Lin611 - Quantitative Methods in Linguistics

Josef Fruehwald

August 2024

Table of contents

1 Key info

1.1 Key Info

Where:	LinCoLab (Breckenridge Room 10)
When:	M-W, 12:00 - 1:15
Credits:	3

1.2 Instructor

	Dr. Josef Fruehwald
email:	josef.fruehwald@uky.edu
office hours:	Thurs, 1pm
office hours location:	1671 POT

---

2 Course at a Glance

Course Webpage:

https://lin611-2024.github.io/

What you'll learn:

the basics of statistical reasoning, linear modelling, data organization & visualization, R

What you'll do:

in-class exercises, R exercises, a midterm project, a final project.

What you'll need:

the course textbook, a computing device with a physical keyboard

The final-est deadline

After **December 16, 2024**.

Attendance Policy

Attendance is crucial for successful completion of the course, but there are no grade penalties.

Late Work Policy

2 day penalty free grace period on all assignments, 5% flat penalty afterwards. See Section 10

3 Course Description

In recent decades, there has been a strong "quantitative turn" in linguistics. Quantitative methods, including statistical analysis, have always been fixtures in some subfields, but there are now few areas of linguistic inquiry where they are completely absent. As a graduate course in quantitative methods, the goals of this course are to help you establish baseline statistical reasoning, and to provide practical experience in data (re)organization and statistical model building. We will be focusing our attention on the most common variety of statistical models (linear models and their generalizations) in the most commonly used programming language (R).

4 Learning Outcomes

After attending class meetings and completion of the coursework, students should be able to

- identify appropriate quantitative analysis procedures for diverse data sets.
- organize data sets tidily
- re-organize untidy data sets in R
- generate exploratory data visualizations
- specify and fit linear or generalized linear models in R
- report the meaningful results of a statistical model

5 Course Materials

Required:

Winter, B. (2019). *Statistics for linguists: An introduction using R.* Routledge. ISBN 978-1138056091.

Recommended:

Wickham, H & G. Grolemund (2022*) *R for Data Science* https://r4ds.had.co.nz/.

6 Course Technology

R/RStudio

We'll be learning how to implement our analyses in the R programming language, specifically using the RStudio IDE. You can install and configure RStudio on your own computer that you bring to class.

Quarto/Quarto Notebooks

Quarto is a program built into RStudio that takes source documents written in Markdown and R, and renders them into various output document formats, including html and pdf. This program is included in RStudio, and won't require additional download or installation.

Git/Github

Git is a "Version Control System" that lets you keep track of changes on software projects. Github is a service that allows online hosting of Git projects. You will need to create a free a Github account for the course.

Canvas

Canvas will be used to make course announcements, and to set & submit assignments.

---

7 Communications

I will respond to emails in a timely manner during normal working hours, but it may take longer if you email me after 5pm on weekdays, or any time during the weekend.

---

8 Course Schedule

The topics and readings listed here are the tentative schedule for the course. We may find, in the room, that some topics will take longer than initially scheduled.

Week 1

Dates:

August 26-30

Topics:

Onboarding

To Do

- Create a Github Account
- Join the course organization

Readings

- If you plan to use R and RStudio on your local laptop, work through this tutorial: Github Onboarding with RStudio
- Tutorial: Hello, Quarto
- Tutorial: Computations
- Tutorial: Authoring

Weeks 2 - 4

Dates:

September 2-20

Notes:

Labor day Monday September 2, no class

Topics:

- Introduction to R
- Plotting Data
- The "Tidyverse"

Readings:

- Winter, Ch 1
- Winter, Ch 2

Supplementary resources

- Hands-On Programming with R
- 2017 LSA Course: Introduction to R
- R4DS, Tidy Data
- R4DS, Data-visualization
- 2017 LSA Course: Data Frames
- 2017 LSA Course: Split-Apply-Combine
- 2017 LSA Course: ggplot2
- tidyr cheat sheet
- dplyr cheat sheet
- ggplot2 cheat sheet

Weeks 5-6

Dates:

September 23-October 4

Topics:

Descriptive Statistics, Probabilities, Distributions, Bayes rule, Beginning Linear Models

Readings

- Winter, Ch3
- Winter, Ch4

Supplementary Material

- R4DS: Exploratory Data Analysis
- R4DS: Model Basics
- 2017 LSA Course: Fitting Linear Models

Weeks 7-9

Dates:

October 7-25

Topics:

Correlation and Transformations, Multiple Regression, Categorical Predictors

Readings

- Winter, Ch 5
- Winter, Ch 6
- Winter, Ch 7

Supplementary Materials

- forcats cheat sheet

Weeks 10-11

Dates:

October 28-November 8

Notes:

  - Fall break October 28, no class.
  - Election Day, November 5

Topics:

Interactions, Inferences

Readings

- Winter Ch 8
- Winter Ch 9
- Winter Ch 10

Weeks 12-13

Dates:

November 11-22

Topics:

Generalized Linear Models

Readings

- Winter Ch 12
- Winter Ch 13

Weeks 14-15

Dates:

November 25-December 6

Notes:

Thanksgiving Break November 27-29, no class

Topics:

Mixed Effects Models

Readings

- Winter Ch 14
- Winter Ch 15

Supplementary Materials

- 2017 LSA Course: Mixed Effects Models

Week 16

Dates:
December 9-13

Notes:
Last day of class: December 11

Topics:
Wrap-up and outlook

9 Course Evaluation

9.1 Grade Components

Exercises	30%
Midterm Project	30%
Final Project	30%
Engagement	10%

9.2 Grading Scale

A	>= 90
B	80 to 89
C	70 to 79
D	60 to 69
E	<= 59

Assignment Submission

Links to Github Classroom Assignments will be posted to canvas. When you accept an assignment, github will create a version of the assignment repository for you. To submit your assignment, you just need to push your commits back to github (this will be covered in class).

Some assignments may have automated code "tests." These are there to provide some feedback while you are working on an assignment. Github classroom may display some "points" or "grades", but these are not final. Only the grades that appear on Canvas are the official grades for the course.

Exercises

There will be frequent, R exercise assignments to reinforce what we've done in class.

Midterm Project

There will be a midterm project to analyze a sample data set utilizing the methods covered in the course up to that point, and to report on your analysis.

Final Project

There will also be a final project in the same format as the midterm project, but to extend your analysis tools to the fuller suite of methods covered in the course.

Engagement

Inspired by Kirby Conrod's approach to Participation Grades

> This portion of the grade is a way for me to give you credit for informal/unstructured collaborative work that you do. Participation and collaboration are strong predictors of success and learning retention, so please make an effort to find a way that works well for you to participate and engage with your colleagues.

A well known process for solving programming problems is "Rubber Duck Debugging." It works by describing how each step of a program is supposed to work to another person or, as the name suggests, a rubber duck. Often the solution to the problem or the typo causing the bug jumps out at you during the process. Having a study buddy or study group could be really helpful if only for this purpose.

10 Late Submissions and Re-submissions

Every graded piece of work will have a due date. After a 2 day grace period, there will be a single, flat 5% deduction from late work, whenever it is submitted between the due date and the The Final-est Deadline

Midterm Grades

I will submit midterm grades on **October 25, 2024**, based on all work that has been submitted at that time.

The Final-est Deadline

The final deadline after which no more work will be accepted is **December 16, 2024**.

Add to calendar (link)

11 Doing Coursework

Group Work and Code Sources

It is acceptable to collaborate and confer with other students in the course. Any collaboration should be indicated in the assignment submission. You may also refer to code sources from elsewhere on the internet, as long as you also document the source, and explain what the code does. You might not receive credit for code which has been copied wholesale from another online source or from another student without credit or documentation.

Large Language Model (a.k.a. AI) Generated Code

There are a number of services that will generate code based on natural language queries. Some words of warning:

LLMs will lead you astray at this learning stage

It is frustrating to run code and not understand why it didn't work. **LLMs like ChatGPT or Github Copilot will only exacerbate this for you in the learning stages.**

In my own experience, Github Copilot will generate code that *looks* like some R code *might* look like, but in fact, references function arguments that and data columns that don't exist. After catching enough of these errors (which I was able to do after years of experience writing R code), I've turned off Copilot suggestions entirely because:

- It got in the way of the code I was *trying* to write.
- The rule-based RStudio hints and autocomplete suggestions are better.

When you're first beginning to learn to write R code, you won't have years of experience at your back to help you correct any LLM errors.

Explain what the code does

As stated above, you should provide credit to any external sources you turned to for code help, and explain what the resulting code does.

Attendance and Engagement

You are expected to attend all scheduled course meetings. It would be helpful, but not necessary, if you let me know in advance if you are going to miss any lectures.

If you feel sick in any way, including but not limited to the well-known symptoms of COVID-19 do not come to class. There are other mechanisms for demonstrating engagement than attending lectures.

I will also expect all of us in the course to treat each other with respect and civility in all aspects of the course, including

- In the audio of a Zoom meeting
- In the text chat of a Zoom meeting
- On any course discussion boards or other forums.

12 Academic Conduct

UK rules on academic offences

Appropriating someone else's work and portraying it as your own is cheating. Collaborating with someone and portraying that work as solely your own is cheating. Obtaining answers to homework assignments or exams from previous semesters is cheating. Using an internet search engine to look up a question and reporting that answer as your own is cheating. Falsifying data or experimental results is cheating. If you are unsure about whether a specific action is cheating, you may check with me.

The minimum penalty for a first offense is a zero on the assignment on which the offense occurred. If the offense is considered severe or if the student has other academic offenses on their record, more serious penalties, up to suspension from the University may be imposed.

When students submit work purporting to be their own, but which in any way borrows ideas, organization, wording or anything else from another source without appropriate acknowledgement of the fact, the students are guilty of plagiarism. Plagiarism includes reproducing someone else's work, whether it be a published article, chapter of a book, a paper from a friend or some file, or something similar to this. Plagiarism also includes the practice of employing or allowing another person to alter or revise the work which a student submits as their own, whoever that other person may be.

Students may discuss assignments among themselves or with an instructor or tutor, but when the actual work is done, it must be done by the student, and the student alone. When a student's assignment involves research in outside sources of information, the student must carefully acknowledge exactly what, where and how they employed them. If the words of someone else are used, the student must put quotation marks around the passage in question and add an appropriate indication of its origin. Making simple changes while leaving the organization, content and phraseology intact is plagiaristic. However, nothing in these Rules shall apply to those ideas which are so generally and freely circulated as to be a part of the public domain.

13 University Academic Policy Statements

Link to University Academic Policy Statements