Statistics/Forestry/Horticulture 572 — Statistical Methods for Bioscience II

Spring 2007

Instructor Information

Bret Larget
1250A MSC (Medical Sciences Center)
`http://www.stat.wisc.edu/~larget/`
`larget@stat.wisc.edu`
262-7979

Office Hours:
Tuesday, 3:00–4:30, Thursday, 1:00–2:30, and by appointment.

Course Information

Lecture:
Tuesday/Thursday 11:00–12:15, Van Hise 114

Discussion:
311: Tuesday, 1:00–2:15, 225 Ingraham
312: Wednesday, 2:30–3:45, 290 Van Hise
313: Wednesday, 1:00–2:15, 10 Ag Hall

Prerequisites:
Statistics 571 or consent of instructor

Course Materials

Required textbook: *Data Analysis and Graphics Using R*, 2nd ed., by Maindonald and Braun.

Course websites: See `http://www.stat.wisc.edu/courses/st572-larget/` for lecture notes and other materials and see `https://uwmad.courses.wisconsin.edu/` for the gradebook.

Lecture Notes: Lecture notes will be on the course web page. It will be useful to print lecture notes to take to lecture so that you can add to them during lecture.

Supplementary Materials: The course web page will also contain other supplementary materials including homework assignments, old exams, and additional notes on statistical concepts and R.

Additional references: While no other content resources are required, students wishing to obtain additional supplementary materials might consider these sources.

- *Course Notes for Statistics/Forestry/Horticulture 571* by Nordheim and Clayton;
- *Linear Models with R* by Faraway;
- *Extending the Linear Model with R* by Faraway;
- *The Statistical Sleuth, 2nd ed.* by Ramsey and Schafer;
- *Regression Analysis by Example, 4th ed.* by Chatterjee and Hadi;
- *An Introduction to Statistical Methods and Data Analysis, 6th ed.* by Ott and Longnecker;
- *Statistical Methods, 8th ed.* by Snedecor and Cochran.

The course will cover in part Chapters 5, 6, 7, 10, and parts of Chapter 8 of the textbook and some additional topics, including experimental design.

Course Objective

The primary course objective is to provide students in the biological sciences with a thorough conceptual and practical understanding of statistical analysis using linear and generalized linear models. Students will develop the ability to select an appropriate statistical model, to understand the underlying assumptions associated with this selection, to use the computer and statistical software to conduct the analysis, to interpret the results correctly, and to evaluate critically the method of analysis and of experimental design. The course does not attempt a comprehensive coverage of all associated statistical methods, but does aim to cover deeply the most important concepts underlying linear and generalized linear models. The goal is to prepare students to analyse data commonly encountered in biological research.

Linear Models

Linear models are a framework for statistical analysis in which the relationships among measured variables and stochastic error is described with a linear function (a combination of addition and scalar multiplication). Statistics 571 included the topics of simple linear regression and one-way analysis of variance, two simple examples of linear models with a single continuous quantitative response variable and a single explanatory variable (quantitative or categorical, respectively) using a normal error model. Statistics 572 explores more extensive linear models that include multiple explanatory variables, possibly of mixed type, whose effects may be modeled as fixed and random. Generalized linear models relax the normal distribution framework and are appropriate for modeling variables from different distributions which is appropriate when response variables are categorical or small integers, as is fairly common in biological settings.

Computing

There are many different possibilities for statistical software. I have selected R to teach in this course for many reasons. First and foremost, it is the first choice of many academic statisticians. R offers superior graphical capabilities, is based on a model of interactive data analysis, is flexible and extensible, and is extremely powerful. R is also free, works on all of the major computing platforms, including Windows, Macintosh, and Linux. R is based on a computing language developed several decades ago, and is under active development by many devoted volunteers. There is a large and vibrant R community and an increasing number of R books and on-line materials aimed at users of different skill levels and background. I have purposely selected a textbook that also serves as a useful reference and instructional manual for the R software. To become proficient in R is to attain a useful skill for life that can be applied long after the course ends. No other software choice offers the same suite of positive attributes.

Some of you may already have experience with another statistical software package and may work in a discipline where there is another standard, such as SAS. I will not require the use of R for exams, but will provide instruction in the use of R to achieve course goals. If you wish to use an alternative statistical software, it is your responsibility to ensure both that the software is sufficient for the requirements of the course and that you are able to use the software effectively without assistance.

Assignments

Work on homework assignments outside of class is likely to be the most effective manner in which you will learn and master course content. Homework assignments are due each week on Friday, in your TA's mailbox, by 4:00 pm. Assignments will be returned the following week in discussion section. Homework assignments should be well organized and reasonably neat. You must show your work to receive credit. Late homework assignments will be penalized unless extenuating circumstances exist. If possible, prior arrangements should be made with me in such cases. My judgment on the penalty for late work and the appropriateness of excuses for late work is final.

Exams

There will be one in-class midterm exam (75 minutes), one take-home midterm exam (one week), and a final exam (2 hours). The exams will cover content from lecture, readings, and homework assignments. All exams will be open-book and open-notes. The in-class midterm will take place on **March 6** (Tuesday), the take-home midterm will be distributed on **April 12** (Thursday) and due on **April 19** (Thursday). The final exam will take place at **10:05 am** on **Wednesday, May 16**. Missed exams will not be permitted except when extenuating circumstances prevail. In such cases, contact me as soon as is possible.

Grading

A final numerical score will be based on homework assignments (20%), the in-class midterm (20%), the take-home midterm (30%), and the final exam (30%). Course grades will be based on this numerical score.

Academic Honesty

All work on exams should be your own. In particular, the only person you may consult for assistance on the take-home midterm is me. In contrast, I encourage you to work collaboratively on homework assignments, provided that each person writes their own assignment solutions and that each person is actively engaged in solving each problem.

Discussion Sections

Assuming there is sufficient space, you may attend any discussion section without changing your registration. Time in discussion section will typically be used to solve problems similar to those on assignments, to ask questions, and to review past assignments. Attendance is strongly encouraged, but not required.

Syllabus

The course web page will include a tentative schedule addressing which topics will be covered during which lectures. As this is my first time teaching this course and the scope of topics differs from the past, this schedule should be treated as an educated guess that is subject to change.

Topics will include: introduction to linear models and generalized linear models, simple linear regression, correlation, outliers and influence, transformations, weighted least squares, estimation an interpretation of regression parameters, assessing accuracy of regression analysis, one-way analysis of variance, matrix description of regression, regression through the origin, multiple linear regression, assumptions and diagnostics of multiple regression, model assessment, model comparison, interpretation of multiple regression parameters, variable selection, multicollinearity, factors, main effects, interaction, analysis of covariance, polynomial regression, one-way random effects model, block designs, complete block designs, balanced incomplete block designs, multi-level experimental design, mixed models, split-plot designs, latin square designs, estimation and assessing predictive accuracy of mixed models, model selection, repeated measures in time, correlated errors, generalized linear models, logistic multiple regression, Poisson and quasi-Poisson models, estimation and model assessment of generalized linear models, and analysis of count data.