Fall 9-1-2022

PUBH 691.01: Data Science and Research Methods Using R

Ethan S. Walker
University of Montana, Missoula, ethan.walker@umontana.edu

UNIVERSITY OF MONTANA
SCHOOL OF PUBLIC AND COMMUNITY HEALTH SCIENCES

Data Science and Research Methods Using R
PUBH 691 – Fall 2022

Tuesday/Thursday 8:00am to 9:20am Mountain Time
Liberal Arts Building Room 106
Zoom: https://umontana.zoom.us/j/96028346473?pwd=dHIzSVNuYVZFOW5sOW8zd0pYLzZqQT09
3 credit hours

Instructor:

Ethan S. Walker, PhD, MPH
Assistant Professor
Center for Population Health Research
School of Public and Community Health Sciences
College of Health
University of Montana, Missoula
ethan.walker@mso.umt.edu

Format:

Hybrid: On-campus/On-line synchronously. PUBH 691 is offered as a hybrid course - an instructional modality that allows students to participate synchronously online or face-to-face. This course will be delivered on campus and simultaneously available via Zoom link. All students, whether enrolled in the on-line or on-campus section, are expected to participate during the scheduled class times and come to class having read the assigned readings. Class periods will be light on lecture and mostly involved hands-on work with R coding and relevant discussion. Students will be assessed on assignments, exams and a comprehensive final project.

Office Hours:

Tuesday and Thursday, 9:30am to 10:30 am MT

I am starting the semester with office hours immediately following class on Tuesday and Thursday, from 9:30am to 10:30 am MT. I will be available to meet with you in-person (office in Skaggs 212) or via Zoom. If possible, send me an email to confirm that you will be attending office hours so that I can coordinate timing with other students.

Email ethan.walker@mso.umt.edu to make arrangements for meeting.

Contact Info:

ethan.walker@mso.umt.edu (Please indicate 'PUBH691' in subject line.)
Skaggs Building Room 212 (office)
Office phone: 406-243-2063

Course Description:

This course is designed for graduate students in public health and other disciplines to learn data science and management techniques using the R programming language. Specific topics covered in the course include exploring, cleaning, and manipulating data, data visualization, summarizing and analyzing data using basic applied statistical methods, reproducible workflows in research, and ethical research practices for working with data. The course is designed to guide students with little to no coding experience through beginner and intermediate level data management techniques in R. However, students with previous coding and/or R experience will also find the course helpful in refining their R skills and learning new and more advanced techniques. The course will prepare students for more advanced courses in statistics and epidemiology, as well as for research-based projects, thesis projects, dissertations, and jobs/careers that require knowledge of data management and analytical techniques.

Course Objectives:

- Understand the basics of working in and troubleshooting R and RStudio.
- Learn how to clean and manipulate datasets, create new variables, prepare datasets to be used in statistical analysis, and produce data visualizations, all while using ethical and reproducible data science techniques.
- Introduction to linear regression analysis techniques and applications in public health and research.
- Produce a codebook that can be used as a reference for future data science projects in R.
- Apply the data science techniques learned throughout the course to a culminating, self-directed final project.

Public Health Competencies

This course will address the following Public Health Competencies and Content Areas for Masters and PhD in Public Health Degrees, as set by the Council on Education for Public Health:

Competencies:

- PhD:
    - Understand current issues and debates in public health research, including multi-cultural dimensions and ethical conduct of public health research.
    - Understand and apply a range of study designs, research methods, and approaches to data management and analysis commonly used in public health and in one's specialized focus area.
    - Design and conduct independent research in a specialized focus area within public health.
- MPH:
    - Select quantitative and qualitative data collection methods appropriate for a given public health context.
    - Analyze quantitative and qualitative data using biostatistics, informatics, computer-based programming and software as appropriate.
    - Interpret results of data analysis for public health research, policy or practice.

PhD Content Coverage:

- Explain the role of quantitative and qualitative methods and sciences in describing and assessing a population's health.
- Explain the critical importance of evidence in advancing public health knowledge.

Course Text and Resources:

A major theme in working with R is that R, RStudio, and many books and online resources are free. There is a great community of resources surrounding R, and this course is modeled after that. You will not have to purchase any books or software for this course. Required and recommended readings will be posted in each week's section on the course Moodle page.

1. **You need a laptop that runs Windows, macOS, or Linux that you can bring to class with you every day!** This course will be difficult or impossible for you if you try to use a tablet, Chromebook, or other machine/operating system.

2. Required Software (available for free online):
   R and RStudio
   https://cloud.r-project.org/
   https://www.rstudio.com/products/rstudio/download/

3. Texts (available for free online):
   a. The Epidemiologist R Handbook
      Batra, Neale, et al.
      https://epirhandbook.com/en/index.html

   b. R for Health Data Science
      Ewen Harrison and Riinu Pius
      https://argoshare.is.ed.ac.uk/healthyr_book/

   c. R for Data Science
      Hadley Wickham and Garrett Grolemund
      https://r4ds.had.co.nz/

   d. Statistical Inference via Data Science: A ModernDive into R and the Tidyverse
      Chester Ismay and Albert Y. Kim
      https://moderndive.com/index.html

4. Other free online resources:
   a. https://education.rstudio.com/learn/
   b. https://psyteachr.github.io/
   c. https://www.statmethods.net/index.html
      i. https://www.statmethods.net/r-tutorial/index.html
   d. https://github.com/iamericfletcher/awesome-r-learning-resources
   e. https://ggplot2-book.org/
   f. https://adv-r.hadley.nz/
   g. https://rstudio-education.github.io/hopr/index.html
   h. Google is your best friend! If you can't figure something out on your own, I promise someone else has had the same question. Chances are, they have asked it online and someone has provided an answer. Don't struggle for too long before checking online. Learn from others who have done it before. Two websites in particular will answer the bulk of your questions:
      i. Stack Overflow
      ii. Cross Validated

Course Grading:

Students will be graded on the following assessment components:

1. Homework assignments (8): 40%
    a. For each assignment, you will be given an example of a dataset, table, or visualization. Your assignment will be to replicate the example and turn in the code you developed to do so.
2. Codebook: 15%
    a. Throughout the semester, you will develop a codebook consisting of notes and useful items you learn throughout the course. This codebook is intended to be a future reference for you specifically. As such, it can be in any format that you wish provided you do your own work and show evidence that you put effort into the project.
3. Midterm: 15%
    a. The midterm will be the first part of the final project, which is described below. For the midterm, you will be expected to have identified a project idea, a dataset, and provide a description of the expected project outcomes. You are encouraged to schedule a meeting with the instructor to develop an attainable final project that aligns with your research interests. Also for the midterm, you need to provide evidence that you have started and made progress on your R codebook.
4. Final: 30%
    a. Your final project will be a semester-long, self-designed, and self-directed project that is based on materials you learn throughout the course. The instructor will provide you a publicly-available data source for the project, or you will have the option to identify your own data source that is related to your specific research interests. You will then make a plan on data processing, analysis, and visualization for the dataset (due for the midterm). Your final deliverable from the project will be a report of the data and the corresponding code that you used to develop it. This project should be reasonable in length and within the scope of the course. You are encouraged to schedule a meeting with the instructor to develop an attainable final project that aligns with your research interests.

Final letter grades will be based on the following categories:

- A = 94 to 100 percent
- A- = 90 to 93 percent
- B+ = 87 to 89 percent
- B = 84 to 86 percent
- B- = 80 to 83 percent
- C+ = 77 to 79 percent
- C = 74 to 76 percent
- C- = 70 to 73 percent
- D+ = 67 to 69 percent
- D = 64 to 66 percent
- D- = 60 to 63 percent
- F = below 60 percent

If you are taking the course as Credit/No Credit (CR/NCR), you must earn a 70% or better in the course to receive credit. CR/NCR is only for special circumstances, so please discuss with me prior to selecting this grading method.

Moodle Course Procedures and Expectations:

Moodle will be used as a supplement in this class. Students are responsible for course information and materials posted to the Moodle course shell. Moodle procedures may be adjusted if necessary. While class content, expectations and assessment will (most likely) not change dramatically, the instructor may make mid-course revisions especially in response to student feedback. Class announcements will be posted by the instructor. Moodle system administrators will sometimes post announcements about the Moodle system.

UMOnline has made available an interactive tutorial for using Moodle as a student. After logging into Moodle using your NetID and password at http://umonline.umt.edu/, you may follow the link "Moodle 101 for Students."

Additional Notes:

The following Indigenous land acknowledgement statement was developed in collaboration with Indigenous scholars and students at the University of Montana:

The University of Montana resides on the traditional lands of many Indigenous peoples including the Selis (Salish), Ksanka (Kootenai), and Qlispe (Kalispel). Many other Indigenous peoples including the Amskapi Pikuni (Blackfeet), Nimiipuu (Nez Perce), Shoshone, Bannock, and Schitsu'umsh (Coeur D'Alene) also relied upon their traditional knowledge and relationships with this land and this space for survival in the past and today.

We acknowledge that education, health, and legal systems have led to the direct removal, oppression, and marginalization of Indigenous people throughout Montana and the nation. The University of Montana strives to improve education, service, and scholarship for all Indigenous peoples through actions aimed at respecting tribal sovereignty, empowering indigenous scholars, and creating safe learning environments for all students to live, work, and learn together in equitable and positive ways.

Accommodations:

Students with disabilities will receive reasonable accommodations in this course. To request course modifications, please contact the instructor as soon as possible. The instructor will work with you and Office for Disability Equity (ODE) in the accommodation process. The ODE is the campus resource for disability-related information. They provide consultation, training, and academic services to advance accessibility and inclusion by taking an intersectional approach to disability. For more information, please visit their website: https://www.umt.edu/disability/.

Student Code of Conduct:

The University of Montana Student Code of Conduct (https://bit.ly/3SMNkft) embodies and promotes honesty, integrity, accountability, and duties associated with citizenship as a student in our community at the University of Montana. This Code exists to protect the interests of the community and dignity of its members, and to challenge those behaviors which are not in accordance with our policies. This Code describes expected standards of behavior for all students, including academic conduct and general conduct, and it outlines students' rights, responsibilities, and the campus processes for adjudicating alleged violations.

Plagiarism:

All students must practice academic honesty. Academic misconduct is subject to an academic penalty by the course instructor and/or a disciplinary sanction by the University. All students need to be familiar with the Student Conduct Code (posted in the previous section).

Plagiarism is the representing of another's work as one's own. It is a particularly intolerable offense in the academic community and is strictly forbidden. Students who plagiarize may fail the course and may be remanded to Academic Court for possible suspension or expulsion. (See UM Student Conduct Code).

Students must always be very careful to acknowledge any kind of borrowing that is included in their work. This means not only borrowed wording but also ideas. Acknowledgment of whatever is not one's own original work is the proper and honest use of sources. Failure to acknowledge whatever is not one's own original work is plagiarism.

You can work with classmates on homework and your codebook. If you do so, document who you worked with. Do not simply copy your classmate's homework/codebook and turn it in. This is plagiarism. It is okay to share code with each other, but you must annotate and turn in your own work. Please talk to me if you have questions about what constitutes plagiarism in this course.

You can ask for help from classmates on your final project, but the idea and content must be your own. This project demonstrates to me how well you know the course content. If you are struggling, reach out to me instead of leaning too much on a classmate.

COVID-19 Notes:

The following is general guidance provided by the UM Office of the Provost:

- *The University encourages COVID-19 vaccines and boosters, which are offered for both students and employees at the Health Services Pharmacy inside Curry Health Center.*
- *Masks are only required inside Curry Health Center and in some medical/research laboratories on campus. This requirement will be clearly posted. Required or not, we respect those choosing to wear a mask to reduce spread of respiratory viruses.*
- *COVID-19 testing for students is available at Curry Health Center. For employees, contact your primary care provider or visit a walk-in clinic. Free at-home tests can be ordered online, or there may be tests available through the Health Services Pharmacy by calling 243-5171.*
- *UM Housing is no longer operating separate quarantine/isolation spaces for students in the residence halls. UM Housing will provide guidance for students isolating in place in the residence halls.*
- *Students who test positive for COVID-19 need to isolate for at least five days, which includes not attending in-person classes. During isolation, students should stay home or follow UM Housing guidance for isolation in place, and, if they must leave for food, medicines or other essentials, wear a high-quality mask.*
- *After five days students can leave isolation if they are symptom-free. If symptoms persist, isolation should continue until students are symptom-free for 24 hours (without use of medications to alleviate the symptoms).*

Changes:

This syllabus is subject to change by the instructor. Any changes will be announced in class and in the announcement section of Moodle.

Course Schedule

Week	Topics	Dates	Assignments
1	Getting started: installing R/RStudio, navigating, scripts, functions, packages, RStudio Cheat Sheets	Aug 30	Practice R using the swirl package
		Sept 1	
2	Loading and exploring data	Sept 6	
		Sept 8	Homework 1 assigned
3	Reproducible research workflows #1: Projects and Markdowns	Sept 13	
		Sept 15	Homework 1 due Homework 2 assigned
4	Cleaning and manipulating data #1	Sept 20	
		Sept 22	Homework 2 due Homework 3 assigned
5	Visualizing data #1	Sept 27	
		Sept 29	Homework 3 due Homework 4 assigned
6	Summarizing and analyzing data #1	Oct 4	
		Oct 6	Homework 4 due Homework 5 assigned
7	Review and practice #1 In-class work time for final projects	Oct 11	
		Oct 13	
8	Reproducible research workflows #2	Oct 18	Homework 5 due
		Oct 20	Midterm due: first half of codebook and final project
9	Cleaning and manipulating data #2	Oct 25	
		Oct 27	Homework 6 assigned
10	Visualizing data #2	Nov 1	
		Nov 3	Homework 6 due Homework 7 assigned
11	Data management and cleaning practices	Nov 8 (no class)	
		Nov 10	Homework 7 due
12	Summarizing and analyzing data #2	Nov 15	
		Nov 17	Homework 8 assigned
13	Review and practice #2 In-class work time for final projects	Nov 22	Homework 8 due
		Nov 24 (no class)	
14	Analyzing and visualizing data #3 Linear regression, continued	Nov 29	R Codebook due
		Dec 1	
15	Present final projects	Dec 6	Final project due
		Dec 8	