Introduction to Quantitative Methods in R

Essex Summer School 2020

Johannes Karreth

Assistant Professor, Department of Politics & International Relations
Ursinus College, USA
Email: jk20769@essex.ac.uk (and jkarreth@ursinus.edu after the workshop)
Course website: http://www.jkarreth.net/r-essex.html (with links to course materials)
Course meetings: M–F, July 13-24, 2020 / 14:15-17:45 BST/UTC+1 via Zoom (core course time; asynchronous options available)
Teaching assistant: Christoph Dworschak, University of Essex (c.dworschak@essex.ac.uk)

Due to Covid-19, this course will run in a virtual format in 2020.

Course description and goals

This course introduces participants to the analysis of quantitative data in the free, open-source software R. R is a highly versatile software environment suitable for introductory and advanced quantitative social science and data analysis. The course offers participants a near-complete foundation to use R for all commonly encountered tasks in social science data analytics.
Specifically, the course will explore the following topics:

- Introduction to the R language and software architecture
- Use of the tidyverse suite of R packages
- Incorporating R code and document production (R Markdown)
- Workflow, reproducibility, and version control in R
- Data import and data management, including working with "messy" datasets
- Descriptive statistics
- Data visualization
- Common techniques for statistical inference, including regression
- R packages for advanced statistical methods, including network analysis and text analysis
- Writing basic functions
- Monte Carlo analysis and simulation

Upon completion of this course, participants will be able to use R for most commonly encountered tasks in social science data analysis, including all of the topics listed above. The course is suitable for researchers at the beginning of their quantitative training as well as researchers with advanced background in quantitative social science wishing to acquire a new, free, open-source, and highly versatile set of tools. Applications from classic statistical methods (such as regression) toward newer tools (such as text analysis) are supported. Participants will also learn to incorporate data analysis and document creation (via R Markdown). A workflow for reproducible data analysis is also a core element of the course. The course content will be reinforced through regular hands-on exercises and frequent feedback from the instructor.

Remote learning setup

Lectures will run on Zoom, and will be recorded for participants to revisit. Participants are invited to interrupt and ask questions any time. To recreate the in-person experience as much as possible, much of the course will consist of 1-on-1 or small-group exercises with the teaching staff. These will take place during the core course time. I will also be available for virtual office hours during every day of the workshop, both during fixed times and by appointment.

Prerequisites

Participants should have a background in introductory statistics or concurrently enroll in an introductory statistics course. Prior initial exposure to statistical techniques up to linear regression (at a fundamental level) is helpful but not required. No background in R or computer programming is required. The course introduces R from a beginner's perspective. At the same time, participants with experience in other tools (e.g. SPSS, Stata, or SAS) will find the course structure helpful to transfer their skillsets into R.

Literature

Participants should have access to:

- Imai, Kosuke. 2018. *Quantitative Social Science: an Introduction*. Princeton: Princeton University Press (abbreviated below as QSS).

The following books are recommended as background companions, depending on participant interest:

- Gandrud, Christopher. 2017. *Reproducible Research with R and RStudio*. Second. Boca Raton, FL: Chapman / Hall/CRC.
- Long, James and Paul Teetor. 2019. *R Cookbook: Proven Recipes for Data Analysis, Statistics, and Graphics*. Sebastopol, CA: O'Reilly Media. Online version: `https://rc2e.com`.
- Moore, Will H. and Siegel, David A. 2013. *A Mathematics Course for Political and Social Research*. Princeton, NJ: Princeton University Press.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2020. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC. Online version: `https://bookdown.org/yihui/rmarkdown/`.
- Chang, Winston. 2018. *R Graphics Cookbook, 2nd edition*. Sebastopol, CA: O'Reilly Media. Online version: `https://r-graphics.org`.
- Healy, Kieran. 2019. *Data Visualization: A practical introduction*. Princeton: Princeton University Press. Online version: `https://socviz.co`.
- Grolemund, Garrett and Hadley Wickham. 2017. *R for Data Science*. Sebastopol, CA: O'Reilly Media. Online version: `https://r-graphics.org`.

Further readings and materials will be made available to participants during the course.

Software and Preparation

Participants will be asked to install R and RStudio on their personal laptops during the first course meeting. We will go over how to use these programs on the first day of the course, using a detailed tutorial with step-by-step instructions. We will also have time to catch up on installation problems on the first day.

Course schedule

For each day, the core reading usually provides substantial details for the units discussed on that day. A typical course period will consist of the following:

- **Lectures** are self-contained mini-units mixing lecture and discussion.
- **Labs** are guided tutorials with documented scripts available to participants.
- **Assignments** are problem sets that participants may complete to reinforce the material learned in the course on that respective day.

Day 1 Monday, July 13

Introduction to the R language and software architecture

1. Why use R? In brief: to make your life easier
2. Foundational elements of the R language
3. Object-oriented programming
4. Software infrastructure
5. Advantages of a text-based workflow

Lab Setting up your first project in R
Assignments Bias in voting turnout, and world population dynamics
Reading QSS, chapter 1

Day 2 Tuesday, July 14

Managing and describing data

1. Importing and managing datasets
2. Manipulating data
3. Describing and summarizing data

Lab Small class sizes and learning outcomes
Reading QSS, chapter 2

Day 3 Wednesday, July 15

Exploring data through visualization

1. Bivariate associations
2. Building simple plots using base-R
3. The `ggplot2` framework
4. Plots for grouped data: small multiples

Lab Optimizing plot content and quality
Reading QSS, chapter 3 (75-96)

Day 4 Thursday, July 16

Common regression-based techniques for statistical inference and their implementation in R

1. Regression: Foundations
2. Regression and causal inference
3. Presentation
4. Diagnostics

Reading QSS, sections 4.2-4.4; also review chapters 2 and 3

Day 5 Friday, July 17

Workflow and reproducibility

- Folder structures
- The Project TIER setup for reproducible research
- Incorporating R code and document production with R Markdown
- From R output to word processing system: tables and graphs

Assignment Your first complete research project in R

Reading The Project TIER protocol (`https://www.projecttier.org/tier-protocol/specifications/`)

Day 6 Monday, July 20

Probability: foundations of inference

1. Debrief on research project assignments
2. Basic rules of probability
3. Randomness
4. Distributions
5. Large sample theorems

Reading QSS, chapter 6

Day 7 Tuesday, July 21

Uncertainty: Regression and hypothesis testing

1. Estimating uncertainty around parameters
2. Interpreting regression estimates for hypothesis tests
3. Critiques of null hypothesis significance testing

Reading QSS, chapter 7

Day 8 Wednesday, July 22

Managing and wrangling data: working with messy scenarios

1. Debrief on research project assignments
2. Concepts for organizing data
3. The `tidyverse` suite of packages

Lab Processing and managing data

Reading Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10): 1–23

Day 9 Thursday, July 23

Text as data

1. Loading texts into R
2. Basic methods for text analysis

Lab Processing and analyzing text

Reading QSS, chapter 5 section 1

Network analysis

1. Managing network data
2. Visualizing networks in R

Lab Network visualizations

Reading QSS, chapter 5 section 2

Spatial data, and R as a GIS tool

1. Loading spatial data into R
2. Visualizing spatial data
3. Basic methods for spatial analysis

Lab Spatial data

Reading QSS, chapter 5 section 3

Day 10 Friday, July 24

Simplifying your life: simulation, functions, and R packages

1. R functions
2. Monte Carlo analysis and simulation
3. R packages

Lab Your first R function

Reading "Functional programming" in Wickham, *Advanced R* (`http://adv-r.had.co.nz/`).

Advanced topics (based on participant interest)

1. Teaching R
2. R for interactive graphics
3. Workflows for collaborative projects

Course wrap-up