Use R for Teaching Statistics in the Social Sciences?!

Wilhelm, Adalbert
Jacobs University Bremen, School of Humanities and Social Sciences
P.O.Box 750 561
28725 Bremen, Germany
E-mail: a.wilhelm@jacobs-university.de

1 Teaching statistics for social scientists

Teaching statistics to students in the social sciences is a challenging task. Instructors of stats classes in the social sciences are all too familiar with students' confessions of the type "unfortunately I was born without the math gene" or "whenever I see a formula, I panic immediately ". Over the last two decades, quantitative literacy projects all over the world have tried to strenghten analytical skills and numeracy among high school graduates. Thus, the background of quantitative knowledge and mathematical skills of social science students has become more and more heterogenous over the last couple of years. Statistics and methods classes now often have a mix of students with no prior stats education, little prior stats education, or a lot of background in statistics, see Rodgers & Manrique (1992). The more international the student body is, the more hetereogenous will be the quantitative background. A globalized world will become even more challenging for university teachers in this respect.

Statistics education without some computer program is hardly imaginable today. It is common practice to introduce statistical concepts by working with example data sets and stressing the interpretation of statistical results instead of focusing on the mathematical derivations. In many places, SPSS is the software of choice for this purpose. Although SPSS has itself established to be the standard statistical software package for the social sciences, it has never been designed as a teaching tool.

Despite the fact, that some students have gone through prior stats courses in their secondary education, there is however a unifying theme for social science students that brigdes their heterogeneity: no matter how much background they have in quantitative methods, the students, typically, lack enthusiasm to enter into the formal world of statistical models and mathematical derivations. They acknowledge the necessity of statistics, because they learn quickly that empirical studies are ubiquitous in the social sciences and that students need active knowledge on statistical methods to successfully accomplish their courses and develop expertise for their future jobs. But they see statistics mainly as a tool that is fundamental for much of empirical research. This understanding defines also students' attitude towards statistical software: their main concern is which menu to select, which option to choose, and which button to press to get the desired result. There is the clear danger of reducing the statistics education to an exclusively utilitarian endeavour, culminating in the students' saying: "I am not here to understand statistics, I'm only here to learn what to do with SPSS."

Quite a number of recommendations have been made in the literature to address the attitude issue of social science students towards statistics. Blalock (1987) mentions six general goals in teaching statistics, namely: (1) overcoming fears, resistance, and overmemorization; (2) stressing the importance of intellectual honesty and integrity; (3) understanding the relationship between deductive and inductive inferences; (4) learning to play the role of reasonable critic; (5) handling complexities systematically; and (6) familiarizing students with current statistically-oriented research and with problems encountered in presenting data to diverse audiences.

Using computers and statistical software contributes to all of these goals, but it mainly addresses goals number one and six. Computers and statistical software are used to ease necessary calculations and to familiarize students with the relevant computer output. They try to decrease the level of complexity, abstraction, and anxiety by introducing concrete problems and their solution strategies in a practical setting. But are computers and statistical software only effective tools to complete the empirical research project or can they be integrated in the learning cylce in such a way as to increase also students' understanding of statistics? Which specifications has a statstical software to fulfill in order to enhance the learning process? This paper will touch upon some important points in choosing a statistical software to be used in courses for the social sciences and particularly aims at a compariosn of SPSS and R (Hornik, 2007).

2 Some General Criteria

2.1 Hardware, License and Maintenance

Teaching computer labs have become state of the art over the last decade in almost all social science departments. In such a lab, students will be walked through a variety of basic computer skills and statistics issues. Working at a rather young university, brings it as a side-effect that there aren't any large scale central computing facilities, but every student has his/her own laptop. Students in the statistics classes will in their first lab course meeting install SPSS on their machines and bring their laptops to each session. The advantage of the laptop set-up is that students can play around with their SPSS installation at any time, they can try out various features and familiarize themselves with SPSS and statistical techniques as they like. No need to go to some central facility, no problems with opening hours, all-computers-currently-in-use, or similar issues.

The backside is that the installation process is time consuming. You need a fairly large number of installation CDs and there are always some individual problems with the specific computer set-up and the SPSS installation routine. Moreover, quite a number of concurrent installations have to be licensed as well as the fact that different SPSS versions are needed for various platforms. Also not to froget is the issue of maintenance. Finally, a computer lab facility provides a highly-controlled environment in which mistakes and problems can be diagnosed and remedied much easier, while individual laptops are used for many other activities that are likely to result in incompatibilities and computer crashes.

2.2 Data handling

Spreadsheets are very convenient tools for data handling and in particular for the novice user they are easy to learn and to operate. The intuitive handling of spreadsheets fosters an interactive approach to data analysis. Thus, a spreadsheet-like data view as in SPSS has a number of apparent benefits: it provides an easy overview of the number and type of variables, a quick and rough idea about the nature of the measurements and it instills some feeling of power into the students. The whole bunch of data is confined in one rectangular array making the notion of a data matrix very transparent.

R on the other hand is based on symbolic notation and one has to know the names of the various objects that one wants to address. This requires a certain familiarity and background knoweldge on the data set at hand. Something that is quite often not the case in a class room situation where different data sets might be chosen to accomodate the various backgrounds and interests of the students but also in order to illustrate particular features and behaviour of the statistical procedures. The idea of integrating R in Microsoft Excel provides a good way to combine the advantages of a spreadsheet with the flexibility of R, see Baier & Neuwirth (2007).

2.3 GUI vs. command-line interface

SPSS as many other software comes with a graphical user interface (GUI) with nice pull-down menus from which the appropriate analysis methods can be easily chosen. While GUI's can be straightforwardly used by anyone who is able to read, command-line interfaces constitute a barrier for the beginners because they require knowledge of a particular syntax. However, managing your way through a sequence of pull-down menus and pop-up windows to start the analysis you aim at is a major effort when using a GUI. Once you are familiar with the software, you'll easily find your way, but beginners often loose track and are bound to learn by trial and error. In the classroom, even when you have access to modern multi-media teaching labs, it is extremely difficult for students to keep track of the instructor clicking her path through the jungle of options and parameter choices. The 'Paste' command in SPSS provides a helpful tool to store the syntax of a command either in order to store the path of analysis or to ease performing repeatedly the same task with different data. However, efective use of this command is only possible if the user is acquainted with the SPSS syntax, something that will not be covered in a regular stats class for social scientists due to time constraints. Moreover, learning the syntax of a menu-driven software is regarded by most students as a complete waste of time and there is little intrinsic motivation of the students to learn some syntax.

In this respect, using a command-line program, such as R, seems to be preferable. Although the learning curve is very steep at the beginning, the students learn the commands step by step and have no reason to question the necessity of knowing the command names and possible otpions. Moreover, the whole analysis is under full control. In SPSS the output usually provides too much details and quite often stipulates questions that reach far beyond the currently learned material. For example, the Pearson correlation coefficient automatically comes with a significance value in SPSS which means that introducing the correlation coefficient as a descriptive technique usually results in the question what the signifcance value means, etc.

2.4 Commercial software vs. open-source

Comparing a commercial software product like SPSS with an open source project like R is an unfair comparison to start with. But it's not only the price of the license that speaks against the commercial solution. The R project with a core development team as backbone and many individual contributors can sustain software of the highest quality while at the same time offering innovative add-on packages for specialized analysis and modelling. Additional modules can be easily downloaded from the web and do not require a time consuming negotation neither with your university's financial department nor with the software vendor.

2.5 Programming language vs. statistical package

A statistical package, such as SPSS, is typically designed for the practitioner who needs a reliable tool for routine and standard tasks of data analysis. It ususally doesn't show any didactic components since the standard user is supposed to have a sufficient knowledge of the theoretical basis. SPSS is designed to serve the needs of a wide range of users, incorporating a plethora of statistical analysis techniques. To meet special needs the statistician must use extensible programming languages.

3 Dsecriptive Statistics and Graphics

The resurgence of statistical graphics since the early seventies was made possible by the event of powerful computer technology, but it was greatly enhanced by the rise of methods of exploratory data analysis that John W. Tukey pioneered. In Tukey (1972) as well as in the monograph *Exploratory Data Analysis* (Tukey, 1977) he introduced a rich palette of data displays that form a critical part of the methodology for exploring data.

In recent years much effort has been put into improving SPSS's graphics capabilities as well as the output quality, in particular to react to the evolution of exploratory data analysis. SPSS now offers two main strands of graphics production: the standard graphics (legacy graphs) and the interactive graphics. The available plot types range from bar charts, histograms, dot and box plots to scatter plot matrices. More specific plots for categorical data like spine plots or mosaic plots are not provided. Also user interaction with the graphics is very limited such that an interactive graphical analysis in the sense specified in Wilhelm (2005) is not possible with these tools.

R provides a broad toolkit of graphics and allows for easy modification and polishing of graphics. Over the last years, a few packages have been created that offer data exploration by user interaction with the graphs, see Urbanek & Theus (2003) and Wickham et al. (2007).

4 Statistical testing and modelling

The main focus of statistical education ofr the social sciences is concerned with statistical hypothesis testing, analysis of variance, and modeling by least squares regression. SPSS and R, both offer a similar suite of statistical procedures and at first glance there might be no difference between the two. However, there are at least three dimensions on which an evaluation can be oriented: first, the flexibility of the output. SPSS usually overburdens the novice user with too much output and it takes some time to instruct students on which parts they have to focus. In addition, there is a clear limit in the amount of options that is available within SPSS and the user can't really enhance the toolkit. Secondly, the modular setup of SPSS (which most likely is only a heritage of the development of SPSS) results in different realizations of identical or closely related procedures. Similarly, in R related procedures might be included in different packages. The differences in the realisations are sometimes subtle and can often only be understood when knowing the numerical details. Thirdly, the choice of default settings is quite often done with regard to either some specific context or a certain state-f-the-art at time of implementation.

The presentation will illustrate the aforementioned points by examples, demonstrate how R and SPSS address the issues differently, and how these differences are supportive in order to motivate and clarify statistical concepts.

REFERENCES (RÉFERENCES)

Baier, T. & Neuwirth, E. (2007), Excel :: COM :: R. *Computational Statistics*, 22 (1), 91-108.

Blalock, H.M. (1987), Some General Goals in Teaching Statistics, *Teaching Sociology*, 15, 164-172.

Hornik, K. (2007), The R FAQ, ISBN 3-900051-08-9, http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

Rodgers, P. H. & Manrique, C. (1992), The Dilemma of Teaching Political Science Research Methods: How Much Computers? How Much Statistics? How Much Methods?. *Political Science and Politics*, 25 (2), 234 – 237.

Tukey, J. W. (1972), Some graphic and semigraphic displays, *in* T. Bancroft, ed., 'Statistical Papers in Honor of George W. Snedecor', Iowa State University, pp. 293–316.

Tukey, J. W. (1977), *Explorative Data Analysis*, Addison–Wesley, Reading, MA.

Urbanek, S. & Theus, M. (2003), iPlots – High Interaction Graphics for R, in Hornik, K., Leisch, F. & Zeileis, A. (eds.) *Proceedings of the 3rd Intl. Workshop on Directions in Statistical Computing*, http://www.ci-tuwien.ac.at/Conferences/DSC-2003/

Wickham, H., Lawrence, M., Swayne, D., Cook, D., Lang, D.T., Buja, A. & Hofmann, H. (2007), The plumbing of interactive graphics. *Proceedings of the 5th Intl. Workshop on Directions in Statistical Computing*, http://had.co.nz/portfolio/2007-dsc-pipeline.pdf

Wilhelm, A.F.X. (2005). Interactive Statistical Graphics: the Paradigm of Linked Views. In C.R. Rao, E.J. Wegman, & J.L. Solka (Eds.), Handbook of Statistics, Volume 24: Data Mining and Data Visualization (pp. 437-538). Amsterdam: Elsevier.