# Linguistics 620. Experimental Linguistics: Methods — Winter 2020

Tuesday/Thursday, 10.05–11:25, Leacock 212

## Instructors

Morgan Sonderegger
morgan.sonderegger@mcgill.ca
1085 Dr. Penfield, 227

**Office hours**: 1–2 PM Thursdays in 1085, or by appointment (calendly.com/morgan-sonderegger).

## 1 Course Goals

This course is an introduction to quantitative analysis for linguists, Our focus will be on tools for exploration and statistical analysis of datasets once they are collected, though aspects of experimental design and annotation will be discussed along the way. By the end of this class, you will have learned fundamental skills for visualization and quantitative analysis of your own data, and for assessing quantitative analyses in research papers. We will begin with exploratory data analysis, basic inferential statistics, and hypothesis tests. Much of the course will be spent on fitting and evaluating different regression models, up to mixed-effects models. Because the methods used for analyzing empirical data vary across subfields and are constantly changing, a primary goal of the course is learning to learn: developing a sufficiently strong basis in quantitative analysis that you can figure out on your own the quantitative methods needed to analyze linguistic data, including methods not covered in this course.

## 2 Overview

The following is a general overview of the course. A detailed schedule with readings and homework dates is available as a Google doc at bit.ly/ling620-W2020. Please check this regularly for updates.

| Weeks | Topic | Note |
|---|---|---|
| 1–2 | Intro, descriptive statistics, exploratory data analysis | |
| 3 | Inferential statistics basics: samples/populations, hypothesis testing, power | |
| 4–5 | Linear regression | |
| 6 | Logistic regression | |
| 7–8 | Practical regression: contrast coding, non-linear effects, etc. | No class Feb. 20 |
| Study break | | |
| 9–10 | Mixed-effects models | |
| 11 | Catch-up/TBA | |
| 12–13 | Wish-list presentations | |

## 3 Prerequisites

An introductory statistics class, such as PSYC 305 or MATH 201, or permission of instructor.

## 4 McGill Policy Statements

**Course Work in French:** In accord with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

**Integrity:** McGill University values academic integrity. Therefore, all students must understand the meaning and consequences of cheating, plagiarism and other academic offences under the Code of Student Conduct and Disciplinary Procedures (see www.mcgill.ca/students/srr/honest/ for more information).

**Copyright:** Instructor generated course materials (e.g., handouts, notes, summaries, exam questions, etc.) are protected by law and may not be copied or distributed in any form or in any medium without explicit permission of the instructor. Note that infringements of copyright can be subject to follow up by the University under the Code of Student Conduct and Disciplinary Procedures.

## 5 Logistics

**Electronic logistics**: The course site is hosted on Piazza. Make sure you are signed up.

In general you should post questions to Piazza (see 'Getting help' below), and communicate with me by Piazza private message. If you do email me, Please include "LING 620" in the subject line. You can generally expect a response within 24 hours (with the exception of weekends).

Classes will be a mixture of lectures and labs, with exercises you'll do using R. Thus, you should bring your laptop if you have one. If not, computers are available in the desks in Leacock 212.

**Materials**:

Readings for this class will be drawn from various sources. Check the google doc for the most up-to-date readings. The main texts we will use are:

1. *R for Data Science* (Grolemund and Wickham, 2017): Weeks 1–3
2. *Regression Modeling for Lingusitic Data* (Sonderegger, 2020): Weeks 4–7
3. *Quantitative Methods for Linguistic Data* (Sonderegger et al., 2018): Weeks 8–10

(1) and (3) are online; (2) is a textbook-in-progress updating (3), from which I will post chapters. There will also be some articles or online resources.

I will usually post a primary reading assignment covering the topic you are responsible for learning, but also recommend similar chapters from other resources.

Many books cover most topics discussed in the course, including some listed below, and I recommend looking at a few different books initially to see which one works best for you (level of math, quantity of R code, conceptual vs. technical explanations). [1]

By/for linguists:

- Winter (2019): Medium math, wide coverage (recommended!)
- Baayen (2008): Little math, wide coverage.
- Gries (2009): Little math, basic coverage.
- Johnson (2008): Medium math, medium coverage.
- Vasishth (2014) Medium math, medium coverage.
- Levy (2012): High math, wide coverage.
- Levshina (2015): Low math, medium coverage

General:

- Chatterjee and Hadi (2012): Medium math, medium coverage, including of some less widely-covered topics (collinearity, regression diagnostics).

[1] All of these books except Chatterjee and Hadi (2012) primarily use R.

- Navarro (2015): Medum/low math, wide coverage, very applied (using R throughout).
- Dalgaard (2008): Medium math, medium coverage.
- Maindonald and Braun (2010): Medium math, wide coverage, extensive R examples throughout.
- Faraway (2015): medium math, wide coverage up to linear regression
- Gelman and Hill (2007): High math, wide (but somewhat unconventional) coverage, mix of math and simulation.

All these books are available either as hard copies or e-books through the library (or elsewhere online), and I have copies that can be made available on request.

**Software:**

All data analysis for this course will be done in R. To prepare, make sure you have R and RStudio (free desktop version) downloaded on your computer.[2] In addition, please install some packages we will be using frequently by entering the following command inside R (while connected to the internet):

```
install.packages(c("languageR","tidyverse","lme4", "arm","zipfR","ggplot2"),
dependencies = TRUE, repos = "http://cran.r-project.org")
```

'Learning R' is inherently atelic! It can take a long time to do a seemingly tiny task, and minor errors can take hours to figure out. Be patient with yourself, and give yourself plenty of time to complete assignments. Learning to debug and find the errors is a valuable process in and of itself. You will be expected to learn functions semi-independently, with the help of online resources. Readings and labs will provide examples of code and practice, but it will take independent practice to gain comfort, flexibility, and generalizability to your own projects. I will keep technical discussion of R to a minimum during class time, so it is important to iron out bugs and make sure you understand the necessary commands in the mini-assignments before class.

After Week 1 I will expect that you have used R before and are comfortable with some basic functions. Some of these will be in a Practice homework to be posted soon. If you haven't worked with R before, don't worry; there are many good tutorials online or in books. Some possibilities are:

- *R for Linguists* Tutorial by Joseph Casillas at Rutgers—I will expect that you are comfortable with all concepts through (and including) the 'Data Structures' section.
- Winter (2019): Chapter 1
- Grolemund and Wickham (2017): first couple chapters

If you would like a more thorough foundation to R, you could enroll in the first 1-2 (free) courses on R at `http://datacamp.com`. This is beyond the Week 2 knowledge prerequisite, but comfort with R will make the homework go faster and will be useful in the future.

There are often multiple ways to do the same thing in R. In this class, I will be asking you to use 'tidyverse' functionality, including `ggplot2` for all graphs, and `dplyr` for most data manipulation. It's fine if you have not used these before–there will be plenty of practice, and Winter (2019) is an excellent tidyverse-centric book for linguists specifically.

**Getting help:**

- Piazza has good functionality for discussions, and we encourage you to use Piazza for asking questions and getting help—especially from each other.[3] You are encouraged to contribute answers to other students' threads, initiate open-ended discussions, or post links of interest. For example, in previous years, students have found and posted helpful readings (that the instructors were not aware of).
- I am happy to discuss any course-related issues, including technical problems with R, in office hours or by appointment (see above).

[2] If you do not have access to a computer on which you can install R, or feel that you would benefit from access to a better computer for assignments for this course, let me know.

[3] Exception: do not post any question that reveals part of the solution to a homework problem or a mini-project!

- I am also available by Piazza message (preferred) or email for logistical/clarification questions. I will check Piazza about once per day, but in general am not able to answer technical questions about R (or content questions about statistics) electronically.
- PhD student James Tanner will be available for technical help, by appointment or online—details TBA.
- Online forum (stackexchange, stackoverflow, quota) are an extremely helpful way to find solutions to problems (particularly errors in R). It takes some practice to learn how to use these (e.g. which search terms are effective to Google), but they will be an invaluable resource.

## 6 Evaluation

**Participation (5%)**

- Active participation in course activities and discussion

**Homeworks (45%)**

- These assignments, typically weekly and due on Mondays, will require you to apply concepts covered in class, and will require some R coding. Some will require write-ups.
- Collaboration is allowed, but you must write your own code (and writeup if applicable), and list your collaborators.
- Two 'super-homeworks' , requiring more conceptual understanding and work, letter-graded: 10% each.
- Six 'regular homeworks', of which you turn in four, graded check-plus/check/check-minus: 25% together
- Each assignment is due, by private Piazza post to the Instructor, at **11:59 PM** on the due date.
- Your homework grade is based on the two 'super-homeworks' plus four regular homeworks. If you submit more than four regular homeworks I'll grade the first four.
- You have three extensions to use of 24 hours for late homework submission, and can't use more than one per homework. To use an exetnsion, you must notify me by 24 hours before the deadline.

**Presentation (10%)**:

- Since we are only covering the basics in class, many topics of potential interest will be left out—especially those relevant for analyzing data common in some subfields but not others
- Groups of 2 (possibly 3) will give a mini-course on a 'wish-list' topic for one-half of a lecture.
- Details TBA

**Final project (40%)**:

- In the final project, you will develop a more in-depth analysis of a real, complex, and ideally unanalyzed dataset.
- You should use skills acquired over the course of the semester, and can optionally bring in extra methods.
- We can help any student who needs a dataset find one, but you are encouraged to analyze a dataset from your own research.
- Due date: **April 16**
- More details TBA
- You may collaborate in groups of 2–3 on the final project. If you do collaborate, you must write up your analysis on your own, list your collaborators, and submit a brief statement of who did what.

# References

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge University Press, Cambridge.

Chatterjee, S. and Hadi, A. S. (2012). *Regression analysis by example.* John Wiley & Sons, 5th edition.

Dalgaard, P. (2008). *Introductory statistics with R.* Springer.

Faraway, J. J. (2015). *Linear models with R.* Chapman and Hall/CRC.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge University Press.

Gries, S. T. (2009). *Statistics for linguistics with R: a practical introduction.* Walter de Gruyter.

Grolemund, G. and Wickham, H. (2017). R for data science. Available at http://r4ds.had.co.nz/.

Johnson, K. (2008). *Quantitative methods in linguistics.* Wiley-Blackwell, Malden, MA.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis.* John Benjamins.

Levy, R. (2012). Probabilistic models in the study of language. Ms in progress. http://www.mit.edu/~rplevy/pmsl_textbook/text.html.

Maindonald, J. and Braun, W. (2010). *Data analysis and graphics using R: an example-based approach.* Cambridge University Press, 3rd edition.

Navarro, D. (2015). Learning Statistics with R. Version 0.5. [Lecture notes] School of Psychology, University of Adelaide, Adelaide, Australia.

Sonderegger, M., Wagner, M., and Torreira, F. (2018). *Quantitative methods for linguistic data.* ebook. http://people.linguistics.mcgill.ca/ morgan/book/.

Vasishth, S. (2014). An introduction to statistical data analysis. Summer 2014 version. Available at https://github.com/vasishth/Statistics-lecture-notes-Potsdam/tree/master/IntroductoryStatistics.

Winter, B. (2019). *Statistics for Linguists: An Introduction Using R.* Routledge.