

ZOL/PLB 851
QUANTITATIVE METHODS IN ECOLOGY AND EVOLUTION
Fall 2008

Instructor: Dr. Ian Dworkin

Office: 56 Giltner Hall
e-mail: idworkin@msu.edu

Office hours: Tuesdays 12:30-1:30
Other times by appointment

Class Info: Tuesday, Thursday 10:20 – 11:40am
145 Natural Sciences Building

Optional class sessions:

In addition to the regularly schedule classes, I will have optional R computer sessions to help people get started. This will take place on the following **Saturday** afternoons (Sept. 6th, Oct. 4th, Oct 25th. And more if need be) between 1-4 pm in room 25 Nat. Sci. If possible bring your laptops to these sessions.

The Course

Course Description

Interpretation and analysis of data in ecology and evolutionary biology. Statistical computer software. (3 credits)

Recommended Background

STT465 (Statistics for Biologists II) or STT814 or an equivalent course.

I assume that you have some familiarity with regression, ANOVA, ANCOVA & have a basic grasp of the utility of probability distributions and are at least vaguely familiar with some of them (normal, binomial, t-distributions, χ^2 , F distributions). While this course will not in any way deal with formal mathematical proofs, I do assume a certain level of comfort with algebraic concepts and a dim recollection of derivatives (we won't be doing any calculus, but we may look at a few derivates with respect to maximum likelihood).

In particular the material that you need to be somewhat familiar with (even if it is rusty):

- A) Basics of probability (in particular some recollection of conditional probability)
- B) Random variables, what are they ? how do we use them in statistics?
- C) Basic understanding of some probability distributions (normal/gaussian, binomial , poisson, t, chi-square, F). Also some understanding of PDFs and CDFs in relation to the above.
- D) Familiarity with ANOVA's, how to set up a basic model, and some recollection of sum of squares, mean squares and F ratios
- E) Regression analysis and interpretation and maybe some dim idea of multiple regression
- F) Analysis of covariance
- G) Maybe a dim recollection of how ANOVA's, ANCOVA and regressions are really all part of the family of general linear models (GLM).

H) While it is no way required if you remember even basic matrix algebra this would not hurt (although we won't be using it a lot).

If you do not have the background as suggested above, I just want to make sure that you are aware well ahead of time so that you are not shocked in the course. I am happy to suggest some remedial reading, but be prepared for a challenging semester. While about 70% of will be the same as the way Andrew McAdam taught it last year (2007), 30% of it will be different including more on likelihood, simulation and Bayesian methods.

Course Justification

Professional biologists interested in ecological and evolutionary questions inevitably deal with variable data. In fact variability is a fundamental concept in both ecology and evolution. The ability to manage, analyze and interpret data collected either through planned experiments or through opportunistic or monitoring programs represents a fundamental skill for all ecologists and evolutionary biologists. This course will further develop your skills and understanding of quantitative methods for the analysis, interpretation and presentation of ecological and evolutionary data. We will spend a fair bit of time considering alternative approaches (Parametric, Likelihood, Bayesian and Resampling) for parameter (point and interval) estimation for statistical models, assessing uncertainty in these estimates and using these to make biological and statistical inferences.

Goals of this course

1. Discuss the philosophical and historical context within which we collect, analyze and interpret data.
2. Overview contemporary techniques for the analysis of ecological and evolutionary data.
3. Introduce R, a powerful and useful programming environment for data analysis and presentation.
4. Provide hands-on experience analyzing data relevant to each student's particular field of study.

What we will not cover in this class: Given the limited time we have together I have decided to exclusively focus on uni-variate statistical models. That is models where there is a single “dependent” variable whose variation we are trying to explain, with one or more “explanatory” variables. While many multivariate techniques (such as MANOVA and other multivariate extensions of GLM, principal components analysis, factor analysis, path analysis and multi-dimensional scaling) are all extremely important methods, we simply will not have the time to cover them in this course. I will happily recommend books that are useful for multi-variate approaches. I apologize for this (and I wish I could discuss multivariate methods, most of my own work utilizes them). The other extremely important topic that we will not formally cover (in lecture) is experimental design, although we will discuss experimental design for some particular examples.

R as a Statistical Platform for the Course

In addition to providing general information on quantitative approaches, I will also provide specific instruction on the use of one particular platform for data management, analysis and graphical presentation. There are many commercial statistical software packages available (e.g. SAS, SPSS, STATISTICA, STATA, S-PLUS) and each has its own pros and cons. The statistical programming environment that we will be using in this course is called R. This statistical computing and graphics package is widely used in statistical methodological research and its use in ecology and evolutionary biology is increasing. One of the advantages of using R is that it is freely available online (www.r-project.org/). This open source platform runs on a variety of operating systems (e.g. Windows, Mac OS X, Linux), which has encouraged the development of many sophisticated statistical packages and an active and generous R-help message board. You will need to download the most recent version of R onto a personal computer, or use one of the microcomputing facilities on campus to complete your assignments. The most recent version is currently version 2.7.1. We will use this version for the entire semester. When discussing specific R commands in R during lecture I will provide hard or electronic copies of the code so that you can follow along more easily and repeat the commands at home.

Your Dataset

It is often easiest to grasp new concepts through their application to familiar problems. In addition to discussing lecture examples from ecological and evolutionary literature, you will also analyze a dataset of your own. As a result, one of your first tasks will be to search out a dataset that you can analyze for your final paper (see below). This can be data that you collected yourself or a dataset that you have acquired from a supervisor or colleague that is related to your thesis (in a general or specific way). It is to your advantage to uncover a dataset relevant to your unique interests but if you are unable to do so, I will provide you with one.

Grading

This course will be graded on a point system as outlined below. Grades may be scaled up if results for the class are low, but they will not be scaled down.

4.0	90-100 points	2.0	70-74 points
3.5	85-89 points	1.5	65-69 points
3.0	80-84 points	1.0	60-64 points
2.5	75-79 points	0	< 60 points

Grading for the course will be based on:

Item	Points	Due Date
Assignment 1	10	Sept. 16th
Assignment 2	10	Sept. 30th
Assignment 3	10	Oct. 21st
Assignment 4	10	Nov. 25th
Final Paper	50	Dec 11th
Class Participation	10	
Total	100	

My Policy on Due Dates

Assignments will be due at the start of class on the date in which they are due. I appreciate that graduate students have many responsibilities in addition to course work. Due dates for assignments can be adjusted individually with prior approval (at least two weeks prior to due date). Grades for late assignments will be deducted at a rate of 10% per day. **The Final paper cannot be handed in late** (since grades are due 3 days later).

Assignments

Assignments are designed to ensure that you have practical working knowledge of the statistical approaches discussed in lecture and to provide you with experience presenting the results of statistical analyses. There will be 4 assignments during this course and each will cover a set of concepts discussed in lecture. An example dataset and relevant biological background or a list of required readings as well as specific instructions will be provided separately for each assignment. Most assignments will consist primarily of annotated output from R, but you will also be expected to present the results of analyses in text, figures or tables and to apply concepts and techniques from the assignments to your own dataset. Assignments are to be completed independently.

Final Paper

At the end of the semester you will submit a final paper of your own, which will resemble a published paper, including an Introduction, methods section, results, discussion and any relevant figures and tables, but will also include an appendix with annotated output from R. The statistical techniques used in your study will come from class, but the specific biological questions and approaches will be entirely determined by you. Your goal should be for this final paper to form a solid portion of a thesis chapter or manuscript for publication.

Participation grade

My goal is to provide a stimulating and interactive learning environment but I cannot do this entirely on my own. It is my belief that students, and graduate students in particular, provide an important component of the learning environment for others in the class. As a result, your participation grade will reflect your contribution to the overall learning environment in the classroom. Clearly you cannot contribute positively to such an environment if you are absent, and overt disruptions of others will be penalized. Full points will be given only to students who communicate insightful questions, perspectives and comments, and encourage similar contributions from others in the class.

Class Outline

Date	Topic	Date	Topic
Aug. 26	Outline of course and objectives. Introductions.	Aug. 28	Approaches to science and statistics
Sept. 2	Presentation of Data and Results using R	Sept. 4	Presentation of Data and Results using R
Sept. 9	The Utility of Probability distributions	Sept. 11	The Utility of Probability distributions
Sept. 16	Hypothesis testing and sampling designs	Sept. 18	Replication – scale, controls, independence
Sept. 23	Simulations for power analyses and inference.	Sept. 25	Resampling methods for estimation and inference
Sept. 30	Maximum Likelihood Estimation	Oct. 2	Maximum Likelihood Estimation & inference
Oct. 7	Bayesian philosophy & estimation I	Oct. 9	Bayesian estimation & MCMC
Oct. 14	Catch up	Oct. 16	Introduction to GLM: ANOVAs and regression as one big happy family.
Oct. 21	General Linear Models – assumptions and diagnostics	Oct. 23	Interactions, SS Types
Oct. 28	ANCOVA, Ratio Variables RMA, Collinearity	Oct. 30	Non-linearities, non-normalities, transformations
Nov. 4	Model Selection I	Nov. 6	Model Selection II , introduction to Generalized linear models (GLiM)
Nov. 11	GLiM 2	Nov. 13	GLiM 3
Nov. 18	Mixed Effects Models	Nov. 20	Mixed Effects Models
Nov.25	Mixed Effects Models	Nov.27	Thanksgiving – No Class
Dec. 2	Make your own statistics: Process models	Dec. 4	Make your own statistics: Process models

Readings

Primary Literature

Just as you need to keep yourself abreast of the developments in your particular field of study (from theory to important empirical studies) it is equally important to keep informed with developments in the statistical methods that inform your area of expertise. Readings from the primary literature are designed to supplement the material presented in lecture and will increase your general understanding and facilitate discussion. All readings will be made available through ANGEL in the appropriate folder for that topic.

Highly Recommended Texts – Most of the topics we cover will be associated with these texts. It is worth your while to have these books. However, in general we will not be working directly out of the texts, and different ones may suit you. In addition some of the recommended texts are expensive (Faraway), so if you choose to get them (which I **highly recommend**), you are best off looking around (amazon, half.com, ebay ,etc...). Please note that several of the books listed below are available as digital copies via the MSU library.

Bolker, B. 2008. *Ecological Models and Data in R*. Princeton University Press.

I think that this book is destined to change the skills of EEBB'ers for the next generation. It presents a fantastic overview of process modeling using Likelihood and Bayesian approaches using examples from R. From the absolute basics to very advanced models. If you put the effort into going through this book you will be well rewarded in all of your future endeavors. We will probably only deal with about half of the chapters in the book (1,2,4,5,6, 9 & some of ten) before we begin to focus on various aspects of linear models (using the Faraway books below) for the course.

As of July 2008, this book is available (and is reasonably priced) : Outdated drafts of chapters may still be available at the authors website, along with a number of resources related to the book.

<http://www.zoo.ufl.edu/bolker/emdbook/index.html>

The book also has a wiki set up for errata and clarifications

<http://emdbolker.wikidot.com/>

Faraway, J. J. 2005. *Linear Models with R*. Chapman & Hall/CRC, Boca Raton, FL.

Faraway, J. J. 2006. *Extending the Linear Model with R: generalized linear, mixed effects and nonparametric regression models*. Chapman & Hall/CRC, Boca Raton, FL.

These two texts cover the majority of the parametric statistical approaches that we will discuss in this course including general (ANOVA, regression & ANCOVA) and generalized (logistic, poisson and so many other) linear models. It does not delve a great deal into statistical theory, but instead focuses on practical aspects of performing analyses

and model evaluation. While they are a bit pricey as a pair, I highly recommend them, as they will be a valuable resource for a long time to come. If either the price or approach is not to your liking, I might suggest Gelman and Hill (2007 – See below) as an alternative, although some advanced approaches in the latter are quite different in their implementation.

These sites link to the authors information about the book (including errata etc..)

<http://www.maths.bath.ac.uk/~jjf23/LMR/>

<http://www.maths.bath.ac.uk/~jjf23/ELM/>

Dalgaard, P. 2004. *Introductory Statistics with R*. Springer. This book is most useful in helping develop your basic skills with using R for performing basic manipulations with your data set, as well as simple plotting and statistical procedures. It will also be useful as a review of some basic statistical concepts, for which you may need to be reminded. It is quite clear, and its examples are easy to follow. The Maindonald & Braun book (see below) is also quite useful and the choice may be a matter of personal preference. **This text is available online VIA the MSU library.**

Links to the book website

<http://staff.pubhealth.ku.dk/~pd/ISwR.htm>

Other recommended texts (in alphabetical order, not importance): These are a variety of texts that may not only serve you well for this class, but for your graduate work and beyond.

Braun, W.J. & Murdoch, DJ. 2008. *A First course in Statistical Programming with R*. Cambridge.

This slim book covers a lot of the concepts with regards to how to efficiently write your code in R. If you plan to use R to do more than run regression models and make very simple graphics, I recommend it. It will be especially useful for people who plan to use R for power analysis (or any other) simulations, and numerical optimization (for maximum likelihood for instance).

Burnham, K.P. & D.R. Anderson. 2002. *Model selection and multi-model inference: A practical information theoretic approach*. Springer.

A thorough introduction to model selection using information theoretic approaches (AIC & BIC). **This text is available online VIA the MSU library.**

Gelman, A. & Hill, J. 2007. *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge.

This book presents an alternative approach to modeling complex data using a GLiM framework, and is quite clear. It teaches concepts using an example based approach, and provides many hints, and “rules of thumbs” to help develop your statistical intuition. It is highly recommended for situations where you are dealing with complexities like repeated measures, time series or error variation that is not constant across treatments. It uses a combination of parametric and Bayesian approaches. The examples are largely not from the biological literature however. One additional and important warning, the multi-level model approach used throughout this book is quite different in implementation from hierarchical models that we will work with in class.

Good, P. 2006. *Resampling methods: A Practical Guide to Data Analysis*. 3rd ed. Birkhauser.

This book serves as a decent introduction to both bootstrapping and permutation/randomization resampling methods (which we will be using in class a fair bit). It is a gentle introduction to the subject with example code for R (and numerous other platforms), but it lacks any great depth. As a book it will get you started with resampling, but not to get you to the high powered tools (which we will not really cover in the course anyways).

Hilborn, R. & Mangel, M. 1997. *The Ecological Detective: Confronting models with Data*. Princeton University Press.

This is a classic text for thinking about how to analyze data for ecological and evolutionary data. While the majority of the nitty-gritty is covered well by Ben Bolker’s book, I still highly recommend reading this book, especially the first few chapters.

Maindonald, J. & Braun, J. 2006. *Data Analysis and Graphics Using R*. 2nd ed. Cambridge.

I own the first edition of this book and it helped me a lot when I was just getting started using R. It will not present a great deal of new statistical material for you, but it will show you how to do all of the basic statistical techniques in R. The book by Dalgaard covers many of the same subjects, and it may be a personal choice (or what you have access to).

McCarthy, M.A. 2007. *Bayesian Methods for Ecology*. Cambridge.

This book provides an extremely gentle introduction to Bayesian thinking and methodology using the Bayesian equivalents of otherwise common statistical methods (ANOVA and regression for instance). It is quite clear, and I recommend it for people who might struggle with some of the ideas of Bayesian methods. One word of caution, it is extremely one-sided with respect to the value of Bayesian methods, and treats all other

methods as if they should be left in the dust-bin. I do not subscribe to this philosophy at all (as you will learn in this course). It exclusively uses the BUGS environment (a language similar to R, but different in a few key ways) for Bayesian analysis.

Pinheiro, J.C. & D.M. Bates. 2000. *Mixed Effect Models in S/S-Plus*. Springer.

This is one of the classic texts describing all of the details (including all of the computational and algebraic gore) required to really understand mixed effect models. It also describes in detail how to use the original package for mixed modeling in R (nlme), which has been largely supplanted by lme4 (also by Douglas Bates). **This text is available online VIA the MSU library.**

Seefeld, K. Linder, E. 2007. *Statistics Using R with Biological Examples*. A free online book (PDF is available on the ANGEL site for the course). This book teaches statistics starting at a reasonably introductory level within a strongly Bayesian and computational framework. In particular the chapters on probability are excellent! http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th edition. Springer, New York, NY.

One of the classic texts used for both R & S (R is derived from S, and is extremely similar). This is not a “how to do/learn statistics” text, but how to use R/S to perform statistics. Still it is extremely invaluable as a resource. **This text is available online VIA the MSU library.**

R Resources

To download R (and the starting point for virtually all things R)
<http://cran.r-project.org/>

The manuals for using & programming in R. “An Introduction to R” & “R Data Import/Export” are the important ones for this class. The rest are largely useful if you really get into programming.
<http://cran.r-project.org/manuals.html>

The contributed documents section on the CRAN site is **REALLY USEFUL**. It has lots of books and tutorials of exceptional quality. Some of my favorites are up on the ANGEL site for the course.
<http://cran.r-project.org/other-docs.html>

The “Task views” section describes some of the packages/libraries in R that are useful for particular tasks (examples include “multivariate”, “Bayesian”, “Ecological and environmental data” & “spatial data” to name a few).

<http://cran.r-project.org/web/views/>

Tips for using R

<http://pj.freefaculty.org/R/Rtips.html>

kick-starting R

<http://cran.r-project.org/doc/contrib/Lemon-kickstart/>

Course notes for “Statistical Programming in R/S”

<http://socserv.mcmaster.ca/jfox/Courses/R-course/index.html>

Programming in R web site with examples

http://www.faculty.ucr.edu/~tgirke/Documents/R_BioCond/R_Programming.html

Examples of high end graphics and plots in R with source codes

<http://addictedtor.free.fr/graphiques/>

The R wiki –lots of useful example code bits. In particular this has useful information on Regression & mixed models in R, and examples from the Ecological Detective for R.

<http://wiki.r-project.org>

Other (statistics) web sites that may be useful to you.

Bruce Walsh’s website including his course notes for biostats, and Quantitative Genetics.

<http://nitro.biosci.arizona.edu/>

Electronic Statistics textbook resource – **EXCELLENT RESOURCE!!!!**

<http://www.statsoft.com/textbook/stathome.html>

Statistical Analysis in Ecology and Evolution – A course with very similar goals to our own. Lecture notes, R code and assignments available on the site.

<http://www.unc.edu/courses/2006spring/ecol/145/001/index.html>