ABSTRACT

The primary goal of the seminar is to teach transparent and reproducible workflows for research data management and statistical analysis using the free R programming language for statistical computing and graphics and the RStudio environment. The basic idea is that transparent data management anticipates the data representation needed for statistical analyses and modeling. A transparent representation of data greatly facilitates the specification of statistical models that are appropriate for the data; in other words, it effectively prevents the specification of incorrect statistical models. The secondary goal of the seminar is to introduce some multivariate statistical analyses. However, the extent and amount of time spent on the secondary goal depends on how fast the primary goal is reached, that is it depends on students' background and success in achieving the primary goal.

PRESENTATION OF THE COURSE

STUDY OBJECTIVES

Implement a complete workflow for an empirical social science project

EXPECTED LEARNING OUTCOMES

- Students know and implement the steps of a data science project: import, clean, transform, visualize, and model data as well as communicate results
- Students are able to formulate goals and research questions about observational and (quasi-)experimental studies
- Students know principles of good scientific practice and learn to document their research in a reproducible format

CONTENT OF THE COURSE

1. R and the RStudio environment
   • Installing R an RStudio
   • Setting up an R project and file structure
   • Navigating the RStudio environment (panes)
   • Basics of RMarkdown

2. Finding, downloading, and reading "interesting" data in the web, also cloning of github accounts.
   The seminar uses real-life observational as well as experimental-research data that are available in the web or from direct contact. Students are expected "to find" data that are of interest to them. As fall-back options the instructor will use, for example:

   • Data available from projects on the discrimination of true and fake news
   • German COVID incidence statistics from Robert-Koch-Institute
   • European COVID incidence statistics from European Commission
   • Results from latest federal election in Germany; statistics available at the city/community level.
   • Other city/community statistics

3. Naming directories, files, variables, style guides for programming

4. Converting "Urdaten" to appropriate formats (wide, long, mixed) of derived files

5. Data representation for
   • within / between, experimental/quasi experimental, fixed/random factors
   • within / between, experimental/quasi experimental, fixed/random covariates
   • distinction between measures and covariates
   • experimental design (w-subj, b-subj) in wide format (usually crossed factors)
   • reflecting hierarchical structure of questionnaire (usually nested factors)
   • multiple-choice and multiple-response formats

- distinction between observations and variables
- wide, long, and mixed format

6. Data wrangling with *tidyverse* syntax of R (taught in parallel to 1 to 5 of above)
   - {dplyr} package
   - {forcats} package
   - {stringr} package

7. Graphics and graphic design considerations -- using the {ggplot2} package
   - Line graphs
   - Scatterplots
   - Boxplots

8. Advanced statistics topics (options)
   - Contrast coding
   - General linear model (t-test, ANOVA, multiple regression): basic R packages
   - Linear mixed models: {lme4} package
   - Signal detection theory: {clmm} package
   - Structural equation models: {lavaan} package

   As students have the option to work on projects and with data of their interest, the projects and type of data chosen will (somewhat) constrain the type of multivariate statistical models covered in the project and may take priority over topics prepared by the instructor. Obviously, the instructor's competence with respect to statistical models are also a constraint.

9. Advanced programming topics
   - control structures for(), while(): basic R packages
   - {purr} package

DISCIPLINE IMPLEMENTATION TECHNOLOGIES

Audience coverage: Your campus

Format of lectures: Lectures will be presented in an online format using the R/RStudio environment; students are expected to execute the commands in parallel on their own computers.

ONGOING ASSESSMENT

1. **Installation of R and RStudio on own computer**
   - **Assessment criterion:** Demonstrate functionality of installation in seminar
   - **Opportunity to retake:** Yes
   - **Comment**: This is a necessary condition for participation.

2. **Homework 1**

3. **Topic and data for student project**
   - **Assessment criterion**

   Submit HTML document created with R Markdown with
   (1) description of research questions (max 200 words)
   (2) description of data [source (web, publication, …), list of variables, number of cases),
   (3) computation of summary statistics for variables
   - **Opportunity to retake:** Yes
   - **Comment:** Feedback on and, if necessary, revision of research questions will be provided

4. **Homework 2**

5. **Report and analyses of student project**
   - **Assessment criterion**

   Submit *Brief Research Report* as PDF document created with R Markdown comprising
   (1) Introduction
   (2) Method
   (3) Results
   (4) Discussion
   - **Opportunity to retake:** Yes
   - **Comment:** Report should be between five and ten pages if printed in (DIN A4)

INTERIM AND FINAL ASSESSMENTS

% **Interim (i.e., homework, midterm) and final assessments are based on Ongoing Assessments**

0.100	Installation of *R* and *RStudio* on own computer – not graded (necessary condition)
0.100	Homework 1: Tutorial exercises covering content from the seminar (graded)
0.300	MIDTERM: Topic and data for personal project (graded)
0.100	Homework 2: Tutorial exercises covering content from the seminar (graded)
0.400	FINAL: Report and analyses of personal project (graded)

Comment: If the task or part of the task with separate deadline was submitted up to an hour after the deadline, the score for it is reduced by 10%, up to 6 hours - by 30%, up to 24 hours – by 60%, after that the task or its part is not accepted resulting in 0 grade.

LITERATURE

Recommended basic literature

Ismay, C., & Kennedy, P.C. (2021-12-30). *Getting used to R, RStudio, and R Markdown*.
https://ismayc.github.io/rbasics-book/

Data Carpentry (2018-2022). *R for Social Scientists.*
https://datacarpentry.org/r-socialsci/

Recommended additional literature

Gelman, A., Hill. J., & Vehtari, A. (2020). *Regression and other stories.* Boston: Cambridge University Press.
https://users.aalto.fi/~ave/ROS.pdf

Ismay, C., & Kim, A.Y. (2020). *Statistical Inference via Data Science: A Modern Dive into R and the Tidyverse.*
https://moderndive.com/index.html#sec:intro-instructors

Rabe, M.M.,

Schad, D.,

Wickham, H., & Grolemund, G. (2017). *R for Data Science.* Boston: O’Reilly.
https://r4ds.had.co.nz/

SOFTWARE

R: https://www.r-project.org/

RStudio: https://www.rstudio.com/products/rstudio/download/

Tidyverse: https://www.tidyverse.org/

R Markdown: https://rmarkdown.rstudio.com/