LINGUISTICS PROGRAM
BOSTON UNIVERSITY - COLLEGE OF ARTS AND SCIENCES

GRS LX 795 - Quantitative Methods in Linguistics

Time:	Tues & Thurs 12:30-1:45p	**Location:**	CAS 220
Professor:	Daniel Erker	**Email:**	danerker@bu.edu
Office:	501a, 718 Comm. Ave	**Office hours:**	M 3:45-4:45p, T/R 11a-noon, and by appt.

Course website: The course will be hosted on Blackboard Learn.

Course description and goals: Modern linguistic research is increasingly making use of quantitative methods to analyze linguistic behavior, including aspects that were (or still are) considered to be categorical. Quantitative methods are helping us answer longstanding questions about the way language works and are also shaping the formulation of new questions. This course will guide students through quantitative approaches to examining linguistic data, including various types of data visualization, hypothesis testing, and data modeling. Along the way, students will gain proficiency in R, an open-source statistical environment. By the end of the course, students will know the logic behind a wide range of data science techniques and the practical skills required to use them appropriately.

Prerequisites for the course: Graduate standing in the Boston University Linguistics program, or consent of instructor.

Learning Outcomes: Students will…

1. Be able to make appropriate methodological choices in all aspects of a research project, including formulation of a question, data management, and statistical analyses.

2. Acquire the ability to summarize, visualize, and otherwise explore data using a variety of methods.

3. Understand the conceptual underpinnings of common statistical tests, and apply them appropriately.

4. Be able to critically evaluate quantitative analyses in a range of linguistic sub-fields.

5. Acquire proficiency in R suitable for moving beyond the course material.

Readings

- Diez, David M., Barr, Christopher D., & Cetinkaya-Rundel, Mine. (2017). *OpenIntro Statistics*. Third edition. https://www.openintro.org/stat/textbook.php

- Wickham, Hadley & Grolemund, Garrett. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* First edition. O'Reilly. http://r4ds.had.co.nz/

- Assorted papers/book chapters evincing recent trends in methodology (I will make these available via Blackboard).

Software

- **R** - a free software environment for statistical computing and graphics (downloadable from https://cran.r-project.org/).

- **R Studio -** software that provides a user interface in which you can send commands to R, edit your code, preview graphics, and see what is going on in your work environment (downloadable from https://www.rstudio.com/products/rstudio/download/).

Course Workflow

Given the nature of the course, it will be essential to devote a substantial chunk of our class time to the *doing* of data science. This means that unlike other courses, in which a lot of time is spent discussing and unpacking the readings, in-class discussion of readings will be kept to a minimum in our course. I nonetheless expect you to do the assigned readings and to internalize the content. I've chosen texts that are maximally supportive to independent learning. *OpenIntro Statistics* (*OIS*) has companion videos that supplement the readings, and its host website provides a number of additional learning resources. *R for Data Science* (*R4DS*) includes R code that can be copied-and-pasted directly into the R console as well as exercises at the end of sections that allow you to practice what you have just read. Our workflow will thus consist of a cycle of *Read-Talk-Practice*, distributed in the following way:

Outside of class

- Read *OIS* and watch associated videos online
- Read *R4DS* and do associated exercises (really do these)
- Read any additional assigned readings and do homework assignments

In class (*nb*: you will need to bring a computer to class)

- Discuss/take questions on key concepts in the readings
- Practice the R techniques that are relevant to the given readings, *e.g.* data visualization, transformation, running statistical tests

The 'glue' holding our *Read-Talk-Practice* cycle together will be *R Markdown* files. It will become much clearer what these are once we get started. For now, it suffices to say that *R Markdown* is a way for you to save and share all of the work that you do. It also makes it possible to author publishable quality reports on your analyses – I use R Markdown to write up the results of my own research.

A note on our order of operations: There is an inherent tension between learning the skills of quantitative data analysis and understanding the logic that underlies them. Because we will often need to do several things to a single data set – visualize, reshape, filter, run statistical tests, etc. – and because our reading about these operations has to proceed in some order, we will sometimes encounter a concept in practice before we have the opportunity to read deeply about its theoretical underpinnings and motivation. For instance, we will be doing linear regression in R much earlier than we will be reading about it in *OIS*. This is okay. I liken it to learning how to drive a car first and understanding how the car works second. Our first priority is getting a license to drive. Our second is to understand what's going on under the hood that's making the car go (*nb*: you are, of course, more than welcome to read ahead of our course schedule should you want to dive into the logic of a particular concept).

Participation

Learning how to do quantitative analysis takes practice. That means that active and constructive participation in class is expected and will be factored into course grades. You are adults and not obligated to inform me of absence, but please be aware that chronic absence is highly likely to affect your final grade by causing you to miss a lot of the laboratory-type work we will be doing in class.

Assignments

In addition to the *Developing a Variable Mindset* exercise and *Course Surveys*, there will be four *R Markdown*-based homework assignments that require you to use quantitative methods to analyze linguistic datasets. Each homework will include detailed instructions and will be accompanied by a reading specific to that dataset. In planning to do these, be sure to give yourself enough time to both read the associated paper and do the work of the assignment itself (i.e. visualization, data transformation, and analysis is R). In your write ups, keep in mind that these are exercises in applying knowledge and using the techniques and rhetoric of the field appropriately. That is, they will be evaluated not only how well you have executed the appropriate R code, but also on how clearly you have communicated to your reader (me) what it is that you have done. Homework write-ups should be submitted as both .html and .rmd files via Blackboard by 5 PM on the due date.

Project

Each of you will propose and complete a final project that demonstrates your grasp of the course material. The project may involve analysis of either: (a) your own dataset that is already collected, (b) a dataset you are planning to collect soon (in which case you can lay out the type of data you might encounter, generate artificial data of this sort, and look at the consequences of analyzing it in several different ways), or (c) someone else's dataset (e.g., any of the datasets in the languageR package). More details on this project will be distributed separately. One-page project proposals are due by 5 PM on February 22, and the final write-up is due by 5 PM on May 1.

Summary of Requirements

- Readings
- Participation

- Assignments
- Project

Grading

- 60% Homework assignments
- 30% Final project
- 10% Participation

Grading standards:

Grading standards			
93-100	A	78-79.99	C+
90-92.99	A-	73-77.99	C
88-89.99	B+	70-72.99	C-
83-87.99	B	60-60.99	D
80-82.99	B-	< 60	F

Academic Integrity

All students are responsible for understanding and complying with the BU Academic Conduct Codes, available at http://www.bu.edu/academics/policies/academic-conduct-code and http://www.bu.edu/cas/students/graduate/grs-forms-policies-procedures/academic-discipline-procedures/

Copyright notice

Course roadmap (subject to adjustment)

Week Dates	Reading for week	Main Topics & Assignments (Due Dates)
1 1-18	*OIS* C1(1st half) *Introduction to Data*	Introductions, syllabus, and course overview
2 1-23 1-25	*OIS* C1(2nd half) *Introduction to Data*	Data basics, numerical and categorical data Getting *R* and *R Studio* up and running Variable Mindset Exercise, Course Survey #1 (1-25)
3 1-30 2-1	*R4DS* I *Introduction & Explore* (C1-8)	Prerequisites, running R code, data visualization, transformation, exploratory data analysis
4 2-6 2-8	*OIS* C2 *Probability*	Defining probability, conditional probability, continuous distributions DE distribute Project Guidelines
5 2-13 2-15	*R4DS* V* *Communicate* (C26-30)	*R Markdown* and graphics for communication **Nb* We are reading R4DS out of order R Markdown Assignment #1 due (2-15)
2-20	**BU ON MONDAY SCHEDULE**	
6 2-22	*OIS* C3 *Random Variables*	Normal and binomial distribution Project Proposal due (2-22)
7 2-27 3-1	*R4DS* II *Wrangle* (C9-16)	Tibbles, data import, tidy R Markdown Assignment #2, Course Survey #2 due (3-1)
BU SPRING BREAK 3-5 to 3-9		

8 3-13 3-15	*OIS* C4 *Foundations of Inference*	Variability in estimates, confidence intervals, and the central limit theorem.
9 3-20 3-22	*R4DS* III *Program* (C17-21)	Pipes, functions, and vectors
10 3-27 3-29	*OIS* C5 *Inference Numerical Data*	Comparison of means, *t*-test, *ANOVA* R Markdown Assignment #3 due (3-29)
11 4-3	*R4DS* IV *Model* (C22-25)	Model building
4-5	**NO CLASS – PROFESSOR ERKER AT CONFERENCE**	
12 4-10 4-12	*OIS* C6 *Inference Categorical Data*	Difference of two proportions, chi-square
13 4-17 4-19	*OIS* C7 *Linear Regression*	Line fitting, residuals, correlation, linear regression R Markdown Assignment #4, Course Survey #3 due (4-19)
14 4-24 4-26	*OIS* C8 *Multiple & Logistic Regression*	Model selection, checking regression assumptions
15 5-1		Final Project due (5-1)