Review of Sociology Vol. 10 (2004) 2, 119–123

SURVIVAL KIT TO THE SPSS

Péter FUTÓ

Corvinus University, Department of Sociology and Social Policy
Budapest, Fővám tér 8. H-1093; e-mail: futo@freemail.hu

Székelyi, M. and Barna, I. (eds.): *Többváltozós elemzési technikákról társadalomkutatók számára.* [On the Multivariate Analytical Techniques for Social Science Researchers.] Budapest: Typotex, 2002.

Multivariate statistical analysis is a relatively new branch of science, the bulk of its theoretical apparatus was elaborated about two generations ago. By now it has been revealed that there is no use speaking separately about the methods of sociometry, psychometry, quantitative market research, biometry and technometry, for whichever method has proved to be good in one area can be used in another one, whereas whichever method proved to be a fad at the implementation in one field, would presumably be useless in the others as well. A universal auxiliary science has evolved and its spread has been greatly promoted by computer technology: while 15 to 20 years ago even leading researchers could have access with difficulty to analytical software running mostly on big computers, by now surveys by questionnaire, extending to over even tens of thousands of respondents, can be analyzed on a multitude of computers of universities, research establishments and of students even in Hungary.

During the past one and half decades the *SPSS program package* has overcome its competitors in Hungary, and it has become the most frequently used tool of multivariate statistical analysis in social science research, in experimental psychology, in market, media and public opinion polls, and even in the field of medical analyses. It is not only the answers of individuals or households that are recorded in the SPSS files. Economic research is also often based on questioning of thousands of companies, and political science draws from surveys covering hundreds of local governments, and currently most people use this software. Though in the workshops processing empirical surveys and experimental data generally specific internal professional jargons have developed and it frequently separates otherwise nearby areas from each other. Under these conditions the methods standardized by the SPSS and the SPSS files sent to each other function as interpreters between the various educational, research, and counseling institutions. Success was enhanced by the early recognition of the makers of the software that even the earlier versions had applied a far wider set of analytical tools and had offered many more statistical services than an average researcher would need. Therefore the subsequent versions of the past years have primarily, though not exclusively moved ahead in the direction of making their use increasingly user-friendly and their handling Windows-compatible.

By now preparing for the use of the SPSS has been solidly built into *higher education in social science*. At the same time the use of the software presupposes a far deeper statistical knowledge than the one a typical student of sociology would possess, when he or she gets acquainted with this software. With the spread of the student version of the program the SPSS is running on an increasing number of PCs, there is an increasing need for manuals that, being based on elemental statistical knowledge only, would show how a survey should be planned, and when it is done how to have the SPSS output tables made, interpreted and how to make them speak, to be understood by outsiders and what is the message of the data bases derived from surveys. Students however, who have been learning SPSS have not yet been pampered with literature in Hungarian. In addition to the leaflets of the company, so far the only useful publication was the guide and educational auxiliary material entitled The Basics of the SPSS for Windows program system written by Dr László Ketskeméty and Dr Lajos Izsó.

The *Survival Kit to the SPSS* is a long-needed assistance to learning and teaching, it is a systematized collection of analytical case studies which presents the chosen procedures mostly on real, but occasionally didactically created plies of data, that can be downloaded from the Internet. The three so-called data-reduction methods of the demonstration extend over: *principal component, factor and cluster analysis, and over six so-called explanatory models: variance analysis, linear regression analysis, path analysis, discriminant analysis, multidimensional scaling and logistic regression.*

The book is much more and at once much less than what its title suggests. It is more because the statistical analytical methods discussed here are in fact not linked to any specific software. The input databases, the computer commands and the outputs formally follow the rules of the SPSS, but other kinds of software available in the market may also be suited for the building and adjustment of the same models, and to answering to the same sociological questions. Such competing software like SAS, STATA and MINISTAT extend over most of the multivariate methods, though with a different syntax. Moreover, the most frequently applied models linear regression calculation and variance analysis – can be realized with Excel, the favored office software, though there it belongs to the most rarely applied services. On the other hand, the book is less than a functioning survival kit in the jungle of multivariate analysis, because it assumes that the most frequently used SPSS analytical procedures are known: practically all that may be the topic of an SPSS course for beginners.

With the help of examples. The authors know the average student of sociology who would patiently listen even to theoretical expositions, but ultimately would understand the framework, the possibilities and dangers of the choice and interpretation of the multivariate statistical methods with the help of specific numerical examples, colorful similes and well-constructed charts. Therefore the book presents the chosen methods with spectacular and fresh style, as if they were good old acquaintances, and speaks about them in a direct voice of the dissemination of knowledge as far as it is possible. Whenever it is possible the authors avoid mathematical detail that would become unintelligible to the targeted readers.

New explanations to statistical indices. A further merit of the Survival Kit is that its explanations present the meaning of a number of such statistical indices about which

the SPSS outputs and the explanations attached to the program (Help and Tutorial) speak only in a few words, if at all. Moreover, the authors themselves innovatively create straightforward indices in order to test the validity of calculations.

Mistakes. It would have been expedient to submit the manuscript to a more careful editorial control, in order to avoid minor mistakes and omissions. For instance, in the title of one of the chapters the concept of the 'Lazarsfeld paradigm' is mentioned, but the book does not explain what it is, neither states what the relationship of the material in the chapter is to the above-mentioned paradigm. Another similar lack of attention is that though a promise is made in the introduction that there would be a subject index attached to the book, one may look for it in vain, and there is no Subject Index at the end of the book either.

Problems of editing. In the introduction the authors indicate that the book should not be read continuously. And it is true that the models presented may be picked almost in any order, and their building, dismantling and reconstruction may be autonomously practiced with the SPSS demonstrative data bases that can be downloaded from the Internet. At the same time it is difficult to understand why the given order of the chapters and the discussed models is the one applied in the book. Perhaps it would have been more expedient to discuss first the more frequently applied regression analysisand variance analysis, and leave the more rarely used models and the ones that are more difficult to interpret to the second part of the book. It would have been the ordering of procedures by increasing level of difficulty. A further shortcoming of editing is indicated by the first chapter of the book: "The Useful Random Error". It is unintelligible for the average reader, and at any rate, it remains unexplained why this exposition was put there and what is its relationship to the other chapters? In addition the reader has to struggle through so many technical details that surely many of them would feel that they are being guided clumsily to the venue of the survival test. Whereas it would be a pity if, for this reason too many of them would drop out, for the important chapters that are easy to understand come right after this phase.

Narrowing the target readership. The most frequently used statistical service of SPSS is frequency calculation, cross-tabulation and linear correlation calculation. The authors assume that these models are known; they rely on them, but do not go deep into their refined aspects. It is visible from this and from other elementary statistical concepts considered as known ones that the reading public targeted by the book is not a beginner on the area of multivariate analysis. For many readers it will be difficult to pay parallel attention to the SPSS command syntax on the one side, and to the mathematical content of the models on the other side. It is true however, that in the Appendix the book explains how these commands are to be generated with the help of a menu system, but it does not alter the fact that the building of the models is done in the SPSS program language throughout the book. Meanwhile it is one of the biggest competitive advantages of the SPSS, in contrast to the other statistical software, that almost all of its services are accessible from the menu as well, and it is due to this quality that the user does not have to know even about the existence of this specific command language if he/she does not want to do so. This is the reason why an ever wider circle of students, used to Windows applications, would use this tool of statistical analysis with self-assurance. For the above reasons *those constitute the*

target group of the book who already know and apply the multivariate techniques on intermediate level.

The operationalization of doubt. The book is not absolutistic about its topic: it does not make it a secret even for a moment that the results obtained by the demonstrated methods would never give an irrevocable and final answer to the analyzed questions of social science. On the contrary, the book often suggests that the validity of the results obtained is limited; therefore it offers several methods to perform experiments testing and challenging the boundaries of validity. Perhaps these are the most valuable pieces of the survival kit offered in the title, and with their help one may 'experiment around' and scan the phenomena and tendencies examined in the multidimensional space. In every chapter of the book there are recipes of how one should doubt the results: one may learn in the case of every method presented how to use numerous tools of experimentation, available in the SPSS, or to be created with ingenuity, sometimes using detours, for the improvement, corroboration and refinement of results, or for their discarding. We learn to be alert so that if there may be trouble we may be able to fend off the always lurking dangers of saying nothing or, on the contrary, of reading too much into the results.

The set of experimental tools, however, is by far not complete: the book remains indebted with the systematization and enumeration of these procedures of validity testing. For instance, in addition to the procedures presented, it would have been the topic of further sensitivity testing to indicate how far the results of certain explanatory models and methods depend on the aggregation level of explanatory categorical variables. For instance, it is daily experience in the case of variance analysis and logistic regression that if the occupation of respondents is taken as one of the explanatory factors, the result, namely the explanation often depends on the fact whether the respondents were classified under 5 or 10 categories by occupation.

Where are the areas of quagmire? The authors give several practical rules of caution in relation to those models where the analyst has the greatest degree of freedom, which may be expressed by different parameters; and which therefore open the door widely to voluntary interpretations, such as in the case of factor and cluster analyses. Yet the reader does not receive a lucidly arranged map on which it is indicated where the quagmires are where one should dare to tread only if it is highly justified. In fact this is one of the most important rules of surviving devastating criticism that may be received for a too daring, or, on the contrary, too commonplace analysis. It is not the authors' fault that some methods have acquired the fame that their results are practically impossible to reproduce, in other words, while using the same method, two different researchers would only rarely produce identical results based on the analysis of the same database searching the answer to the same research question. At the same time, it could be expected to assess the offered models with the help of a uniform sensitivity test from the angle of the extent of the danger of reading too much into the results or of saying nothing.

Analysis: is it an algorithm or an art? Perhaps the book should have taken a more resolute stand that it is neither. Naturally, there is no exact recipe of applying multivariate statistical methods, which would consider all possibilities, which would prescribe the course of action as an 'expert's system' that can be programmed how to

proceed in the case of a given data base and research question – although there is a great demand for it among students of sociology. The other extreme is not true either, according to which the multivariate analysis is nothing else but a test of survival in a terrain difficult to survey, and only those may be successful who are able to apply flexibly the smart tricks permitted by well-informed masters. The truth is between the two extremes: multivariate methods are suitable for acquiring some insight into the structure of the data, but often they allow a too broad space for speculation. The best inference is promised if the researcher first arrives to preliminary insights by using descriptive methods, which should be subsequently supported in many ways by explanatory models, primarily by regression and variance analyses, experimenting with various schemes of re-coding and aggregation. During this process the researcher should frequently return to the original elementary data.

The multivariate analytical methods can be applied most effectively when they are adequately coupled with the tools of qualitative analysis. One should know what kind of questions are to be worded already before touching the empirical data. The nature of the correlations explored, and whether there are causal relations or merely 'moving together' may never be learned by studying solely the numerical results from the SPSS output. It is just these issues of integrating qualitative and quantitative methods that may be the subject of the next book to be written by the authors.

The *Survival Kit* will undoubtedly help lots of students and researchers of sociology in surviving the vicissitudes of multivariate analysis, and the libraries of university departments do the right thing if they order a number of copies at a time, for it will be much in demand.