2nd ECPR Winter School in Methods and Techniques, 17-22 February 2013
University of Vienna, Austria
Course Description Form

Course title

C1. Introduction to R

Instructor details

First name, last name: **Zoltán Fazekas**
Department/Unit: Department of Political Science and Public Management
Institution: University of Southern Denmark
Full postal address for ECPR correspondence: Campusvej 55, 5230 Odense M, Denmark
Phone: +45-6550 4495
Fax: +45 6550 2280
E-mail : zoltan.fazekas@gmail.com

Short Bio

Zoltán Fazekas is a Post-Doctoral Researcher in Political Behavior and Individual Differences at the Department of Political Science and Public Management, University of Southern Denmark. He earned his PhD in political science at the Department of Methods in the Social Sciences, University of Vienna where he was an Early Stage Researcher in the Marie Curie Initial Training Network in Electoral Democracy (ELECDEM). He holds a BA in Economics, MA in European Affairs and an MA in Political Science. His fields of interest are: political psychology, quantitative methods and statistics. Among others, his work appeared in the Journal of Theoretical Politics, Social Science Quarterly and European Journal of Industrial Relations. He taught introductory statistics courses and offered consultancy on using R for quantitative research. He prefers simultaneous estimation, truly dislikes heteroskedasticity and the quest for R-square maximization.

Short course outline

This introductory R class focuses on making students familiar with data manipulation and basic modelling in R. R has become a widespread statistical package that offers both flexibility and power, but it also needs careful attention by the user. After this course, students should able to (1) set up R, load in any data formats and prepare them for analysis; (2) carry out descriptive and inferential analysis (up to the most commonly used regression models) on the prepared data; (3) write R script that incorporates simple custom functions and the split-apply-combine strategy; (4) specify and run simple statistical models and present their results both in tabular, but mostly visual formats that are both informative and aesthetically pleasing (publication ready); (5) easily navigate through the myriad of further resources on R and extend their knowledge on their own efficiently. This course is designed to be a hands-on practical introduction to R for quantitative researchers.

Long course outline

Any course description about R starts with pointing out the advantages of using R. To be fair, if you – as participant – are interested in this course, it means that you either want to use R or you have to use R. In both cases, there is a good reason to start setting up your research workflow around R, an open-source (and hence free) statistical package that offers easy implementation of a vast array of statistical procedures and models. Along the accessibility, it is also the most rapidly growing statistical package with the number of additional implementations growing day-by-day. Furthermore, R offers nice integration with text editors, making it easy to carry out writing and analysis in the same framework.

The other most frequent description of R is a negative one: it has a steep learning curve, the error messages are unintelligible, and some people on the R forums are very condescending. Normally,

deciding whether one wants (or needs) to use R weighs these pros and cons. Throughout this course we will not present R as the best possible solution for everything, but we will try to use R in a manner that also informs how we generally look at our data, describe it, and ultimately analyse it.

After completing this course students should be able to:

1. Comfortably set up R, load in any data formats and prepare them for analysis.
2. Carry out descriptive and inferential analysis (up to the most commonly used regression) on the prepared data.
3. Write R script that incorporates custom functions, loops, or the split-apply-combine strategy.
4. Specify statistical models and present their results both in tabular, but mostly visual formats that are both informative and aesthetically pleasing (publication ready).
5. Easily navigate through the myriad of further resources on R and extend their knowledge on their own efficiently.

In order to meet these goals we will spend 5 days on building systematic knowledge about how R thinks and works and how students should make the best out of this setup of the R language. Our first day will offer a comprehensive overview of how one installs, starts and customizes R for an easy use. We will work in R GUI alternating between the script editor and the console. We will cover how one should write syntax (with meaningful comments), what are R packages, and how symbols and objects differ and operate. On the second day we turn our attention to data, from the point of loading (or creating) data, accessing elements of the data frame, handling and recoding different types of variables, and finally looking at summary descriptive statistics for different variable types.

Building on this general background we will designate our third day to things that make life easier when working with R. Writing simple customized functions is the first step, as most of the operations we want to carry out repeatedly are based on multiple pre-written functions in R. Furthermore, it is customary that we need to carry out the same transformations multiple times or only under certain conditions. R has the great advantage of working with vectorized structures and hence we will learn to use the apply() function family and its extensions that make data reformatting efficient, clean, and fast (the split-apply-combine strategy). These are very basic functions in R that work far better than loops. Avoiding the use of these simple functions can be the reason for some frustration with R.

The fourth day is fully designated to modelling. We will cover linear regression and models for dichotomous. All these examples will be expressed in the generic lm() and glm() forms, but there are easier ways to deal with these relatively simple models. For this, we will go through **zelig**, an extremely flexible and overarching package that offer intuitive specifications and further simulation procedures for easily understandable results. These are easy and essential tools for those who want to communicate their results in substantive terms.

Our last day will focus on one of the main advantages of R: visual display. A well-designed plot might tell your readers more than a crowded table. Also, it is essential if there is a more general public that one aims to communicate with. There are many possibilities to produce figures that are both aesthetically pleasing and informative. In our course we will use the **ggplot2** package that is built on the principles of the *grammar of graphics*. Its power lies in following the communication aims that one has in mind, but also its tremendous flexibility. Last, but not least it produces informative plots that look very nice and fit well with the rest of the text in a paper. We will use these tools for both descriptive and inferential statistics.

Our meetings are set up as follows. I will use a split screen that incorporates a presentation material (with code) and R window in which I will go through the script accompanying the analysis. All these materials will be available before the start of the course. Students are required to follow and implement the examples that we are discussing on their computer. Generally, we will spend 5-10 minutes on one sub-topic on the slides and equal amount of time on the review of the practical implementation. At the end of each lecture slightly modified examples will be given to students to

complete them on their own. Ideally, students are required to spend an additional 30 minutes on reviewing the examples and completing the example codes. These will be reviewed in the first 10 minutes of the next lecture. At the end of the course I will supply all correct solutions.

Students who take this course because they are interested in a specific implementation in R (usually more complex models) should contact me either before the lecture or latest on the first course day. I am available for separate meetings on very specific questions related to particular models or solutions in R for a problem. We can go through the available packages or solutions in R. In this case students are required to prepare their data (and research questions) prior to the meeting.

Day-to-day schedule (Monday 18 February to Friday 22 February)

	Topic(s)	**Details [NB : incl. timing of lecture v/s lab or fieldwork etc. hours]**
Day 1	Setting up R, packages, syntax, objects and symbols	An introduction of how R works and what are the key elements for setting up an easy working environment in R. After these general aspects we will cover the elements of the R language that are essential for understanding how R is set up and how one should carry out quantitative analysis in R.
Day 2	Loading in data, working with data frames in R, variables and descriptive statistics	We will cover how to load in various data formats into R (.csv, .dta, .spss, .tab delimited) that include already predefined variables. We will go through how data frames are set up and how one can access its elements and what are the basic variable types. Finally, we review basic descriptive statistics and handling of missing data in these operations.
Day 3	Custom functions, loops, and the split-apply-combine principle for data handling	We will cover the generic rules of function writing in R in order to increase modularity in the workflow. We will introduce a powerful application of working with data that relies on split-apply-combine strategy implemented in the **plyr** package.
Day 4	Modelling in R using **zelig**	We will cover simple OLS linear regression and logit specifications. We will rely on the **zelig** package for this because it is easy and intuitive when it comes to the results.
Day 5	Visual display of analysis: **ggplot2** for descriptive statistics and model results	We will go through the principles of good visual display of descriptive results from histograms to more informative scatter plots. After that we will cover visual representations of model results: coefficient plots, slopes with confidence intervals, marginal plots.

Day-to-day reading list

	Readings (please list at least the compulsory reading for the scheduled day)
Day 1	Adler, Joseph. (2009/2010). R in a Nutshell. O'Reilly Media. Chapters 1,2,5,6,7 (Total number of pages: 68) [Adler, 2010 from now on] and Lecture Notes.
Day 2	Adler, 2010 Chapters 12,13 (Total number of pages: 64) and Lecture Notes.
Day 3	Adler, 2010 Chapter 9, but mostly Lecture Notes for this course. Hadley Wickham. 2011. The Split-Apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1): 1–29, April 2011.
Day 4	Kosuke Imai, Gary King, and Olivia Lau. 2008. "Toward A Common Framework for Statistical Analysis and Development" *Journal of Computational and Graphical*

	Statistics, Vol. 17, No. 4 (December), pp. 892-913 [as general reading] Zelig user manual (http://r.iq.harvard.edu/docs/zelig.pdf) pages 7-53. Lecture notes.
Day 5	Lecture notes and ggplot2 website (http://had.co.nz/ggplot2/).

Requested prior knowledge

As this is an introductory R course, no prior knowledge with R is expected. Students are required to be familiar with statistics at the introductory level (including multivariate regression) and have some prior (applied) experience with data manipulation and analysis.

Software/hardware information

This course will be taught in R, preferably in the latest stable version available at the time of the winter school (but not older than R 2.15). There will be a list of packages (such as plyr, zelig, ggplot2, etc.) that will be crucial for the topics covered during this course. This list will be communicated to the organizers well before the start of the winter school. There are no specific hardware requirements.

Literature

R in general:

Online resources are very good when it comes to R. However, these books (with different difficulty levels) are useful for more detailed descriptions or more advanced applications:

Spector, Ph. 2008. Data Manipulation with R (Use R!). Springer US. [General]

Conway, D. and J. M. White. 2012. Machine Learning for Hackers: Case Studies and Algorithms to Get You Started. O'Reilly Media. [Highly recommended for applications because it has clear code and very nice examples. Not too advanced, but with stronger pace]

Wickham, H. 2009. **ggplot2:** Elegant Graphics for Data Analysis (Use R!). Springer US. [visualization]

Jones,O., Maillardet, R., and A. Robinson. 2009. Introduction to scientific programming and simulation using R. CRC Press. [overarching]

And two excellent **statistics books** that have accompanying R-code (in book and authors' webpage):

Gelman, A. and J. Hill. 2006. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Keele, L. 2008. Semiparametric Regression for the Social Sciences. Wiley Publishing.

Lab requirement

Ideally, each participant should have his/her own computer during the lab sessions. When this is not possible a maximum of 2 students can be paired up. Students can decide to work on their own laptops (as R is free) but they will need internet access throughout the classes.