LingPhil PhD course "Statistics for linguistics with *R* bootcamp"

Place: UiT The Arctic University of Norway

Time: August 3-7, 2020

Credits: 5 ECTS

Course instructors:
Professor Stefan Th. Gries — University of California, Santa Barbara & Justus Liebig University Giessen

Course organizers:
Jason Rothman — UiT The Arctic University of Norway
Stefanie Wulff — University of Florida & UiT The Arctic University of Norway

Course content:
Statistics for linguistics with R is a hands-on introduction to statistical methods for both graduate students and seasoned researchers and is based on the second edition (2013) of Gries' textbook *Statistics for linguistics with R*. The course is mainly intended for linguists who already have a basic knowledge in statistics and some experience using *R*, and who wish to improve their proficiency in statistical analysis of linguistic data. Participants who are new to statistics and/or *R* should prepare beforehand by working through the readings listed below. The course puts a particularly strong emphasis on various kinds of fixed- and mixed-effects regression modeling as well the use of other predictive modeling techniques such as classification/conditional inference trees and (random) forests. The course features:

- a brief recap of basic aspects of statistical evaluation as well as several descriptive statistics insofar as they facilitate later predictive modeling approaches;
- a selection of monofactorial statistical tests for frequencies, means, and correlations and how they constitute special (limiting) cases of regression methods;
- an exploration of different kinds of multifactorial regression modeling approaches as well as other techniques on the basis of both observational and experimental, published and unpublished data.

For all modeling methods to be explored, we will discuss how to test their assumptions and visualize their results with instructive annotated statistical graphs. There also will be in depth discussion of different model selection strategies, how to interpret predictive modeling results (such as different kinds of interactions and contrasts), threats to the validity of modeling, etc.

Learning methods:
This is a five-day intensive course that requires:

- the reading assignment (see Required Readings listed below) to be completed prior to the start of the course;
- downloading and installing the software (which you will have been emailed about) via links and emailed instructions prior to course start;
- testing that the software packages are functional on your computer prior to class.

The course will be taught in English and grading is done on a pass/no pass basis. The course will feature lecture-style teaching, with about half of the instructional time each day being hands-on work on a variety of different data sets. Data sets and (thousands of lines of) code will be provided to the participants, as will be a variety of helper functions that participants will be able to use for their own statistical applications. Also, we will discuss queries that were sent to *R* newsgroups as well as reviews of papers under review with an eye to help participants understand what mistakes to avoid. The course will consist of a morning and an afternoon teaching module from Monday through Friday of one week. It will run much longer than the typical "class", hence the name bootcamp, starting at 9am and finishing at 5pm with a 1.5 hour break for lunch at midday, and 30-minute coffee breaks in the morning and afternoon.

Course schedule:
Day 1: 3-4 hours lecture: linear fixed-effects modeling; 2-3 hours practice
Day 2: 3-4 hours lecture: generalized linear fixed-effects modeling; 2-3 hours practice
Day 3: 3 hours lecture: linear mixed-effects modeling; 3 hours practice
Day 4: 3 hours lecture: generalized linear mixed-effects modeling; 3 hours practice
Day 5: 3-4 hours lecture: tree-based approaches; 2-3 hours practice
(approx. 16 hours of teaching and 14 hours tutoring in total, yielding 5 classes and 30 hours)

Learning outcomes:
At the end of the course, participants will be able to understand any discussion of a regression model they come across in research literature and will be able to conduct their own fixed- and mixed-effects modeling analyses; time permitting, there will be a small section on how to write small statistical/visualization functions yourself.

Course requirements:
PhD students will be awarded 5 ECTS if they

- read the required texts and download and test the software prior to the course;
- attend all teaching sessions;
- provide a written question each evening to the instructor, a selection of which will be used in the course to go over common queries;
- complete one practical assignment of a data set provided by the instructor as the final assessment.

Any student with an interest in statistics for empirical research is encouraged to attend.

Course evaluation by students:
The students will be expected to evaluate the overall quality of the lectures, relevance of the reading materials, student-instructor interaction and learning outcomes achieved. All course evaluation reports provided by students will be submitted to the Norwegian Graduate Researcher School in Linguistics and Philology (LingPhil) after the course. The template for course evaluation by students can be found at https://www.ntnu.edu/lingphil/course-proposals.

Contact persons:
Professor Jason Rothman (jason.rothman@uit.no) or Professor Stefanie Wulff (swulff@ufl.edu)

Course plan

Time	Day 1: Monday August 3	Day 2: Tuesday August 4	Day 3: Wednesday August 5	Day 4: Thursday August 6	Day 5: Friday August 7
9.00–9.15	Coffee & cookies	Coffee & cookies	Coffee & cookies	Coffee & cookies	Coffee & cookies
2 academic hours: 9.15–10.00 10.15–11.00	Lecture 1	Lecture 2	Lecture 3	Lecture 4	Lecture 5
11.00–12.00	Lunch	Lunch	Lunch	Lunch	Lunch
2 academic hours: 12.15–13.00 13.15–14.00	Lecture 1 and Practice	Lecture 2 and Practice	Lecture 3 and Practice	Lecture 4 and Practice	Lecture 5 and Practice
14.00–14.15	Coffee break	Coffee break	Coffee break	Coffee break	Coffee break
2 academic hours: 14.15–15.00 15.15–16.00	Practice	Practice	Practice	Practice	Practice

Modules
There are 5 modules in this course. Each day introduces a different module.

MODULE 1

This module will introduce participants to linear modeling, in particular interpretations of coefficients; the concepts of interactions and multifactoriality, and how to select appropriate models; how to set contrasts properly; how to do model diagnostics; and how to interpret curvature in model visualizations. The data we will use for illustration are reaction time and word duration data.

MODULE 2

This module will introduce generalized linear modeling, in particular odds, logits, and probabilities; and how to interpret coefficients and perform model performance metrics. The data used for illustration will be corpus data of clause-ordering.

MODULE 3

In Module 3, we will discuss linear mixed-effects modeling, with particular attention to the meaning and purpose of varying intercepts and varying slopes; how to select models for mixed-effects models; how to interpret models and perform model diagnostics; how to do model performance metrics; and how to interpret curvature in model visualizations. The data used for illustration are acquisition of determiners data from a corpus.

MODULE 4

Module 4 discusses generalized linear mixed-effects modeling, with particular attention to the meaning and purpose of varying intercepts and varying slopes; how to select models for mixed-effects models; how to interpret models and perform model diagnostics; how to do model performance metrics; and how to interpret curvature in model visualizations. The data used for illustration are corpus data of subject realization in Japanese and variable complementizer realization in second language learner language.

MODULE 5

In module 5 we will become familiar with tree-based approaches, specifically classification, regression, and conditional inference trees. We will discuss both advantages and potential weaknesses of these approaches. The data used for illustration are selections of data used previously in the course to compare results from tree-based and regression approaches, respectively.

Reading list

All registered applicants will receive a link to an on-line page where all readings can be downloaded. The reading assignment is to read all required readings and be familiar with recommended readings.

Required readings

Gries, Stefan Th. 2013. Statistics for linguistics with R. 2nd rev. and ext. ed. Berlin & Boston: De Gruyter Mouton, chapters 1-4 and 5.1.

Recommended readings

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.