Using Statistical software packages in Statistical Data Analysis

Mohammad Ehsanul Karim <wildscop@yahoo.com>,
Institute of Statistical Research and Training
University of Dhaka, Dhaka – 1000, Bangladesh

Many statisticians use several of the statistical packages at the same time. The core statistical capabilities exist in each of the packages. Each of the packages has its own strengths and ease of use features for the different types of analysis. We used some of the most popular statistical packages, which were available to us.

✓ *SAS*

Some features of SAS:

a) A very complete package for statistical analysis.

b) Can link to Access Database via ODBC.

c) Data can be put in several different locations including mainframes.

d) Can be run on several different platforms i.e. large mainframe computers down to PC's.

e) SAS programming is used as a means to access the data for several other purposes besides statistical analysis. SAS is an integrated suite of software tools used for purposes such as data warehousing, executive information systems, data visualization, application development, etc. besides statistical analysis.

f) We have a staff of highly experienced, dedicated SAS programmers around the world.

g) There is a long term commitment to dedicate staff to the SAS environment. There is probably nothing, which cannot be done with SAS but getting it done is always very difficult. There is so much literature in it that manuals take up entire shelf! May be that's why there are a lot of jobs out there for SAS programmers!

h) It is expensive (unless you get an academic discount). SAS has this trick. They sell you base SAS, which is quite limited. Then, they sell you SAS/STAT, and other modules. These prices add up.

i) SAS requires a lot of support, by someone who knows computers very well and knows SAS very well. It is a pain to learn. This is partially mitigated by the fact that SAS is very modular. Basically, there is the data step, where SAS data sets are made, manipulated and saved. Then there is a long list of procedures (called PROC's), which can be learned one at a time, for the most part.

j) The graphics in SAS are horrible. They came out with SAS/GRAPH, which has (theoretically) vast capability, but is almost impossible to use.

However, SAS appears to be more appropriate for an enterprise solution, where the data may reside in many different formats and SAS is the tool used to get at the data

and perform Statistical analysis. Where we are looking only for a single user PC somewhat limited use solution the other packages appear more appropriate.

✓ *SPSS*

Some features of SPSS:

a) SPSS is owned by the same company, which owns some other statistical package copyright such as SYSTAT; that is, this experience of statistical programming makes it more advanced and user friendly.

b) SPSS owes a lot to the days when the manual (the plum one) was one of the best introductory books around on statistics.

c) Non programmers find it easier to use than some other statistical packages; more menu driven versus programming

d) Slower performance

e) Tends to be used more in the professional field.

f) Training available everywhere.

g) It seems that SPSS is not the software of choice in industries that need data analyst. But SPSS has its audience too. It depends on our area of expertise.

SPSS and SAS

SPSS is a nice program for doing social research. SAS is a wonderful tool for doing data mining. They both have their place. I can tell you that in smaller databases, such as market research, SPSS is a far better tool for some quick and painless analysis. When you need to get some numbers flying around and are doing transformations like a maniac you had better know how to do some SAS. In terms of creating easily exportable, attractive output SPSS blows SAS away. It is not even close. SAS/Graph is well, not so good. The reason SAS is more valuable is because SPSS cannot really handle large databases very well. If I need to manipulate thousands and thousands of cases, SAS has the power and reliability to get you there. It also exists in a UNIX/mainframe environment. SPSS does not.

✓ *STATA*

Some features of STATA:

a) Interactive, very fast performance. Analysis can be done iteratively or through programming (called ado files). STATA loads the entire data set into RAM (which makes it much faster than SPSS in the normal case) which means the performance degrades considerably when data set size exceeds available memory. Whether this matters depends on the exact size of the data sets, and how much RAM you have, and how effectively you can subset the data. You may find that you can extract subsets with programs like stat/transfer (cheap and very effective) and avoid the problem.

b) Fast performance requires large memory, would need memory and possibly system upgrade. Model could be constrained by lack of memory.

c) Offers “NetCourses”; less expensive training accomplished via E-mail.

d) Some love the analysis tools and support, both from the STATA company and from other users on the STATALIST (many of whom are STATA employees who

answer questions and take input back into the company. Even the president of the company participates in their listserv.).

e) STATA does not play as nicely with other programs as SAS does. Also, STATA does not have many of the modules available in SAS.

f) STATA can handle extremely large datasets, both in terms of number of cases and variables. We can take a table of results from STATA and using the Copy Table command, paste results directly into Excel, which we can then very easily format into a Word document.

g) Updates are easy to install, and there are many very good user-written plug-ins available.

h) STATA's documentation is superior to those available from SAS.

i) STATA is much more competitively priced considering you get the entire package rather than getting pieces of it. They even have a deal where students can get a 160 page manual and a one year license for a negligible amount of money (Small STATA).

STATA and SPSS

STATA has many advantages from a statistics point of view, depending on what you are doing. They have excellent support for complex samples (cluster, stratified, weighted) and offer lots of statistics you can't get from SPSS--especially for limited dependent variables. SPSS is ahead of STATA in terms of the friendly interface and we think it is easier to do data management with SPSS. If you're looking for a single program to handle big data sets and do relatively simple analyses, SPSS is certainly up to the job.

✓ *MINITAB*

MINITAB is very quick to learn. Nearly 450 textbooks (up to now) and textbook supplements reference MINITAB Statistical Software, making it easy to use MINITAB in academic courses. There are a lot of macro's written for MINITAB. MINITAB Statistical Software is available for PC and Macintosh microcomputer systems. MINITAB is also available for mainframes, minicomputers, and workstations including VAX and other DEC platforms, Sun, IBM, Prime, Data General, Hewlett-Packard, and others.

MINITAB , SPSS, SAS

MINITAB is much more suited to statistical analysis than Excel, but definitely has some significant disadvantages compared to SAS or SPSS. It has a good interface and decent graphics, a reasonable macro facility, and decent support.

✓ *MATLAB*

MATLAB is a (non-statistics-specific) mathematical programming language which happens to contain some statistical routines. MATLAB, of course, is wide open in terms of adding new capabilities, but requires that you have the time, energy and knowledge to construct the necessary code. MATLAB is very powerful numerical computational package. We can do the same analysis in several ways.

✓ *S-PLUS and R*

S is considered a very high-level language and an environment for data analysis and graphics since its evolution in the mid-70's at Bell Labs. The evolution of the S language is characterized by four books by John Chambers and coauthors, which are also the primary references for S. There is a huge amount of user-contributed code for S.

S-PLUS is a value-added version of S sold by Insightful Corporation (previous MathSoft. Inc) since 1994 providing professional support to user-end. Based on the S language, S-PLUS provides functionality in a wide variety of areas, including robust regression, modern non-parametric regression, time series, survival analysis, multivariate analysis, classical statistical tests, quality control, and graphics drivers. Add-on modules add additional capabilities for wavelet analysis, spatial statistics, GARCH models, and design of experiments.

R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored inscript files. R was initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics of the University of Auckland in Auckland, New Zealand. The name suggests that the authors consider R to be a pre-stage of S. In addition, a large group of individuals has contributed to R by sending code and bug reports. The reason behind this might be its free availability including source codes. The design of R has been heavily influenced by two existing languages: Becker, Chambers & Wilks' S and Sussman's Scheme. Whereas the resulting language is very similar in appearance to S, the underlying implementation and semantics are derived from Scheme.

Since almost anything we can do in R has source code that we could port to S-PLUS with little effort there will never be much we can do in R that we couldn't do in S-PLUS if you wanted to. However, R is considered superior to S-PLUS in context of drawing graphs since several graphics features that R has; S-PLUS does not have.

All the Statistical Packages yields almost same results (number of decimal places or some decimal points may change at best) although it follows different algorithm in each cases. In addition, some more information is being provided in each package compared to others. Thus, according to need, users should choose their packages carefully[1].

[1] This document is a part of a chapter of one of my technical documentation that was intended to demonstrate some differences among the results of these packages for one particular Statistical analysis, which was first written in June, 2003. Since these stuffs are subject to change time to time, I would really appreciate being informed about any correction, comment, update, suggestion via electronic mail at <wildscop@yahoo.com> or <ehsan@isrt.ac.bd> with appropriate words in the subject field.