General information

This course is a hands-on introduction to fundamentals of statistical and data mining / machine learning methods in linguistics, it is based largely on the second edition (2013) of my textbook *Statistics for Linguistics with R: a practical introduction*. We begin by looking at a few basic notions of statistical analyses (e.g., variables, hypotheses, significance etc) and then discuss the logic of quantitative studies using the null-hypothesis falsification approach as well as how data should be set up for subsequent statistical evaluation. Then, we will explore data preparation and processing with the open-source programming language and environment R. The largest part is concerned with a variety of classification and regression tools such as different kinds of regression models, classification and regression trees, random forests, and unsupervised learning.

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 (preparatory and follow-up) readings so that you can discuss them and/or ask about things you have not understood;
   - work on (in pairs) and submit two small assignments (by email, see below);
iii. one office hour visit in the quarter during which we talk about how statistical methods may be applied to something you are working on or something you find interesting;

All assignments are due as executable R scripts called <104_assignment0#_lastname.r>; 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. submitting all assignments in good quality and in a timely fashion (each assignment is worth max. 45 points; all assignments are due 11 Dec at 20:00 PST). Each assignment is supposed to be an analysis of a linguistic data set (I can provide those) that we haven't analyzed in the same way in the course; the assignments submitted by a student need to involve different statistical methods. An assignment can be submitted early once to get feedback before the final submission as an executable R script called <104_assignment0#_lastname_draft.r>.
ii. coming to the office appointment to discuss your assignment choices, to ask questions regarding course contents, or talk about how course contents could maybe be applied in your own research (worth 10 points).

Good participation and/or homework assignments can result in up to 10 bonus points.

Contact Office hours: Wed 14:30-15:30 in my office and upon appointment
Web: <http://tinyurl.com/stgries>
Email: <stgries@linguistics.ucsb.edu>

Course plan

(1) 09/27: Fundamentals of statistical methods
Read as follow-up: <104_01_intro-stat.pdf> and SFLWR 1.1-1.3
Read for next time: SFLWR 2.1-2.5 (note in particular exercises 1-19!)

(2) 10/04: R: functions, arguments, data structures
Read for next time: SFLWR 2.6-2.7 (note in particular exercises 20-23!)
SFLWR 3.1.1-3.1.3.6, 3.2

(3) 10/11: R: programming (conditionals, loops, apply, functions) and plotting
Read for next time: SFLWR 5.1-5.2

(4) 10/18: linear regression
Read for next time: SFLWR 5.3

(5) **10/25: binary logistic regression**
Read for next time: SFLWR 5.4.2

(6) 11/01: multinomial regression
Read for next time: review SFLWR 5.1-3, 5.4.2

(7) 11/08: regression practice
Read for next time: Torgo (2011: Chapter 2, but mainly Section 2.6)

(8) 11/15: classification (and regression) trees
Read for next time: Torgo (2011: Section 3.1-3.3)

(9) 11/22: random forests
Read for next time: SFLWR 5.6, Torgo (2011: Section 2.5)

(10) 11/29: missing data imputation and cluster analysis

Preparation: you should make sure you have the following software installed (in this order):
- R (<https://mran.revolutionanalytics.com/download/#download>;
- the math library (<https://mran.revolutionanalytics.com/download/#download>;
- RStudio (<http://www.rstudio.com/products/rstudio/download/>).

Then start R (with administrator access!) and run the following line:

```
install.packages(c("amap", "Amelia", "arules", "arulesViz", "car", "cluster",
      "DMwR", "effects", "fpc", "gbm", "ggplot2", "MASS", "nnet", "party",
      "polytomous", "pvclust", "randomForest", "rgl", "rms", "rpart",
      "tree"), dependencies=TRUE)
```

References / Bibliography

Books, articles, websites

Statistics for linguists (with R)

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

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

Johnson, Keith. 2008. *Quantitative methods in linguistics*. Malden, MA & Oxford: Blackwell.

General statistics and general R

Adler, Joseph. 2009. *R in a nutshell*. Sebastopol, CA: O'Reilly.

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.

Good, Philip I. & James W. Hardin. 2012. *Common errors in statistics (and how to avoid them)*. 4th ed. Hoboken, NJ: John Wiley and Sons.

Horton, Nicholas J. & Ken Kleinman. 2011. *Using R for data management, statistical analysis, and graphics*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

R Development Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.

Spector, Phil. 2008. *Data manipulation with R*. New York: Springer.**[l, o]**

Statsoft Inc. 2012. *Electronic Statistics Textbook*. Available online at <http://www.statsoft.com/textbook>.

Teetor, Paul. 2010. *R cookbook*. Sebastopol, CA: O'Reilly.

Statistical learning/data mining

Hastie, Trevor, Robert Tibshirani, & Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Berlin & New York: Springer. **[l, o]**

James, Gareth, Daniela Witten, Hastie, Trevor, & Robert Tibshirani. 2013. *An introduction to statistical learning, with applications in R*. Berlin & New York: Springer. **[l, o]**

Torgo, Luís. 2011. *Data mining with R: learning with case studies*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC. **[l, o]**

Yu-Wei, Chiu. 2015. *Machine learning with R cookbook*. Birmingham: Packt Publishing.

R and graphics

Keen, Kevin J. 2010. *Graphics for statistics and data analysis with R*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Murrell, Paul. 2011. *R graphics*. 2nd ed. Boca Raton, FL, London, New York: Chapman & Hall / CRC. **[l, o]**

Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with R*. New York: Springer.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. 2nd ed. New York: Springer.

<http://gallery.r-enthusiasts.com/>

<http://www.datavis.ca/gallery>

R and programming

Braun, W. John & Duncan J. Murdoch. 2008. *A first course in statistical programming with R*. Cambridge: Cambridge University Press.

Chambers, John M. 2008. *Software for data analysis: programming with R*. New York: Springer.

Gentleman, Robert. 2008. *R programming for bioinformatics*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Rizzo, Maria L. 2008. *Statistical computing with R*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC.

Wickham, Hadley. 2015. *Advanced R*. Boca Raton, FL, London, & New York: Chapman & Hall / CRC. **[l, o]**

Fun stuff

Best, Joel. 2001. *Damned lies and statistics: untangling numbers from the media, politicians, and activists*. Berkeley, Los Angeles, London: The University of California Press.

Gonick, Larry & Woolcott Smith. 1993. *The cartoon guide to statistics*. New York: HarperCollins.

Huff, Darrell. 1954. *How to lie with statistics*. New York: W.W. Norton.