DATA SCIENCE FOR SOCIAL SCIENCES

Reinhold Kliegl (University of Potsdam)

ABSTRACT

The primary goal of the workshop 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 workshop 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 participants' background and success in achieving the primary goal.

WORKSHOP OBJECTIVE

Implement a complete workflow for own empirical social science project

EXPECTED LEARNING OUTCOMES

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

CONTENT OF THE WORKSHOP

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. Prepare your own data set --
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 from R, also {MixedModels.jl} package from Julia
   • Generalized linear mixed model: {lme4} package from R, also {MixedModels.jl} package from Julia
   • Structural equation models: {lavaan} package from R
   • Signal detection theory: {clmm} package from R

   As participants 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

ONGOING ASSESSMENT

1. **Installation of R and RStudio on own computer**
   • **Assessment criterion:** Demonstrate functionality of installation in workshop
   • **Comment**: This is a necessary condition for participation.

2. **Topic and data for participant 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
   • **Comment:** Feedback on and, if necessary, revision of research questions will be provided

3. **Report and analyses of participant project**
   • **Assessment criterion**
   Submit *Brief Research Report* as PDF document created with R Markdown comprising
   (1) Introduction
   (2) Method
   (3) Results
   (4) Discussion
   • **Comment:** Report should be between five and ten pages if printed in (DIN A4)

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., Vasishth, S., Hohenstein, S., Kliegl, R., & Schad, D. (2020). hypr: An R package for hypothesis-driven contrast coding. *The Journal of Open Source Software.* https://joss.theoj.org/papers/10.21105/joss.02134

Schad, D., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on *a priori* contrasts in (linear) mixed models: A tutorial. *Journal of Memory and Language, 110,* 104038. https://www.sciencedirect.com/science/article/pii/S0749596X19300695?via%3Dihub

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/