Data Analysis with R

Rodrigo Rodrigues-Silveira

Introduction

The adoption of R is becoming a new standard in the social sciences around the world. There are some excellent reasons for this wide diffusion of this "tool without aesthetical attraction," to put it in some gentle words. Firstly, it is free and open-source. Proprietary licenses from most of the solutions used by social scientists are expensive, although there are some discounts for students. The open-source nature of R makes innovation easier and faster. In SPSS or Stata, for instance, users need to wait for new versions to assess new methods. So, they offer a small room for customized solutions. Just the opposite occurs in R, with customizable packages and user-made functions.

Secondly, R is versatile. You can perform both quantitative and qualitative data analysis, data visualization. You can also be benefited by many functionalities that usually go beyond statistical packages and are typical of programming languages. Some possibilities available are discourse or content analysis, text mining, statistical analysis, image and sound processing, maps. R also includes many other applications that would be virtually impossible to achieve by the traditional statistical packages available.

Thirdly, and, maybe, most importantly, R provides a native platform for replication. The combination of R code with markdown and latex languages, offers a productive alternative to face the increasing need to document datasets and analytical procedures, both in academia and in firms.[1] Top Sociology and Political Science journals are requiring more and more that authors provide, alongside their manuscripts, the original datasets and the scripts employed to perform the analysis. These requirements help reviewers and readers to understand the analytical process in detail, reproducing the results and expanding from a more solid basis.

The tool is capable of handling the data analysis needs from a small data project by a graduate student to large international corporations with terabytes of data. One would wonder, what is the trick? Why there are so few people working with R? The first to blame is the script. People get nervous around a black screen with a blinking bar. It goes beyond rational. R is ugly as hell, and he needs to be programmed to work using a language that is all but intuitive.

Our purpose is to provide a review of fundamental aspects of R and a programming language in general. To lose the fear of R, we will understand what a script is and how it works. Next, we continue with fundamental tools as data opening, cleaning, and filtering, followed by

[1] By the way, Microsoft has launched its version of R (the Microsoft R) that is integrated as part of SQL Server, connecting the power of relational databases of terabytes to a language built to facilitate both data analysis and visualization.

applied (systems of) concepts such as the grammar of graphics, textual analysis, and replicability.

We divide the course into two days and seven sessions. The first day introduces some basic features of R and makes a refreshment for those who already are familiar with the program. The second day focuses on more applied techniques widely employed in the Social Sciences, such as model building, content analysis, and replication.

You can find the full contents of this brief introductory course below.

Contents

The topics of this brief course are structured as follows:

Day 1

- Making the most of R (2 hours)
- Data: types, files, and filtering (2 hours)
- Working with survey data in R (1 hour)
- The grammar of graphics (1 hour)

Day 2

- Model estimation in R: OLS, Probit and Logit (1 hour)
- Content Analysis in R: Coding and Sentiment Analysis (2 hours)
- Replicability and productivity using rmarkdown (2 hour)

Detailed structure of the course

Making the most of R

- The R working environment
- Working with scripts
- Functions
- Conditionalities
- Repetitions

Working with data

- Data types
- Opening data files (xlsx, csv, fixed width, sav, dta, json, xml, etc.)
- Filtering and selecting data

Surveys in R

- The discreet charm of weights
- Tabulations, ranks and means
- Visualizing different scales

The grammar of graphics

- How to create an effective graph
- Visual elements and how to join them in a chart
- The right chart for my precious data

Model estimation in R

- When to use a OLS model
- Logit and Probit models

Content analysis in R

- Corpora
- Tokenization
- Data cleaning
- Coding
- Sentiment analysis

Replicability and productivity

- Replicate, Replicate
- Rmarkdown basics
- Using templates

Contact

Rodrigo Rodrigues-Silveira
University of Salamanca
rodrodr@gmail.com