A subtle introduction to R

Jorge Cimentada & Basilio Moreno Peralta
Research and Expertise Centre for Survey Methodology (RECSM)

Course description

R is a free programming language designed to apply statistical analysis and produce data visualization. Among its endless capabilities, R is famous among statistical software for its flexibility in creating functions, packages, data visualization, for being open source and for offering a wide range of statistical tools. The aim of this seminar is to introduce you to the R language and give you hands-on experience on how to use it.

R has a steep learning curve. This means that it takes a lot of work and persistence to achieve fluency. Just as when you learn a new language, the process requires practice, repetition and familiarity. It is impossible to learn the necessary R skills that will help you achieve your day-to-data tasks in a single class. For that reason, this seminar will concentrate on two core objectives: understanding the building blocks of R and learn to understand R code. The former will allow you to start writing code right away, and in addition, it will give you the basic fundamentals to understand the latter.

Because many of the seminars at the RECSM summer school will be conducted in R, we want to make you sure you understand the most in the least time possible. For that reason, we'll spend a great deal of time reading and adapting R code. We'll do that by presenting you with actual R code from previous RECSM seminars, running through each step and adapting it to our own needs.

The seminar will consist of two sessions. In the morning session we will spend most of the time understanding the basics of R while we use some toy datasets. The syllabus for the first session is the following:

- Basic objects: Vectors
  - What are vectors?
  - Types of vectors
  - Generation of regular sequences
  - Operations with R objects

- Basic objects: Matrices and lists
  - Creation and modification of matrices
  - Creation and modification of lists
  - Selecting rows and columns in matrices
  - Selecting slots in lists

- Basic objects: Data frames
  - Creation of data frames

- Basic data manipulation:

- Importing data from SPSS, EXCEL and STATA.
- Creating new variables
- Recoding variables
- Renaming variables
- Selecting and removing variables
- Filtering and subsetting variables
- Grouping variables
- Merging data frames

- Introduction to functions:

- Functions as programming objects
- Defining user-made functions
- Already-made functions: packages.

- Introduction to loops:

- For loops.
- “If” statements.
- “If else” statements.

In the afternoon section we will go through 2 or 3 R scripts taken from actual seminars that professors will use. Here you will be able to practice what you learn. As an exercise, the scripts will be mined with errors already programmed by us. Your task will be to read the files, understand what the script is trying to do and fix any errors that are encountered in the process. The errors will not simply be semantical, but you will have to get your hands dirty and reason about the problem at hand.

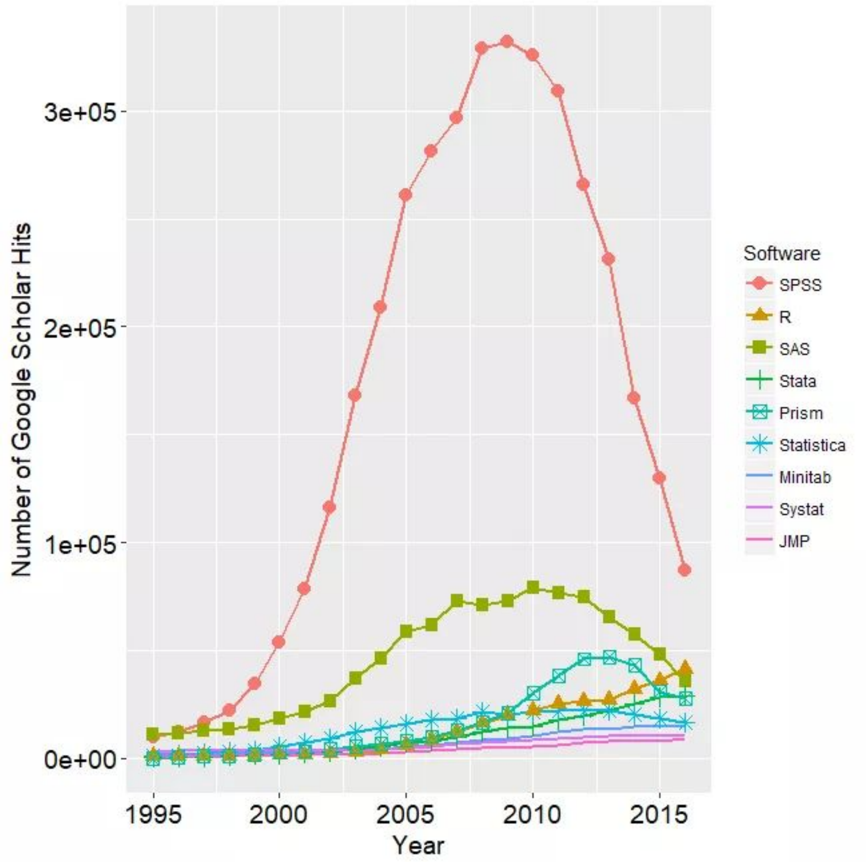
As an additional resource, we have programmed a series of exercises which you can complete outside the course. The exercises can be accessed here with an explanation on how to install R and Rstudio. We recommend you complete these exercises whenever you can before going into your other seminars. The exercises can be used for both learning and refreshing your skills from the seminar.

Why should you attend the course?

R is one of the biggest statistical programming language with an estimated community of between 250,000 to 2 million users. As you can imagine, R forums are packed with questions, tutorials and guides[1].

Have a look at the number of academic articles using different software.

[1] http://www.r-bloggers.com/r-is-the-fastest-growing-language-on-stackoverflow/



The graph above shows that R has been growing exponentially since 2007 and there's no sign of slowing down. One thing we can take the risk of inferring from this graph is that the more a software is being used for academic papers, the more the students are going to be educated with that software. Please notice that SPSS is used much more than all of these different software and was excluded from this graph. However, recent statistics show that the usage of SPSS among academics is decreasing steeply while R became the second most common statistical language in 2016 in a clear upward trend.

On top of everything else, R is completely free and will not bring any licensing costs to anyone ever!

Here we have listed some of the reasons why we think you should consider learning R.

- To augment the commercial software you are used to; i.e. to be able to perform statistical analyses that are not available in other software, but which are available in R
- For example: Sequence Analysis, Latent Class Analysis and Bayesian Modelling are some of the techniques which were never developed for Stata and only recently there have been updates that allow the usage of these techniques. In R, they've been around before 2006. Also, tools for machine learning techniques are very limited in Stata.[2]

[2] For for those interested in comparing all the techniques available in R, SPSS, Stata and SAS, visit this website: http://stanfordphd.com/Statistical_Software.html

- Everyone wants to stay up-to-date. Leading statisticians which are developing new techniques are almost always using R

- Taking advantage of best of both worlds: cleaning data, for example, in Stata, and performing analysis in R

- R is directly accessible from SPSS, SAS and Stata (this last one is still under work.)

- If you go out of academia or conduct independent research, R is free and always will be.

- It can make your life easier! You can automate and design exercises, exams, tests, classes, reports and PDF presentations within R.

- Students can access exams online in R from wherever they are. There is even auto grading where you don't have to revise these exams. (Swirl package)

Software

We will be using the R software together with the Rstudio interface. No laptop is required as the seminars will take place in the RECSM facilities. Any packages we plan to use will be already downloaded previous to the session.

Prerequisites

- No knowledge of any statistical software is required. We know some of you will be familiar with other commercial software such as Stata, SPSS or SAS, but regardless of this, we will teach the contents starting with the basics. Do not worry if you're unfamiliar with other statistical software.

- Minimal statistical background is recommended. We will not touch upon formal statistics in the course, but as R is a statistical programming language we will use examples that contain concepts such as mean, median and basic descriptive statistics. No interpretation is needed but some familiarity is recommended.

Schedule

Time	**Topic**
09.15-10.45	Introduction to R objects
10.45-11.15	*Break*
11.15-13.00	The basics of data manipulation (subsetting and functions)
13.00-14.00	*Lunch break*
14.00-15.45	Introduction to R functions and loops
15.45-16.15	*Break*
16.15-18.00	Special cases of R functions and data import

References

These books are really useful to understand basic R structures as well as advanced R code:

- Kabacoff, Robert. R in Action: Data Analysis and Graphics with R. Manning Publications Co., 2015.

- Matloff, Norman. The Art of R programming: A tour of statistical software design. No Starch Press, 2011.

- Muenchen, Robert A. R for SAS and SPSS users. Springer Science & Business Media, 2011.

Aside from these books, there are hundreds of free resources online. We point you towards the ones we find more interesting:

- http://swirlstats.com/students.html
- https://www.datacamp.com/courses/free-introduction-to-r (This resource has a fee but many of its lessons have free chapters. If you're planning to invest in some online resource, we definitely suggest this one.)
- https://www.edx.org/course/introduction-r-data-science-microsoft-dat204x-2
- https://www.coursera.org/learn/r-programming
- http://tryr.codeschool.com/ (now PluralSight)

About the instructors:

Jorge Cimentada is a Data Scientist at *Kernel Analytics* and a PhD candidate in Sociology at Pompeu Fabra University. He belongs to the Research and Expertise Centre for Survey Methodology (RECSM) and his main research interests are the study of achievement inequality, the role of schools in reproducing these inequalities and early education as a remedy to achievement inequalities. He's developed several R packages and is very passionate about statistical programming with R, data visualization and statistics. You can check out his blog at cimentadaj.github.io and if you wish to contact him, send him an email to cimentadaj@gmail.com or Github.

Basilio Moreno Peralta is a Data Analyst at *Rebold* and MA in Political Science Research by the Pompeu Fabra University. He holds a Bachelor's degree in Political Science from Universidad Pablo de Olavide and the University College of London. He also took part at the Research and Expertise Centre for Survey Methodology (RECSM). Basilio has been working on different projects using R in order to manage survey data from non-standardised sources or to predict electoral outcomes with Bayesian modelling. If you wish to contact him, you can send him a message through email at basilio.moreno@upf.edu, GitHub or Twitter.