

An Introduction to R including Data Management and Programming: 29 September 2008

R is an open-source (i.e. free) piece of statistical software that is the default statistical language in Statistics departments and is growing in popularity in the social sciences due to its flexibility and relative ease of use. The crash course assumes no prior knowledge of **R** and will take participants from installation of the software to opening and editing data to running models. The course will also deal with the graphical display of data and model results. Additionally, the course will deal with different data structures and some basic data management tasks. Further, the course will offer an introduction to basic programming in **R**, including loops and other mechanisms for automating computations. The programming topics will be of particular interest to scholars attempting to pool data from a number of different sources as will be demonstrated in the workshop the following day.

Topics to be covered:

- What is **R**?
- Dealing with Data
 - inputting data
 - reading data in from other software packages (e.g., Stata, SPSS)
 - editing data
- Running models
 - linear and poisson as well as binary, ordered, conditional and multinomial logit
 - accessing model results
 - generating quantities of interest (e.g., predicted probabilities)

- Graphical Displays
 - making basic graphs (scatterplots, histograms, line graphs and dot plots)
 - graphing model results
- Data structures
 - understanding vectors, matrices, arrays and lists
 - referencing and indexing different data structures
- Programming
 - loops (e.g., for and while) along with a discussion of when these are useful
 - various versions of the 'apply' command and how to efficiently automate computations

Workshop on Automating Cross-National Statistical Modelling: 30 September 2008

This workshop introduces methods for combining and analyzing a large number of different datasets with an eye toward testing hypotheses about individual- and survey-level political and social phenomena in different contexts. Traditional techniques for accomplishing this goal (e.g., multilevel modelling) require the all of the datasets to have the same variables present. This often results in throwing a lot of information away. The set of techniques we discuss will attempt to use all of the relevant data from all of the different data sources to generate results. As more and more organizations are undertaking the task of generating election and other social surveys (e.g., WVS, CSES, BES, NES, LatinoBarometer, EuroBarometer, AfroBarometer), these techniques become increasingly useful. We have collected roughly 500 such datasets and have successfully implemented an automated strategy to simulate quantities of interest at the individual and survey levels over all of the datasets. We will provide step-by-step instruction on how to go about this type of work from theoretical issues, to data collection and preparation to generating and presenting the results. We use as a running example, the work done by Duch and Stevenson (2008) which uses these different datasets from different contexts to learn about the prevalence and

magnitude of economic voting. In particular, these authors wanted to know how a change in economic perceptions might affect the likelihood of voting for the prime minister's party in the next election.

The workshop will talk about simulation as a method for obtaining quantities of interest with measures of uncertainty. By simulation, we are talking about a CLARIFY-type simulation (<http://gking.harvard.edu/clarify/docs/clarify.html>) where random draws are repeatedly taken from the coefficient vector and quantities of interest (like predicted probabilities) are calculated for each draw (King, Tomz and Wittenberg 2000; Tomz, Wittenberg and King 2003). The result is a distribution of the predicted probability that can be assessed for its difference from zero or from any other distribution.

Students attending this workshop should be comfortable with linear and some basic generalized linear models (e.g., logit, probit). Further, students should have taken the course "Introduction to R including Data Management and Programming" offered the previous day or have familiarity with that material in R. In the interest of full disclosure, we are not offering the roughly 500 datasets compiled by the instructors. We will offer some data for practice, but these will be smaller versions of a few of the datasets used in these large projects.

Duch, Raymond M. and Randolph T. Stevenson. 2008. The Economic Vote: How Political and Economic Institutions Condition Election Results. Cambridge: Cambridge University Press.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 347-61

Tomz, Michael, Jason Wittenberg and Gary King. 2003. CLARIFY: Software for Interpreting and Presenting Statistical Results. Version 2.1. Stanford University, University of Wisconsin and Harvard University. January 5. Available at: <http://gking.harvard.edu>