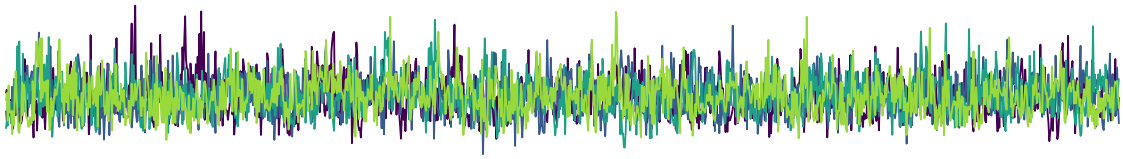
Introduction to Statistics for Ecology and Evolutionary Biology (EEEB UN3005/GR5005)

Columbia University - Fall 2020
UPDATED: 09 September 2020



Staff

Instructor

Steffen Foerster (sf2041@columbia.edu); office hours after class or by appointment

Teaching Assistants

Rachel O. Cohen (roc2109@columbia.edu); office hours Monday, 2-3 pm, Zoom

Dean M. Bobo (deanmbobo@gmail.com); office hours Thursday, 9-11 am, Zoom

Meeting Times

Lecture: Monday, 6:10-7:25 pm, Zoom

Lab: Monday or Wednesday, 7:40-8:55 pm, Zoom

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 many 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, 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 and performing statistical analyses 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 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 recommend you work step-by-step through the book material in R as you complete your readings each week.**

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 critical that you do the required reading *before* the relevant lecture.** Lectures will closely follow readings and are meant to solidify your understanding of the material. In lab sections, you will address problems in statistics and 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, with portions of the work due throughout the semester. Project details will be provided separately.

Statement on Inclusivity

I strive to create a classroom environment that is welcoming of students with various identities and abilities and is conducive to everyone's intellectual growth. Specific pedagogical elements of the course are meant to support this goal. For example, there are no timed exams in this course, and our grading practices are intended to reduce bias. If there are aspects of the class or your classroom interactions that hinder your learning, please do not hesitate to begin a dialogue with myself or the course TAs, and we will work to resolve these issues.

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. Many users never interact with R except through RStudio. The version of RStudio intended for personal use (RStudio Desktop) is also free. Install that software here (scroll down to "All Installers" list and select your operating system).

While not required, having a personal computer for this class will be a 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.**

Students who do not have access to a personal computer or who are unable to install the required software and packages can take advantage of RStudio Cloud, which provides access to R and RStudio on any computer with internet access. Unfortunately, the free version will not have enough time or computing power to meet our needs, though the premium version

is available at a steep 75% discount for academic use, which comes to $25/month for the duration of use. Another option is Codio, which is freely available and will be added to Canvas, though it is not as easy to use. Please alert your TA if you need to make use of either option.

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	65%	45%
Lab Attendance and Participation	15%	15%
Final Exam	20%	20%
Research Project	N/A	20%

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 their work. 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 responses, all students involved will be penalized.**

In other words, feel free to collaborate 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 (Sep 14) - Introduction to R and RStudio

Required reading:

- *Statistical Rethinking*: Preface, Chapter 1

Week 02 (Feb 21) - Data Cleaning and More Advanced R

Required reading:

- Wickham, H. 2014. Tidy data.
- "Introduction to dplyr." Article link

Week 03 (Sep 28) - 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 (Oct 05) - 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 (Oct 12) - Statistical Distributions and Summary Statistics

Required reading:

- *Statistical Rethinking*: Chapter 3

EEEB GR5005 Students: Research project prospectus due Oct 15

Week 06 (Oct 19) - Gaussian Regression Models

Required reading:

- McElreath, R. 2017. "First World modeling problems." Article link
- *Statistical Rethinking*: Chapter 4

Week 07 (Oct 26) - Multiple Regression

Required reading:

- *Statistical Rethinking*: Chapter 5

Week 08 (Nov 2) - Model Selection

Required reading:

- *Statistical Rethinking*: Chapter 6
- Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods.

Optional reading:

- Bolker, B. M. 2018. "Multimodel approaches are not the best way to understand multifactorial systems." Article link
- McGill, B. 2015. "Why AIC appeals to ecologist's lowest instincts." *Dynamic Ecology*. Article link (And be sure to see the many interesting comments.)

EEEB GR5005 Students: Research project rough drafts due Nov 5

Week 09 (Nov 9) - Interactions

Required reading:

- *Statistical Rethinking*: Chapter 7

Week 10 (Nov 16) - Link Functions and Poisson Regression

Required reading:

- *Statistical Rethinking*: Chapters 9-10

Optional reading:

- *Statistical Rethinking*: Chapter 8

Week 11 (Nov 23) - Binomial Regression

Week 12 (Nov 30) - Introduction to Multilevel Modeling

Required reading:

- McElreath, R. 2017. "Multilevel regression as default." Article link
- *Statistical Rethinking*: Chapter 12

EEEB GR5005 Students: Research projects due Dec 3

Week 13 (Dec 7) - More Multilevel Modeling

Required reading:

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

Week 14 (Dec 14) - 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.