STATISTICAL PROGRAMS AND MISSING DATA

KM 2(1981)2
pag 56-65

by

Douwe M. van der Sluis

Computer Center University of Groningen, The Netherlands

Samenvatting: Veel statistische programma's en pakketten zijn in omloop. Bij presentaties wordt over het algemeen slechts aandacht besteed aan de fraaie uitvoer en de vele methoden. Zo ook op COMPSTAT 1978 in Leiden, waar een potentiële gebruiker met deze informatie moeilijk een keus kon maken. Een van de facetten, die een potentiële gebruiker moet weten, is hoe worden ontbrekende gegevens ingelezen en hoe wordt er mee gerekend. In dit artikel wordt op deze zaken ingegaan.

Trefwoorden : Statistische pakketten, ontbrekende gegevens, BMDP, GENSTAT, PSTAT, SAS, SPSS, WESP.

Paper presented at ISSUE '79
3rd annual SPSS Users and Coordinators Conference

This paper contains two parts:
1. A general introduction
2. A more technical comparison of six general statistical packages, especially with respect to the possibilities of missing data handling.

Six general statistical packages have been selected from a collection of more than hundred packages, from all over the world.
It forms a multi coloured collection.

OSIRIS
SCSS
SURVO
STATAL
SURVEY70
GUMMA
RUMMAGE
SIS
W E S P
INTERSTAT
STATICS
S P S S
DATABLE
TSP
STATPACK
NISAN
G E N S T A T
PSST
XDS2/3
P S T A T
DAEDAL
OMNITAB
S A S
DONOVAN
SPS
B M D P
COI
CUMULUS
ASCOP
BBC
PMMD
RGSP
BATCHSTAT
FAKAD
GLIM
NESSA
TTWESP
CADA

The six packages in the centre of the picture are not selected for quality reasons, but because at the COMPSTAT 1978 (COMPutational STATistics) congress in Leiden a paper or workshop has been given about each of these packages. Selection on quality would be impossible, because the number of statistical computer programs and packages has increased enormously in the last ten years, resulting in an impossibility to evaluate all of them. A general statistical package happens to be a tool for social-scientific research, which cannot be disregarded nowadays. In the last ten years the number of users of statistical packages has increased enormously, also in Holland.

The use of statistical programs in Groningen
The University of Groningen has at its disposal several statistical programs and packages on a CDC CYBER 74-18 computer. The amount of students is about 15000.
These programs are used about 80.000 times a year; the general statistical packages WESP, a package developed in Groningen, SPSS and TTWESP, the interactive version of WESP, about 62.500 times. 60 to 70 percent of the WESP and SPSS jobs are run correctly, that means correct in the sense of syntax and executing without looking at the relevance of the statistical method used.

An overview:

SPSS	24043
WESP	31737
TTWESP	6834
others	17026
	79640

which means 13.5 percent of all the jobs running on the computer.

Others implies: Multivariance
Clustan
Distor - library of statistical and operations research routines
Cada
Stap - statistical appendix of SPSS
and
other programs and libraries

Especially with respect to the future, it will be clear, that the quality of the technical support of statistical programs will be an important factor. Only high quality will give some guarantee for continuity.
Our impression is that in general the users of statistical packages are at present not very quality-conscious.
Each general statistical package has its own "fanclub" for which it has a nearly inviolable monopoly position.

The wellknown - most used - packages are presented by commercial institutions, with too much attention for maintaining a monopoly position and to appeal to the computer users in the future. This ambition finds expression in continuous efforts to make the access easier and to add new features; more and new methods can be used with increasing comfort.

Generally improvements of the frame work and the quality have been considered unnecessary, a loss of time or impossible. At presentations and other advanced activities future users and other interested people have been one-sidedly informed about the ease of access, the many features and other strong points of the package. It is impossible for the audience to make a selection with these information. On the COMPSTAT congress in Leiden last year many people have had this experience. It is clear, that this situation could only have arisen and can only continue through a lack of interest in evaluation activities within the several users groups related to statistical packages.

The groups concerned with statistical programs can be divided into 4 categories:
1 - Sales organisations
2 - Computer Centers
3 - Users
4 - Statisticians

The users can be divided into 3 categories:
3 - A Education
    B Research
    C Production

Each of these groups has its own special interests. It is very unfortunate that in general these groups have no or only marginal contact.

At this moment the use of statistical methods is affected by the availability of many statistical programs, resulting in an increasing use of statistical methods.
The quality of the use of statistical methods is affected through:
- program errors.
- bad default settings, which means values assigned to parameters by the programmer and which in some cases can be altered by the user.
- lack of robustness of the implemented statistical techniques.
- occurrence of missing values and the insufficient possibilities for missing values handling.

While the majority of computer center personnel and practicioning statisticians knows sufficiently well, that the design and the quality of most packages is generally insufficient or at least doubtful, the common user does not know this fact and the conclusion is, that this user is the victim and he is unable to improve his situation.

Evaluation of statistical programs is a must.
Francis et al (1975) have drawn up a list of criteria and rules for the evaluation of statistical packages. This paper can be found in "The American Statistician".

With respect to the possibilities of handling data structures, a general statistical package is very limited. Looking at a correlation matrix as an intermediate result, most packages can only read a rectangular data matrix and no other formats. J.A. Nelder points out this limitation in his invited paper at the last COMPSTAT congress. In general it is difficult to handle an imcomplete data matrix in any other way than by listwise and pairwise deletion of missing values.

The rest of my presentation is devoted to the representation and the different ways of handling missing values in six general statistical packages.

The first question, that will arise is:

How will missing values originate?

Verbeek (1979) mentioned three possibilities in the Dutch VVS-Bulletin:
1 - the questioner forgets to pose a question
2 - the respondent refuses to answer a question
3 - the registration of the answer is unreadable, lost or obviously wrong.

I owe to A. Verbeek the remark, that an important phase in research is generally skipped, namely the question: Does the lack of scores show a certain structure or in other words, is the lack of scores systematical or random? BMDP has a feature to investigate this, but in other packages the user can not answer this question in an easy way. A. Verbeek gives an example, that he has programmed in SPSS.

I will review the following six general statistical packages with respect to a number of missing values aspects.

WESP - Waarlijk Eenvoudig Statistisch Pakket
- University of Groningen

SPSS - Statistical Package for the Social Sciences
- SPSS inc.

PSTAT - Statistical program package
- PSTAT inc.

GENSTAT - A general statistical program
- Rothamsted experimental station

SAS - Statistical Analysis System
- SAS institute corporation

BMDP - Biomedical computer programs
- University of California

Some background information about these packages
In general the packages have been written in Fortran except SAS, which is written in Pl/1. They have usually a small nucleus of assembler routines to work with own file names, to increase the [illegible], etc.

BMDP, PSTAT, SAS and SPSS have been developed in the United States of America, GENSTAT in the United Kingdom and WESP in the Netherlands.

SAS is only available on IBM equipment and WESP only on CDC. The other four packages run both on IBM and CDC. SPSS is even running under more than 25 different operating systems.

PSTAT and SAS can also be used conversationally. WESP has a conversational version TTWESP - teletype version of WESP, that also runs on a DEC10 system. TTWESP and WESP have an identical instruction language and the same system files, SPSS has a conversational version SCSS - the SPSS conversational statistical system - that has an instruction set different from SPSS. Also the system files are not equally structured.

Descriptive statistics can be performed with all these packages. In BMDP the main accent is laid on advanced statistical methods; BMDP is more a set of programs with similar instruction language and data structures, but little data management features. The data management is strongly represented in PSTAT.

The philosophies underlying these packages diverge. WESP is simple in use with conscious limitations. 80 or 90 percent of the users can do all their statistical analysis with WESP. It is considered appropriate for a first contact with data analysis with the computer. SPSS will cover all possible techniques with a detailed output.

The instruction language of SPSS contains some format limitations, for instance column 16 and one and only one blank between keywords. The instruction language in WESP is nearly free formatted. The instruction language of GENSTAT looks like Fortran.

GENSTAT has the possibility for adding new statistical models with a system of macro's. It is very simple to add new modules to PSTAT and WESP, but this is nearly impossible in SPSS.

I. Francis has studied the presence and the extensiveness of statistical methods in several statistical packages. The results have been published in the proceedings of the 41st session of ISI (International Statistical Institute, 1977). Although the division in categories is not very clear, the surveys gives a certain impression about each of the packages concerned:

	BMDP77	GENSTAT	PSTAT	SAS76.5	SPSS
multiple regression	3	2	2	3	1
anova/linear model	2	3	1	3	1
linear multivariate	3	3	2	3	1
multi-way tables	3	3		1	1
other multivariate	2	2	2	2	2
time series				2	
non parametric	2	1	1	1	3
exploratory	1			1	
robust	1	2		1	
curve fitting	3	1		2	
bayesian	1			1	
econometrics				2	

BLANK - no facilities
1 - present, but brief, often as a by-product of an other program
2 - present
3 - present with all the facilities

The list dates from 1977 and in the last few years some packages have been strongly extended, see SPSS.

Internal representation of missing values

The internal representation of missing values is machine dependent. I shall consider one machine for each package, but the representation on other machines will be similar:

WESP	CDC	10^{300}
SPSS		original values
PSTAT	IBM	-123456E20
		-123457E20
		-123458E20
GENSTAT	CDC	1H*
SAS	IBM	. 80.00.00.00.00.00.00.00
		.A thru .Z
		._
BMDP	IBM	16^{31}
		2.16^{31}
		-2.16^{31}

All the packages, except SPSS have reserved values for missing values, which are easy to recognize.
SPSS has the philosophy, that data can be regarded both as missing value and as nonmissing value. The missing values in SPSS can only be recognized by flags. The SPSS program has to check a list of values in order to identify a value as missing value. The consistency of this method is discussed in the next part of my presentation.

Input of blanks and other missing data

	BLANK default	BLANK with provisions	other missing values could be
WESP	MV	irr	irr
SPSS	0	MV	max 3 values (alfanumerical) ----<, <---, <----<
PSTAT	MV	other value	+,-,1 value (alfanumerical) ----<, <----
GENSTAT	MV	0	&,-,M (any word beginning with character M)
SAS	MV	irr	A thru Z, underline
BMDP	0	MV	1 value

irr - irrelevant
MV - value for missing data

Except in SPSS and BMDP a blank will be always regarded as missing value. Values assigned to blanks are written in the output of BMDP. With provisions it is also possible to regard blanks as missing values in SPSS and BMDP. PSTAT and GENSTAT have provisions to change a blank into an other value. Other values can be interpreted as missing values, which is not possible in WESP. Missing values can be reported in GENSTAT. In BMDP a user can give minimum and maximum bounds for variables; all values outside these bounds are left out of the computations. Each program checks on these bounds; the bounds can be altered at any time.

For all the packages the ways of handling the case where the data do not pass the input checks are different.

WESP - lists errors and the individual concerned is not written to the system file.
PSTAT - a missing value is assigned to errors. The program stops if the number of errors exceeds a maximum amount, specified by the user.
SAS - a missing value is assigned to errors.
GENSTAT - by default, the program stops when the first error is encountered. A missing value is assigned. The user can specify a maximum amount of errors.
SPSS - the value 0 is assigned to errors, representing a missing value if specified as such. The program stops if the number of errors exceeds a maximum amount, specified by the manager of the package, in contrast with PSTAT and GENSTAT.
BMDP - the program stops with a Fortran error message if an alpha-numerical character is encountered in a numerical field.

Example with blanks in input in SPSS

Visitors of a toy-shop have been asked about their age and the distance between the shop and their home. A blank has been filled in, if the question has not been answered.

A crosstable has been made with SPSS as implemented on the CDC computer of the University of Groningen. SPSS reads a blank as zero if we do not make provisions. The results can be read in the first table.

Table 1:

Age \ Dis	0	1	2
0	1	1	1
1	1	1	0
2	4	1	0
	6	3	1

Take provisions: MISSING VALUES Age(BLANK)/Dis(BLANK)

The results when provisions are made can be read in the second table; the differences with the first table will be clear.

Table 2:

Age \ Dis	0	1
0	1	0
1	0	1
2	0	1
	1	2

As there are few individuals, the results of the computation is striking, but what to do in case of many individuals? How can one know whether the results are reliable or not. In this case, the user has to know that he has to take provisions, but a lot of them do forget it or do not know any better.

Default values of newly created variables

Newly created variables are assigned to missing values in WESP, PSTAT, SAS and BMDP, which value is overwritten when an assignment takes place. GENSTAT stops if an assignment is invalid. In SPSS the variable is set to zero, which can give confusion with a real code zero. In these two packages it is necessary for the user to initiate new variables.

In all of the packages under study it is possible to replace a missing value by a real value and vice versa.

Example with the creation of a new variable

A constructed example has been run with SPSS. In a clumsy way the variable AGE has been divided into four classes in the new variable AGECAT. The user has made some mistakes, which means that some categories have been overlooked.

```
IF    (0 ≤ AGE ≤ 18)    AGECAT = 0
IF    (10 ≤ AGE ≤ 35)   AGECAT = 1
IF    (40 ≤ AGE ≤ 65)   AGECAT = 2
IF    (70 ≤ AGE ≤ 99)   AGECAT = 3
```

In table 3 too many ages are set to code zero in AGECAT.

table 3:

CASE NO	AGE	AGECAT
1	6	0
2	26	1
3	36	0
4	46	2
5	66	0
6	76	3

Some provisions have to be made.

```
ASSIGN MISSING     AGECAT (999)
```

Table 4 contains the result. The code 999 is absurd and is consequently striking.

table 4:

CASE NO	AGE	AGECAT
1	6	0
2	26	1
3	36	999
4	46	2
5	66	999
6	76	3

Data-transformation

If in a computation a missing value for a variable is encountered, by default all packages except SPSS assign a missing value to the result. On account of the internal representation SPSS performs the computation as if a real value has been read. However in SPSS, the result will be missing value, if the ASSIGN MISSING statement has been used.

Invalid computations like dividing by zero, logarithm of zero or a negative value result in all these packages, except SPSS and GENSTAT, in a missing value. For SPSS hold the same rules as above. If no provisions are made, GENSTAT stops. In GENSTAT, the result can be a special value, given by the user, by example zero, the value for missing value.

Example with data transformation

A constructed example in SPSS to compute the Y-coordinate of a parabola.

```
MISSING VALUES          X(BLANK,6,9)
COMPUTE                 Y=X*X+10
```

The result using the defaults in table 5.

Table 5:

CASE NO	X	Y
1	-0	10
2	0	10
3	1	11
4	2	14
5	5	35
6	6	46
7	8	74
8	9	91

It is not possible to distinguish between results of computations with missing values or with real values.

With provisions

```
ASSIGN MISSING          Y(9999)
```

the results are clear. The values in cases with a missing value are absurd and the results of computations with missing values are recognizable. See table 6.

Table 6:

CASE NO	X	Y
1	-0	9999
2	0	10
3	1	11
4	2	14
5	5	35
6	6	9999
7	8	74
8	9	9999

Missing data in modules for statistical analysis

In all packages under study, it is, by default, not possible to calculate with the missing values. In SPSS there exists an option for handling missing values as real numbers.

Listwise deletion is often superior to pairwise deletion. Listwise deletion is default in all modules only in WESP and GENSTAT. Pairwise deletion is default in PSTAT, strangely enough. Both options can be default in the other packages. Dependent of the kind of module pairwise deletion is chosen mostly in the computation of rank correlation coefficients.

An overview:

	Default pairwise deletion	Default listwise deletion
WESP		X
SPSS	X	X
PSTAT	X	
GENSTAT		X
SAS	X	X
BMDP	X	X

Pairwise-listwise deletion

A constructed example has been run with WESP, our own statistical package. Listwise deletion is default. The results are given in table 7.

Table 7:

```
CORRELATE
NUMBER OF INDIVIDUALS IS 3
VAR   CORRELATION COEFFICIENT (*1000)
 1  |  1000
 2  |   982   1000
 3  |   982   1000   1000
 4  |   500    327    327   1000
          1      2      3      4
```

With the parameter MISDAT the correlation matrix can be computed with pairwise deletion. Some correlation coefficients change remarkably. See table 8.

Table 8: CORRELATE

VAR	CORRELATION COEFFICIENT (*1000)			
1	1000			
	9			
2	368	1000		
	9	10		
3	- 164	152	1000	
	7	8	8	
4	629	576	272	1000
	4	4	4	4
	1	2	3	4

In the output is indicated whether listwise or pairwise deletion of missing values is used.

The lack of consistency in the use of missing values in these packages is due to lack of a theory about the use of missing values.
The user should realise all the time whether he treats his missing values in the right way.

References

- N.C. Alvery et al - GENSTAT, a general statistical program, Rothamsted Experimental Station, october 1977.

- J.A. Barr et al - A user's guide to SAS, SAS institute, 1976.

- COMPSTAT 1978 - Proceedings in computational statistics.

- W.J. Dixon et al - BMDP, Biomedical Computer Programs P-series, UCLA, 1977

- I. Francis et al - Criteria and considerations in the evaluation of statistical program packages, the american statistician, 29, 1975.

- I. Francis et al - Features of a statistical program or package: ratings of programs by their developers, Proceedings of the 41th session of the ISI, 1977.

- N.H. Nie et al - SPSS, Statistical package for the social sciences, second edition, McGraw Hill.

- S.H.J. Veling et al - Handleiding bij het statistische programmapakket PSTAT, Akzo Arnhem, april 1977.

- A. Verbeek - Ontbrekende scores bij het schatten van covariantie matrices, VVS-bulletin no. 3, maart 1979.

- L.Th. van der Weele et al - WESP, Waarlijk Eenvoudig Statistisch Pakket, RC-Publikatie 8, Rijksuniversiteit Groningen, november 1977.