Statistical Espresso for Biologists:
Extending the linear model
BIOL/CAS 4XXX/6XXX—Spring 2018

Short summary of course

Starting with a brief review of linear regression, an overview of multiple regression (including ANCOVA), generalized linear models, and an introduction to mixed effects, time series, and spatial models. An overview of many tools with an emphasis on application and an extended coursework exercise.

Justification

This class has arisen from direct requests within Biology and the Ecology Center for a series of short, 'espresso' courses that cover advanced statistical topics from an applied angle. This course, like the other 'Statistical Espresso for Biologists' course I have proposed, differs from other, more traditional courses, in that:

- It address identified skills-gaps within the Climate Adaptation Science graduate program within the Ecology Center. Each course is tailored to give students within that program the skills they have identified they need (multivariate statistics, regression analysis of complex data, and applied spatial statistics) to complete their internships.
- It has a strong emphasis on application. In the one-credit version of the class, students are assessed on the basis of their application of concepts in R. In the two-credit class, students write on a dataset related to their thesis. This class represents a way for students to get support working on a statistical problem that is holding back their thesis.
- They address a broad variety of topics at an introductory level. Other Biology classes do not introduce students to cutting-edge statistical techniques, focusing instead on traditional methods such as ANOVA. Conversely, classes in math-stat often have a strong emphasis on theoretical foundations, making them inaccessible to biologists. This class can, if desired, act as a transition into those more advanced classes.

Pre-requisites

MATH-1210, BIOL–3760, STAT-3000, and instructor permission. Basic statistical classes (i.e., knowledge of how to perform a t-test and simple linear regression) are

required for this class. No prior experience with R whatsoever is required, as the class will focus entirely on how to interpret R output and students will have no need to perform any programming in the class. Students who are concerned about their level of R experience are encouraged to study the USU baseR website (http://learnr.usu.edu/), which covers more detail than is required for this course. The starting statistical requirements for this course are very low, but, nevertheless, students who are concerned are encouraged to contact Dr Pearse directly (will.pearse@usu.edu).

Course description and learning objectives

Fitting a line through a scatter-plot—a 'linear regression'—is the simplest example of the family of statistical models known as generalized linear models. These form the basis of most statistical tests, and are probably the most important part of any biologist's toolkit. By the end of this course, you will:

- Improve your fundamental statistical knowledge, and understand the most fundamental 'family' of statistical models: the generalized linear model
- Be able to test for the influence of multiple explanatory variables on a single response variable
- Be able to model data that are not Normally distributed, such as yes/no surveys, abundances of species (count data), proportions (e.g., 12 out 36 individuals survived), and percentages.
- Be able to account for non-independence of data (e.g., surveys conducted at the same site) using mixed effects models
- Be able to identify and control for temporal and spatial auto-correlation.

The class emphasizes both theoretical concepts and practical applications equally: students are encouraged to 'follow-along' with the examples given in the lecture in the R statistical environment. This will give you the practical skills you need to make use of the statistical approaches covered in the class in your own work.

As a graduate student, the majority of your grade will be determined by a written project, using the skills you develop in the class to investigate a dataset of your choosing. Through one-on-one meetings with the instructor during and after the class. This is an excellent opportunity to get help with your thesis...

Course materials

No text-books are required for this course, but a laptop computer (Windows, Mac, or Linux) with RStudio version $\geq$ 1.0.143 and R version $\geq$ 3.4.0 installed is required.

Advice and help getting these tools installed will be provided ahead of the start of class.

Attendance & participation

Attendance of all lectures is compulsory; we are covering a lot of ground in a compressed period of time and students who miss one session may not be able to catch back up. Students who miss more than one session may be dropped from the course, unless they can give a good reason (with evidence if necessary) for missing that session.

Assessment

In the one credit version of the class, you will be assessed through a combination of coursework and a short exam delivered a few weeks after the end of the class. Coursework will be started in the classroom, and will consist of completing an analysis of one or more datasets on which the instructor will give guidance, and then analyzing and interpreting additional dataset(s) in a similar vein. The exam will be short, and test the retention of key statistical concepts—it will not require detailed calculations or rote memorization of arcane statistical fact. Example questions (and answers) will be given to the students ahead of the exam.

In the two credit version of the class, you will be required to complete a short paper describing the use of the statistical techniques developed within this class in a novel dataset. Graduate students are *strongly* encouraged to use this as an opportunity to develop one of their thesis chapters.

The marking breakdown for the one- and two-credit versions of the class is as follows:

Section	1 credit	2 credit
Coursework	55	27.5
Exam	35	17.5
Paper	0	45
Participation	10	10

The “Honor System” of Utah State applies to the coursework and paper. I strongly advise you *not* to copy-paste answers to exercises from the Internet or from the work of other students. Not only are solutions from outside sources (and other students) often wrong, they are also quite simple for me to detect and those who cheat will be punished in accordance with Utah State regulations.

Schedule

The course is split into three main sections. In the first (weeks 1–2), you will learn the main concepts of frequentist statistics and basic commands in the R statistical language, in the second (weeks 3–4) we will place these basics within the framework of the Generalized Linear Model and outline the basics of mixed-effects modeling, and in the third (weeks 5-6), we will cover controlling for autocorrelation through time and space. In week 7, we will review everything that has been covered in a series of review exercises and, if the class wishes, students can request additional topics to be covered (but not assessed in the exam) within this period. No coursework will be due until the end of the class, but students are *strongly* encouraged to complete exercises as the class proceeds.

Week	Title	1st session	2nd session
1	Fundamentals	Frequentist definitions	t-tests
2	Regression	Regression	ANOVA & multi regression
3	GLMs	Link functions	Logistic & binomial regression
4	Hierarchy	Variance decomposition	Mixed effects
5	Time	Detecting temporal auto-correlation	Controlling for temporal auto-correlation
6	Space	Detecting spatial auto-correlation	Controlling for spatial auto-correlation
7	Synthesis	Over-dispersion	Synthesis

Each week follows the same format: two sessions, each comprised of one 40-minute lecture and a 40-minute session to work on coursework and practical examples. Each lecture will give equal weight to core theoretical concepts and practical demonstrations of those within R: this is not a theoretical class, it is a practical class where you will pick up skills and experience. It is likely that students will want to have whatever they use for note-taking to hand and their laptops to try things with the instructor in real-time in R. The coursework practicals will be an opportunity for

students to continue working through the examples given in the preceding lectures, and also to begin work on their coursework exercises and receive feedback and help with those exercises.

Students have the option to take a two-credit version of this class. If so, after week 7 they will have a series of meetings with the instructor to work on a long-term project (ideally related to their thesis) using the skills they have developed during the class.

Miscellanea

- ADA compliance: Students with physical, sensory, emotional or medical impairments may be eligible for reasonable accommodations in accordance with the Americans with Disabilities Act and Section 504 of the Rehabilitation Act of 1973. All accommodations are coordinated through the Disability Resource Center in Room 101 of the University Inn, 797-2444 voice, 797-0740 TTY, or toll free at 1-800-259-2966. Please contact the DRC as early in the semester as possible. Alternate format materials (Braille, large print or digital) are available with advance notice.

- Sexual harassment is defined by the Affirmative Action/Equal Employment Opportunity Commission as any "unwelcome sexual advances, requests for sexual favors, and other verbal or physical conduct of a sexual nature." If you feel you are a victim of sexual harassment, you may talk to or file a complaint with the Affirmative Action/Equal Employment Opportunity Office located in Old Main, Room 161, or call the AA/EEO Office at 797-1266.

- Students whose religious activities conflict with the class schedule should contact me at the beginning of the semester to make alternative arrangements.

- Course scheduling and structure may change at short notice with no warning.