Introduction to R
ICPSR Summer Program
May 16-20, 2022
University of Michigan
Ann Arbor, MI

Instructor	Contact Information
Ryan Kennedy, PhD Associate Professor Department of Political Science University of Houston	Office: Philip G. Hoffman Hall (PGH), room 447 Email: rkennedy@central.uh.edu Office Phone: 713-743-1663 Cell Phone: 713-855-4811

Course Overview

R has become the software of choice for a great deal of social science analysis. The fact that it is free and open source, along with the wide variety of libraries available for use in analysis, have contributed to this popularity. And a number of graduate programs, including my own, emphasize skills in R to the exclusion of traditionally popular statistical software (e.g. SPSS, Stata, SAS).

Yet, for many students, their first foray into using R is an exercise in frustration. Part of the problem is that R is usually taught in a cookbook-like manner. Students are shown how to do basic statistics using R by copying a series of commands – basically like using dropdown menus, but without the convenience of the dropdown menus. There is little emphasis on understanding how R works. For example, students are taught to run a regression using the command lm(y ~ x, data = dataset), but are never taught that lm() is a function – much less, how they can write their own functions to do almost anything they want. In other words, R is often taught in a way that emphasizes its weaknesses (no dropdown menus, flexible command structure) and barely mentions its strengths (ease of programming, openness to new techniques, quality of replication files). Predictably, this results in frustrated students who learn to rely on copying formulas from cookbooks rather than developing their own unique projects. Instead of seeing the new vistas R opens for analysis, students primarily see R as a more difficult way to cover the same ground.

This course takes a very different approach. Instead of teaching students a series of models or summary statistics commands (which they can easily look up online), it focuses on using R as an elegant and approachable programming language, through which students can do a variety of tasks, just one of which is running statistical models.

Moreover, the course presents a “tidy” version of R programming, utilizing the “tidyverse” group of tools developed by Hadley Wickham and others to make R programming more unified and easier to understand. These tools produce a “grammar of analysis” that will allow students

to quickly produce outstanding research. It also produces code that will be more easily understandable to others (and yourself) and allows more consistency in process.
This is not to say we will be completely ignoring some aspects of base R, rather the goal is to get students working in a system that will be more comfortable, readable, and consistent. It will also be more in line with current publication expectations (e.g., producing plots using ggplot2, rather than the base R graphics).

The goal of the course is to produce students who are confident users of R in a range of situations and can easily expand their knowledge to new fields.

Schedule

Morning session: 10:00-12:00
Lunch: 12:00-1:00
Afternoon Session: 1:00–3:00

[Note: We will be splitting the class instruction into regular intervals within each session to avoid fatigue and allow personalized attention. All online participants will be able to join on Zoom.]

Textbook

This course is based on *Introduction to R for Social Scientists: A Tidy Programming Approach* (I2RSS) by Ryan Kennedy and Philip D. Waggoner (2021, CRC Press). You are not required to purchase this book for the course. The instructor will provide detailed handouts, a pdf working copy of the book, and labs. For those who would like a copy of the book for reference, it is available on the publisher's website or on Amazon. For another book that is a good follow-up to this 3-day workshop, I recommend Hadley Wickham's *R for Data Science*. This book is also available for free online and a link to it will be provided on the course Canvas site.

All of the needed materials are available on the course website on Canvas.

Outline

- Day 1 – We will start with the setup and basic use of R and RStudio (and RTools for Windows users). Students should already have these installed before class. Links to useful startup materials can be found on the I2RSS website (https://i2rss.weebly.com/resources.html) and will also be made available on Canvas. There will also be some work troubleshooting to ensure everyone is set and ready to go. We will also cover some of the basic parts of R: objects, functions, and libraries. We will also go over some of the contemporary tools provided by RStudio to assist with collaboration and replication. We will finish by going through the "tidyverse" tools for data setup and management.

- Day 2 – We will continue discussion of the “tidyverse” tools for data setup and management, as well as for some basic analyses. We will then go through an introduction to visualization in R using “ggplot2”.

- Day 3 – Students will be introduced to basic programming, including conditional logic, loops, and functions. We will work through some applications, where students will be asked to design their own functions to assist in processing unstructured data.

- Day 4 – We will start by discussing data exploration and statistical modeling in R. Students will be introduced to markdown language to share their work with others (and even automate some of the process of book/article/blog writing). They will also be introduced to tools for automatic generation of tables and figures to make creating and updating manuscripts easier.

- Day 5 – We will cover an example of application creation using Shiny, and we will cover topics specifically chosen by class participants to address their particular research needs. (If you would like to recommend topics of interest prior to the course, you can reach out to me at rkennedy@central.uh.edu.) Past topics have included making maps (including interactive maps), R-based software for teaching undergraduate courses, text analysis, web scraping, survey analysis, database interaction, and many more.