SYLLABUS: SPAN 589
Title: Data Science for Linguists
01:940:589:01 - Spring 2018
Meetings: AB 5141, 09:50–12:50

Professor: Joseph V. Casillas, PhD
Email: joseph.casillas@rutgers.edu
Office: AB 5174
Office hours: by appointment

Course description

In this course students examine the fundamental principles of doing experimental research in linguistics. Specifically, the focus is on developing an in depth understanding of the experimental paradigms and statistical procedures used in sociolinguistics, phonetics, psycholinguistics, syntax, and corpus linguistics. While a large part of the class will be spent on basics of how to analyze quantitative data, another goal of the class is to examine the statistical analyses which appear in actual published linguistic literature, and to discuss how students' current and future research might be analyzed statistically. Students will learn advanced techniques used to explore, tidy, visualize, and analyze data. We will also focus on how to make the aforementioned procedures reproducible and shareable. Students will develop a foundation in programming in R, as well as learning the most common tools at the disposal of todays data scientist (i.e. GitHub, Knitr, etc.).

Prerequisites: No prior experience with statistics or programming is necessary.

Materials

Class websites

- Sakai: https://sakai.rutgers.edu/portal/site/c0add82e-3c74-45df-8f34-dcb0fff07bf8
- Class website: www.jvcasillas.com/ru_teaching/ru_spanish_589/589_01_s2018/

Books

- Wickham, H. and G. Grolemund (2016). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.
- Johnson, K. (2011). *Quantitative Methods In Linguistics.* Wiley.
- Lewis-Beck, M. (1980). *Applied Regression: An Introduction.* Sage University Paper Series on Quantitative Applications in the Social Sciences - 22. Newbury Park, CA: Sage. ISBN: 9781483381497.
- Berry, W. and S. Feldman (1985). *Multiple Regression in Practice.* Sage University Paper Series on Quantitative Applications in the Social Sciences - 50. Newbury Park, CA: Sage. ISBN: 9780803920545.
- Schroeder, L, D. Sjoquist and P. Stephan (1986). *Understanding Regression Analysis: An Introductory Guide.* Sage University Paper Series on Quantitative Applications in the Social Sciences - 57. Newbury Park, CA: Sage. ISBN: 9780803927582.
- Hardy, M. (1993). *Regression with Dummy Variables.* Sage University Paper Series on Quantitative Applications in the Social Sciences - 93. Newbury Park, CA: Sage. ISBN: 9780803951280.

Weekly readings

Students will receive a package of readings to be distributed in electronic format (.pdf).

Coursework

Evaluation

Component	Percentage	Grade distribution	
Preparation and participation	15%	A	92–100
Programming assignments	40%	B+	87–91
Online presentation	10%	B	80–86
Midterm exam	15%	C+	77–79
Research project	20%	C	70–76
- paper (8%)		D	65–69
- presentation (6%)		F	0–64
- peer review (6%)			

Preparation and participation

Students are expected to attend class prepared and to actively participate. Part of this grade is derived from reading summaries. Students will occasionally be required to write a brief summary (max. 1 page) of the weekly readings and answer assigned questions.

Programming assignments

Students will complete 4 programming assignments over the course of the semester. These assignments are designed in a way so that the student must demonstrate adequate knowledge of basic programming and statistical principles covered in class. The skills required in each assignment are cumulative, each building on the material learned in the previous weeks. All statistical programming assignments must be completed in RMarkdown and will be handed in via GitHub unless otherwise noted.

Online presentation

The presentation will be on the statistical analyses used in some published paper. This presentation must be hosted on GitHub and in HTML format using RMarkdown. Aside from creating an online presentation, students will also be required to read and comment on the presentations of two classmates.

Midterm exam

There will be an in-class exam during the 11th week of the semester (April 3rd). Details will be provided beforehand.

Research Project

Overview Each student will complete a research project in which they put in practice the tools learned over the course of the semester. The primary focus will be on managing the project in an automatic and reproducible way so that it can be shared with other collaborators. The project will be hosted on GitHub and will include the following:

- slides
- manuscript
- r code

- data (raw and tidy)

Students have two options regarding the type of project they do:

1. Personal project (real)
    - For advanced students working on their own data
    - Ideal for QP, thesis, other projects
2. Hypothetical project (simulated)
    - For students w/o data
    - Ideal for students in proposal phase (IRB, NSF, ect.)

All projects require the prior approval of the professor. Project due dates will be established in class, but can be expected to be due several days before the university assigned final exam (though there is no exam).

Paper The manuscript will be a write up of the methods/results sections of a research article. The focus is on clearly and accurately explaining the statistical analyses used in the project and appropriately interpreting the results. The paper must be a literate document written in RMarkdown using `papaja`. We will demo this in class.

Presentation Students will present their work in a semi-finished state during the final week of the semester (10 min. presentation + 5 min. for questions). The slides of this presentation are part of the project and must also be hosted on GitHub.

Peer review Each student is required to evaluate the project of two other students. They will fork the project in order to evaluate the reproducibility of the code and the statistical validity of the analysis. Students will write up two evaluations for each peer: one for the professor (not to be seen by anybody else), and one for the owner of the project in the form of an issue/comment on GitHub. The evaluation written for the professor should be longer and more in depth. Both evaluations should be written in the style of a peer review for an academic journal, thus they should include comments, questions, suggestions, and *constructive* criticisms (#BeReviewer1). The point of this excersize is to help the author make the final product better.

Department rules and course policy

The course is designed to satisfy the learning goals of the Department of Spanish and Portuguese. More information available at: http://span-port.rutgers.edu/learning-goals

Communication

All course communication will be via Slack. You should have received an email with an invitation link to join the course Slack. Some rules for using Slack:

- Use an identifiable username and add your picture to your profile.
- Only the professor is allowed to use the @channel and @here mentions.
- While this is an informal communication channel, all rules of academic discourse apply.
- Ask and answer questions on the appropriate channel.
- Create channels as needed, especially for study groups.

Attendance

Regular class attendance is essential for successful completion of the course. More than 1 absences will have a negative effect on your final grade. The 2nd absence and every subsequent absence after that will result in the loss of 5% point off the final overall course grade, regardless of reason. Keep in mind that while you have 1 "free" absence, on the day/s you miss you will not be able to earn participation points, you will miss the material given in class and you might miss your own presentation. If you are absent, contact a classmate immediately to get the assignments and to keep up with the material scheduled in the syllabus. The instructor is not responsible for catching you up. Do not send emails to the instructor asking for updates if you missed class.

Any planned absence that you are aware of ahead of time, such as religious holidays recognized by Rutgers University or Dean's excuses, should be made up before the absence occurs. If you know that you will be absent, it is your responsibility to let the instructor know ahead of time. All holidays or special events observed by any religion will be honored for those students who show affiliation with that particular religion. Absences pre-approved by the RU Dean of Students (or Dean's designee) will be honored.

Code of academic integrity

The professor will initiate an academic integrity case against students suspected of cheating, plagiarizing, or aiding others in dishonest academic behavior. Students are responsible for reading and understanding the Code of Academic Integrity.

Examples of academic dishonesty include, but are not limited to, plagiarism, cheating, and aiding and abetting dishonesty. An example of plagiarism would be to submit a written sample which in part or in whole is not the student's own work without attributing the source. Cheating includes allowing another person to do your work and to submit the work under one's own name. Any work which is submitted for a grade must be 100% the student's own work. If you are not sure when it is appropriate to seek help, please see the professor.

Plagiarism is the use of another person's words, ideas, or results without giving that person appropriate credit. Do not plagiarize.

Rutgers University Academic Integrity Policy, p. 2: http://academicintegrity.rutgers.edu/files/documents/AI_Policy_2013.pdf

---

For more information

- http://academicintegrity.rutgers.edu/
- http://www.libraries.rutgers.edu/avoid_plagiarism.

Students with disabilities

Rutgers University welcomes students with disabilities into all of the University's educational programs. In order to receive consideration for reasonable accommodations, a student with a disability must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: https://ods.rutgers.edu/students/documentation-guidelines.

If the documentation supports your request for reasonable accommodations, your campus's disability services office will provide you with a Letter of Accommodations. Please share this letter with your instructors and discuss the accommodations with them as early in your courses as possible. To begin this process, please complete the Registration form on the ODS web site at: https://ods.rutgers.edu/students/registration-form.

Week	Date	Topic
1	01/16	Intro, Setup, GitHub, PA1 assigned
2	01/23	Stats: Variables, distributions Programming: RStudio projects, RMarkdown, ggplot
3	01/30	Stats: Hypothesis testing, comparing means Programming: PA2 assigned, R basics, getting/tidying data, ggplot
4	02/06	Stats: Bivariate correlation Programming: R basics, tidying/transforming data, ggplot
5	02/13	Stats: The linear model: regression Programming: PA3 assigned, R basics, GitHub Pages
6	02/20	Stats: The linear model/regression Programming: R basics, HTML presentations, slidify
7	02/27	Stats: The linear model/regression Programming: PA4 assigned, HTML presentations, xaringan
8	03/06	Stats: The linear model/regression Programming: papaja
	03/13	**Spring break**
9	03/20	Stats: General Linear Model Programming: modeling tips and tricks
10	03/27	Stats: General Linear Model Programming: modeling tips and tricks
11	04/03	**Exam**
12	04/10	Stats: Generalized Linear Model **Online presentations due**
13	04/17	Stats: Generalized Linear Model **Online presentation comments due**
14	04/24	Research project presentations, residuals
15	05/01	Reading day
16	05/??	Final exam day (Projects due)