Western University
Department of Psychology
PSYCHOL 9041
Introduction to Data Management and Linear Modeling Using R
Fall 2024
Lecture: Monday from 9AM to 12PM
Lab: Friday from 10AM to 11AM

Enrollment Restrictions

Enrollment in this course is restricted to graduate students in the Department of Psychology, as well as any student that has obtained special permission to enroll in this course from the course instructor as well as the Graduate Chair (or equivalent) from the student's home program.

Instructor and Teaching Assistant Information

Instructor: Dr. John Sakaluk ("Sack-uh-luck") (He/Him/His)
Office: SSC 6312
Office Phone: 519-661-2111 ext. 87755
Office Hours (by appointment) : booking link
Email: professor.sakaluk@gmail.com (for day-to-day class inquiries); jsakaluk@uwo.ca (for emergencies)

Teaching Assistant	Anthony Cruz (He/him/his)
Email	acruz27@uwo.ca
Office	(Online meeetings only)
Office Hours	Mondays, 1:30-3:30 (email or Slack to schedule a zoom)

Course Description

The last decade has seen a rapid increase in the popularity of the statistical programming language R as a data analysis workhorse of the social sciences. This popularity is largely owing in R's status as both an open source, cross-platform, and all-inclusive alternative to other proprietary, operating-specific, and analysis-specific statistical software (e.g., SPSS, SAS, Mplus, CMA), as well as its tremendous capacity for data visualization and reproducible document creation.

With R, it is often the case that fitting one's desired statistical model is the quickest and most straightforward part of the data-analysis process; where useRs experience a sometimes challenging learning curve is the before and after steps. In this course, we will therefore focus on increasing your fluency with getting the data you need into R and appropriately wrangled, and to also calculate and extract the pieces of quantitative information you need from your datasets and place them into compelling formats for dissemination.

Most days will include a lecture and a live-coding session (including discussion and Q&A). There is also a lab each week, during which you will have time to apply coding

techniques taught during the course. Throughout the course you will also be expected to complete several self-guided tutorials (some of which include video content). See the Schedule of Dates and Activities for a complete list of topics.

Course Format

The course will be taught synchronously--initially in person, with the flexibility to pivot to online via zoom should public health circumstances dictate a shift.

Course Learning Outcomes/Objectives

Upon completion of this course, students should be able to:

1. Understand the value (to themselves and others) of creating and maintaining a "mindlessly reproducible" coding workflow for data analysis
2. Understand at a "birds-eye" level how R handles information (e.g., assignment to objects, object types, common operators), enabling them to write (and read) sensible code
3. Apply features of a mindless reproducible workflow to their analyses (e.g., absolute file paths, coherent and documented Quarto files and/or Scripts)
4. Apply common "tidyverse" functions for the purpose of data management (e.g., importing data, selecting cases and variables, creating or altering variables and their values, transforming and/or merging across data sets)
5. Apply the psych package for computation (and evaluation) of more complex psychometric scores
6. Calculate simple descriptive statistics (e.g., frequency, central tendency, dispersion, missingness)
7. Create dynamically populated tables with reproducible values of descriptive statistics and/or model output
8. Create (and understand) effective visualizations of of raw data, descriptive statsistics, and/or inferential trends
9. Understand the major elements of the general linear model (e.g., observed vs. predicted scores, intercepts, slopes, variables, residuals) and how they can be combined to conduct a variety of commonly used analyses in psychological research (e.g., *t*-tests, analyses of variance, regression, moderation and mediation)
10. Apply linear models to data sets in R, and correctly interpret and/or extract diagnostic, model, and/or parameter estimate information
11. Dynamically report general linear model output (i.e., reproducibly in text, tables, and/or visualizations)
12. Appreciate how modifications to the general linear model enable more specialized forms of data analysis (e.g., categorical data analysis, analysis of dependent observations, multivariate modeling)
13. Apply and understand simple loops and user-written functions to expedite computationally- or coding-intensive tasks

Course Materials

The course will require you to regularly consult several open-access resources to learn about how to do different things in R (e.g., basic data management, visualization, Quarto). Most of these have hard-copy versions that you can purchase, if interested (I own/regularly use many of these),

but the open-access versions are entirely sufficient (and oftentimes more recently updated than the hardcopy versions).

Books

1. R for Data Science ("R4DS", Grolemund & Wickham)
    - Basics of data importation, management, and visualization

2. R Graphics Cookbook ("The Cookbook", Chang)
    - "Recipes" for ggplot syntax for virtually any type of plot imaginable

3. Fundamentals of Data Visualization (Wilke)
    - Less of a "how-to?" and more of a "what and why?" resource for good data visualization practices

4. Learning Statistics with R ("LSR", Navarro)
    - An excellent introductory grad-stats textbook with application in R

5. Beyond Multiple Linear Regression" Applied Generalized Linear Models and Multilevel Models in R
    - An in-development (but still great-looking) resource for modeling non-normal, categorical, and multilevel observations

6. Big Book of R
    - Not an independent book, but organized like a book with a Table of Contents to link to open source books (incl. Some mentioned here) for various tasks (from general-specific, basic-advanced) in R

Software

Please have the following software downloaded and installed/accounts created before the start of the first day of class.

1. *R* (Download/Install)
    - *R* is the program that actually does the programming/analytic "work"

2. Positron (Download/Install)
    - Positron is the new go-to IDE for programming in *R* (with good python support too!) and adds a graphic interface to R (including a script editor and Quarto document editor) that greatly improves *R*'s functionality and user-friendliness. I wouldn't attempt to try to learn/use *R* without it!
3. Quarto (download/install)
    - Quarto is the new go-to reproducible report document format preferred for *R* coding, and is used to "knit" documents including text, code, and output elements.

4. A LaTeX TeX distribution (Download/install [different options for Mac and Windows)
    - LaTeX is a system for rendering text, math, and symbols (imagine: how painful it would be to specify complex equations having to go into Word's "symbol" list one at a time…). This is necessary for R Markdown to be able to fully knit your documents.

5. Slack (Download/install)
    - Slack will be useful for everyone (including myself and the TA) to more expediently ask/answer questions, share resources, ask for help, etc., as a community of learners, than email/office hours (which should be reserved for more complex/personalized matters). Please join the channel here.

Other Helpful Online Resources

1. Posit Recipes
    - The Posit (formerly "R Studio") team has already created several online tutorials called "Recipes" to help ease your learning into particular types of tasks in *R*.

2. Posit Cheat-Sheets
    - The Posit team has created "cheat-sheets" for key processes and tidyverse packages. These provide nice overviews of the key concepts, functions, and arguments needed to perform common tasks in R, and are a great place to start troubleshooting problems you come across.

3. R for Psychological Science
    - Created by Dr. Dani Navarro (of LSR fame), this is an incredible set of videos and tutorials for many of the skills you will be developing throughout the class

4. R Package documentation .pdfs and (when possible) vignettes (see the documentation and vignette for metafor:: as an example)
    - For most other packages, you will often need to do your own detective-work to identify the functions you need, the arguments they will demand, and how to navigate the output they will supply. Mainstream packages have a standard-looking documentation .pdf that will provide this information to you; other (often well-developed) packages will also have more user-friendly vignette (i.e., tutorial) papers demonstrating the core functionality of a package.

5. StackOverflow and CrossValidated
    - Community Q&A websites for programming (e.g., R) and statistics questions (those that are not programming-related), respectively. When you can't solve your own problems, the answers are often already waiting for you in a post here, and if

not, you can ask the community who are often very responsive/effective (but your mileage with friendliness/warmth may vary…).

Methods of Evaluation

Assignment	Date of Evaluation (if known)	Weighting
Coding Capacity Building Assignments (x6)	Throughout Semster	36% (6% each)
R Language Fluency Test	Oct. 21	15%
Data Viz Assignment	Nov. 11	15%
Capstone Project	Dec. 6 (11:59 PM)	34%
Total		100%

Coding Capacity Building Assignments (x6)

These capacity building assignments are designed to incrementally increase your ability navigate the coding environment in R and add to your ability to manage, describe, and analyze data in R. These will typically be graded according to a rubric that assigns points on the basis of assignment reproducibility AND correctness.

You are permitted to submit *one* of the six Mini-Assignments up to two days late without explanation (life happens); it will be graded without penalty (however, you will not receive a grade/feedback at the normal pace, and we cannot guarantee it will be returned in advance of the next mini-assignment submission). Submissions more than two days late will not be graded without a formal/excused absence (i.e., with official documentation).

R Language Fluency Test

This test will occur during class time, and contains roughly 30 questions designed to assess your ability to read, interpret, and anticipate the output of R code.

For the R Fluency Test, a formal/excused absence (i.e., with official documentation) is necessary should you miss the day of the test.

Data Viz Assignment

This assignment will assess your ability to describe a “story” you wish to tell with data visualization and apply it in code. You will ned to make 2-4 visualizations in 2x styles: 1) befit for publication in a peer-reviewed journal; and 2) befit for a more dynamic dissemination outlet (e.g., a science knowledge translation outlet, a blog, social media). You will (like the Coding Capacity Building Assignments) be graded on the basis of both reproducibility AND correctness, with the correctness points being affected by a “complexity modifier” based on the difficulty of the visualization techniques you have applied.

For the Data Viz. Assignment, late work (defined as beginning once the submission portal closes) will be penalized by a 5% deduction from your assigned grade per day of

lateness, up to a maximum of 5 days, at which point you will be assigned a grade of 0. For example, if your submission was originally graded as 8.5/10, but you submitted it two days late, you would receive 8.5*(1.00 - [.05*2]), you would receive an adjusted grade of 7.65/10.

Capstone Project

For the Capstone Project, you will be asked to create a new/full "mindlessly reproducible" analytic workflow for a data set and analysis(/analyses) of your choosing. You will be able to bring your own data set into the course (e.g., from your honours thesis of yore [as long as you have not analyzed it in R before], a data set from your advisor [used with permission], an open data set available online), and write a dynamic research paper, including a short introduction, full (and reproducible) methods and results sections, and a short discussion section. Like the Data Viz Assignment, you will be graded on the basis of both reproducibility AND correctness, with the correctness points being affected by a "complexity modifier" based on the difficulty of the analytic techniques you have applied.

For the Capstone Project, late work (defined as beginning once the submission portal closes) will be penalized by a 10% deduction from your assigned grade per day of lateness, up to a maximum of 5 days, at which point you will be assigned a grade of 0. The increased penalty is a result of the strain lateness will introduce for timely submission of final grades. For example, if your submission was originally graded as 36/40, but you submitted it three days late, you would receive 36*(1.00 - [.15*3]), you would receive an adjusted grade of 25.2/40.

On the Use of AI-based Technologies in the Course

The use of AI-based Technologies like ChatGPT and Co-Pilot (for which you are eligible for a free license, through a free-to-join program offered [while you are a student] via GitHub Education) are changing the ways many courses are designed, and by which evaluation of course assessments occur.

In the statistics courses I teach (e.g., 9041, 9545), these technologies can provide incredible assistance for some of the more parsnickity tasks in coding, including function-writing, iteration, and parallelization; they are also useful to help with debugging (e.g., deciphering cryptic error messages). As such, I fully encourage students to make use of these technologies, *as appropriate*; it is one thing to use these technologies as but one tool in your toolbelt, and it is quite another to use them outsource all of your creative work. Further, I've found these technologies to be error-prone (laughably so) in other instances. However, I have little interest (at the graduate level) or ability (given fickleness of AI-detecting-AI tools) to police your judicious use of AI to a high degree. And so, my formal policy is this:

You, and you alone, are responsible for the coding products you submit. You commit to—if using ChatGPT and/or Co-Pilot or other LLMs to enrich your workflow—ensuring that a critical mass of your coding is originally generated through your own keystrokes and considerations. Further, you accept all liability of using AI technologies to any degree. This includes penalties as stark as:

- *a grade of "0", should you forego your commitment and depend entirely on AI to generate your code, and I find that your submission is mostly/entirely incorrect.*
- *conversely, should you instead feel inclined to plead that your grade should be considering because you used AI to generate most/all of your code, or I otherwise learn that your submission was generated by AI, I will consider this as evidence of academic misconduct, in the form of plagiarizing an AI-generated script and attempting to pass it off as your own work*

My bottom line advice: AI is a tool worth using in your *R* workflow, but make sure *you* remain in control of *its work* (and not the other way around). A heuristic I might recommend is to ensure you could independently (i.e., without AI) and at some other time:

- explain to me what each line of syntax is doing,
- explain to me why you made the coding choices you did, and
- you could reproduce most or all of the code, given the opportunity

Course Timeline

Week	Date	Topics/Content Areas/ Learning Activities	Suggested Resources
1	Mon., Sept. 9	Introduction to the class, R + Positron, and how *R* processes information	• Gandrud (2015), Chapter 1: *Introducing Reproducible Research* • R4DS: 1: Introduction, 4: Workflow basics, 6: Workflow scripts, 8: Workflow projects, 10: Tibbles, and 11: Data Import • Navarro's R for Psychological Science: File System, Workspaces, Packages, Getting Started, and Scripts • Navarro Video: Say Hello to Your Data
2	Mon., Sept. 16	Data Management in R, 1: Importing and Wrangling Data ***Capacity Building Assignment #1 Due***	• R4DS 5: Data transformation • Navarro, Manipulating Data • Navarro Videos: Filtering Data, Data Arranging, Variable Selection • dplyr cheatsheet
3	Mon., Sept. 23	Data Management in R II: Advanced Restructuring and Composites ***Capacity Building Assignment #2 Due***	• R4DS 5: Data transformation, 12: Tidy Data • Navarro Video: Mutate

Week	Date	Topics/Content Areas/ Learning Activities	Suggested Resources
4	Mon., Sept. 30	Summarizing Your Data I: Descriptive Statistics, Missingness, and Tables ***Capacity Building Assignment #3 Due***	• Navarro: Describing data, Visualizing data • R4DS: Data Transformation (see Grouped Summaries) • 10 Guidelines for Better Tables • 10 Guidelines with gt
5	Mon., Oct. 7	Summarizing Your Data II: Visualizations ***Capacity Building Assignment #4 Due***	• R4DS: Visualizing Data • Wilke: 29: Telling a Story and Making a Point • From data to viz • R Graph Gallery • R Graphics Cookbook • ggplot2 cheatsheet • ggplot2 extension gallery
6	Mon., Oct. 14	**READING WEEK/THANKSGIVING**	
7	Mon., Oct. 21	**R "Fluency" Test**	
8	Mon., Oct. 28	Modeling I: GLM Fundamentals and Preparing Data for GLMs ***Capacity Building Assignment #5 Due***	• Cohen, Cohen, West, & Aiken (2003; Chapter 4, 8, and 10) • MacCallum et al. (2002) • Simmons et al. (2011)
9	Mon., Nov. 4	Modeling II: Fitting, Interpreting, and Reporting GLMs in R	• Dienes (2008, Chapter 3)

Week	Date	Topics/Content Areas/ Learning Activities	Suggested Resources
			• Funder & Ozer (2019) • Bakker et al. (2016) • Lakens (2014) • Simmons (2014)
10	Mon., Nov. 11	Modeling III: Advanced Applications of the GLM (Moderation and Indirect Effects) ***Data Viz Assignment Due***	• Judd et al. (2017) • Schoemann et al. (2017) • West et al. (1996) • Götz et al. (2021)
11	Mon., Nov. 18	Modeling IV: Causal Inference with Experimental and Non-Experimental Data	• Rohrer (2024) • Rohrer (2018) • Rohrer et al. (2022)
12	Mon., Nov. 25	Advanced Programming: Iteration and Simulations of GLM Principles	• R4DS: Functions • Advanced R: Functions • Morris et al. (2019) • Siepe et al., (2023)
13	Mon., Dec. 2	Capstone Workday ***Capacity Building Assignment #6 Due*** ***Capstone Project Due (Dec. 6)***	

Offline Readings, By Topic

Note: I do not formally assess your completion of these readings. However (especially for later topics) these are among the most practically helpful readings I know of, in terms of helping you to develop mastery of planning for, fitting, interpreting, and reporting useful linear models.

Introduction to the class, R + Positron, and how *R* processes information

Gandrud, C. (2015). *Reproducible research with R and R-Studio*. CRC Press: Boca Raton, FL.

- Chapter 1: "Introducing reproducible research"

Modeling I: GLM Fundamentals and Preparing Data for GLMs

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum Associates, Inc.: Mahwah, NJ.

- Chapter 4: “Data Visualization, Exploration, and Assumption Checking: Diagnosing and Solving Regression Problems I”
- Chapter 8: “Categorical or Nominal Independent Variables”
- Chapter 10: “Outliers and Multicolinearity: Diagnosing and Solving Regression Problems II”

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19-40.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359-1366.

Modeling II: Fitting, Interpreting, and Reporting GLMs in R

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Palgrave MacMillan: New York, NY.

- Chapter 3: “Neyman, Pearson and Hypothesis Testing”

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168.

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers’ intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069-1077.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701-710.

Simmons, J. (2014). MTurk vs. The Lab: Either Way We Need Big Samples. Retrieved from https://datacolada.org/18

Modeling III: Advanced Applications of the GLM (Moderation and Indirect Effects)

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*(1), 601-625.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science*, *8*(4), 379-386.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, *64*(1), 1-48.

Götz, M., O'Boyle, E. H., Gonzalez-Mulé, E., Banks, G. C., & Bollmann, S. S. (2021). The "Goldilocks Zone":(Too) many confidence intervals in tests of mediation just exclude zero. *Psychological Bulletin*, *147*(1), 95.

Modeling IV: Causal Inference with Experimental and Non-Experimental Data

Rohrer, J. M. (2024). Causal inference for psychologists who think that causal inference is not for them. *Social and Personality Psychology Compass*, *18*(3), e12948.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27-42.

Rohrer, J. M., Hünermund, P., Arslan, R. C., & Elson, M. (2022). That'sa lot to process! Pitfalls of popular path models. *Advances in Methods and Practices in Psychological Science*, *5*(2), 25152459221095827.

Advanced Programming: Iteration and Simulations of GLM Principles

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102.

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A., Heck, D. W., & Pawel, S. (2023, October 31). Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting. https://doi.org/10.31234/osf.io/ufgy6

Statement on Academic Offences

Scholastic offences are taken seriously and students are directed to read the appropriate policy, specifically, the definition of what constitutes a Scholastic Offence, at the following Web site:
http://www.uwo.ca/univsec/pdf/academic_policies/appeals/scholastic_discipline_grad.pdf

All required papers may be subject to submission for textual similarity review to the commercial plagiarism-detection software under license to the University for the detection of plagiarism. All papers submitted for such checking will be included as source documents in the reference database for the purpose of detecting plagiarism of papers subsequently submitted to the system. Use of the service is subject to the licensing agreement, currently between The University of Western Ontario and Turnitin.com (http://www.turnitin.com).

Health/Wellness Services

Students who are in emotional/mental distress should refer to Mental Health@Western http://www.uwo.ca/uwocom/mentalhealth/ for a complete list of options about how to obtain help.

Accessible Education Western (AEW)

Western is committed to achieving barrier-free accessibility for all its members, including graduate students. As part of this commitment, Western provides a variety of services devoted to promoting, advocating, and accommodating persons with disabilities in their respective graduate program.

Graduate students with disabilities (for example, chronic illnesses, mental health conditions, mobility impairments) are strongly encouraged to register with Accessible Education Western (AEW), a confidential service designed to support graduate and undergraduate students through their academic program. With the appropriate documentation, the student will work with both AEW and their graduate programs (normally their Graduate Chair and/or Course instructor) to ensure that appropriate academic accommodations to program requirements are arranged. These accommodations include individual counselling, alternative formatted literature, accessible campus transportation, learning strategy instruction, writing exams and assistive technology instruction.