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 by the end of 17 June 2025 and must have the following file name structure: <202_lastname_assignment0#.html> (as in <202_**smith**_assignment0**2**.html>). The final grade will depend on your number of points, which are a function of (i) the difficulty levels of the assignments you picked, (ii) the statistical quality of your analyses (does the code work? did you explore and prepare the data? choose the right method? visualize properly? summarize the findings in writing? etc.) and (ii) the form in which you submit it (on a scale from a haphazard piece of code crap most of which doesn't even work to a nicely formatted HTML knitted from Quarto).

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

Course plan

(1) 04/03: Linear modeling 1
Read as follow-up: SFLWR[3] 5.2.1-5.2.3 (without 5.2.3.3)
Read for next time: SFLWR[3] 5.1, 5.5

(2) 04/10: Model formulation & selection
Read for next time: SFLWR[3] rest of 5.2

(3) 04/17: Linear modeling 2
Read for next time: SFLWR[3] 5.3.1-5.3.3

(4) 04/24: Binary logistic regression modeling 1
Read for next time: SFLWR[3] 5.3.4-5.3.5

(5) 05/01: Binary logistic regression modeling 2
Read for next time: SFLWR[3] 5.4.1

(6) 05/08: Multinomial regression
Read for next time: SFLWR[3] 5.4.2

(7) 05/15: Ordinal regression
Read for next time: SFLWR[3] 5.6-5.7

(8) 05/22: Model assumption & diagnostics

(9) 05/29: Similarity-based prediction and clustering
Read for next time: SFLWR[3] 7.1

(10) 06/05: 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.