Using R and SPSS simultaneously in teaching statistics: A tricky business

Vandewaetere, M., Van den Bussche, E. and Rosseel, Y.
Department of Data Analysis
Ghent University
Henri Dunantlaan 1, B-9000 Gent, Belgium.

Applied statistics and data analysis are crucial courses in a psychology department. The authors are involved in teaching introductory statistics and data analysis to both undergraduate and graduate students. Naturally, statistical software packages play an important role in our courses, aiding students in understanding and applying frequently used statistical issues. In most course material, theoretical sections describing a particular topic (say, ANOVA) are immediately followed by practical examples, with detailed instructions (even screenshots) on how the procedure should be carried out by a computer program, and how the output should be interpreted.

The statistical software package of choice in our department is, and will probably always remain, SPSS. However, given the growing popularity of R, we wanted to give our students the opportunity to use this non-commercial, free and open-source alternative. For various reasons, it was not possible to replace SPSS by R. Consequently we opted for the simultaneous use of SPSS and R in our computer practica, giving the students the choice of using the computer program they prefer.

We have been using R in this fashion for a few years now, and we have encountered major advantages, but also many difficulties in the simultaneous use of R and SPSS. In most cases, these difficulties are due to subtle statistical differences between the two packages, obvious to experienced statisticians, but often hard to explain to students in a psychology department with little mathematical background. Typical issues are the following:

- The '`anova`' command in R produces sequential (or Type I) sum of squares, while SPSS uses Type III sum of squares per default.
- Testing a linear hypothesis of the form $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{c}$ (where $\boldsymbol{\beta}$ is a vector of model parameters, and $\mathbf{L}$ is a hypothesis matrix), especially in a linear model with categorical predictors is a tricky business for the unexperienced user:
  - a custom command is not availabe in R (per default); it is available in several packages (eg the '`car`' package)
  - in the 'GLM' procedure in SPSS, redundant parameters (typically corresponding to the last level of a factor) are fixed to zero but still included in the parameter vector; in R, redundant parameters are dropped from the parameter vector. As a result, the parameter vectors differ in length, and the $\mathbf{L}$ matrix has a different number of columns in R and SPSS.
- Testing a linear hypothesis of the form $H_0 : \mathbf{L}\boldsymbol{\beta}\mathbf{M} = \mathbf{C}$ in the context of MANOVA or (the multivariate approach to) repeated measures is currently not possible in R without writing your own code.
- Several basic procedures (calculating the a priori power for t-tests, getting a correct one-sided p-value for a t-test, etc. ...) are currently not possible in SPSS.

The discrepancies between the two programs make it hard for us teachers to create concise and comprehensive course material integrating both software packages (SPSS and R). During the presentation, we will illustrate these issues with some practical examples, and discuss various ways of how we have dealt with them.