Linguistics 620. Experimental Linguistics: Methods **Fall 2014**

Time and Location

	Lecture:	F 11:35 AM–2:25 PM, Leacock 212
	R practicum:	W 12:05–12:55, Arts 260

Instructor

Morgan Sonderegger
morgan.sonderegger@mcgill.ca
1085 Dr. Penfield, 227
Office hours: By appt

1 Course Goals

This course is an introduction to experimental research methods 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, culminating in mixed-effects models. By the end of the course you should have a sufficiently strong basis in quantitative analysis to figure out on your own many other types of analyses that we will not cover.

The course will take a hands-on approach to data and statistical modelling, using the free programming language R. You will learn to use R for visualizing and analyzing data, including basic programming. Because the best way to learn statistical tools is by using them a lot, this class will be comparatively heavy on homework assignments. There will be weekly programming practica as needed to address questions about R, but you should expect to primarily spend time independently (and with classmates) figuring out how to do things in R based on the readings and online resources.

2 Prerequisites

PSYC 305 or equivalent

3 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.

4 Logistics

Electronic logistics: The course site is hosted on Piazza (https://piazza.com/mcgill.ca/fall2014/ling620). Make sure you are signed up.

Please include "LING 620" in the subject line of emails to the instructor. (This will put your emails in a folder which helps me respond to them more quickly.)

Materials: Readings will be drawn from a number of sources.

Depending on what type of presentation of quantitative methods works best for you (level of math, quantity of R code, conceptual vs. technical explanations), you may find a different source more useful for you than a given week's reading for understanding that week's topics. Many books cover most topics discussed in the course including:[1]

- For linguists:
  - Gries (2009): Little math, basic coverage.
  - Baayen (2008): Little math, wide coverage.
  - Johnson (2008): Medium math, medium coverage.
  - Vasishth and Broe (2011) Medium math, medium coverage.
  - Levy (2012): High math, wide coverage.
- General:
  - Chatterjee and Hadi (2012): Medium math, medium coverage, including of some less widely-covered topics (collinearity, regression diagnostics).
  - Dalgaard (2008): Medium math, medium coverage.
  - Maindonald and Braun (2010): Medium math, wide coverage, extensive R examples throughout.
  - Gelman and Hill (2007): High math, wide (but somewhat unconventional) coverage, mix of math and simulation.

All these books are either on reserve in HSSL, or available as an e-book through the library.

Because an important goal of this course is for you to be able to learn about new methods for your future research, I encourage you to try out different sources, and find one (or more) which suits you. To facilitate this process, I will often assign reading from more than one source covering similar material.

Software: You should install R on your computer as soon as possible.[2] Enter the following command inside R to install packages some packages that you will need (while connected to the internet):

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

Note that the default R interface is very minimal: you work largely from the command line. If you prefer a richer user interface (e.g. a window showing what data is in your workspace, the option of executing some commands via menus instead of the command line), one popular option is `RStudio`.

We will briefly discuss R basics on the first day of class. However, if you have not used R before, I highly recommend doing a basic tutorial during the 1.5 weeks of term (such as Ch. 1 of Baayen, 2008, or one of many tutorials available online).

R practica: There will be a weekly informal R practicum Wednesdays from 12–1.on Wednesday Attendance is optional, and the format will be the instructor showing how to do various things in R, including in response to questions. Our priority will be answering R questions which have come up doing the week's assignment (due a few days after the practicum), followed by covering a list of set topics (hopefully posted 1–2 few days before).

Lunch: Feel free to bring a lunch to class. There will be a short break in the middle of class. (Note that consuming food or beverages is not allowed in Leacock 212 itself.)

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

[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, contact the instructor.

5 Evaluation

Participation in discussion (5%):

- It is important that you actively participate in class. Ask questions whenever you like, and don't be afraid to ask something that seems trivial.

5 short homework assignments (3% each, 15% total)

- These assignments will require you to apply concepts covered in class, and will require some R coding.
- Collaboration is allowed, but you must write your own code and writeup, and list your collaborators.
- It is important to do all assignments, since later parts of the class will build on the skills you learned in earlier homework assignments.
- These assignments will be graded on a check-minus/check/check-plus basis. If you turn in more than 5 assignments, only the 5 with the highest grades will count.

2 mini-projects (40%)

- These longer assignments will be more involved: you will develop an analysis of a real dataset, and submit R code which performs your analysis, as well as a writeup describing and justifying it.
- You are encouraged to collaborate on the mini-projects.
- If you do collaborate, group members may turn in the same code, but should submit independent writeups.

For both short homeworks and mini-projects, you may work in groups of 2–3, and the assignment is due (by email to the instructor) at **9 AM** on the due date.

Final project (40%):

- In the final project, you will develop a more in-depth analysis of a real, complex, and ideally unanalyzed dataset, using skills acquired over the course of the semester.
- I will provide a dataset to any students who need one, but I encourage you to analyze a dataset from your own research for the final project, if possible. If you have a dataset you would like to use, please discuss with me around mid-term, so that I can assess whether it will work for the final project, and help you structure it in the right away if it does.
- As with the mini-projects, 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.

6 Course schedule (subject to change)

Readings will be announced in class.

Date	Topics	Assignments	
		Out	Due
Sep 5	R basics Probability and distributions I	HW 1	—
Sep 12	Probability and distributions II Summarizing and visualizing data I	HW 2	HW 1
Sep 19	Summarizing and visualizing data II	HW 3	HW 2
Sep 26	Sample and population statistics Hypothesis testing	HW 4	HW 3
Oct 3	Linear regression I	HW 5	HW 4
Oct 10	Linear regression II	MP 1	HW 5
Oct 17	Analysis of variance (ANOVA)	—	—
Oct 24	Variable selection Collinearity Categorical data analysis 1	HW 6	MP 1
Oct 31	Logistic regression 1	—	HW 6
Nov 7	Logistic regression 2 Contrast coding	HW 7	—
Nov 14	Mixed models 1	MP 2	HW 7
Nov 21	Mixed models 2	—	—
Nov 28	Mixed models 3	—	MP 2

Final project due: Dec 15

7 Course outline (subject to change)

A more detailed outline of what I plan to cover (not necessarily in this exact order) is below. Not all topics listed can be covered in depth, and some topics may be dropped depending on students' backgrounds or if we're short on time, particularly those followed by an asterisk. Our priority will be building up to mixed models, which will be the main tool used in the final project.

1. **Probability and distributions**
   (a) Basic probability theory
   (b) Discrete and continuous random variables, probability distributions

2. **Summarizing and visualizing data**
   (a) `ggplot`
   (b) Continuous data: histograms, density plots, smooths, etc.
   (c) Categorical data: bar charts, mosaic plots, etc.
   (d) Dimension reduction: Principal components analysis, multidimensional scaling *

3. **Hypothesis testing**
   (a) Background: sample and population statistics, test statistics, significance, confidence intervals
   (b) Parametric tests: t tests, Wald tests, etc.
   (c) Nonparametric tests *

4. **Linear models**
   (a) Covariance and correlation
   (b) Single and multiple linear regression
      - Model assumptions
      - Practical issues in fitting regression models in R
      - Model interpretation, visualization; model selection
   (c) Analyis of variance
      - Decomposition of variance, F tests
      - Relationship of ANOVA to linear models
      - ANOVA terminology & common analyses: one/two-way, repeated measures, etc.

5. **Categorical data analysis**
   (a) Contingency tables, tests of association: χ^2, Fisher
   (b) Single and multiple logistic regression
      - Model assumptions, practical issues, etc. (same as 4b)
      - Comparison with other methods (e.g. ANOVA on transformed proportions) *
   (c) Decision trees *

6. **Further regression issues**
   (a) Predictor transformations: nonlinear effects, contrast coding
   (b) Type I and Type II error issues: multiple comparisons, post-hoc tests, power
   (c) Model criticism (collinearity, outliers, residuals) and validation (bootstrapping, overfitting) *

7. **Mixed-effects regression**
   (a) Motivation: grouped data, fixed and random effects
   (b) Mixed-effects linear regression
      - Single, multiple grouping levels ('crossed random effects')
      - Varying intercepts, varying slopes
      - Model assumptions, practical issues, etc. (same as 4b)
   (c) Mixed-effects logistic regression

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.

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.

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

Levy, R. (2012). Probabilistic models in the study of language. Ms in progress. http://idiom.ucsd.edu/~rlevy/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.

Vasishth, S. and Broe, M. B. (2011). *The foundations of statistics: A simulation-based approach*. Springer, Berlin.