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 R reports, i.e. as self-contained HTML files generated with RStudio and must have the following name (structure): <104_lastnames_assignment0#.html>; as in <104_smith_miller_assignment02.html> assignments that do not conform to these requirements 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 15 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. Each assignment can be submitted early once to get feedback before the final submission; this, too, would be an R report called <104_lastnames_assignment0#-draft.html>.
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) 09/28: 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, bring this cheatsheet to classes from now on)

(2) 10/05: R: functions, arguments, data structures
Read for next time: SFLWR 2.6-2.7

(3) 10/12: R: programming (functions, conditionals, loops, apply)
Read for next time: SFLWR 5.1-5.2

(4) 10/19: linear regression
Read for next time: SFLWR 5.3

(5) 10/26: regression practice
Read for next time: SFLWR 5.4.2

(6) 11/02: binary logistic regression
Read for next time: review SFLWR 5.1-3, 5.4.2

(7) 11/09: binary logistic regression practice
Read for next time: Torgo (2011: Chapter 2, but mainly Section 2.6)

(8) 11/16: classification and regression trees
Read for next time: Torgo (2011: Section 3.1-3.3)

11/23: no class (Thanksgiving)

(9) 11/30: no class (STG in Moscow) but I prepared exercise data for you
Read for next time: SFLWR 5.6

(10) 12/07: cluster analysis

Preparation: you should make sure you have the following software installed (in this order):
- R (<https://cran.r-project.org/>);
- RStudio (<https://www.rstudio.com/products/rstudio/download/#download>).

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

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