID' ◊ SSRESID
     ('B:LEP2',H[COUNT+4],'.AID') ◊ ('B:LEP2',H[COUNT+4],'.AID')⎕NCREATE(-(COUNT+4))
     XX←HOLDXX-⌊/⌊/HOLDXX ◊ LARXX←⌈/⌈/XX ◊ DIGXX←⌈(10⍟LARXX)
     XX←XX÷(10*(DIGXX-5)) ◊ HOLDXX←XX
     XX[1 101 ;] ◊ XX←('4G(99999,),G(99999)' ⎕FMT XX)
     XX←XX,⎕TCNL ◊ XX←XX,⎕TCLF
     XX ⎕NAPPEND(-(COUNT+4)) ◊ ⎕NUNTIE(-(COUNT+4))
     'APLBETA' ◊ APLBETA←HOLDXX[;1]⌹HOLDXX[; 2 3 4 5] ◊ SSRESID←+/((HOLDXX[;1]-HOLDXX[; 2 3 4 5]+.×APLBETA
     APLBETA
     'SSRESID' ◊ SSRESID
     COUNT←COUNT+1 ◊ '' ◊ ''
     →(COUNT≤(⍴SPEC)[1])/7
    ∇
     ∇

    ∇DATASET[⎕]
    ∇ DATASET SPECROW
     ⍝ LN-KISQARE LN-KJSQUARE RKSQUARE L M
     Y0←(1÷P*0.5)×U+.×(P⍴1)
     RK2←SPECROW[3]
     Y←((RK2*0.5)×Y0)+(OZ[;ZCOUNT]×((1-RK2)*0.5))
     ZCOUNT←ZCOUNT+(RK2<1)
     Y←Y×SPECROW[4]×SPECROW[5]
     D[1;1]←(*SPECROW[1])*0.25
     D[3;3]←1÷D[1;1]
     X0←U+.×D+.×(⍉V)
     KJ2←*SPECROW[2]
     L←SPECROW[4]
     AJ←(KJ2*0.5)×(N*100)
     AJ←AJ÷(((KJ2)+1)×10×(N*0.5))
     X0[;1]←(AJ×(AJ≥0)×X0[;1])+((((1-AJ))>0)×(1-AJ)÷N)*0.5)
     X←(N,P+1)⍴0
     X[;1]←N⍴1
     X[;1+⍳P]←L×X0
     XX←(N,1+(⍴X)[2])⍴0
     XX[;1]←Y
     XX[;1+⍳((⍴X)[2])]←X
     Y SOLUTIONCHECK X
    ∇
     ∇
```

```
    ∇SOLUTIONCHECK[⎕]
  ∇ DEPEN SOLUTIONCHECK INDEPEN
    APLBETA← DEPEN ⌹ INDEPEN
    'APLBETA'
    APLBETA
    ''
    BETA0←+/(V+.×(SD))÷(P*0.5)
    BETA←((RK2)*0.5)×BETA0×SPECROW[5]
    BETA[1]←(BETA[1]÷AJ)
    BETA←(-(((1-AJ)×((1-AJ)>0)÷N)*0.5)×L×BETA[1]),BETA
    'BETA BY FORMULA'
    BETA
  ∇
 ∇

    ⍝ SAVE TO DISK AID APL 1
    )SAVE
0 SEAMAN SAVED 10/2/83 02:18:10
```

R. LaPage (Rem DDNAE 7)
6-19-83

```
DATASETS TTT
0 9.950330853E)3 1 10 1
  .BETA
 .116714331E)18 0.78867528 4.999796866E)3 0.6147896427

BETA BY FORMULA
0 0.78867528 4.999796866E)3 0.6147896427
2 4 0.01 100 15
APLBETA
)0.2846938773 3.339129451 )6.673329132E)3 1.508911985

BETA BY FORMULA
)0.2846938773 3.339129451 )6.673329132E)3 1.508911985
4 8 0.01 100000 1000
APLBETA
)124.3261562 1272.987842 10.32630553 161.6630558

BETA BY FORMULA
)124.3261562 1272.987842 10.32630553 161.6630558
      IDATASETS[]
    I DATASETS SPEC
[1]   P3
[2]   N101
[3]   CREATEU
[4]   CREATEV
[5]   CREATED
[6]   COUNT1
[7]   SPEC[COUNT;]
      SPEC[COUNT;]NAPPEND )1
[9]   ZCOUNTP+1+COUNT
[10]  DATASET SPEC[COUNT;]
[11]  XX NAPPEND )1
[12]  BETA NAPPEND )1
[13]  SSRESID+/((Y-X+.BETA)*2)
[14]  SSRESID NAPPEND )1
[15]  'SPACE' NAPPEND )1
[16]  COUNTCOUNT+1
[17]  (COUNTs((SPEC)[1])/7
     I
[18]
```

- PARAMETER VECTOR (THIS ONE IS THE "BASE DATA")
- APL BETA BY LEAST SQ.
- Ck. APL to EXACT FORMULAS
- EXACT β
- NEW PARAMETERS + β̂, β.
- You can ignore the APL particulars here.

CREATION OF THE PC DOS NATIVE FILE LEPI.AID

- $(\log_e \chi_i^2, \chi_j^2, R^2, L, M)$ APPEND 5-DIM VECTOR OF PARAMETERS AS pg 86 VELLEMAN, SEAMAN, ALLEN
- APPEND 101×5 DIM MATRIX OF DATA, THE FIRST COLUMN OF WHICH IS Y, THE SECOND COLUMN OF WHICH IS CONSTANT, ETC. AS PER THE ABOVE PARAMETERS.
- APPEND THE ACTUAL REGRESSION COEFFICIENTS FOR COL 1 (i.e. Y) ON COLUMNS 2, 3, 4, 5.
- APPEND SS RESIDUALS (EXACT)
- APPEND THE WORD "SPACE"

REPEAT ABOVE CYCLE FOR THREE CHOICES OF THE PARAMETER VECTOR.

LEPI.AID

Multiple Linear Regression Data

In testing the regression routines, we focused on accuracy when faced with (1) moderate ill-conditioning, (2) severe ill-conditioning, and (3) perfect multicollinearity. Particularly in the latter case, a good routine will halt and alert the user to the collinearity. As we will report, this was not always the case with the evaluated packages.

We note once again the vast disparity between the packages as regards their input restrictions. AIDA, by restricting input to numbers requiring 16 bits or less, imposes obvious burdens. SPS's restriction to 7 digits is less severe but still may cause problems for some data sets.

The breadth of MLR output also varies greatly. For instance:

(a) AIDA and ABSTAT do not report standard errors of coefficients, or t-statistics and related probability values (p-values). AIDA does give the partial correlation coefficient and F-values of coefficients and ABSTAT gives the correlation of each independent variable with the dependent variable.

(b) NUMBER CRUNCHER provides standard error estimates of coefficients.

(c) SPS will give t-statistics, p-values, and even confidence intervals for coefficients, but only if an extra run is made with a subroutine of the regression routine.

(d) A-STAT gives standard errors and p-values, while MICROSTAT gives these along with the t-statistics.

The data sets used in the accuracy testing are described in Tables 2.3 and 2.4. Table 2.5 reports the theoretical solution for each regression.

Table 2.3. Parameters and Number of Digits for the MODERATELY ILL-CONDITIONED Data Sets

Data Set	$\log_e K_i^2$	K_j^2	R^2	L	M	Number of Digits
4A	0	.00995033	1	10	1	14
4B	2	4	.01	100	15	14
4C	4	8	.01	100,000	1,000	14
5	10	8	1	1,000	10	14
6	10	8	.99	1,000	10	14
7	10	8	.1	1,000	10	14
15	10	8	1	1,000	10	5
16	10	8	.99	1,000	10	5
17	10	8	.1	1,000	10	5
15A	10	8	1	1,000	10	4
16A	10	8	.99	1,000	10	4
17A	10	8	.1	1,000	10	4
21	5	4	.9	10	10	14
22	5	4	.01	10	10	14
23	5	4	.5	10	10	14
25	5	4	.9	10	10	5
26	5	4	.01	10	10	5
27	5	4	.5	10	10	5
25A	5	4	.9	10	10	4
26A	5	4	.01	10	10	4
27A	5	4	.5	10	10	4

Table 2.4. Parameters and Number of Digits for the SEVERELY ILL-CONDITIONED Data Sets

Data Set	$\log_e K_i^2$	K_j^2	R^2	L	M	Number of Digits
41[a/]	32	.16	.5	2	200,000	14
61	32	.01	.5	2	200,000	14

[a/] Wampler analogue.

Table 2.5. Theoretically Determined Regression Coefficients and Sums of Squared Residuals of Data Sets

Data Set	B_0	B_1	B_2	B_3	SSRES
4A	0	.5285472992	.3105626869	.790056055	7.232216016E30
4B	-21.22526262	2.489477476	1.109691854	1.181243306	2227500
4C	-8586172.114	879.1466773	132.8458417	98.64165118	9.9E15
5	-1339.18305	116.1030197	-46.54896375	-52.8310843	2.873631112E17
6	-11282.34468	115.521046	-46.36538352	-52.56626517	1000000
7	-3585.764523	36.71499854	-14.73588621	-16.70665576	90000000
15	-12307165.18	560.3904406	-62.00607974	-26.73555982	6.439089415E¯15
16	-12485597.99	557.5814482	-61.69527037	-26.60154614	100000000
17	-4379893.062	177.2110171	-19.60804408	-8.454526354	9000000000
15A	-12307165.18	560.3904406	-62.00607974	-26.73555982	6.439089412E¯15
16A	-12485597.99	551.5814482	-61.69527037	-26.60154614	100000000
17A	-4379893.062	177.2110171	-19.60804408	-8.454526354	9000000000
21	-17.30718515	20.29932369	-18.86020729	-3.621694465	1000
22	-1.824337499	2.139736594	-1.988040406	-0.3817601166	9900
23	-12.90001417	15.13022255	-14.05756852	-2.699451672	5000
25	63735.77716	20.29932369	-18.86020729	-3.621694465	1000000000
26	22005.10965	2.139736594	-1.988040406	-0.3817601166	9900000000
27	54057.40871	15.13022255	-14.05756852	-2.699451672	5000000000
25A	63735.77716	20.29932369	-18.86020729	-3.621694465	1000000000
26A	22005.10965	2.139736594	-1.988040406	-0.3817601166	9900000000
27A	54057.40871	15.13022255	-14.05756852	-2.699451672	5000000000
41	-1520711.709	135525782.7	201582556.5	18830226.9	8.000000001E10
61	6500.831174	-92817097.88	-210962149	-78234925.79	7.999999999E10

Table 2.6 reports typical calculation times of the regression routines using the test data sets. These times generally represent the actual calculation time from the completion of input procedures to the initiation of printing. A special case, however, is A-STAT, which requires a two-step procedure, for which we give combined times. Notably, MICROSTAT gave essentially all of the basic output we desired but took only 1:25 to complete the regression run (2nd best). SPS required the longest time of 5:05, and even this underestimates the wait since SPS also took far longer than any of the others to read in data and get it ready for the regression routine.

Table 2.6. Approximate Calculation Time for Regression

Package	Time
ABSTAT	1:50
AIDA	1:00
A-STAT	3:35
MICROSTAT	1:25
NUMBER CRUNCHER	3:20
SPS	5:05

A few significant problems were encountered in trying to obtain results for all packages:

(a) MICROSTAT could not run two of the fifteen moderately ill-conditioned data sets because of an apparent internal overflow error or print formatting problem. For five of the remaining thirteen data sets, MICROSTAT could not calculate sums of squares of the residuals because of problems in calculating or printing the p-value of the F-statistic for the regression.

(b) A-STAT could not run two of the fifteen moderately ill-conditioned data sets through the correlation routine, the first step of its two-step regression procedure.

(c) AIDA's output was very poorly formatted. Decimal points of coefficients were not vertically aligned, and the misalignment differed from one data set to another. A worse problem, however, was that numbers were run together and could only be picked apart by knowing that AIDA always reports three decimal places.

Tables 2.7A, 2.7B, and 2.7C report the results of the moderately ill-conditioned data sets. Regression errors are represented as standardized errors of the calculated coefficients and the sum of squared residuals, where:

$$\text{Standardized Error of } B_0 = \frac{|B_0 - \hat{B}_0|}{|B_0|}$$

and similarly for the SSRESID.

Table 2.7A. Precision Results of Regression Runs for the MODERATELY ILL-CONDITIONED Data Sets

Package	Data Set	St. Error of B_0	St. Error of B_1	St. Error of B_2	St. Error of B_3	St. Error of SS Res.
ABSTAT	4A	1.683 E¯9 [a]	0	0	0	1.563 E¯13 [a]
A-STAT	4A	1.632 E¯9 [a]	0	0	0	4.656 E¯8 [a]
N. CRUNCHER	4A	1.00 E¯8 [a]	0	0	0	5.183 E¯12
ABSTAT	4B	0	0	0	0	0
A-STAT	4B	0	0	0	0	0
MICROSTAT	4B	0	0	0	0	0
N. CRUNCHER	4B	4.711 E¯7	2.000 E¯7	0	0	0
ABSTAT	4C	0	0	0	0	0
A-STAT	4C	4.176 E¯7	0	0	0	1.010 E¯9
MICROSTAT	4C	2.329 E¯10	0	0	0	0
N. CRUNCHER	4C	1.456 E¯5	1.445 E¯5	7.528 E¯7	1.622 E¯6	1.010 E¯7
ABSTAT	5	0	0	0	0	424 E¯5 [a]
A-STAT	5	5.283 E¯5	5.167 E¯5	2.145 E¯5	0	867.69 [a]
N. CRUNCHER	5	1.764 E¯6	2.584 E¯6	4.292 E¯7	1.439 E¯5	1.048 E¯2 [a]
ABSTAT	6	0	0	0	0	0
A-STAT	6	5.335 E¯5	5.193 E¯5	2.157 E¯5	1.902 E¯5	9.126 E¯4
MICROSTAT	6	0	0	0	0	[b]
N. CRUNCHER	6	1.241 E¯6	8.656 E¯7	0	2.854 E¯7	3.660 E¯5
ABSTAT	7	0	0	0	0	0
A-STAT	7	5.355 E¯5	5.447 E¯5	0	3.951 E¯5	9.889 E¯7
MICROSTAT	7	2.789 E¯8	0	0	0	0
N. CRUNCHER	7	1.032 E¯5	1.035 E¯5	3.393 E¯6	3.591 E¯6	1.111 E¯7

[a] Actual values since denominator is equal to (or approximately) 0.

[b] Did not report because of format problems.

Table 2.7B. Precision Results of Regression Runs for the MODERATELY ILL-CONDITIONED Data Sets

Package	Data Set	St. Error of B_0	St. Error of B_1	St. Error of B_2	St. Error of B_3	St. Error of SS Res.
ABSTAT	15	1.043 E¯3	1.179 E¯3	6.196 E¯4	1.404 E¯3	3.212 E¯6[a/]
AIDA	15A	8.8 E¯2	8.779 E¯2	8.380 E¯2	9.402 E¯2	c/
MICROSTAT	15	1.237 E¯3	1.178 E¯3	6.196 E¯4	1.404 E¯3	b/
N. CRUNCHER	15	1.237 E¯3	1.178 E¯3	6.196 E¯4	1.404 E¯3	3212528[a/]
SPS	15	1.238 E¯3	1.179 E¯3	6.135 E¯4	1.347 E¯3	3495612.67[a/]
ABSTAT	16	4.453 E¯3	3.835 E¯3	4.180 E¯3	3.806 E¯3	1.43 E¯2
AIDA	16A	6.696 E¯2	6.577 E¯2	5.670 E¯2	7.028 E¯2	c/
A-STAT	16	3.256 E¯3	3.123 E¯3	1.860 E¯3	3.739 E¯3	2.147 E¯1
MICROSTAT	16	3.801 E¯3	3.835 E¯3	4.180 E¯3	3.806 E¯3	b/
N. CRUNCHER	16	3.801 E¯3	3.835 E¯3	4.180 E¯3	3.806 E¯3	1.422 E¯2
SPS	16	3.801 E¯3	3.840 E¯3	4.147 E¯3	3.681 E¯3	2.900 E¯2
ABSTAT	17	5.112 E¯2	3.051 E¯2	2.664 E¯2	3.301 E¯2	2.222 E¯3
AIDA	17A	6.197 E¯3	6.619 E¯3	1.163 E¯2	6.686 E¯3	c/
A-STAT	17	3.464 E¯2	3.403 E¯2	2.968 E¯2	3.684 E¯2	2.317 E¯3
MICROSTAT	17	3.107 E¯2	3.51 E¯2	2.664 E¯2	3.301 E¯2	2.199 E¯3
N. CRUNCHER	17	3.107 E¯2	3.051 E¯2	2.664 E¯2	3.301 E¯2	2.199 E¯3
SPS	17	3.106 E¯2	3.051 E¯2	2.668 E¯2	3.311 E¯2	2.153 E¯3
ABSTAT	21	0	0	0	0	0
A-STAT	21	0	0	0	0	0
MICROSTAT	21	0	0	0	0	0
N. CRUNCHER	21	0	0	0	0	0
ABSTAT	22	0	0	0	0	0
A-STAT	22	0	0	0	0	0

[a/] Actual values since denominator is equal to (or approximately) 0.

[b/] Did not report because of problems in printing p-value for F-statistic.

[c/] Did not report as part of regression routine.

Table 2.7C. Precision Results of Regression Runs for the MODERATELY ILL-CONDITIONED Data Sets

Package	Data Set	St. Error of B_0	St. Error of B_1	St. Error of B_2	St. Error of B_3	St. Error of SS Res.
MICROSTAT	22	0	0	0	0	0
N. CRUNCHER	22	0	0	0	0	0
ABSTAT	23	0	0	0	0	0
A-STAT	23	0	0	0	0	0
MICROSTAT	23	0	0	0	0	0
N. CRUNCHER	23	0	0	0	0	0
ABSTAT	25	3.440 E¯4	1.736 E¯4	7.992 E¯5	3.712 E¯4	0
AIDA	25A	3.751 E¯3	6.071 E¯4	4.124 E¯4	1.020 E¯3	a/
A-STAT	25	3.378 E¯4	1.637 E¯4	6.401 E¯5	0	6.249 E¯4
MICROSTAT	25	3.432 E¯4	1.736 E¯4	7.992 E¯5	3.574 E¯4	6.459 E¯4
N. CRUNCHER	25	3.432 E¯4	1.741 E¯4	8.045 E¯5	3.709 E¯4	6.450 E¯4
SPS	25	3.371 E¯4	1.737 E¯4	8.091 E¯5	3.718 E¯4	6.459 E¯4
ABSTAT	26	4.367 E¯4	1.857 E¯3	9.857 E¯4	3.808 E¯3	1.010 E¯4
AIDA	26A	4.320 E¯3	2.927 E¯3	4.507 E¯3	8.487 E¯3	a/
A-STAT	26	4.392 E¯4	1.992 E¯3	0	3.248 E¯3	6.950 E¯6
MICROSTAT	26	4.355 E¯4	1.852 E¯3	1.086 E¯3	3.772 E¯3	6.890 E¯6
N. CRUNCHER	26	4.353 E¯4	1.859 E¯3	1.106 E¯3	3.809 E¯3	6.970 E¯6
SPS	26	4.340 E¯4	1.857 E¯3	1.107 E¯3	3.818 E¯3	6.889 E¯6
ABSTAT	27	3.553 E¯4	6.360 E¯4	4.957 E¯4	1.364 E¯3	2.000 E¯4
AIDA	27A	5.468 E¯3	2.658 E¯3	2.672 E¯3	5.724 E¯3	a/
MICROSTAT	27	3.548 E¯4	6.360 E¯4	4.957 E¯4	1.353 E¯3	2.184 E¯4
N. CRUNCHER	27	3.548 E¯4	6.360 E¯4	4.957 E¯4	1.365 E¯3	2.186 E¯4
SPS	27	3.642 E¯4	3.359 E¯4	4.959 E¯4	1.370 E¯3	2.134 E¯4

a/ Did not report as part of regression routine.

The precision errors for the moderately stressful data sets (Tables 2.7A through 2.7C) are in general very similar for all SSP, with none doing poorly. Relative errors range from 10^{-5} to 10^{-2} in magnitude and in our opinion are not significant. It appears that all six SSP packages can handle the modest degree of ill-conditioning of data that might be encountered in practice, with the exceptions noted above.

Results for the severely stressful data sets (Table 2.8) are another matter, with some standardized errors even greater than one. Notice that AIDA and SPS were not tested because input restrictions made it a meaningless exercise.

Table 2.8. Precision Results of Regression Runs for the SEVERELY ILL-CONDITIONED Data Sets

Package	Data Set	St. Error of B_0	St. Error of B_1	St. Error of B_2	St. Error of B_3	St. Error of SS Res.
ABSTAT	41	.007	.007	.007	.007	.005
A-STAT	41	Required correlation routine would not run				
MICROSTAT	41	.511	.551	.551	.551	a/
N. CRUNCHER	41	.114	.114	.114	.114	.049
ABSTAT	61	.001	.001	.001	.001	.0005
A-STAT	61	1.001	1.001	1.001	1.001	.334
MICROSTAT	61	.635	.635	.635	.635	a/
N. CRUNCHER	61	.049	.104	.104	.104	.010

a/ Did not report because of format problems.

As part of the MLR fidelity analysis, we also tested how each SSP handled perfect multicollinearity. We noted whether multicollinearity was detected by the routine, possibly by an error message or halt in operation. If multicollinearity was not detected, we examined the reported coefficients to see if they were in fact a solution of the regression. We introduced two types of multicollinearity. In the first type, we placed two constant columns of different magnitudes in the regression matrix. In the second type, one column in the matrix was set equal to the sum of two non-constant columns. This is similar to tests performed by Velleman et al.

As with other data sets, we prepared two versions. One version consisted of 14-digit data for the four packages (MICROSTAT, ABSTAT, A-STAT, and NUMBER CRUNCHER) which accept this precision. The other was rounded to 5 digits, acceptable to all the SSP (except AIDA which, as stated earlier, requires numbers between -32,767 and +32,767). A description of two data sets is given in Table 2.9. Each of these data sets (81 and 85) was then transformed to create the two aforementioned forms of multicollinearity. Table 2.10 reports the results.

Table 2.9. Parameters and Number of Digits for the PERFECT MULTICOLLINEARITY Data Sets

Data Set	$\log_e K_i^2$	K_j^2	R^2	L	M	Number of Digits
81	10	8	1	1000	10	14
85	10	8	1	1000	10	4

Note: The above data sets were transformed to generate two types of perfect multicollinearity, as described in the text.

Table 2.10. Results of Regression Runs for the PERFECT MULTICOLLINEARITY Data Sets

Package	Data Set	Regression Coefficient B_0	Regression Coefficient B_1	Regression Coefficient B_2	Regression Coefficient B_3
ABSTAT	81	-18938.8	188.081	3.92328	5.83517
A-STAT	81	-18960.749	5.183	188.958	4.448
MICROSTAT	81	-18964.9038	180.7556	-3.8392	13.4279
N. CRUNCHER	81	-19160.63	542294.3	542107.8	-542098.1
ABSTAT	85	-4484000	170.497	-14.4542	24.2051
AIDA	85[a/]	-252.133	0	-.87	.364
A-STAT	85	b/	b/	b/	b/
MICROSTAT	85	-4477837.2021	183.4143	-1.5412	11.1461
N. CRUNCHER	85	-4477837	178.8937	-6.061788	15.66667
SPS	85	-4477837.41	-3354660.44	-3354845.39	3354855

[a/] Modified 85 in order to be acceptable to the input constraints of AIDA.

[b/] Did not run and gave an error message.

The results with multicollinearity caused by two constant columns were quite interesting. Only MICROSTAT diagnosed and reported the problem, while SPS ran the data set with no indication of any problem. Each of the other four packages gave some indication, ranging from "blowing up" or error message to an explicit diagnostic indicating presence of multicollinearity. NUMBER CRUNCHER and AIDA gave a "divide by zero" error message, while ABSTAT gave output consistent with setting the second constant vector equal to zero. Finally, A-STAT gave a meaningless message of "y = 133, LINE = 0" repetitiously until the computer was turned off.

For the multicollinearity caused by one independent variable being the sum of two other independent variables, the results were quite different. With the exception of A-STAT and AIDA, most SSP gave very little direct indication. A-STAT gave value 0 to all quantities related to variable 3 (variable 5 = variable 3 + variable 4). It should be noted that the data set used for AIDA was modified to make it acceptable to AIDA's data input constraints. The regression coefficients as output by the SSP for the second type of multicollinearity are given in Table 2.10. According to the results, we see what appears to be high instability of NUMBER CRUNCHER on data set 81 and of SPS on data set 85.

It should be mentioned that some of the packages do report certain statistics which might be used in some cases to help determine if a high degree of multicollinearity is present. A-STAT lists the determinant of the quadratic form of the regression matrix and NUMBER CRUNCHER gives (for each variable) the increase in the sum of squares if that variable is deleted from the full model. A knowledgeable user could use such output to help detect significant degrees of multicollinearity.

ANOVA

Tests were run on each of the SSP one-way ANOVA routines for the purpose of testing fidelity. We generated data for a one-way ANOVA with two treatments, adapting the methods used to generate data sets for regression. In each data set, there were 101 cases. Once again, we prepared two groups of data sets (one with 14 digits for MICROSTAT, ABSTAT, A-STAT, and NUMBER CRUNCHER and one group with only 5 digits to be used on all SSP). Further description of the data is given in Table 2.11.

Table 2.11. Description of ANOVA Data

Data Set	Number of Digits	Group Means	Group Standard Deviations	Number of Observations
101	14	55.46434	144.8605	48
		-50.23185	137.8582	53
102	14	-71.27045	147.8665	46
		59.60800	135.2282	55
103	14	-71.46736	146.2704	47
		62.20302	136.4609	54
105	5	21659.79	1448.601	48
		20602.72	1378.579	53
106	5	20950.35	1478.662	46
		22259.13	1352.279	55
107	5	24506.96	1462.703	47
		25843.69	1364.609	54

ANOVA output for the different SSP were in general similar, with a few important differences. All of the SSP print a standard ANOVA table, along with appropriate F values, the number of cases per group, and the group means. All except A-STAT and MICROSTAT also gave the group standard deviations or variances. SPS also gave the value for the Bartlett Chi Square. SPS also exhibited a serious output formatting problem which resulted in numbers being run together. The only way we could attempt to pick numbers apart was to count decimal places, which was not always possible. AIDA has the less severe formatting problem of jagged columns, making it sometimes difficult to determine the correct column for a given number.

Fidelity for all the SSP were very good. There were no errors by any of the packages at the decimal precisions of their respective output.

DESCRIPTIVE STATISTICS

Data

In testing the descriptive statistics routines, we calculated means and standard deviations on each of five different sets of data for all six SSP. These data sets consisted of a sequence of 1,000 cases of the same number with alternating + and - signs (e.g., +100, -100, +100, etc.). The five numbers used to generate the five data sets were 100, 1000, 3000, 10000, and 30000. A descriptive statistics routine which directly sums the squares would have to deal with a sum of 9×10^{12} (for the squares of 1,000 numbers of absolute value 30000) which might cause trouble. A better method is to recursively calculate the variance as each number is brought in. A summary of these data sets is given in Table 2.12. All six packages passed these tests without error.

For an unknown reason, AIDA read -10,000 as -10,001 and -30,000 as -30,001. Such numbers are well within the stated data value restrictions given in the AIDA manual. We did not check if this was a problem for other routines. A different peculiarity is exhibited by SPS, which always computes skewness and kurtosis with the mean and variance. This can cause annoying delays, since some of the test data were too extreme for the skewness and kurtosis routines to handle. As regards calculation times (reported in Table 2.13), the ratio of slowest to fastest was around 2.3 to 1. AIDA was fastest at one minute.

Table 2.12. Summary of Data Sets Used for Descriptive Statistics Test

Data Set	Number of Cases	Magnitude	Description
31	1000	100	Alternating positive and negative values
32	1000	1000	Alternating positive and negative values
33	1000	3000	Alternating positive and negative values
35	1000	10000	Alternating positive and negative values
35	1000	30000	Alternating positive and negative values

Table 2.13. Results of Descriptive Statistics Tests

Package	Min:Sec[a/]	Error	Comments
ABSTAT	1:45	0	
AIDA	1:00	0[b/]	c/
A-STAT	2:35	0	
MICROSTAT	2:15	0	
NUMBER CRUNCHER	2:10	0	
SPS		0	d/

[a/] Includes printing of output.

[b/] For altered data, see (c).

[c/] Subtracted one from -10,000 and -30,000 when reading in data.

[d/] Requires calculation of skewness and kurtosis along with calculations of mean and standard deviation. Kurtosis was negative for one data set.

SUMMARY AND CONCLUSIONS

The procedures developed in this evaluation have severely tested the descriptive statistics, ANOVA, and multiple linear regression routines of the selected microcomputer statistical packages. No errors of consequence were encountered in either the calculation of standard deviation (descriptive statistics) or the sum of squares (ANOVA).

Our principal evaluative effort was directed towards exposing possible flaws in the Multiple Linear Regression (MLR) routine. Using procedures adapted from Velleman et al. (1977), data were generated which subjected the MLR routines to moderate ill-conditioning, severe ill-conditioning, and perfect multicollinearity. Except for the implications of certain program errors and value restrictions of some packages, results from the moderately ill-conditioned data indicate that the selected SSP can handle the degree of ill-conditioning likely to be encountered in normal research applications. As we endeavored to break down the SSP, four packages were then subjected to severely ill-conditioned data, with results indicating a more marked differentiation between SSP. A-STAT shows high instability and MICROSTAT demonstrates instability and formatting problems. Both NUMBER CRUNCHER and ABSTAT responded quite well to the severe testing, although ABSTAT was clearly superior in precision of output.

Tests conducted with perfectly multicollinear data demonstrate that the selected SSP in general do not provide the data diagnostics normally expected of a mainframe statistical package. Of particular concern was the failure by most SSP to provide any indication of multicollinearity caused by one independent variable being the sum of two other independent variables.

The procedures developed by Velleman et al. (1977) and modified here provide a framework to evaluate the accuracy of microcomputer and mainframe SSP regression routines. At least for the six evaluated SSP, it appears that the fears expressed by some researchers concerning the accuracy of statistical computing on microcomputers have not been corroborated. The evaluation exposed some weaknesses in the presence of severely ill-conditioned data, but the more obvious problems of the evaluated routines are input value restrictions, poor output formatting, and lack of data diagnostic procedures.

BIBLIOGRAPHY

Allen, D. "Computational Methods for Estimation of Linear Combinations of Parameters." Proceedings Statistical Computing Section, Amer. Stat. Assoc. (1977): 1-3.

Boehm, W.T. and D.J. Menkhaus. "Accuracy of Least Squares Computer Programs: Another Reminder." Amer. J. Agr. Econ. 58 (1976): 757-760.

Diaconis, P. and B. Efron. "Computer-Intensive Methods in Statistics." Sci. Amer. 248 (1983): 116-130.

Flynn, B. "Numerical Accuracy of Least Squares Computer Programs." Access, July/August (1983).

Francis, I.; R.M. Heiberger; and P.F. Velleman. "Criteria and Considerations in the Evaluation of Statistics Program Packages." Amer. Statist. 29 (1975): 52-56.

Frane, J.W. "Detecting and Describing Statistical and Numerical Ill-conditioning." Proceedings Statistical Computing Section, Amer. Stat. Assoc. (1978): 68-77.

Golub, G. and C. Reinsch. "Singular Value Decomposition and Least Squares Solutions." Numer. Math. 14 (1970): 403-420.

Kennedy, W.J. and J.E. Gentle. Statistical Computing. Marcel Dekker, N.Y. and Basel (1981).

Knuth, D.E. Seminumerica and Algorithms, Vol. II, 2nd Edition, Addison Wesley (1981).

Longley, J.W. "Least Squares Computations and the Condition of the Matrix." Comm. Statist. B-Simulation Compute. 10 (1981): 593-615.

Longley, J.W. "An Appraisal of Least Squares Programs for the Electronic Computer from the Point of View of the User." J. Amer. Statist. Assoc. 62 (1967): 819-829.

Muller, M.E. "Aspects of Statistical Computing: What Packages for the 1980's Ought to Do." Amer. Statist. 34 (1980): 159-168.

Platt, G.W. and C.A. Platt. "Microcomputer Statistical Packages and Regression Performances." Proceedings Statistical Computing Section, Amer. Stat. Assoc. (1981): 309-312.

Velleman, P.F.; J. Seaman; and E. Allen. "Evaluating Package Regression Routines." Proceedings Statistical Computing Section, Amer. Stat. Assoc. (1977): 82-92.

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

ANNEX 3.

TABLES OF CONTENTS OF PROGRAM MANUALS

ABSTAT™

USER MANUAL

RELEASE 3.0

ANDERSON - BELL
P.O. Box 191
Canon City, CO 81212
(303) 275-1661

TABLE OF CONTENTS

Apple Interactive Data Analysis

AIDA

>+< >+< >+< >+<

Version 9/82

Action-Research Northwest
11442 Marine View Drive, S.W.
Seattle, Washington 98146

(206)241-1645

s/n: 2-106

TABLE OF CONTENTS

>+< >+< >+< >+< >+< >+< >+< >+<

Table of Contents

AIDA UPDATES

The PROGRAM DISK contains a listing of changes and updates made in AIDA since this documentation was published. To use it: 1) boot the PROGRAM DISK, and 2) select the UPDATES option on the menu. You may view the text on the screen, and/or route it to your printer. (The two other options run either the AIDA CHANGES program or the FILE PROGRAMS.)

CREDITS

The DOS MOVER routine used to increase the size of AIDA's data set for 64K machines was written by Cornelus Bongers, and was published in "Call-A.P.P.L.E." magazine, with whose permission it is used here.

A-STAT 79.6

Language Reference Manual

by

Gary M. Grandon

A subset of P-STAT 78 (tm) for the Apple II (tm) Computer.

Table of Contents Page

MICROSTAT

TABLE OF CONTENTS

*** Appendices ***

NUMBER CRUNCHER

TABLE OF CONTENTS

U S E R 'S M A N U A L

STATISTICAL PROCESSING SYSTEM VERSION 4.2

(S. P. S. Version 4.2)

FOR APPLE II+

BY

Gregory J. Buhyoff[1]

H. Michael Rauscher

R. Bruce Hull IV

Kevin Killeen

Rodney C. Kirk

INDEX

ANNEX 4.

PUBLICIZED CHANGES IN NEW VERSIONS OF EVALUATED PROGRAMS

ANNEX 4. PUBLICIZED CHANGES IN NEW VERSIONS OF EVALUATED PROGRAMS

ABSTAT Version 3.03

Reports standard errors and t-tests of regression coefficients.

AIDA

Improved output formatting.
Improved speed.

A-STAT 83

Price: $200
All routines support missing values and weight variables.
Output may be routed to disk file.
Output print width is user definable.
Choice of menu- or command-driven operation.

MICROSTAT Release 4.0

Price: $375
Interface with fixed-field ASCII and dBASE II files.
Output may be routed to disk files.
Recoding, grouping transformation capability.
More transformation codes.
Analysis of subsets of cases through selection by index variables.
Written in APC BASIC (comes with run-time interpreter).
Selection of subset of variables for regression analysis.
Dynamic formatting.

NUMBER CRUNCHER

Improved error-trapping capability.
Analysis of subsets of cases through selection by index variables.

SPS Version 4.0

Available for CP/M and IBM Personal Computer.
Price: $395, from DATABASIC Software, Inc.
102 S. Main Street
Mt. Pleasant, MI 48858

MSU INTERNATIONAL DEVELOPMENT PAPERS

		Price
IDP No. 1	Carl K. Eicher and Doyle C. Baker, "Research on Agricultural Development in Sub-Saharan Africa: A Critical Survey," 1982 (346 pp.).	$ 8.00
IDP No. 2	Eric W. Crawford, "A Simulation Study of Constraints on Traditional Farming Systems in Northern Nigeria," 1982 (136 pp.).	$ 5.00
IDP No. 3	M.P. Collinson, "Farming Systems Research in Eastern Africa: The Experience of CIMMYT and Some National Agricultural Research Services, 1976-81," 1982 (67 pp.).	$ 4.00
IDP No. 4	Vincent Barrett, Gregory Lassiter, David Wilcock, Doyle Baker, and Eric Crawford, "Animal Traction in Eastern Upper Volta: A Technical, Economic and Institutional Analysis," 1982 (132 pp.).	$ 5.00
IDP No. 5	John Strauss, "Socio-Economic Determinants of Food Consumption and Production in Rural Sierra Leone: Application of an Agricultural Household Model with Several Commodities," 1983 (91 pp.).	$ 5.00

MSU INTERNATIONAL DEVELOPMENT WORKING PAPERS

WP No. 1	Daniel Galt, Alvaro Diaz, Mario Contreras, Frank Peairs, Joshua Posner and Franklin Rosales, "Farming Systems Research (FSR) in Honduras, 1977-81: A Case Study," 1982 (48 pp.).	Free
WP No. 2	Edouard K. Tapsoba, "Credit Agricole et Credit Informel dans le Region Orientale de Haute-Volta: Analyse Economique, Performance Institutionnelle et Implications en Matiere de Politique de Developpement Agricole," 1982 (125 pp.).	Out of Print
WP No. 3	W.P. Strassmann, "Employment and Construction: Multicountry Estimates of Costs and Substitution Elasticities for Small Dwellings," 1982 (48 pp.).	Out of Print
WP No. 4	Donald C. Mead, "Sub-contracting in Rural Areas of Thailand," 1982 (52 pp.).	Out of Print
WP No. 5	Michael T. Weber, James Pease, Warren Vincent, Eric W. Crawford and Thomas Stilwell, "Microcomputers and Programmable Calculators for Agricultural Research in Developing Countries," 1983 (113 pp.).	$ 5.00
WP No. 6	Thomas Stilwell, "Periodicals for Microcomputers: An Annotated Bibliography," 1983 (70 pp.).	$ 4.00
WP No. 7	W. Paul Strassmann, "Employment and Housing in Lima, Peru," 1983 (96 pp.).	Out of Print
WP No. 8	Carl K. Eicher, "Faire Face a la Crise Alimentaire de l'Afrique," 1983 (29 pp.).	Free
WP No. 9	Thomas C. Stilwell, "Software Directories for Microcomputers: An Annotated Bibliography," 1983 (14 pp.).	$ 3.00
WP No. 10	Ralph E. Hepp, "Instructional Aids for Teaching How to Use the TI-59 Programmable Calculator," 1983 (133 pp.).	Out of Print
WP No. 11	Michael L. Morris and Michael T. Weber, "Programmable Calculator (TI-59) Programs for Marketing and Price Analysis in Third World Countries," 1983 (105 pp.).	Out of Print
WP No. 12	Valerie Kelly, Robert D. Stevens, Thomas Stilwell, and Michael T. Weber, "An Annotated Directory of Statistical and Related Microcomputer Software for Socioeconomic Data Analysis," 1983 (165 pp.).	$ 7.00
WP No. 13	Chris Wolf, "Guidelines for Selection of Microcomputer Hardware," 1983 (90 pp.).	$ 5.00
WP No. 14	Eric W. Crawford, Ting-Ing Ho, and A. Allan Schmid, "User's Guide to BENCOS—SuperCalc Template for Benefit-Cost Analysis," 1984 (35 pp.).	$ 3.00
	Copy of BENCOS Template in IBM PC-DOS 1.1 Format, on single sided double density diskette (readable on most MS-DOS systems).	$15.00
WP No. 15	James W. Pease and Raoul Lepage with Valerie Kelly, Rita Laker-Ojok, Brian Thelen, and Paul Wolberg, "An Evaluation of Selected Microcomputer Statistical Programs, 1984 (187 pp.).	$ 7.00