

General information

This course is a selective introduction to predictive modeling applications in linguistics. We start with a one-session intro of predictive modeling with an emphasis on regression modeling, which will survey model formulation, model selection, multifactoriality, and validation. Then, we work our way through a variety of regression modeling applications: linear regression, binary logistic regression, multinomial, and ordinal regression models. Then, one session will be concerned with model diagnostics and, perhaps, model validation. Finally, there is a session on classification and regression trees. Like its prerequisite course Ling 201, this course is based on the third edition of my textbook *Statistics for linguistics with R: a practical introduction (2021)* and uses the open source [programming language R](#) and, as an IDE, [RStudio](#).

Course requirements and grading

Attendance is not required and will not be monitored. For course credit, pick two course-final take-home assignments from [these 10](#) and analyze the data comprehensively (as if they were your own); note the difficulty levels, which also correspond to weights: if you do equally well on two assignments with different difficulty levels, you'll get more points for the one with the higher difficulty level. Note: I expect you to do these yourself and alone; check the [UCSB Student Conduct Code](#) and/or with me if you're unsure what's permitted. You can score extra/bonus points by good class participation (in sessions that you attend). Assignments must be sent to me as R code files (.r), R Markdown (.rmd) or Quarto documents (.qmd), or as fully self-contained HTML reports and must have the following file name structure: `<202_lastname_assignment0#.html>` (as in `<202_smith_assignment02.html>`). The final grade will depend on your number of points, which are a function of the difficulty levels of the assignments you picked and the quality of your analyses.

Contact (STG)

Office hours: Zoom, upon appointment
Web: [<https://www.stgries.info>](https://www.stgries.info)
Email: [<stgries@linguistics.ucsb.edu>](mailto:stgries@linguistics.ucsb.edu)

Course plan

(1) 04/05: Overview of modeling fundamentals

Read as follow-up: go over the slides and the HTML report
Read for next time: SFLWR³ 5.2.1-5.2.3 (without 5.2.3.3)

(2) 04/12: Linear modeling 1

Read for next time: SFLWR³ 5.1, 5.2.3.3, 5.2.4-5.2.7, 5.5

(3) 04/19: Linear modeling 2

Read for next time: SFLWR³ 5.3.1-5.3.3

(4) 04/26: Binary logistic regression modeling 1

Read for next time: SFLWR³ 5.3.4-5.3.5

(5) 05/03: Binary logistic regression modeling 2

Read for next time: SFLWR³ 5.4.1

(6) 05/10: Multinomial regression

Read for next time: SFLWR³ 5.4.2

(7) 05/17: no class ([ICAME](#) and [ICCG](#))**(8) 05/24: Ordinal regression**

Read for next time: SFLWR³ 5.6

(9) 05/31: Model assumption & diagnostics

Read for next time: SFLWR³ 7.1

(10) 06/07: Trees

Preparation: you should make sure you have (up-to-date! versions of) the following software installed (in this order):

- R (<https://cran.r-project.org/>);
- RStudio (<https://posit.co/products/open-source/rstudio/>).

If you're likely to do statistical analyses for your own data, I'd also recommend you install [Quarto](#).