BISHOP GROSSETESTE UNIVERSITY

BG Research Online

Dunn, T.J. (2011) *The use of 'R' statistical software in psychology research.* PsyPAG Quarterly, 81. pp. 10-13.

This is the final, published version of a document / article published by The British Psychological Society in its final form on Decmber 1, 2011 at http://www.psypag.co.uk/

PSYPAG
Psychology Postgraduate Affairs Group
Quarterly
Issue 81 DECEMBER 2011

Recession-beating ways to research and collaborate

Tips on getting your first grant

Publish or perish

Also in this issue:

Making enough time in the day

Workshop, symposium and conference reviews

Using 'R' in psychology research

Thomas Dunn

The following article is designed to bring to the attention of researchers the possibilities of using statistical software other than SPSS, which has become something of a convention in psychology. It will briefly highlight the considerable advantages of using R (Ihaka & Gentleman, 1996) as well as offer directions for pursuing a more comprehensive 'R' education.

YOU MAY or may not have heard of R, but it is certainly a word, or more correctly a letter, being bandied around with some frequency within psychology departments. It is now becoming acknowledged as the *lingua franca* of statistical programming languages (Vance, 2009). For those of you unfamiliar with it, it is a statistics programme that relies on syntactical input by the user to manipulate and analyse data, known as a command-line interface. It is based on a very successful language called 'S' which was developed by John Chambers at Bell Labs in America (Chamber & Hastie, 1992). The major difference between R and existing popular programmes such as SPSS© is that the programming language itself is available to be manipulated directly (i.e. it can be seen by the user). This is in comparison to programs such as SPSS which have set functions (defined by 'pieces' of programming called 'code') that are carried out on the user's behalf when a button or tab is clicked (this is known as a GUI – a graphical user interface).

Over recent years there has been a growing trend, particularly in the social sciences, to switch from such user-interface based statistical programs such as SPSS to command-line based ones such as R (Fox & Andersen, 2005). It seems prudent to discuss why we have seen this increase and why researchers would want to switch from clearly laid out and aesthetically pleasing types of software to a program which, at first, appears to over-complicate analyses with various pieces of syntax and code.

Advantages of using 'R'

Cost and support

There are a number of advantages from using R as one's primary statistics program. Chiefly, R is open-source, which means it is available to anyone who wants it, for free. This means any change in working environment will not affect a researcher's access to a familiar and practiced statistics program. Additionally, because R is an extension of the highly commercial S programming language (Chamber & Hastie, 1992) it could be considered more commercially advantageous, particularly for people who are considering research-based work or perhaps applying their academic skills in a commercial domain. However, being open-source has more benefits than purely cost and job prospects. Firstly, open-source software tends to breed very comprehensive and friendly online support networks which are widely accessible (e.g. forums, websites, guides, etc.), and R is no exception to this. A number of core online resources are instantly available with plenty of additional and supplementary materials in the forms of guides, articles, papers, presentations, and lecture notes (see Resources below). This makes it easy to find a wide variety of different approaches and explanations to one's R objective, allowing one to choose a style to suit one's own understanding of a problem. This can be considered somewhat serendipitous when dealing with statistical problems, which can often be expressed in many guises. Secondly, being open-source means it is extremely easy to obtain and is available on many different platforms and

operating systems, such as Macintosh, Linux and Windows).

Flexibility

R employs a command-line (interpreter) whose job it is to communicate with the computer's operating system (OS). This means if one were to type '2+2' into R and hit enter, the computer would reply with '4' – simple. This enables the user to define their own functions which may be a likened to creating a personal button or tab in SPSS to perform the exact command intended. This is accomplished in R through the use of some basic functions and programming arguments which can easily be learnt from the resources available. Although at first this may appear to complicate the analytical process, it actually makes it much easier in the long term. The ability of R to accept the user's syntactical input is what lies at the heart of its raw simplicity; there is nothing unnecessary about the options presented to the user because there are none. It will only perform the types of analyses one is interested in. There are no options to confuse, as one would see in GUI-based programs, particularly where the options in such programs are labelled and defined by the software developers themselves. R on the other hand offers nothing to the user at the commencement of their R education, which means all procedures are understood from the ground-up. Thus, it can appear slow to get going on R but in the long term, understanding the 'language' of statistics through R allows one to converse rather creatively with the data.

Admittedly, creating specific functions in R via the source command (command-line interpreter) can be a somewhat arduous task for the novice programmer. However, R is yet to reveal its true *pièce de résistance* – 'packages'. R can be supplemented with additional programs that are included as 'packages' using the package manager. Most packages are directly available through the CRAN repository (see Resources below) and with the global appeal and application of R the probability of finding a package to suit your personal task demands are very high. Packages are specific add-ons which the user will need to choose and install themselves. Each one contains a discrete set of functions which have been tailored to suit 'types' of statistical methods usually employed within a particular domain. For example, a common package used for psychometric and personality-based research is 'psych' (developed at Northwestern University by William Revelle). This package includes a number of functions which organise and manipulate data derived from questionnaires as well as calculating many routinely used reliability estimators (i.e. Cronbach's Alpha and MacDonald's Omega) and factorial analyses. However, there are also a number of built in statistical functions that will also be very useful to psychologists, including the function 'lm' (linear model). This will simply carry out a regression analysis on whatever data the user specifies (see below for an example). The number of packages that are available to the R user is constantly expanding, with there now being somewhere in the region of 1,600 differing ones (Vance, 2009). Finding the appropriate package can easily be achieved via a simple Google search and is nearly all cases each package will be accompanied by a very comprehensible guide (usually in PDF format) illustrating the order of syntax and all analytical options available to the user.

The knock-on effect of such built-in flexibility is that the pure breadth of analyses which can be performed by R out-classes most other statistics software. This is not suggesting that R can do *everything* better than *any* other program, because some programs are specifically designed to perform one type of analyses and are very efficient at carrying out such tasks. However, it does mean that researchers do not need to swap and learn additional software to perform a different type of analysis they may not be initially familiar with. The basic 'arguments' and data manipulation skills that are acquired in R will remain constant for whatever analyses one chooses to perform.

Beyond these skills the only other necessity is to recognise which package is suitable for one's research requirements and become familiar with the order in which the syntax is entered.

Enhancing understanding of statistical processes

'Languages shape the way we think, and determine what we can think about' (Whorf, as cited in Maindonald, 2008, p.i)

The above quotation sums up that which is a very important aspect of using R over GUI-type statistics programs. R will ultimately enhance one's understanding of statistics itself due to its inline syntax method of input. The syntax is specifically designed to be reflective of the statistical formula behind the procedure. Therefore the R language, as a form of rhetoric (i.e. 'arguments', etc.), cogently spells out the process behind a statistical procedure. This convention involves the user manipulating variables in a rudimentary formulaic fashion. Thus, it is not surprising that by using R one is able to 'see' what is being performed on your data set. A good example of some R syntax which illustrates this nicely is that of a regression model with a dependent variable (A) predicting three independent variables (B, C, D). The R code for this is lm(A~B+C+D, data = your dataset) where 'lm' is the function 'linear model' which is already built in to 'R', and the symbol tilde '~' signifies predicting. Accordingly, this code is telling 'R' to model variables B, C and D as predictors of variable A (note: 'data =' is simply informing 'R' of the data set from which to retrieve the variables). This is very similar to the basic regression equation ($y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x \beta_3$) where y is the dependent variable and the x's represent the independent variables.

It seems that as researchers become more reliant on GUI-based programs such as SPSS there is a risk that the important underpinnings of statistics and method (i.e. the maths behind the stats) get lost through the process of repetitive 'button-pressing' analyses. As researchers we might often find that we employ the same 'type' of statistical analyses over and over again for certain pieces of research. Under these circumstances we are perhaps at risk of losing a basic understanding of exactly what it is that we are doing to the data. The impact of this may be twofold. Firstly, it can impact our interpretation of the statistical outcomes of our analyses. It will inevitably enhance one's interpretation of what the results mean in the real world if one has a better understanding of what has actually been 'done' to the data. Secondly, it can leave a researcher rigid in their statistical approach (i.e. forcing one's 'usual' methods to define the research

Resources

The Comprehensive R Archive Network (2011). R-project. Retrieved 15 July 2011, from http://cran.r-project.org/index.html (for downloading the R program; you will need 'base' if you are installing for the first time; packages can also be found here).

Verzani, J. (2002). simpleR – Using R for introductory Statistics. Retrieved 10 July 2011 from http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf (overall introduction to R, including how to get your data into the program).

The Personality Project (2011). Using R for psychological research: A simple guide to an elegant package. Retrieved 5 July 2011 from http://www.personality-project.org/r/ (excellent introduction to 'R' in psychology, particularly experiment and personality-based research).

Statmethods.net (2011). Quick-R accessing the power of R. Retrieved 7 July 2011 from http://www.statmethods.net/index.html (this is a great quick reference guide about how to perform basic t-tests and histograms to more advanced bootstrapping and probability plots).

Maindonald, J.H. (2008). Using R for aata analysis and graphics: Introduction, code and commentary. Retrieved 9 July 2011 from http://cran.r-project.org/doc/contrib/usingR.pdf (introduction to using graphics in 'R').

question, rather than allowing the question to dictate the method). In light of this, R offers a range of possibilities that researchers, particularly within psychology, can make good use of. Flexibility in quantitative methods allows one to take on more varied psychological endeavours, which are not constrained by stylistic rigidity.

Conclusion

As more researchers turn to R it is proving to be a most powerful weapon in the armoury of any quantitative psychologist. Admittedly it is presented in a very different and non-aesthetic form to what many GUI-based users are familiar with. However, function over form should be the priority of any researcher wishing to remain malleable in their endeavour to carry out rigorous scientific research within psychology. Importantly, the nature of the inline-command method of R encourages researchers to be more hands-on with their analysis and take an active interest in the statistical structure behind data manipulation. It is hoped that this article has illustrated that quite often program that look quite technical and complex can be very accessible and offer a flexibility and ultimate ease of use that alternative 'get-started-quick' programs (i.e. GUIs) can never offer. The intrinsic structural characteristics of the GUI force one to adhere to a software developer's constraints and eliminate any learning processes from the ground-up. R may be a little tricky to begin with, simply because it is unfamiliar, but a little bit of leg-work initially will allow for a much more fruitful approach to analysing quantitative data in the long run. Like learning any language, the basics can get you a long way and with the online support and masses of free supplementary material one can quickly increase your R competency very efficiently. Basic data set manipulation and understanding the functions and arguments of R will allow a researcher to begin constructing their own 'sentences' and express their research ideas in a much more eloquent and robust manner.

Correspondence

Thomas Dunn, Division of Psychology, Nottingham Trent University.
E-mail: Thomas.Dunn@ntu.ac.uk
http://nottinghamtrent.academia.edu/Thomas-Dunn.

References

Chamber, J.M. & Hastie, T.J. (1992). *Statistical models in S.* Pacific Grove CA: Wadsworth and Brooks Cole Advanced Books and Software.

Fox, J. & Andersen, R. (2005). *Using the R statistical computing environment to teach social statistics courses.* Department of Sociology, McMaster University.

Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*, 299–314.

Maindonald, J.H. (2008). *Using R for data analysis and graphics: Introduction, code and commentary.* Centre for Mathematics and Its Applications, Australian National University.

Vance, A. (2009). *Data analysts Captivated by R's Power.* New York Times, B6.