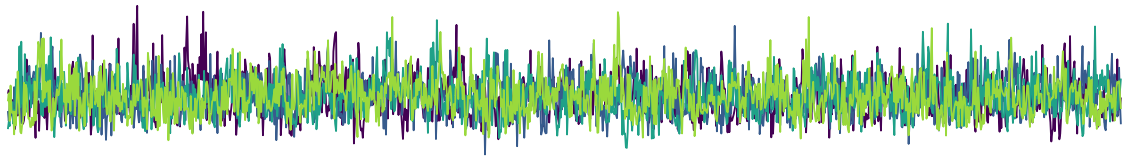


Introduction to Statistics for Ecology and Evolutionary Biology (EEEB UN3005/GR5005)

Columbia University - Spring 2019



Staff

Instructor

Evan Eskew (cae2144@columbia.edu); office hours after class or by appointment

Teaching Assistants

Krishna Anujan (ka2602@columbia.edu); office hours by appointment

Alexandra Huddell (amh2284@columbia.edu); office hours by appointment

Course Description

The goal of this course is to introduce data manipulation, visualization, and statistical modeling with an emphasis on applications in ecology and evolution. In the first few weeks of class, we'll become familiar with using the R programming language for data handling and visualization. For the remainder of the semester, we'll explore the statistical workhorse of ecology and evolution: linear regression. The primary aim is to develop a firm grasp of how this method works “under the hood” so we have a strong conceptual basis for understanding other, more complex statistical methods as extensions of the basic linear regression strategy. For example, when we encounter new data types, instead of adopting entirely new statistical procedures or performing transformations to make our data better fit our statistical approach,

we will simply use the same regression modeling framework with new likelihood distributions that are consistent with our outcome data (i.e., generalized linear modeling). Late in the course, we will begin to implement models with multiple layers that fit different levels of the data, thereby leading to improved parameter estimates and predictions (i.e., generalized linear mixed modeling). Along the way, there will be a primer in Bayesian statistical philosophy, and we'll touch on related considerations like model selection.

At the conclusion of this course, you should be much more comfortable working with data in R. You will be able to read the ecology and evolution literature with a critical eye towards statistical procedures and interpretation. You should be well-positioned to use your knowledge regarding the fundamentals of regression modeling to learn more specialized, complicated types of statistical models.

Course Texts

The primary text for this course is *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* by Richard McElreath. This is an incredibly accessible statistics book, and it will serve us well in establishing the foundations of regression modeling. It is worth checking online retailers for the book since used copies can sometimes be found at a discount. The book contains extensive R code, so it will also assist us in strengthening our R skills throughout the semester. **I highly encourage you to work step-by-step through the book material in R as you complete your readings.** In coding, as in most things, practice makes perfect.

In addition to chapters from *Statistical Rethinking*, there will be some additional assigned readings from the scientific literature and blogs. I will post these materials on the class CourseWorks site, or they are freely available online and linked in the syllabus.

Course Organization

The course topic for each week will be introduced through that week's readings. Therefore, it is imperative that you do the required reading *before* the relevant lecture. **If you do not complete the readings, you will be lost!** Lectures will closely follow readings and should help solidify your understanding of the material. In lab sections, you will address problems in statistical coding in an environment where help is readily available from myself and the course TAs. Finally, homework assignments are your opportunity to demonstrate mastery of the material through independent work. Homework assignments will be released on Wednesday of each week and will be due the following Tuesday at 5 pm.

Those enrolled in the graduate course will also complete an additional research project assignment throughout the semester. Project details will be provided separately.

Computers and Software

This course will be taught entirely in the R programming language. R is widely used by researchers in ecology and evolution, has a diverse ecosystem of packages that provide specialized functionality, and is free. R can be downloaded for all major operating systems [here](#) (choose the “precompiled binary distribution” option). In addition, we’ll be using the RStudio software to work more easily with R. Essentially, RStudio spruces up the basic R software with a number of convenient features, and many users never interact with R except through RStudio. The version of RStudio intended for personal use (RStudio Desktop) is also free. Install the software [here](#) (scroll down to “Installers for Supported Platforms” and select your operating system).

While not required, having a personal computer for this class will be a significant benefit given that R coding will be a major focus of teaching and all homework assignments will require access to a computer. If you have one, I encourage you to bring your personal computers to lecture and follow along with the class material on your own machine. If you exercise this option, do not be a distraction to other students by using your computer for anything not course-related during class time. **Please attempt to install R and Rstudio on your personal computers prior to the start of the course, but if you have trouble, we will be able to help you with software installation during the first week of class.** Computers will be available in the lab space during lab sections for all students, and these machines can also be used to complete homework assignments as needed.

Course GitHub

In addition to using the CourseWorks site, I will maintain a [GitHub repository for this course](#). This will allow others to freely access materials and also provides an opportunity for you to become familiar with git and GitHub, if you’re so inclined.

Grading

Grading Breakdown

Assignment	UN3005	GR5005
Homework	70%	55%
Lab Attendance and Participation	10%	10%
Final Exam	20%	20%
Research Project	N/A	15%

Grading Scale

Percentage	Grade
97.50 - 100	A+
92.50 - 97.49	A
90.00 - 92.49	A-
87.50 - 89.99	B+
82.50 - 87.49	B
80.00 - 82.49	B-
77.50 - 79.99	C+
72.50 - 77.49	C
70.00 - 72.49	C-
60.00 - 69.99	D
< 60.00	F

Grading Policies

In this course, **late assignments will receive no credit.** However, before final grades are calculated, **I will drop your lowest homework grade.** This is intended to provide all students with protection against any circumstances that might affect the timeliness or quality of classwork. Your lab grade will be determined by your weekly attendance and engagement with the material. Similar to my homework grading policy, **I will grant all students one lab absence for the semester without penalty.**

If a serious medical, family, or other emergency may prevent you from completing an assignment on time, please talk with me as soon as possible about the circumstances, and we will discuss plans for helping you to complete your work. I will grant assignment extensions only in cases of extreme emergency, which are determined at my discretion.

For homework assignments, I encourage you to discuss the problem sets and potential solutions with your classmates. **However, if it becomes obvious that any students are directly**

copying each other's work, all students involved will be penalized. In other words, feel free to work with your classmates to figure out *how* you should address a given problem, but the coding and any written interpretation should be your own work. This course's final exam will be a take-home exam that must be completed independently.

Course Schedule

Week 01 (Jan 28) - Introduction to R and RStudio

Required reading:

- *Statistical Rethinking*: Preface, Chapter 1

Week 02 (Feb 04) - Data Cleaning and More Advanced R

Required reading:

- Wickham, H. 2014. Tidy data.
- “Introduction to dplyr.” [Article link](#)

Week 03 (Feb 11) - Data Visualization

Required reading:

- Wickham, H. 2010. A layered grammar of graphics.

Optional reading:

- Martin, L. J. 2015. Mathematizing nature’s messiness: graphical representations of variation in ecology, 1930-present.

Week 04 (Feb 18) - Bayesian Basics

Required reading:

- McElreath, R. 2017. “There is always prior information.” [Article link](#)
- *Statistical Rethinking*: Chapter 2

Optional reading:

- McElreath, R. 2011. “The tyranny of Fisher.”

Week 05 (Feb 25) - Statistical Distributions and Summary Statistics

Required reading:

- *Statistical Rethinking*: Chapter 3

Week 06 (Mar 04) - Gaussian Regression Models

Required reading:

- McElreath, R. 2017. “First World modeling problems.” [Article link](#)
- *Statistical Rethinking*: Chapter 4

EEEB GR5005 Students: Research project prospectus due Mar 07

Week 07 (Mar 11) - Multiple Regression

Required reading:

- *Statistical Rethinking*: Chapter 5

Week 08 (Mar 18) - No Class (Spring Break)

Week 09 (Mar 25) - Model Selection

Required reading:

- *Statistical Rethinking*: Chapter 6
- Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods.
- Bolker, B. M. 2018. “Multimodel approaches are not the best way to understand multifactorial systems.” [Article link](#)

Optional reading:

- Whittingham, M. J., et al. 2006. Why do we still use stepwise modelling in ecology and behaviour?
- McGill, B. 2015. “Why AIC appeals to ecologist’s lowest instincts.” *Dynamic Ecology*. [Article link](#) (And be sure to see the many interesting comments.)

Week 10 (Apr 01) - Interactions

Required reading:

- *Statistical Rethinking*: Chapter 7

Week 11 (Apr 08) - Link Functions and Poisson Regression

Required reading:

- *Statistical Rethinking*: Chapters 9-10

Optional reading:

- *Statistical Rethinking*: Chapter 8

EEEB GR5005 Students: Latest research project rough drafts accepted Apr 11

Week 12 (Apr 15) - Binomial Regression

Week 13 (Apr 22) - Introduction to Multilevel Modeling

Required reading:

- McElreath, R. 2017. “Multilevel regression as default.” [Article link](#)
- *Statistical Rethinking*: Chapter 12

Week 14 (Apr 29) - More Multilevel Modeling

Required reading:

- Bolker, B. M., et al. 2009. Generalized linear mixed models: a practical guide for ecology and evolution.

EEEB GR5005 Students: Research projects due May 02

Week 15 (May 06) - The Wide World of Statistical Models

Required reading:

- *Statistical Rethinking*: Chapter 11

Optional reading:

- Freedman, D. A. 1991. Statistical models and shoe leather.

Supplementary Resources

There are an overwhelming number of resources on R coding and statistics, both in print and online. Any attempt to summarize this material is woefully incomplete, but I've tried to point out a few items that might be of particular interest. The R programming and data visualization resources may be of use during this class if you want additional treatment of the material we're covering (as a plus, a couple of them are free online). In contrast, the statistical modeling books would be good to pick up at the conclusion of this course if you're looking to dive into more advanced statistical models.

R Programming

Wickham, H. and G. Grolemund. 2017. *R for Data Science*. [Book link](#)

Data Visualization

Tufte, E. R. 2001. *The Visual Display of Quantitative Information* (Second Edition). Graphics Press LLC.

Wilke, C. O. 2018. *Fundamentals of Data Visualization*. [Book link](#)

Statistical Modeling

Gelman, A., and J. Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Royle, J. A., and R. M. Dorazio. 2008. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press.

Wood, S. 2017. *Generalized Additive Models: An Introduction with R* (Second Edition). CRC Press.

Course Bibliography

Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *The Journal of Wildlife Management* **66**: 912-918.

Bolker, B. M. 2018. "Multimodel approaches are not the best way to understand multifactorial systems." [Article link](#)

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution* **24**: 127-135.

Freedman, D. A. 1991. Statistical models and shoe leather. *Sociological Methodology* **21**: 291-313.

"Introduction to dplyr." [Article link](#)

Martin, L. J. 2015. Mathematizing nature's messiness: graphical representations of variation in ecology, 1930-present. *Environmental Humanities* **7**: 59-88.

McElreath, R. 2011. "The tyranny of Fisher."

McElreath, R. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.

McElreath, R. 2017. "First World modeling problems." [Article link](#)

McElreath, R. 2017. "Multilevel regression as default." [Article link](#)

McElreath, R. 2017. "There is always prior information." [Article link](#)

McGill, B. 2015. "Why AIC appeals to ecologist's lowest instincts." *Dynamic Ecology*. [Article link](#)

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology* **75**: 1182-1189.

Wickham, H. 2010. A layered grammar of graphics. *Journal of Computational and Graphical Statistics* **19**: 3-28.

Wickham, H. 2014. Tidy data. *Journal of Statistical Software* **59**: 1-23.