Introduction to R

Data analysis and statistical computing for quantitative social science research

Laboratorio de Investigación para el Desarrollo del Ecuador

Instructor: Daniel Sánchez, MA

Module length: 15 hours

Course level: Intermediate

Prerequisite knowledge: Basic computer skills, Git, shell.

Corequisite knowledge: **Introduction to Statistics** module.

GitHub repository: https://github.com/laboratoriolide/intro-to-r

1 Module overview

This module introduces the R programming language basics for data analysis and statistical computing. The material is designed for those who aim to use R as a tool for quantitative research methods, such as statistics, econometrics, and (to a lesser degree), data science/machine learning. I assume little to no prior coding experience, but I expect students to be familiar with basic computer skills, Git and the command line. Further, the course will be relatively fast-paced, combining self-study with the DataCamp platform and live coding in lecture. We will cover many important concepts in the language, however, the focus is not on technical programming but rather on understanding the language as a tool for statistical analysis.

This is the companion module to the *Introduction to Statistics* module, which provides the foundations of statistics for social science research.

2 Module contents

The following is a planned outline of the course. This may change depending on the pace of the class. Each lecture will have one or two assigned readings, which will all be academic articles submitted to the course's GitHub repository. See below for details on these readings.

- **Lecture 1**: Intro
  - Introduction to the R language: history, uses, advantages and applications in the social sciences.
  - The development environment: R, RStudio, IDEs, and R Packages.
  - The RStudio interface: R scripts, file viewer, console, and environment.
  - Ensuring reproducibility: Filepaths, working directories, and project organization; using R Projects with RStudio.
  - Installing packages, loading libraries; the `install.packages()` and `library()` functions.
  - Basic R syntax: variables, assigners, data types, structures, and (logical) operators.
  - Data frames: definition and basic manipulation with base R syntax.
  - Looking at data: `head()`, `tail()`, `str()`, `summary()`
  - Base R graphics: `plot()`, `hist()`, `boxplot()`, `barplot()`.
  - Getting help in R/RStudio, the internet and AI tools.
  - **Complete before class**: installation of R and RStudio, cloning of the course's GitHub repository and completion of the first DataCamp course, *Introduction to R for Finance.*
- **Lecture 2**: Introduction to the tidyverse and importing data
  - Base R and the *tidyverse*
  - Introduction to the *tidyverse*: the tidy process, tidy data, the pipe (`%>%`) operator
  - Importing data with base R functions
  - Importing data with `readr`, `haven` from different file formats
  - Importing Excel files and preliminary manipulation with `readxl`
  - Downloading data from R packages, GitHub and the web/URLs
  - Tibbles: the tidy data frame; using `glimpse()` to understand data
  - dplyr: Selecting and renaming columns, pipe workflows, `transmute`.

  - Complete before class: DataCamp course *Introduction to importing data in R* and *Introduction to the Tidyverse* (select modules) and make sure you have access to the module's `data` folder in the GitHub repository.
- **Lecture 3**: More on data manipulation
  - Filtering rows, sorting, grouping, and summarizing data.
  - Using `mutate()` to create new columns: `if_else`, `case_when`, `case_match()`.
  - Joins, semijoins and antijoins: applications for researchers.

  - The `bind_rows()` function for appending data frames.
  - The `across()` function for applying functions to multiple columns.

- **Lecture 4**: Data visualization with ggplot2
  - Introduction to the `ggplot2` package: the grammar of graphics.
  - The `ggplot()` function: aesthetics, geoms, and layers.
  - Scatter plots, line plots, bar plots, histograms, and box plots.
  - Faceting: `facet_wrap()` and `facet_grid()`.
  - Themes and customization: `theme()`, `labs()`, `scale_x_continuous()`, `scale_fill_manual()`.
  - Saving plots, changing plot size, and exporting to different formats.

- **Lecture 5**: Advanced topics in data cleaning
  - Understanding long vs. wide data formats; reshaping data with tidyr.
  - Expanding data: `expand()` and `complete()` functions.
  - Missing data: identifying, handling, and imputing missing values.

- **Lecture 6**: Reporting and reproducibility
  - The concept of reproducible research.
  - Introduction to R Markdown: the R Markdown document, YAML header, code chunks.
  - Knitting documents: HTML, PDF, Word, and slides.
  - Customizing R Markdown documents: themes, templates, and CSS.
  - Quarto: the next generation of R Markdown.
  - Math formulas: incorporating LaTeX into R Markdown.

- **Lecture 7**: Select topics in R programming
  - Control structures, conditional statements, and logical operators; relationships to prepackaged functions.
  - Functions: when to use, how to define functions, arguments, return values.
  - The `apply` family of functions: `apply()`, `lapply()`, `sapply()`, `mapply()`.
  - Loops: while and for loops, the `map` family of functions.

- **Lecture 8**: Further advanced topics with R
  - Time and date management: `lubridate`.
  - String manipulation with `stringr`.
  - Fuzzy joins: string distance and the `fuzzyjoin` package.
  - Time-permitting: regex, text analysis.

- **Lecture 9**: High-performance computing in R
  - Why high performancec? The need for speed in data analysis.
  - Loading very large datasets: `data.table` and `fread()`.

  - Reading many files at once: `purrr` and `map()` with `fread()`.
  - Grouping and summarizing large datasets: `dplyr` and `data.table` comparison
  - Other alternatives: bigmemoryr, ff, and SQL databases.
  - Introduction to parallel computing: the `parallel` package.

- **Lecture 10**: Advanced data access with R
  - Web scraping: the `rvest` package.
  - Accessing twitter data: the `rtweet` package.
  - Accessing financial data: the `quantmod` package.
  - Accessing spatial data: the `sf` package.
  - Time-permiting: APIs, applications with open data.

The module will be graded following the same weighting as the one proposed by the program's regulation handbook. The final grade will be calculated as follows:

Component	Percentage
Attendance & Participation	15%
Assignments	50%
Case Study	35%

2.1 Attendance and participation

I do not have any special requirements for attendance nor participations other than the requirements set by the program. Consult the program's regulation handbook for more information. I encourage you to participate in class and ask questions, as this will help you understand the material better. Statistics typically inspires frustation, so it is important to ask questions when you are confused - as any other quantitative course, the material builds up on itself, so better to understand things sooner rather than later.

2.2 Assignments

There will be weekly DataCamp assignments that reinforce the material taught in this module as well as that of the companion module *Intro to Statistics* module. The assignments will focus on R implementations of statistical methods, statistical theory or other relevant topics. These will be either DataCamp courses, projects or exercises, and are graded on a pass/fail basis. Please check the DataCamp platform for the due dates of each assignment.

Apart from the DataCamp assignments, you may need to complete some self-study or autonomous practice work **before** each lecture. This is necessary to make the most of the class time, as programming is a skill that requires practice. Please see above for all tasks to be completed before each lecture.

2.2.1 Tentative DataCamp assignments

- Introduction to R for Finance
- Introduction to importing data in R
- Introduction to the Tidyverse
- Data manipulation with dplyr
- Data visualization with ggplot2

2.3 Case study

The case study is due the last day of class. The case study involves using the skills gained in this module as well as the *Intro to R* module to analyze the provided case study dataset in groups of 5-6 students. Each group will present their findings in 10-minute academic-style presentation. followed by a 5-minute questions and answers (Q&A) session by the instructor and the rest of the class. A handout will be provided with the case study details and rubric.

3 Course materials

All course materials will be provided in the course's GitHub repository. This includes lecture slides, readings, datasets, assignments and any other relevant material. The repository will be updated regularly, so please check it often for new material. I recommend using a Git client, such as GitHub Desktop or GitKraken, to keep your local repository up-to-date.

3.1 Readings

A reading will be assigned for each lecture. These can be academic or general articles that will be used as to illustrate the concepts covered in class. Students are expected to read the assigned readings before class and be prepared to discuss them. Files for all readings will be provided in the course's GitHub repository.

When readings are academic articles, students should not expect to fully understand the methods used in the paper. This module is an introduction to statistics, so students will progressively learn the methods used in the readings as the course progresses. Students should focus on the specific part the reading was assigned for, as well as the paper's objective, a general overview of the methods and any applicable public policy implications. For example, for @ahmed_etal13, the focus is on understanding Table 1.

3.2 Textbook

There is no required textbook for this course, as I will provide slides for all lectures However, I recommend the following books for those who want to delve deeper into the material. These were used as references for the course.

- *Statistics for Business and Economics*, @anderson_etal20.
- The statistics review in *Introductory econometrics: A modern approach*, @wooldridge20.
- *Using R for Introductory Econometrics*, @heiss20[1]
- *Discovering statistics using R*, @field_etal12.
- The Effect, @huntington-klein22[2].

3.3 Software

The course will mainly use R [@rcoreteam24] for practical implementations of the statistical concepts covered in class. R is a free and open-source software which has growingly become the standard for advanced statistical analysis. The sister module to this course, *Intro to R*, introduces the language in depth, focusing on technical aspects, while here we mostly use it as a tool to understand statistical concepts.

Further, the course will also feature demonstrations of statistical concepts using Stata, a widespread software package across academic environments. Further, I will provide practice exercises to be solved with this language. These will not be graded, but feedback will provided to those that want it. I will not introduce the use of Stata, as it is assumed all students have already taken and succesfully passed the introductory course on Stata from our last module, *Intro to Stata*. However, I will answer questions and provide personalized help to those who need it, conditional on available time. It is recommended that students review @huntington-klein22 for comparisons on R-Stata syntax in statistical analysis.

Finally, I will use Git and GitHub for version control and to distribute course materials. I will not introduce or evaluate the use of Git, as it is assumed all students have already taken and succesfully passed the introductory course on Git.

[1] This book contains R implementations of Wooldridge's textbook, with code uploaded here.

[2] This textbook, while focused on causality, describes much of the modern R development environment for statistics and econometrics. It also contains Stata and Python code.

4 Communication

All communications to the instructor or teaching assistant (TA) should be made through the course's Slack channel. We hope to respond to questions within 72 hours, but please be patient if we take longer. I do not monitor email regularly, so please use Slack for all communications if you need a timely response.

5 References