INSTITUTE OF LAW, POLITICS AND DEVELOPMENT

Scuola Superiore Sant'Anna

Introduction to R
AY 2024/25 - 20h

Prof. Kevin Koehler

Sant'Anna School of Advanced Studies
kevin.koehler@santannapisa.it

Course description

This course is an introduction to the programming language R. It focuses on hands-on applications, rather than statistical or mathematical background and is therefore open to students at all levels of quantitative skills. R is not a software package in the traditional sense, but a programming language with a wide range of potential applications, from standard statistical analysis, to data management and visualization, to web scraping and text analysis. The course will use R Studio, a free user interface for R. In a first part, we we will cover issues of data management and data wrangling using base R and the `tidyverse` family of tools, the basics of R programming, as well as data visualization using the grammar of graphics package (`ggplot2`). Each session consists of an introduction to the topic, followed by practical exercises. Participants should thus make sure to install R and R Studio before the first class. In the second part of the course, we will move to more specific applications, focusing in particular on ways in which R can complement other software packages with which students might be familiar (in particular STATA). While R has a somewhat steep learning curve, it does provide more flexibility and adaptability (in addition to being free for users). We will start the second part by replicating published analyses employing standard regression approaches, learning how to implement such techniques in R, but also how to efficiently present the results in graphical as well as tabular form. We then move to basic web scraping and text as data approaches as an example for how R can be employed beyond the remit of traditional statistics. By the end of the course, participants will be able to independently implement basic data management and analysis tasks in R.

The main textbook used in this class is Kennedy, Ryan and Philip D. Waggoner. 2021. *Introduction to R for Social Scientists: A Tidy Programming Approach*, CRC Press; we will also use sections of Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2019. *R Markdown: The Definitive Guide*, CRC Press and Silge, Julia and David Robinson. 2017. *Text Mining with R*, O'Reilly.

Prerequisites

- Install R (here) and R Studio (here) before the first class
- Bring your laptops with R and R Studio installed to each class

Assessment

Participants are assessed based on a final research note in which they implement an analysis of their choice in R. The note should be written as an R markdown file including all necessary code to reproduce the analysis. **Please make sure to agree the content of your research note with me**.

Overview of sessions

1	Tue 8 April, 10-12	Introduction: Getting set up
2	Mon 14 April, 10-13	Visualization and R programming
3	Tue 22 April, 10-13	Exploring data and reporting with R markdown
4	Tue 29 April, 10-13	Replication I: U.S. Soft-Power and Foreign Policy Behavior (OLS)
5	Mon 5 May, 10-13	Replication II: Does Counterbalancing Prevent Military Coups? (Logistic Regression)
6	Mon 12 May, 10-13	Basic Web-Scraping with R
7	Thu 15 May, 14-17	Text as Data

Detailed description

Session 1: Getting set up

- **Aim**: What is R and why should you care about it? What can you expect from this course (and what can't you)? In addition to approaching these questions, we will cover some basic tools of data management in R.
- **Readings**: Kennedy and Waggoner 2021, Introduction and Chapters 2 and 3.

Session 2: Visualization and R programming

- **Aim**: Data visualization is important for exploring data, but also for communicating findings. In this session, we cover plotting with the `ggplot2` (Grammar of Graphics) package and also explore the basics of R programming.
- **Readings**: Kennedy and Waggoner 2021, Chapters 4 and 5.

Session 3: Exploring data and reporting with R markdown

- **Aim**: Exploring your data is an essential first step preceding data analysis. This session introduces tools and procedures for data exploration and explains how you can use R markdown documents for reproduceable reporting.
- **Readings**
  - Kennedy and Waggoner 2021, Chapter 6.
  - Xie, Yihui, J.J. Allaire, and Garrett Grolemund. 2019. *R Markdown: The Definitive Guide*, CRC Press, Installation and Chapter 2.

Session 4: Replication I - U.S. Soft-Power and Foreign Policy Behavior (OLS)

- **Aim**: Understanding the logic of linear regression is essential for understanding more complex regression approaches. This session reviews the linear regression model and its implementation in R by replicating published research.
- **Readings**
  - Kennedy and Waggoner 2021, Chapter 7 (specifically focus on section 7.5 Ordinary Least Squares Regression).
  - Goldsmith, Benjamin E. and Yusaku Horiuchi. 2012. "In Search of Soft Power: Does Foreign Public Opinion Matter for US Foreign Policy?" *World Politics*, 64 (3): 555-585. doi:10.1017/S0043887112000123
  - Replication data can be found here

Session 5: Replication II - Does Counterbalancing Prevent Military Coups? (Logistic Regression)

- **Aim**: Many outcomes we are interested in as social scientists are binary: Whether an event does or does not occur (for example, civil war onset, military coups, revolutions, etc). This session discusses how the linear regression approach discussed before can be generalized to deal with such outcomes. We then implement the logistic regression in R by replicating Erica De Bruin's work on counterbalancing and coups.
- **Readings**
  - Kennedy and Waggoner 2021, Chapter 7 (specifically focus on section 7.6 Binary Response Models).

  - **Readings**: De Bruin, Erica. 2018. "Preventing Coups d'état: How Counterbalancing Works." *Journal of Conflict Resolution* 62 (7): 1433–58. doi: 10.1177/0022002717692652.
  - Replication data can be found at the journal website.

Session 6: Basic Web-Scraping with R

- **Aim**: We frequently want to collect data from online sources such us online newspapers, institutional websites, or similar sources. In many cases, we can save significant time and effort if we automate online data collection.
- **Readings**:
  - Bradley, Alex, and Richard J. E. James. 2019. "Web Scraping Using R." *Advances in Methods and Practices in Psychological Science* 2 (3): 264–70. doi: 10.1177/2515245919859535.
  - Schweinberger, Martin. 2022. Web Crawling and Scraping using R. Brisbane: The University of Queensland. (Version edition = 2022.11.15).

Session 7: Text as Data

- **Aim**: How can we automatically analyze large amounts of text? In this session, we look at the *tidytext* approach to automatic text analysis in R and learn how to run basic topic models.
- **Readings**: Silge, Julia and David Robinson. 2017. *Text Mining with R*, O'Reilly, Chapters 1, 3 and 6.