General information

This course is a hands-on introduction to more advanced statistical methods to analyze observational and experimental data. After a small introduction to the processes of modeling or model selection and your recap of monofactorial methods and graphs, we systematically extend monofactorial tests to their multifactorial and multivariate counterparts. We begin with the linear model and extend correlations and *t*-tests to multiple linear regression, ANOVAs, and ANCOVAs. We then broaden the scope to the powerful methods included in generalized linear modeling by extending chi-squared tests to binomial logistic regression for binary dependent variables and Poisson regression for dependent variables that are counts and then proceed to ordinal logistic and multinomial regression. There is also one session on tree-based methods (classification and regression trees as well as random forests). In addition to these modeling techniques, we also discuss the exploratory method of hierarchical cluster analysis to find structure in large, potentially messy data sets. We use the open source software tool the open source software tool R and Gries (2013).

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 work on the code files so you can discuss them and/or ask about things you have not understood;
   - submit two of the four assignments as presentable HTML reports called <202_assignment[12]_lastname.html> (!!) by 19 March 2021 @ 07:00 PST by email.

The final grade will depend on the number of points you score. You can get 100 points by

i. attending, and participating in, all classes;
ii. submitting two assignments in good quality and in a timely fashion (each assignment is worth max. 50 points); the assignments will be graded on exploration of the data, conducting the/one right kind of analysis, diagnostics/validation, visualization, overall code 'elegance, and a succinct write-up of the interpretation of the results. For each assignment, you can send me one draft solution for feedback before the final submission.

Exceptionally good participation or homework assignments can result in max. 15 bonus points.

Contact Office hours: Zoom by appointment
Web: <http://www.stgries.info>
Email: <stgries@linguistics.ucsb.edu>

Course plan

(1) **01/04: multifactoriality and aspects of models, their selection, and validation**
Read for next time: slides, SFLWR: Section 5.1-5.2, Crawley (2013:388-401),
Do for next time: install R (≥4.0.3) and RStudio (≥1.3) on your computer

(2) 01/11: multiple linear regression 1: the mtcars data
Read for next time: Zuur, Ieno, & Elphick (2010); Hilpert & Blasi (2021)

(3) 01/18: multiple linear regression 2: the dative alternation
Recommendation for follow-up: Field, Miles, & Field (2012: Ch. 7, 10-12, 14) ;-)
Read for next time: SFLWR: Section 5.3 (optional: Speelman 2014)

(4) 01/25: binary logistic regression 1: the genitive alternation
Recommendation for follow-up: Field, Miles, & Field (2012: Section 8.1-8.8)

(5) 02/01: binary logistic regression 2: another dative alternation
Read for next time: SFLWR: Section 5.4.3

(6) 02/08: count/Poisson regression
Read for next time: SFLWR: Section 5.4.1

(7) 02/15: exercise for count/Poisson regression / ordinal logistic regression
Read for next time: SFLWR: Section 5.4.2

(8) 05/22: multinomial regression
Recommendation for follow-up: Field, Miles, & Field (2012: Section 8.9)
Read for next time: Levshina (2021), Crawley (2013:768:784)

(9) **03/01 tree-based approaches**
Recommendation for follow-up: Gries (to appear)
Read for next time: SFLWR: Section 5.6, Moisl (2021)

(10) 03/08: similarity-based predictions and cluster analysis

Nearly everything but Gries (2013) – e.g., data files, additional readings, code – will be available on the course website; Crawley (2013) is available in the library as an e-book.

References / Bibliography

Statistics for linguists (with R)

Baayen, R. Harald. 2008. *Analyzing linguistic data*. Cambridge: Cambridge University Press.

Desagulier, Guillaume. 2018. *Corpus linguistics and statistics with R: Introduction to quantitative methods in linguistics*. Berlin & New York: Springer.

Gries, Stefan Th. 2013. *Statistics for linguistics with R: a practical introduction*. 2nd rev. and ext. ed. Berlin & New York: De Gruyter Mouton.

Hilpert, Martin & Damián E. Blasi. 2021. Fixed-effects regression modeling. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*. Berlin & New York: Springer.

Levshina, Natalia. 2021. Conditional inference trees and random forests. In Magali Paquot & Stefan Th. Gries (eds.), *A practical handbook of corpus linguistics*. Berlin & New York: Springer.

Moisl, Hermann. 2021. Cluster analysis. In Magali Paquot & Stefan Th. Gries (eds.), A practical handbook of corpus linguistics. Berlin & New York: Springer.

Speelman, Dirk. 2014. Logistic regression: a confirmatory technique for comparisons in corpus linguistics. In Dylan Glynn & Justyna A. Robinson (eds.), *Corpus methods for semantics: quantitative studies in polysemy and synonymy*, 487-533. Amsterdam & Philadelphia: John Benjamins.

General statistics and/or general R (good to catch up on things)

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

Field, Andy, Jeremy Miles, Zoë Field. 2012. *Discovering statistics using R*. Los Angeles & London: Sage.

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, & Graham M. Smith. 2007. *Analysing ecological data*. Berlin & New York: Springer. **[UCSB web]**

(Regression/linear) modeling

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

Fox, John. 2016. *Applied regression analysis and generalized linear models*. 3rd ed. Los Angeles, CA et al.: Sage Publications.

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

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