Issue 55: Spring 2009

UNIVERSITY OF SURREY

social research Update

R and Quantitative Data Analysis

Paul Webb
<paulwebb@praxiscare.org.uk>

Paul Webb is a Research Officer at Praxis Care—a major provider of services for adults and children with a learning disability, mental ill health, acquired brain injury and for older people, including people with dementia. Paul was previously a Research Officer at the Northern Ireland Association for the Care and Resettlement of Offenders (NIACRO).

- **R is a powerful and free statistical environment and programming language.**
- **R's extensive feature set can be extended by installing additional packages.**
- **R has an active user base and detailed documentation.**
- **R may be used as a supplement to or as a replacement for proprietary statistical programs.**

For many social researchers, quantitative data analysis is done by using the Statistical Package for the Social Sciences (SPSS). Although SPSS is powerful and comparatively easy to use, R (R Development Core Team, 2007) may be a possible alternative for researchers who do not have access to SPSS but who wish to do quantitative work. R may also be a viable alternative for current SPSS users.

R: Benefits

R is a powerful statistical environment and programming language which can be used to perform statistical analyses, to manipulate datasets and to produce high quality graphics (Murrell, 2005). Although initially quite difficult to use if migrating from an environment where data is manipulated via a Graphical User Interface (GUI) rather than from the command line, a competent R user can extend R's functionality via a series of packages or by writing code to implement a new procedure.

R is actively supported, can be learnt with reference to a plethora of free documentation (Hornik, 2008), integrates well with Excel and SPSS and can be used in the social sciences. Examples of packages which may complement the work of social scientists who use R include the "survey" package which is used to analyze data from complex surveys and the "psych" package for personality and psychological research. Another plus point is that R is a cross platform application and runs on Windows, Mac OS X and GNU/Linux.

R is also free which may be an attractive feature for those researchers who cannot afford a licence for proprietary software.

Obtaining R and other Programs

In order to obtain R, go to the Comprehensive R Archive Network (CRAN http://www.cran.r-project.org) and download the binary which is specific to your machine. If you have a Windows machine, download `R-a.b.c-win32.exe` from the base folder from one of the web sites near you (a.b.c in the name of the R executable corresponds to the version number). In order to install, double click on the executable and follow the instructions. Pre-compiled binaries are also available for GNU/Linux machines running Debian, SUSE, Ubuntu, and Redhat although installation instructions may vary depending on the Linux distribution. A universal binary can be obtained for users of Mac OS X.

R Commander

Although the R learning curve can be steep because R is driven from the console, it is possible to work with R by using the R Commander GUI. An installation guide for Windows, GNU/Linux and Mac OS X is available from http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/installation-notes.html. The R Commander interface should be more intuitive for users without experience of command line driven applications. You will see that the interface is divided into two main windows: the script window for R commands and scripts, and the output window for the results of your commands or scripts. R commands are executed by pressing the Submit button and you can choose sections of your code to submit by highlighting code with your mouse. If your R commands are incorrect, there is a third Messages window outputting helpful information. R Commander should be very intuitive for SPSS users but if you need more information go to the R Commander menu bar and select `Help/Introduction to the R Commander`.

SPSS R Integration Package

If you work with SPSS 16 or above, download the SPSS R Integration Package from http://www.spss.com/devcentral, extract the plugin from the folder with Winzip (http://www.winzip.com) or the equivalent for your platform and follow the installation instructions. The package is exceptionally powerful and includes the ability to generate pivot table output and to write results to new SPSS data sets. Go to the SPSS menu bar and select `Help/Programmability` for additional information.

XEmacs

An extensible editor like XEmacs may also be beneficial if you intend to write lengthy R code. You can obtain XEmacs itself, the Emacs Speaks Statistics Package (ESS) in order to extend XEmac's capabilities and a configuration file. These applications, together with information on setting up XEmacs for Windows, may be obtained by referring to http://socserv.mcmaster.ca/jfox/Books/Companion/ESS/. XEmacs binaries for Linux and for Mac OS X can be obtained from http://www.xemacs.org and http://www.macports.org respectively and you should follow the installation instructions for your operating system.

XEmacs may be thought of as a complete working environment inside which you can write and edit documents and issue and execute commands without having to open different software to perform separate tasks. XEmacs uses modes which means that some of its features may be toggled on or off depending on the work done. For our purposes, the most relevant mode when working with R is the ESS[S] mode which will be loaded automatically in the upper of two windows if opening an R file with file name extension `r`. The lower window is in iESS mode and within the window you should see the R prompt. The ESS Mode is useful because it toggles on special features like syntax colouring thereby making your code easier to read and to write. You can now open an R file in the upper window or type commands directly into it before passing the commands from the upper to the lower window by selecting from a range of buttons on the XEmacs menu bar.

Cameron et al (2004) give a comprehensive introduction to Emacs, much of which is relevant to the XEmacs editor.

With the exception of SPSS, all the software referred to in the article is free.

Using R: An Example

To use R, double click on the R icon on your desktop or type `R` at the command line prompt. If you are more accustomed to manipulating and analyzing your data by using a series of menus, the R interface seems sparse. This is because experienced users drive R by entering instructions in the area after the "greater than" sign or prompt which you will see on the R Console.

It is possible to enter data directly into R's Data Editor, but it is more likely that you will be working with data held in another format. This *Update* deals with reading and analyzing data which is stored in SPSS 'sav' format, but you can import data in text or comma separated value (CSV) formats and you should type `?read.table` or `?read.csv` at the R prompt for further information. R can read data in SPSS format, but it is necessary to extend R's base capability by downloading and installing the `foreign` package. In order to download the package from CRAN, select `Packages/Install Package(s)` from the R menu bar, your nearest CRAN mirror and the "foreign" package. Then load the package by selecting `Packages/Load Package` from the R menu bar. If your copy of R does not come bundled with a GUI, type `install.packages("foreign")` and

`library(foreign)` at the R prompt in order to load the package.

If you check which packages are available on your system by typing `search()` at the prompt, you should see `package:foreign` in the list.

We are now in a position to read an SPSS 'sav' file into R and conduct an initial exploration of some data. In order to illustrate the process, we are going to use data obtained from Praxis Care's Staff Survey which is a triennial omnibus survey of health professionals who primarily work with people with mental health problems and/or learning difficulties.

Commands which you have to type are in **`bold`** and output, where generated, is shown in `thus`.

Our file may be read into R by using the `read.spss` function.

```
staffdata<-read.spss('\\path_to_file\\staffdata.sav', to.data.frame=TRUE, reencode=NA, use.missings=99)
```

The `read.spss` command loads a file with the name and pathname which we have specified into a dataset called "staffdata". (The way in which you describe the path to the file will differ depending on your operating system). In addition to the path to the file in the parentheses, there are three further options. We first set the `to.data.frame` option to TRUE in order to create a data set. Missing values in the SPSS file have been coded to 99 but R assigns missing values to `NA` so we have set the last two options so that R will remap any instance of 99 in the SPSS file to `NA` in the new R file.
We will now create a second dataset with the subset command by selecting three variables from staffdata. The second dataset will be called "sample".

```
sample<-subset(staffdata, select=c(Q1g,Q10a,Q13))
```

We use the `attach` command to attach the sample dataset to R's path so that we can refer to variables within the dataset without a full pathname. `names(sample)` outputs a list of variables within the sample dataset.

```
attach(sample)
names(sample)
"Q1g"   "Q10a"  "Q13"
```

Q1g asks respondents to rate their response to the statement "I feel I am happy helping people in my job" on the scale Strongly Agree, Agree, Disagree, Strongly Disagree.

Q10a asks respondents to rate their commitment to their job on a ten point scale where 10 represents "as committed as I could be" and 1 represents "not at all committed."

Q13 refers to the respondents' main working environment, divided into two categories: Central Office or administrative staff and Scheme or clinical staff who work with clients.

If we want to obtain a frequency count of the number of workers by working environment, we use the `table` command

```
table (Q13)
Central office as base           Scheme as base
                    43                      398
```

We can produce a barchart with the `barplot` command and a number of options which create a title, a labelled y axis and hatched bars enclosed within a blue border. The barplot will be saved to the specified directory as a PNG file.

```
png(filename='\\path_to_file\\barQ13.png')
barplot(table(Q13), main='Where do you work?', ylab='Frequency',
density=c(10,20), border='blue', ylim=c(0,400))
dev.off()
```

The `table` command may also be used to cross tabulate two variables by enclosing both variable names within parentheses:

```
table(Q1g, Q13)
                    Q13
Q1g                 Central office as base        Scheme as base
  strongly disagree                      0                     2
  disagree                               1                    11
  agree                                 30                   188
  strongly agree                        11                   196
```

To create a table which is easier to read, we can calculate column percentages with the `prop.table` command where the second argument to `prop.table` may be '1' for 'proportions of row totals' or '2' for 'proportions of column totals.'

```
colpercents<-table(Q1g,Q13)
prop.table(colpercents,2)*100
                    Q13
Q1g                Central office as base          Scheme as base
  strongly disagree              0.0000000               0.5037783
  disagree                       2.3809524               2.7707809
  agree                         71.4285714              47.3551637
  strongly agree                26.1904762              49.3702771
```

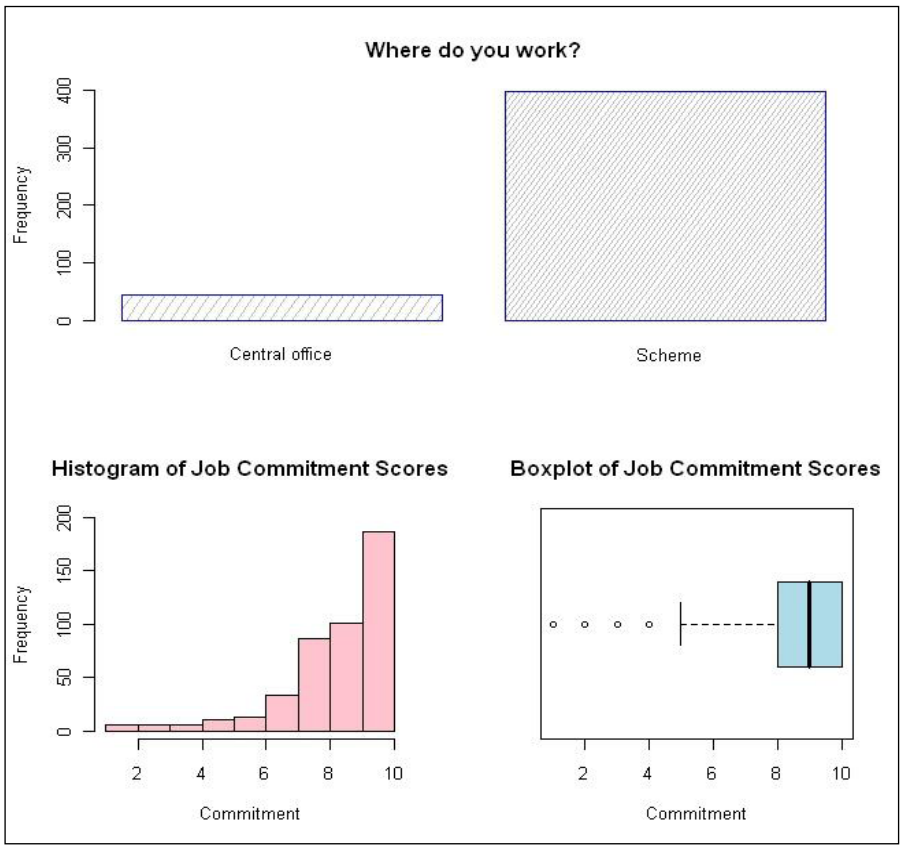
We may also generate frequency counts for the continuous variable Q10a and list the minimum, first quartile, median, mean, third quartile and maximum values by using the `summary` command before charting the distribution of Q10a with the `hist` and `boxplot` commands

```
table(Q10a)
Q10a
  1   2   3   4   5   6   7   8   9  10
  4   1   6   6  10  13  33  86 101 187
summary(Q10a)
Min.    1st Qu.  Median    Mean    3rd Qu.    Max.
1.000   8.000    9.000     8.667   10.000     10.000
```

```
png(filename='\\path_to_file\\histQ10a.png')
hist(Q10a,main='Histogram of Job Commitment Scores',
xlab='Commitment', ylim=c(0,200), col=c('pink'))
dev.off()
```

The `hist` command produces a histogram for Q10a with a title, pink columns and a labelled x axis.

```
png(filename='\\path_to_file\\boxQ10a.png')
boxplot((Q10a),main='Boxplot of Job Commitment  Scores',xlab='Commitm
ent',col=c('lightblue'),horizontal=TRUE)
dev.off()
```



The `boxplot` command produces a light blue boxplot which is rotated horizontally because the option for horizontal is toggled to `TRUE`.

During the course of the session, we have generated three charts which have been saved to a specified directory. Textual output from the R console may be copied and pasted into an application of your choice.

In order to end the session, we now remove our created variables, detach datasets so that the dataset variables are no longer accessible to R and exit the application by using the `q()` command.

```
detach(sample)
rm(list=ls())
q()
```

Conclusion

R may seem to be a difficult application to get to grips with but any effort expended will be amply rewarded by R's power and its ability to work seamlessly with a number of equally powerful applications.

References

Cameron, D., Rosenblatt, B., Raymond, E. (2004) *Learning GNU Emacs*. O'Reilly and Associates, Sebastopol, CA.

Hornik, K, (2008) *The R FAQ*. http://CRAN.R-project.org/doc/FAQ/R-FAQ.html.

Murrell, P (2005) *R Graphics*. Chapman and Hall, Boca Raton, FL.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

social research UPDATE is distributed without charge on request to social researchers in the United Kingdom by the Department of Sociology at the University of Surrey as part of its commitment to supporting social research training and development.

Contributions to *social research UPDATE* that review current issues in social research and methodology in about 2,500 words are welcome. All UPDATE articles are peer-reviewed.

social research UPDATE (ISSN: 1360-7898) is published by the Department of Sociology, University of Surrey
Guildford GU2 7XH, United Kingdom. tel: +44 (0)1483689450
Edited by Nigel Gilbert (n.gilbert@surrey.ac.uk)