General information

This course is a more advanced course on statistical modeling with an emphasis on various kinds of regression modeling; it presupposes a good understanding of the third edition (2021) of my textbook *Statistics for Linguistics with R: a practical introduction*. We begin with a first recap of linear and generalized linear regression modeling. We then discuss the use of contrasts and general linear hypothesis tests for linear and generalized linear regression models, followed by some ideas on how to explore curvature in data (regressions with breakpoints, polynomial regressions, and generalized additive models). This is followed by a larger chunk on linear and generalized linear mixed-effects (or multilevel) modeling, where we reanalyze published data and discuss numerical and visual exploration of regression results. The last session is then devoted to random forests. Obviously, we use the open source software tool R.

Course requirements and grading

i. regular attendance in class;
ii. preparation for, and active participation in, class. That is, I expect you to do the readings and/or assignments so that you can discuss them and/or ask about things you have not understood;
iii. a review of the statistical methods of a paper/ms;
iv. my current idea is to have you do a comprehensive (!) analysis of a data set (ideally, one of yours, but it could be something else), using what you have learned in this class.

The review is due as a pdf called <204_review_lastname.pdf> together in one email with the paper you reviewed called <204_paper_lastname.pdf>. The comprehensive-analysis assignment is due as a fully executable R script called <204_assignment_lastname.r> (.rmd or .qmd is of course also fine) together in one email with the data file that you analyzed as a tab-separated .csv file called <204_assignment_lastname.csv>; reviews/assignments that do not conform to this will be considered as not submitted! The final grade will depend on your number of points. You can get 100 points by

i. active participation in class (20 points);
ii. submitting the review in good quality and in a timely fashion (30 points);
iii. submitting the assignment in good quality and in a timely fashion (50 points).

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

Course plan

(1) 10/02: Recap: (fixed-effects linear) and binary logistic regression models
Follow-up: SFLWR[3] 5.1-5.3 but not 5.2.1.3 and 5.2.3.3 (this looks like a lot but most of it is of course recap from 202)
Homework: the exercises/homework questions in the html report you can answer

(2) 10/09: Effects. coefficients, general linear hypothesis tests 1
Follow-up: the html report
Homework: the exercises/homework questions in the html report you can answer

(3) 10/16: Effects. coefficients, general linear hypothesis tests 2
Follow-up: the html report, SFLWR[3] 5.2.3.3

(4) 10/23: Breakpoints & curvature
Follow-up: the html report, SFLWR[3] 5.2.1.3

(5) 10/30: no class (colloquium in OR) → statistical reviewing
As 'class activity': review (statistically) Parviainen & Fuchs (2018) (or something else, ask me)
Homework: SFLWR[3] 6.-6.3

(6) 11/06: Mixed-effects modeling, part 1
Follow-up: the html report

(7) 11/13: no class (grant evaluations for FWO)
Homework: analyze the data from SFLWR[3]: 6.4 but don't dichotomize `GIVENNESS`

(8) 11/20: Mixed-effects modeling, part 2
Follow-up: the html report
Homework: Gries (2015)

(9) 11/27: Mixed-effects modeling, part 3
Homework: Gries (2021: Sections 7.1-7.2)

(10) 12/04: Random forests
Follow-up: Strobl, Malley, & Tutz (2009)

References

Statistics for linguists (with R)

Gries, Stefan Th. 2021. *Statistics for linguistics with R: a practical introduction*. 3rd rev. and ext. ed. Berlin & New York: De Gruyter Mouton, pp. 496

Sonderegger, Morgan. 2023. *Regression modeling for linguistic data*. Cambridge: MA: The MIT Press, pp. 440.

Speelman, Dirk, Kris Heylen, & Dirk Geeraerts (eds.). 2018. *Mixed-effects regression models in linguistics*. Berlin & New York: Springer, pp. 146.

Winter, Bodo. 2019. *Statistics for linguists: an introduction using R*. London & New York: Routledge, pp. 310.

General statistics and/or general R and/or (regression) modeling applications

Baayen, R. Harald & Maja Linke. 2020. Generalized additive mixed models. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 562-591. Berlin & New York: Springer.

Baguley, Thom. 2012. *Serious stats: a guide to advanced statistics for the behavioral sciences*. Basingstoke & New York: Palgrave Macmillan.

Crawley, Michael J. 2013. *The R book*. 2nd ed. Chichester: John Wiley and Sons.

Egbert, Jesse & Luke Plonsky. 2020. Bootstrapping techniques. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 592-610. Berlin & New York: Springer.

Faraway, Julian J. 2015. *Linear models with R*. 2nd ed. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Faraway, Julian J. 2016. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. 2nd ed. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Fox, John & Sanford Weisberg. 2019. *An R companion to applied regression*. 3rd ed. Los Angeles & London: Sage.

Gelman, Andrew & Jennifer Hill. 2008. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2015. The most underused statistical method in corpus linguistics: Multi-level (and mixed-effects) models. *Corpora* 10(1). 95-125.

Gries, Stefan Th. 2020. On classification trees and random forests in corpus linguistics: […]. *Corpus Linguistics and Linguistic Theory* 16(3). 617-647

Gries, Stefan Th. 2021. (Generalized linear) Mixed-effects modeling: a learner corpus example. *Language Learning* 71(3). 757-798.

Gries, Stefan Th. & Allison S. Adelman. 2014. Subject realization in Japanese conversation by native and non-native speakers: exemplifying a new paradigm for learner corpus research. *Yearbook of Corpus Linguistics and Pragmatics 2014: New empirical and theoretical paradigms*, 35-54. Cham: Springer.

Harrell, Frank. 2015. *Regression modeling strategies: […]*. 2nd ed. New York: Springer.

Matloff, Norman. 2017. *Statistical regression and classification: from linear models to machine learning*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Makin, Tamaer R. & Jean-Jacques Orban de Xivry. 2019. Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* 2019;8:e48175.

Parviainen, Hanna & Robert Fuchs. 2018. 'I don't get time only': an apparent-time investigation of clause-final focus particles in Asian Englishes. *Asian Englishes* 21(3). 285-304.

Schäfer, Roland. 2020. Mixed-effects regression modeling. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*, 534-561. Berlin & New York: Springer.

Wilkinson, Leland & the Task Force on Statistical Inference. 1999. Statistical methods in psychology journals: guidelines and explanations. *American Psychologist* 54(8). 594-604.

Zuur, Alain, Elena N. Ieno, & Chris S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1. 3-14.

Zuur, Alain, Elena N. Ieno, & Neil Walker, Anatoly A. Saveliev, & Graham M. Smith. 2009. *Mixed effects models and extensions in ecology with R*. Berlin & New York: Springer.