LINGUISTICS DEPARTMENT
BOSTON UNIVERSITY - COLLEGE OF ARTS AND SCIENCES

GRS LX 795 - Quantitative Methods in Linguistics – Spring 2020

Time:	MWF 10:10-11:00a	**Location:**	CAS 427
Professor:	Daniel Erker	**Email:**	danerker@bu.edu
Office:	501a, 718 Comm. Ave	**Office hours:**	M 1:30-2:30p, W 9-10a, 11:05a-12:05p, by appt.

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

Course Description

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. Students will gain proficiency in *R*, an open-source statistical environment. By the end of the course, students will understand 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 quantitative data analysis, including formulation of questions, data manipulation, and statistical tests.

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 and reproduce quantitative analyses in a range of linguistic sub-fields.

5. Acquire proficiency in *R* suitable for independent work beyond the course material.

Readings

- Diez, David M., Barr, Christopher D., & Cetinkaya-Rundel, Mine. (2019). *OpenIntro Statistics*. Fourth edition. https://www.openintro.org/book/os/

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

- **R Studio –** 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/).

- **LaTeX** – document preparation software. We will not need this until the last few weeks of the course (https://www.latex-project.org/get/).

Course Workflow

Given the nature of the course, it will be essential to devote a substantial amount of in-class time *doing* linguistically-oriented 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, slides, and labs 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
- Read *R4DS* and do associated exercises (really do these; you will be tempted not to, fight it)
- Read any additional assigned readings and do homework assignments

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

- Discuss/ask 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 that I have created to serve as companions to the course. 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 (and will demo this in the last weeks of the course).

A note on our order of operations: There is 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, subset, 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 this 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 the laboratory-type work we will be doing in class is expected and will be factored into course grades (10%). You are adults and not obligated to inform me of an unexpected absence, but please be aware that multiple (more than 3) unexcused absences over the course of the semester will result in a failing participation grade (a grade of 0).

Assignments

In addition to the *Developing a Variable Mindset* exercise and (ungraded and anonymous) *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 in 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 .html and .rmd files via *Dropbox* by 11 PM on the due date.

Assignment 4 – Optional analysis of your own data

For the final assignment of the course, students may choose between analyzing a data set that I provide or one of their own. The latter option may involve analysis of: (a) your own dataset that is already collected (but that you have not previously analyzed), (b) a dataset you are planning to collect (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. Details for this assignment will be distributed in Week 7.

Summary of Requirements

- Readings
- Participation
- Assignments

Grading

- 90% Homework assignments (22.5% each)
- 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 Code, available at https://www.bu.edu/cas/current-students/undergraduate/academic-conduct-code-2/. The academic conduct code specific to GRS can be found here https://www.bu.edu/cas/files/2017/02/GRS-Academic-Conduct-Code-Final.pdf.

A note on collaboration: We will spend a great deal of in-class time working together, both as an entire class and also in small groups. However, homework assignments must be completed individually. This is essential to building a sense of independence and confidence with quantitative methods. You gain very little by copying a classmate's R code or problem solution.

Copyright notice

All class materials are copyrighted and may not be reproduced for anything other than personal use without written permission from the instructor. This means that no course materials distributed, in hard copy or electronically, in conjunction with this course (including markdown files, html files, slides from presentations, and assignments etc.), may be shared with any note-sharing website.

Course schedule (subject to adjustment)

Week Dates	**Reading for week**	**Main Topics / Assignments and Due Dates**
1 1-22 1-24	*OIS* C1-2 *Intro and Summarizing Data*	Introductions, syllabus, and course overview The Variable Mindset
2 1-27 1-29 1-31	*R4DS* I *Introduction & Explore* (C1-4)	Data basics, numerical and categorical data Getting *R* and *R Studio* up and running Variable Mindset Exercise, Course Survey 1 due 1-29
3 2-3 2-5 2-7	*R4DS* I *Introduction & Explore* (C5-8)	Prerequisites, running R code, data visualization, transformation, exploratory data analysis
4 2-10 2-12 2-14	*OIS* C3 *Probability*	Defining probability, conditional probability, continuous distributions
2-17	**No Classes – Presidents' Day**	
5 2-18 2-19 2-21	*R4DS* V* *Communicate* (C26-30)	2-18 a BU Monday *R Markdown* and graphics for communication ***NB* - We are reading R4DS out of order**
6 2-24 2-26 2-28	*OIS* C4 *Dist. Random Variables*	Assignment 1 due 2-24 via *Google Drive* Normal and binomial distributions Principles of Visualization
7 3-2 3-4 3-6	*R4DS* II *Wrangle* (C9-16)	Assignment 4 Guidelines posted 3-2 Tibbles, data import, tidy Introduction to multivariate statistical tests
BU Spring Break 3-7 to 3-15		

8 3-16 3-18 3-20	*OIS* C5 *Foundations for Inference*	Variability in estimates, confidence intervals, and the central limit theorem. Multivariate tools continued. Option 1 Assignment 4 Proposals due 3-20
9 3-23 3-25 3-27	*R4DS* III *Program* (C17-21)	Course Survey 2 due 3-23 Pipes, functions, and vectors
10 3-30 4-1 4-3	*OIS* C6 *Inference for Categorical Data*	Assignment 2 due 3-30 Comparison of means, *t*-test, *ANOVA* Data wrangling
11 4-6 4-8 4-10	*R4DS* IV *Model* (C22-25)	Model building Mixed effects regressions
12 4-13 4-15 4-17	*OIS* C7 *Inference for Numerical Data*	Difference of two proportions, chi-square Mixed effects regressions continued
4-20	**No Classes – Patriot's Day**	
13 4-22 4-24	*OIS* C8 *Intro to Linear Regression*	Line fitting, residuals, correlation, linear regression Rbrul Writing papers with R Studio Assignment 3 due 4-24
14 4-27 4-29	*OIS* C9 *Multiple & Logistic Regression*	Model selection, checking regression assumptions Writing papers with R Studio continued

Tues May 5: Assignment 4 due